Efficient Learning for Crowdsourced Regression

Jungseul Ok∗, Sewoong Oh†, Jinwoo Shin∗, Yunhun Jang∗, and Yung Yi∗

March 1, 2017

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, one of the fundamental questions is how to identify more reliable workers and exploit this heterogeneity to infer the true answers. Despite significant research efforts for classification tasks with discrete answers, little attention has been paid to regression tasks where the answers take continuous values. We consider the task of recovering the position of target objects, and introduce a new probabilistic model capturing the heterogeneity of the workers. We propose the belief propagation (BP) algorithm for inferring the positions and prove that it achieves optimal mean squared error by comparing its performance to that of an oracle estimator. Our experimental results on synthetic datasets confirm our theoretical predictions. We further emulate a crowdsourcing system using PASCAL visual object classes datasets and show that de-noising the crowdsourced data using BP can significantly improve the performance for the downstream vision task.

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 low cost. To a data analyst, such systems provide unprecedented accesses to get training dataset at a scale and budget that was not previously feasible. Thus obtained training dataset can then be seamlessly integrated into downstream machine learning tasks together with the state-of-the-art classification and regression methods. 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 noisy data, we add redundancy which is a common and powerful strategy widely used in real-world crowdsourcing. We assign each task to multiple workers and aggregate these responses by some inference algorithm. For classification tasks, where each task asks a worker to find the best label from a finite set, the fundamental question of how to model the worker behavior [Dawid & Skene (1979); Zhou et al. (2015); Shah et al. (2016), how to assign tasks [Karger et al. (2011)], and how to aggregate the responses [Smyth et al. (1995)] to efficiently use the given budget and achieve the best accuracy, has been extensively studied. The key insight to achieving budget-optimal performance is to identify the good workers by comparing a worker’s responses with those of others on the same task and appropriately weighting the worker’s responses according to estimated reliability. Although the optimal inference algorithm is computationally intractable, various efficient approaches have been proposed with provable guarantees [Karger et al. (2011); Zhang et al. (2014)].

On the other hand, there are little principled approaches for crowdsourced regression tasks, where each task asks a worker to provide the best answer in a form of a real valued vector. While numerous machine learning tasks routinely done on crowdsourcing platforms require continuous valued evaluation of a training dataset, e.g., the location of an object [Everingham et al. (2015); Su et al. (2012), center of a galaxy, or the center of a marker in a Cryo-EM image, we lack systematic study of how to tackle the human noise in thus collected data. For this crowdsourced regression problem, we address 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

∗Jungseul Ok, Jinwoo Shin, and Yung Yi are with the Department of Electrical Engineering at Korea Advanced Institute of Science and Technology, Daejeon, South Korea (e-mail: {ockjs, jinwoos, yhjang, yiyung}@kaist.ac.kr)
†Sewoong Oh is with the Department of Industrial and Enterprise Systems Engineering at University of Illinois at Urbana-Champaign, Illinois, USA (e-mail: swoh@illinois.edu)
we pay for each response, and hence the budget per task is proportional to the redundancy: how many times a task has been assigned.

**Contribution.** We propose a probabilistic model for crowdsourced regression, where each worker provides a noisy observation on the (continuous valued) ground truths with his/her own noise level. We formulate the inference problem to minimize mean squared error (MSE). We propose a novel inference algorithm based on belief propagation (BP) [Pearl (1982)] to estimate workers’ noise levels, and consequently estimate the ground truths. We provide a tight upper bound on MSE of this estimator, and show that we achieve optimal performance for a broad regime of parameters by comparing it to an oracle estimator who knows all the noise levels of all workers (see Theorem 1). However, this oracle estimator fails to give a tight lower bound when each worker is assigned only a small number of tasks; it is impossible to learn worker noise levels accurately.

To provide a tighter lower bound, we directly compare the optimal inference algorithm by borrowing ideas from a long line of work in BP, e.g., [Mossel et al. (2014)]. We show the optimality of our estimator even when each worker is assigned a finite number of tasks (see Theorem 2), but under additional assumptions on the noise level. We provide experimental results suggesting that the optimality of our algorithm generally holds for all regimes, i.e., without the assumptions. We further demonstrate the superiority of our estimator on downstream tasks of training the state-of-art neural network based regressor, coined SSD [Liu et al. (2016)], on PASCAL visual object classes. Under our emulated crowdsourcing system, SSD trained using crowdsourced datasets by BP provides up to 5% mAP gain in its prediction, compared to that by baseline algorithms.

**Related work.** Crowdsourcing systems are widely used in practice for a variety of real-world tasks ranging from classification [Sheng et al. (2008)] and regression [Marcus et al. (2012)] to more complex tasks such as protein folding [Peng et al. (2013)], searching videos [Bernstein et al. (2011)], ranking [Lee et al. (2012)] and natural language processing [Wu et al. (2012)]. However, systematic approaches have been developed only for 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, 2014); Liu et al. (2012); Ghosh et al. (2011); Zhou et al. (2012); Dalvi et al. (2013); Zhang et al. (2014); Ok et al. (2016)]. In this paper, we formally define a *crowdsourced regression problem* and study these fundamental questions.

Belief propagation is a widely used heuristic for solving inference problems on probabilistic graphical models, such as our crowdsourcing regression model. Although it enjoys numerous empirical successes, theoretical analysis of BP has been limited to a few instances including community detection [Mossel et al. (2014)], error correcting codes [Kudekar et al. (2013)] and combinatorial optimization [Park & Shin (2015)]. The analysis of Theorem 2 follows the proof strategy initially introduced in [Mossel et al. (2014)] where BP is shown to be optimal for community detection, and generalized in [Ok et al. (2016)] for crowdsourced classification problems.

# 2 Crowdsourcing for Regression

In this section, we present our probabilistic model capturing heterogeneity of the workers and the corresponding optimal estimator minimizing the mean squared error. Since this is computationally intractable, we introduce a tractable estimator using belief propagation.

## 2.1 Crowdsourcing Model

The task requester has a set of $n$ regression tasks, denoted by $V = \{1, \ldots, n\}$. As a running example, consider object detection where a worker is asked to locate the position of an object of interest, e.g., the center of a galaxy or the center of a marker in a Cryo-EM image. In the $i$-th task, we denote the true position by $\mu_i \in \mathbb{R}^d$. To estimate these unknown true positions, we assign the tasks to a set of $m$ workers, denoted by $W = \{1, \ldots, m\}$ 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 $u$, worker $u$ provides his/her estimation/guess $A_{iu} \in \mathbb{R}^d$ for the true location $\mu_i$. Each worker $u$ is parameterized by his noise level $\sigma_u^2$, such that the response $A_{iu}$ suffers from an additive spherical Gaussian noise with variance $\sigma_u^2$. Precisely, conditioned on $\mu_i$ and $\sigma_u^2$, $A_{iu}$ is independently distributed with Gaussian...
where we define
\[ f_{A_i}(x \mid \mu_i, \sigma_i^2) = \phi(x \mid \mu_i, \sigma_i^2) \]
defined as
\[ \phi(x \mid \mu_i, \sigma_i^2) := \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left( -\frac{1}{2} \frac{\|x - \mu_i\|^2}{\sigma_i^2} \right) \]

We assume each worker \( u \)'s variance \( \sigma_u^2 \) is independently drawn from a finite set \( S = \{\sigma_1^2, \ldots, \sigma_S^2\} \) uniformly at random. We further assume that the true position \( \mu_i \) is independently drawn from a Gaussian prior distribution \( \phi(x \mid \mu_i, \tau^2) \) for given mean \( \mu_i \in \mathbb{R}^d \) and variance \( \tau^2 \in (0, \infty) \), i.e., we have a prior knowledge that the true position \( \mu_i \) is near by \( \nu_i \). We are interested in sufficiently large \( \tau^2 \) only assuming marginal knowledge on \( \mu_i \).

### 2.2 Crowdsourcing Regression

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

\[
\min_{\hat{\mu}} \frac{1}{n} \sum_{i \in V} \text{MSE}(\hat{\mu}_i(A))
\]

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

\[
\hat{\mu}_i^*(A) = \sum_{\sigma_i^2 \in S} \mathbb{E}[\mu_i \mid A_i, \sigma_i^2] \prod_{u \in M_i} \mathbb{P}[\sigma_u^2 \mid A]
\]

where we let \( A_i := \{A_{iu} : u \in M_i\} \) and for the last equality, we use the conditional independence between \( A_i \) and \( (A \setminus A_i) \) given \( \sigma_i^2 \). Computing the marginal posterior \( \mathbb{P}[\sigma_u^2 \mid A] \) in general requires summing over the rest of (exponentially many) \( \sigma_u^2 \)'s, making this optimal estimator intractable. We make this intractable estimator explicit in the following, which leads to a tractable estimator based on belief propagation in Section 2.3.

We first show that the posterior density of \( \mu_i \) given \( A_i = y_i := \{y_{iu} \in \mathbb{R}^d : u \in M_i\} \) and \( \sigma_i^2 \) is a Gaussian density and compute its mean and variance:

\[
f_{\mu_i}[x \mid A_i = y_i, \sigma_i^2] = f_{\mu_i}[x] f_{A_i}[y_i \mid \mu_i = x, \sigma_i^2] \]

\[
= \frac{1}{\sqrt{2\pi\tau^2}} \exp \left( -\frac{1}{2} \frac{\|y_i - \mu_i\|^2}{\tau^2} \right) \prod_{u \in M_i} \mathbb{P}[\sigma_u^2 \mid A]
\]

where we define \( \sigma_i^2 : S \rightarrow \mathbb{R} \) and \( \hat{\mu}_i : \mathbb{R}^{d \times M_i} \times S \rightarrow \mathbb{R}^d \) as follows

\[
\sigma_i^2 (\sigma_i^2) := \frac{1}{\tau^2 + \sum_{u \in M_i} \frac{1}{\sigma_u^2}}
\]

\[
\hat{\mu}_i (A_i, \sigma_i^2) := \sigma_i^2 (\sigma_i^2) \left( \frac{\nu_i}{\tau^2} + \sum_{u \in M_i} \frac{A_{iu}}{\sigma_u^2} \right).
\]

The Gaussian posterior density \( f_{\mu_i}[x \mid A_i = y_i, \sigma_i^2] \) follows from:

\[
f_{\mu_i}[x] f_{A_i}[y_i \mid \mu_i = x, \sigma_i^2] \]

\[
= \phi(x \mid \nu_i, \tau^2) \prod_{u \in M_i} \phi(y_{iu} \mid \nu_i, \sigma_u^2) \]

\[
= C_i (y_i, \sigma_i^2) \cdot \phi(x \mid \hat{\mu}_i (y_i, \sigma_i^2), \sigma_i^2 (\sigma_i^2))
\]

---

1 To make the joint density normalizable, it is necessary to assume a prior distribution of \( \mu_i \). It doesn’t have to be the Gaussian, and our results generalize to other prior distributions, such as a uniform distribution on a Euclidean ball.
and \(f_{A_i}|y_i|\sigma_{M_i}^2| = C_i(y_i, \sigma_{M_i}^2)\), where

\[
C_i(A_i, \sigma_{M_i}^2) := \left(\frac{2\pi \cdot \sigma_{M_i}^2}{2\pi \sigma_i^2 \prod_{u \in M_i} (2\pi \sigma_u^2)}\right)^\frac{1}{2} \cdot \exp(D_i(A_i, \sigma_{M_i}^2)),
\]

\[
D_i(A_i, \sigma_{M_i}^2) := -\frac{1}{2} \sigma_{M_i}^2 \left(\sum_{u \in M_i} \frac{||A_{iu} - \nu_i||^2}{\sigma_u^2} + \sum_{v \in M_i \setminus \{u\}} \frac{||A_{i} - A_{iv}||^2}{\sigma_v^2}\right).
\]

Equation (4) leads to the posterior mean, which is weighted average of the prior mean and the worker responses, each weighted by the inverse of its variance:

\[
E[\mu_i|A_i, \sigma_{M_i}^2] = \bar{\mu}_i(A_i, \sigma_{M_i}^2).
\]

Thus, the optimal estimator \(\bar{\mu}_i^*(A)\) in (2) is given by

\[
\bar{\mu}_i^*(A) = \sum_{\sigma_{M_i}^2 \in S_{M_i}} \bar{\mu}_i(A_i, \sigma_{M_i}^2) \prod_{u \in M_i} \mathbb{P}[\sigma_u^2|A].
\]

The marginal probability of \(\sigma_u^2\) given \(A\) in (7) can be calculated by marginalizing out \(\sigma_{-u}^2 := \{\sigma_v^2 : v \in W \setminus \{u\}\}\) from the joint probability of \(\sigma^2\), i.e.,

\[
\mathbb{P}[\sigma_u^2|A] = \sum_{\sigma_{-u}^2 \in S_{W \setminus \{u\}}} \mathbb{P}[\sigma^2|A],
\]

where for given \(A = \{y_{iu} \in \mathbb{R}^d : (i, u) \in E\}, \mathbb{P}[\sigma^2|A = y] \propto f_A[y|\sigma^2] = \prod_{i \in V} f_{A_i}[y_i|\sigma_{M_i}^2]\) and

\[
\mathbb{P}[\sigma^2|A] \propto \prod_{i \in V} C_i(A_i, \sigma_{M_i}^2).
\]

The summation in (8) is taken over exponentially many \(\sigma_{-u}^2 \in S_{m-1}\) with respect to \(m\). Thus, in general, the optimal estimator \(\bar{\mu}_i^*(A)\) in (7), requiring the marginal probability of \(\sigma_u^2\) given \(A\) in (8), is computationally intractable.

### 2.3 Belief Propagation

We note that the joint probability of \(\sigma^2\) given \(A\) in the product form of (9) forms a factor graph [Jordan 1998] where each worker \(u\)'s variance \(\sigma_u^2\) and each task \(i\) correspond to a variable and a local factor \(C_i(A_i, \sigma_{M_i}^2)\) on the set of workers, \(M_i\), to whom task \(i\) is assigned, respectively. This probabilistic graphical model motivates to use the popular (sum-product) belief propagation (BP) algorithm [Pearl 1982] on the factor graph of \(\mathbb{P}[\sigma^2|A]\) for approximating the marginalization in (8), which is intractable. BP typically is an efficient heuristic with little known provable guarantees.

First, we give explicit iterative BP update rules on the messages \(m_{u \rightarrow i}\) and \(m_{i \rightarrow u}\) between task \(i\) and worker \(u\) and belief \(b_{iu}\) on each worker \(u\):

\[
m_{u \rightarrow i}^{t+1}(\sigma_u^2) \propto \prod_{j \in N_u \setminus \{i\}} m_{j \rightarrow u}(\sigma_u^2)
\]

\[
m_{i \rightarrow u}^{t+1}(\sigma_u^2) \propto \sum_{\sigma_{M_i}^2 \in S_{M_i}} C_i(A_i, \sigma_{M_i}^2) \prod_{v \in M_i \setminus \{u\}} m_{v \rightarrow i}^{t+1}(\sigma_v^2)
\]

\[
b_{iu}^{t+1}(\sigma_u^2) \propto \prod_{i \in N_u} m_{i \rightarrow u}^{t+1}(\sigma_u^2)
\]
where the belief \( b_u(\sigma_u^2) \) denotes the estimated marginal probability of \( \sigma_u^2 \) given \( A \). We initialize messages with a constant \( \frac{1}{\sqrt{|S|}} \) and normalize messages and beliefs, i.e.,

\[
\sum_{\sigma_u^2} m_{0 \rightarrow i}(\sigma_u^2) = \sum_{\sigma_u^2} m_{i \rightarrow u}(\sigma_u^2) = \sum_{\sigma_u^2} b_u(\sigma_u^2) = 1.
\]

At the end of \( k \) iterations, one can estimate \( \hat{\mu}_i^{BP}(A) \) from (7) by substituting \( b_u^0(\sigma_u^2) \) into \( \mathbb{P}[^2 | A] \). Formally,

\[
\hat{\mu}_i^{BP(k)}(A) := \sum_{\sigma_{M_i}^2 \in S_{M_i}} \hat{\mu}_i(A_i, \sigma_{M_i}^2) \prod_{u \in M_i} b_u(\sigma_u^2).
\]

We note that if the factor graph is a tree, i.e., having no loop, then it is well known that BP can calculate the exact marginal probability [Pearl 1982], i.e.,

\[
b_u^t(\sigma_u^2) = \mathbb{P}[\sigma_u^2 | A] \text{ for all } t \geq m.
\]

However, for general graphs having loops, BP has no performance guarantee, i.e., BP may output \( b_u(\sigma_u^2) \neq \mathbb{P}[\sigma_u^2 | A] \), and even the convergence of BP is not guaranteed, i.e., the value of \( \lim_{t \rightarrow \infty} b_u(\sigma_u^2) \) may not exist.

Even though BP doesn’t have the performance and convergence guarantees, it has been applied to many applications having loops with empirical successes [Murphy et al. 1999; Liu et al. 2012; Yanover et al. 2006]. We propose BP for crowdsourced regression under our model assuming the finite set of the worker variance \( S = \{\sigma_1^2, \ldots, \sigma_k^2\} \). If the support \( S \) is a finite set, e.g., a continuous interval, running BP becomes computationally intractable since the messages become functions on the infinite support. To address the issue, several methods approximating messages have been studied [Minka 2001; Wald & Globerson 2014; Noorshams & Wainwright 2013; Moallemi & Roy 2009]. However, such an extra layer of approximation renders their performance analysis significantly more challenging.

## 3 Performance Guarantees on BP

In this section, we provide the theoretical guarantees of BP estimator under our model for the crowdsourced regression. We first describe our proposed task assignment.

\( (\ell, r) \)-regular task assignment. In general, the performance of an estimator in our model differs depending how tasks are assigned to workers. We propose a simple assignment scheme, referred to as \( (\ell, r) \)-regular task assignment, popularly adopted in crowdsourcing [Dalvi et al. 2013; Liu et al. 2012; Karger et al. 2011, 2013, 2014; Ok et al. 2016]. The assignment graph \( G \) is generated as a random \( (\ell, r) \)-regular bipartite graph, i.e., \( G \) is drawn uniformly at random out of all \( (\ell, r) \)-regular graphs, where each task is assigned to \( \ell \) workers and each worker is assigned \( r \) tasks.

The presentation of our main results is two-fold. First, in Section 3.1, we provide a sharp upper bound on MSE achieved by BP. This implies that BP approaches optimal MSE if both the number of tasks assigned to one worker (i.e., \( r \)) and the total number of tasks (i.e., \( n \)) increase. However, our simulations suggest BP is near optimal even for finite \( r \). We make this precise in Section 3.2 where we compare BP directly to the optimal estimator, quantifying the relative gap. We show that under some mild assumptions on the model parameters, this gap vanishes even if we maintain finite \( r \), proving a stronger notion of optimality.

### 3.1 Quantitative Performance of BP

We first present a performance guarantee of the BP estimator in Theorem 1 whose proof is given in Section 4.1

**Theorem 1.** Consider the crowdsourced regression model with \( S = \{\sigma_1^2, \ldots, \sigma_k^2\} \) and a random \( (\ell, r) \)-regular graph \( G \) consisting of \( n \) tasks and \( (\ell/r)n \) workers. For given \( \epsilon, \sigma_{\min}^2, \sigma_{\max}^2 > 0 \) and \( \ell \geq 2 \), if (i) \( |\sigma_2^2 - \sigma_s^2| > \epsilon \) and \( \sigma_{\min}^2 \leq \sigma_s^2 \leq \sigma_{\max}^2 \) for all \( 1 \leq s \neq s' \leq S \), and (ii) \( r, k \leq \log \log n \), then for sufficiently large \( n \), \( k \) iterations of BP achieves

\[
\mathbb{E} \left[ \frac{1}{n} \sum_{i \in V} \text{MSE}(\hat{\mu}_i^{BP(k)}(A)) \right] \leq \frac{d}{n} \sum_{i \in V} \mathbb{E} \left[ \sigma_{M_i}^2 \right]
\]
+ \mathcal{E}_t \cdot S \cdot \sqrt{\exp \left( - \frac{\varepsilon^2}{8(8\varepsilon + 1)\sigma_{\text{max}}^2} \cdot r \right) + 2^{-k}} \quad (15b)

where the expectation is taken with respect to the distribution of G and A and we define

\[ \mathcal{E}_{t,S} := 6\ell \cdot S^{2t} \cdot \sqrt{\frac{d\sigma_{\text{max}}^6}{\varepsilon^3\sigma_{\text{min}}^2} \left( 2 + \frac{d\sigma_{\text{max}}^6}{\varepsilon^3\sigma_{\text{min}}^2} \right)} \quad (16) \]

We provide three interpretations of Theorem 1. First, consider an oracle estimator that knows the hidden variances \( \sigma_u^2 \)'s and makes optimal inference as follows:

\[ \hat{\mu}_i^\text{ora}(A, \sigma^2) := \mathbb{E}[\mu_i | A, \sigma^2] = \hat{\mu}_i(A, \sigma_{\text{max}}^2). \quad (17) \]

This gives the MSE of \( \hat{\mu}_i^\text{ora}(A, \sigma^2) \):

\[ \mathbb{E} \left[ \frac{1}{n} \sum_{i \in \mathcal{V}} \text{MSE}(\hat{\mu}_i^\text{ora}(A, \sigma^2)) \right] = \frac{n}{d} \sum_{i \in \mathcal{V}} \mathbb{E} \left[ \sigma_i^2(\sigma_{\text{max}}^2) \right]. \]

Note that the oracle estimator \( \hat{\mu}_i^\text{ora} \) always outperforms even the optimal estimator \( \hat{\mu}^* \) in (7), providing a lower bound on the MSE of any estimator. This coincides with (15a) in our bound, implying that the gap to oracle performance is (15b). We emphasize oracle here as, without an access to an oracle, the analysis of the actual optimal estimator should give a tighter lower bound than (15a). This is made precise in Theorem 2.

Second, for sufficiently large \( n \), the number \( r \) of per-worker tasks and the total iterations \( k \) grow with \( n \), BP's performance approaches that of the oracle estimator, as (15b) vanishes, i.e.,

\[ \lim_{n \to \infty} \mathbb{E} \left[ \frac{1}{n} \sum_{i \in \mathcal{V}} \text{MSE}(\hat{\mu}_i^{\text{BP}(k)}(A)) \right] = \frac{d}{n} \sum_{i \in \mathcal{V}} \mathbb{E} \left[ \sigma_i^2(\sigma_{\text{max}}^2) \right]. \]

This is because under \((\ell, r)\)-regular task assignment, for increasing \( r \) with the total number of tasks \( n \), BP estimator accurately infers all workers’ variances and thus optimally estimates the true positions \( \mu \). Note that the above performance limit holds for any \( r = o(1) \), implying that a reasonable number of tasks per worker is enough in practice to achieve BP’s optimality.

Third, we compare BP with a simple average-based estimator \( \hat{\mu}_i^{\text{avg}}(A) := (1/|M_i|) \sum_{u \in M_i} A_{iu} \):

\[ \mathbb{E} \left[ \frac{1}{n} \sum_{i \in \mathcal{V}} \text{MSE}(\hat{\mu}_i^{\text{avg}}(A)) \right] = \frac{1}{n} \sum_{i \in \mathcal{V}} \frac{d}{|M_i|} \cdot \mathbb{E} \left[ \sum_{u \in M_i} \sigma_u^2 \right] \]

where the expectation is taken with respect to the distribution of \( A \). \( \text{MSE}(\hat{\mu}_i^{\text{avg}}(A)) \) increases proportionally to the arithmetic mean of variances of workers assigned to each task, while \( \text{MSE}(\hat{\mu}_i^{\text{BP}(k)}(A)) \) is proportional to the harmonic mean of variances of workers and prior, i.e., \( \mathbb{E} [\text{MSE}(\hat{\mu}_i^{\text{avg}}(A))] \geq \mathbb{E} [\text{MSE}(\hat{\mu}_i^{\text{BP}(k)}(A))] \). 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 task \( i \) has high accuracy, i.e., \( \sigma_u^2 \simeq 0 \), and the others’ variances are \( x \)'s, then \( \text{MSE}(\hat{\mu}_i^{\text{avg}}(A)) \simeq (d/|M_i|) x \) but \( \text{MSE}(\hat{\mu}_i^{\text{BP}(k)}(A)) \simeq 0 \). 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.

### 3.2 Relative Performance of BP

We provide the relative performance of BP estimator by comparing with the optimal estimator, in particular, when the quantitative guarantee in Theorem 1 is not tight, i.e., \( r \) is small, in Theorem 2 whose proof is provided in Section 4.2.

**Theorem 2.** Consider the crowdsourced regression model with \( S = \{\sigma_{\text{min}}^2, \sigma_{\text{max}}^2\} \) and a random \((\ell, r)\)-regular graph \( G \) consisting of \( n \) tasks and \((\ell/r)n \) workers. For given \( \varepsilon > 0 \) and \( \ell \), there exists a constant \( C_{\ell, \varepsilon} \), depending on only \( \ell \)
and ε, such that if (i) \( \sigma_{\text{min}}^2 + \varepsilon \leq \sigma_{\text{max}}^2 \leq 2\sigma_{\text{min}}^2 \) (ii) \( C_{\ell,\varepsilon} \leq r \leq \log \log n \), and (iii) \( k \leq \log \log n \), then for sufficiently large \( n \),

\[
E \left[ \frac{1}{n} \sum_{i \in V} \left| \text{MSE}(\hat{\mu}_i^*(A)) - \text{MSE}(\hat{\mu}_i^{\text{BP}(k)}(A)) \right| \right] \leq \mathcal{E}_{\ell, S} \cdot 2^{-k}
\]

where the expectation is taken with respect to the distribution of \( G \) and \( A \) and \( \mathcal{E}_{\ell, S} \) is defined in \( (16) \).

As a corollary, it follows that when we set \( k \) increasing with \( n \), e.g., \( k = \log \log n \), we have an asymptotic optimality of BP estimator:

\[
\lim_{n \to \infty} E \left[ \frac{1}{n} \sum_{i \in V} \left| \text{MSE}(\hat{\mu}_i^*(A)) - \text{MSE}(\hat{\mu}_i^{\text{BP}(k)}(A)) \right| \right] = 0.
\]

This result is not directly comparable to Theorem 1 as they apply to different regimes of the parameters. In particular, the oracle optimality gap \( (15b) \) does not vanish for finite \( r \). We believe this is because the oracle is too strong to compete against when \( r \) is small. Hence, we need to compare against a more practical lower bound on the optimal estimator as described in \( (8) \) that does not rely on the oracle.

Such a comparison can be made rigorous by constructing the following lower bound on the fundamental limit. We use the fact that the random \((\ell, r)\)-regular bipartite graph has a \emph{locally tree-like structure} with depth \( k \leq \log \log n \) [Bollobás (1998)] and BP 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 \emph{weaker} oracle estimator that gives a tighter lower bound. Directly analyzing the performance of such a weaker oracle is hard. Instead, we show that the gap between our estimator (that does not have the ground truths at the boundary of local tree) and the weaker oracle decreases as the depth of the tree increases. This is made clear by establishing \emph{decaying correlation} from the information on the outside of the local tree to the root of the tree.

However, for the analytic tractability, we need a constant lower bound of \( r \) and \( S = 2 \), i.e., \( r \geq C_{\ell, \varepsilon} \) and \( |S| = 2 \) as similar conditions are required in the analysis [Ok et al. (2016), Mossel et al. (2014)]. In addition to \( S = 2 \), our analysis further requires the assumption on \( S \), i.e., \( \sigma_{\text{min}}^2 + \varepsilon \leq \sigma_{\text{max}}^2 \leq 2\sigma_{\text{min}}^2 \). However, the condition on \( \sigma_{\text{max}}, \sigma_{\text{min}} \) is the most challenging/important regime for inference algorithms since the ratio of the maximum and minimum variances is bounded by some constant, i.e., it is hard to distinguish the workers’ variances. If the ratio is large, most inference algorithms would provide outputs of enough quality to use in practice. The experimental results in Section 5.2 indeed confirm that BP is optimal for general regimes violating the conditions assumed in Theorem 1.

### 4 Proofs of Theorems

#### 4.1 Proof of Theorem 1

We start with a bound on the conditional expectation of MSE of \( \hat{\mu}_i^{\text{BP}(k)}(A) \) conditioned on \( \sigma^2 = \hat{\sigma}^2 \in S^W \). Let \( E_{\hat{\sigma}^2} \) be the conditional expectation given \( \sigma^2 = \hat{\sigma}^2 \). Using Cauchy-Schwarz inequality for random variables \( X \) and \( Y \), i.e., \( |E[XY]| \leq \sqrt{E[X^2]E[Y^2]} \), and some calculus, it is not hard to obtain that

\[
E_{\hat{\sigma}^2} \left[ \left\| \hat{\mu}_i^{\text{BP}(k)}(A) - \mu_i \right\|_2^2 \right] \\
\leq d E[\sigma_i^2(\hat{\sigma}^2_m)] + \frac{\mathcal{E}_{\ell, S}}{2E[S]} \sum_{u \in M_i} \sum_{\sigma_u^2 \neq \hat{\sigma}_u^2} \sqrt{E_{\hat{\sigma}^2} [b_u^k(\sigma_u^2)]}
\]

where the detailed steps for (18) is provided in Appendix A.4.

Let \( \rho \in W \) denote a worker chosen uniformly at random for given \( G = (V, W, E) \). Then it is enough to show that for \( \sigma_u^2 \in S^W \) such that \( \sigma_u^2 \neq \hat{\sigma}_u^2 \) for every \( u \in W \),

\[
E \left[ \sqrt{E_{\hat{\sigma}^2} [b_u^k(\sigma_u^2)]} \right] \leq \sqrt{4 \frac{\varepsilon^2_{\rho}}{8(8\varepsilon+1)\hat{\sigma}_{\text{max}}^2} + 2^{-k}}
\]

(19)
where the first expectation is taken with respect to $G$. It is known that a random $(\ell, r)$-regular bipartite graph $G$ is a locally tree-like. Formally, from Lemma 5 in [Karger et al. (2014)], it is straightforward to check that
\[
\mathbb{P}[G_{\rho, 2k} \text{ is not a tree}] \leq \frac{3}{n} (\ell r)^{2k+1}
\]  
(20)
where we let $G_{\rho, 2k} = (V_{\rho, 2k}, W_{\rho, 2k}, E_{\rho, 2k})$ denote the subgraph of $G$ induced by all the nodes within (graph) distance $2k$ from root $\rho$. From (20), it follows that for any given $\sigma^2 \in \mathcal{S}^W$, 
\[
\mathbb{E} \left[ \mathbb{E}_{\sigma^2} \left[ b^k_{\rho}(\sigma^2) \right] \right] 
\leq \mathbb{E} \left[ \mathbb{E}_{\sigma^2} \left[ b^k_{\rho}(\sigma^2) \mid G_{\rho, 2k} \text{ is a tree} \right] \right] + \frac{3}{n} (\ell r)^{2k+1}
\]  
(21)
where from the choice of $r, k \leq \log \log n$ and constant $\ell$, it follows that for sufficiently large $n$, 
\[
\frac{3}{n} (\ell r)^{2k+1} \leq 2^{-k}.
\]  
(22)
Recalling (14), it follows that if $G_{\rho, 2k}$ is a tree, 
\[
b^k_{\rho}(\sigma^2) = \mathbb{P}[\sigma^2 = \sigma^2_{\rho} \mid A_{\rho, 2k}]
\]  
(23)
where we let $A_{\rho, 2k} = \{A_{iu} : (i, u) \in E_{\rho, 2k}\}$. Hence it is enough to show that if $G_{\rho, 2k}$ is a tree, $b^k_{\rho}(\sigma^2_{\rho}) = \mathbb{P}[\sigma^2_{\rho} \mid A_{\rho, 2k}]$ is concentrated at $\sigma^2 = \hat{\sigma}^2$. We present Lemma 1 providing the concentration formally whose proof is given in Appendix A.1.

**Lemma 1.** Suppose $G_{\rho, 2k} = (V_{\rho, 2k}, W_{\rho, 2k}, E_{\rho, 2k})$ is induced from $(\ell, r)$-regular bipartite graph $G = (V, W, E)$ with $\ell \geq 2$ and $r \geq 1$ and it is a tree with depth $2k \geq 2$. For given $\varepsilon, \sigma^2_{\min}, \sigma^2_{\max} > 0$, consider $\mathcal{S} = \{\sigma^2_1, \ldots, \sigma^2_S\}$ such that (i) $|\sigma^2_s - \sigma^2_s'| > \varepsilon$ and $\sigma^2_{\min} \leq \sigma^2_s \leq \sigma^2_{\max}$ for all $1 \leq s \neq s' \leq S$. Then, for any given $\hat{\sigma}^2 \in \mathcal{S}^W$, 
\[
\mathbb{E}_{\sigma^2} \left[ \mathbb{P}[\sigma^2 \neq \hat{\sigma}^2 \mid A_{\rho, 2k}] \right] \leq 4e^{-\frac{\varepsilon^2}{8(8\varepsilon+1)\sigma^2_{\max}}}.
\]  
(24)
where the expectation is taken with respect to $A$ from the crowdsourced regression model given $\sigma^2 = \hat{\sigma}^2$ with $G$.

Since the choice of $\hat{\sigma}^2$ is arbitrary, combining (21) and (22) with (24) in Lemma 1 it implies (19) and completes the proof of Theorem 1.

### 4.2 Proof of Theorem 2

Recalling the calculations of $\mu^*_i(A)$ and $\mu^{{\text{BP}(k)}}_i(A)$ in (7) and (13), the only difference between them is the estimation on $\sigma^2_{\alpha}$, i.e., BP uses $b^k_{\alpha}(\sigma^2_{\alpha})$ instead of $\mathbb{P}[\sigma^2_{\alpha} \mid A]$. Using Cauchy-Schwarz inequality and some calculus, similarly as (18), we can quantify an upper bound on the expectation of the gap between MSE’s of $\mu^*_i(A)$ and $\mu^{{\text{BP}(k)}}_i(A)$ in terms of the difference between $\mathbb{P}[\sigma^2_{\alpha} \mid A]$ and $b^k_{\alpha}(\sigma^2_{\alpha})$ as follows:
\[
\mathbb{E} \left[ \frac{1}{n} \sum_{i \in V} \text{MSE}(\hat{\mu^*_i}(A)) - \text{MSE}(\hat{\mu^{{\text{BP}(k)}}_i}(A)) \right] 
\leq 3\ell \cdot S^2 \cdot \frac{\varepsilon L S}{2} \sum_{u \in M_i} \mathbb{E} \left[ \left| \mathbb{P}[\sigma^2_{\alpha} \mid A] - b^k_{\alpha}(\sigma^2_{\alpha}) \right| \right].
\]  
(25)
where the detailed steps for (25) is provided in Appendix A.5. Hence it is enough to show that for sufficiently large $m = \frac{L}{\ell} n$ with constant $\ell$ and $r, k \leq \log \log n$, 
\[
\frac{1}{m} \sum_{u \in W} \mathbb{E} \left[ \left| \mathbb{P}[\sigma^2_{\alpha} \mid A] - b^k_{\alpha}(\sigma^2_{\alpha}) \right| \right] \leq 2^{-k+1}
\]  
(26)
where the expectation is taken with respect to $G$ and $A$. 8
Lemma 2. completes the proof of Theorem 2. it is a tree with depth \( \ell \geq 2 \). Let \( \rho \in W \) denote a worker chosen uniformly at random for given \( G = (V, W, E) \) and fix \( \sigma^2 \in S^W \). Recalling (20) and (22), it follows that for sufficiently large \( n \),

\[
E_{\sigma^2} \left[ \left( |P[\sigma^2 | A] - b_{\rho}^k(\sigma^2) | \right) \right] \\
\leq E_{\sigma^2} \left[ \left( |P[\sigma^2 | A] - b_{\rho}^k(\sigma^2) | \right) G_{\rho,2k} \text{ is a tree} \right] + 2^{-k}.
\]

We present Lemma 2 providing an upper bound on the last term of (27), which implies (26) with (27) and (23) and completes the proof of Theorem 2.

Lemma 2. Suppose \( G_{\rho,2k} = (V_{\rho,2k}, W_{\rho,2k}, E_{\rho,2k}) \) is induced from \((\ell,r)\)-regular bipartite graph \( G = (V, W, E) \) and it is a tree with depth \( 2k \geq 2 \). For given \( \varepsilon, \sigma^2_{\text{min}}, \sigma^2_{\text{max}} > 0 \), consider \( S = \{\sigma^2_{\text{min}}, \sigma^2_{\text{max}}\} \) such that \( \sigma^2_{\text{min}} + \varepsilon \leq \sigma^2_{\text{max}} \leq 2\sigma^2_{\text{min}} \). Then, for any given \( \sigma^2 \in S^W \), there exists a constant \( C_{\ell,\varepsilon} \) such that if \( r \geq C_{\ell,\varepsilon} \), then

\[
E_{\sigma^2} \left[ \left| |P[\sigma^2 | A] - \bar{P}[\sigma^2 = \bar{\sigma}^2 | A_{\rho,2k}]| \right| \right] \leq 2^{-k}
\]

(28)

where the expectation is taken with respect to \( A \) from the crowdsourced regression model given \( \sigma^2 = \bar{\sigma}^2 \) with \( G \).

A rigorous proof of Lemma 2 is given in Appendix A.2. Here, we briefly provide the underlying intuition on the proof. As Lemma 1 states, if there is the strictly positive gap \( \varepsilon > 0 \) between \( \sigma^2_{\text{min}} \) and \( \sigma^2_{\text{max}} \), one can recover \( \sigma^2 \in \{\sigma^2_{\text{min}}, \sigma^2_{\text{max}}\} \) with small error using only the local information, i.e., \( A_{\rho,2k} \). On the other hand, \( A \setminus A_{\rho,2k} \) is far from \( \rho \) and is less useful on estimating \( \sigma^2 \). In the proof of Lemma 2, we quantify the decaying rate of information with respect to \( k \).

5 Experiment Results

In this section, we present experimental results that support our analytical findings and the superiority of BP for the crowdsourced regression, where we consider the task of locating objects of interest in images.

5.1 Tested Algorithms

We test four algorithms: BP, Oracle, Average, and Simple as implemented in what follows:

- **BP**: We implement BP without any use of prior information on true positions by taking limit \( \tau \to \infty \), i.e., it outputs \( \mu_{\text{BP}}(A) \) in [10]–[12] with \( \lim_{\tau \to \infty} C_t(A, \sigma^2_{\text{2d}}(\tau)) \). We terminate BP at the maximum of 100 iterations or after checking convergence of messages.

- **Oracle**: For comparison, we consider an artificial estimator having free access to workers’ variances \( \sigma^2 \). It outputs \( \mu_{\text{ora}}(A, \sigma^2) := \lim_{\tau \to \infty} \bar{\mu}_t(A, \sigma^2_{\text{2d}}(\tau)) \).
In our model, for mathematical rigorousness, we assumed that we have some information of the true position $\mu$ advance as specified by the density of $\mu$. Simple $\circ$ Average $\circ$ BP

Figure 2: An example of detections of SSD trained by the crowdsourced VOC-07/12 datasets by Average, Simple, BP, and Oracle.

- **Average**: Without learning workers’ variances, this just takes the average of workers’ observations, i.e., $\mu_{i}^{\text{avg}}(A) := \frac{1}{|M|} \sum_{u \in M} A_{iu}$.
- **Simple**: This has a simple mechanism to estimate each worker’s variance by comparing the worker’s answers and the others’ answers, and then it aggregates the workers’ answers based on the estimated variances. Formally, it outputs $\mu_{i}^{\text{nav}}(A) := \mu_{i}^{\text{avg}}(A, \sigma_{\text{sim}}^{2})$ where $\sigma_{\text{sim}}^{2} := \frac{1}{|N_{i}|} \sum_{u \in N_{i}} \| A_{iu} - \mu_{i}^{\text{avg}}(A) \|^{2}$.

In our model, for mathematical rigorousness, we assumed that we have some information of the true position $\mu_{i}$ in advance as specified by the density of $\mu_{i}$ as the spherical Gaussian with mean $\nu_{i}$ and variance $\tau$. Since the exact knowledge of such statistical information might not be easy to obtain in practice, we implement BP with no prior information on true positions by taking the limit of BP as $\tau \to \infty$. Note that our theoretical guarantee on BP still holds in this regime. To obtain a lower bound on the minimum MSE that any estimator can achieve, we use Oracle for computational tractability, instead of the optimal estimator which is computationally intractable while it provides a tighter lower bound.

### 5.2 Synthetic Datasets

We first test synthetic datasets generated by the set of random $(\ell, r)$-regular bipartite graphs, having 200 object detection tasks, where each task $i$ is associated with the true position $\mu_{i}$ chosen uniformly at random in a $100 \times 100$ image. We randomly choose each worker’s variance using $S_{\text{small}} = \{10, 100, 1000\}$ or $S_{\text{large}} = \{10, 100, 5000\}$. The simulation results with varying either $r$ or $\ell$ are plotted in Figures 1a-1b and Figures 1c-1d, respectively, where we take the average of 50 random samples.

**Optimality of BP.** As discussed in Section 3.1, in Figures 1a and 1b we observe that MSE of BP matches with that of Oracle when each worker is assigned just 5 or more tasks. In addition, Figures 1c and 1d show that when $\ell$ increases, MSE of BP decreases at the optimal rate of Oracle and the gap between MSE’s of BP and Oracle is negligible. The other two algorithms decrease at much slow rate but also the MSE differences between them and Oracle increase. For example, in order to make MSE less than 100 with $S_{\text{small}}$, BP and Oracle require only $\ell = 3$, but Simple and Average require $\ell = 4$ and 9, respectively, i.e., Simple and Average need to hire more workers than BP.

**Tolerance to high variance worker.** Comparing Figures 1c and 1d we observe that with the minimum of workers’ variance fixed, for small and large maximum variances of workers, BP sustains good performance, whereas Average performs bad for the large maximum variance. In particular, the performance of Average is extremely degenerated by increasing the worst workers’ variance to 5000 from 1000, while BP is not. This is because BP is able to identify good workers and exploit their answers as the oracle estimator so that BP is tolerant to spammers who have large variances. It is interesting to see that MSE of Simple estimating workers’ variance decreases as $r$ increases, similarly as BP.

### 5.3 Visual Object Classes Datasets

In this section, we provide the experiment results demonstrating the impact of crowdsourced regression on real-world machine learning tasks. To do so, we investigate how much an efficient estimator, refining the crowdsourced training dataset, improves the performance of convolutional neural network (CNN) for the object detection problem. The...
vision task requires a huge amount of the training datasets often obtained by the crowdsourcing system, e.g., Amazon’s mechanical turk [Deng et al. (2009)].

**Emulating a crowdsourcing system.** We use two PASCAL visual object classes (VOC) datasets from Everingham et al. (2015): VOC-07 and VOC-12 consisting of 12,608 and 27,450 annotated objects in 5,011 and 11,540 images, respectively. Each object is annotated by a rectangular bounding box expressed by two opposite corner points. We emulate the crowdsourcing system with a random \((\ell = 3, r = 10)\)-regular bipartite graph between images and virtual workers each of which has variance drawn uniformly at random from support \(S = \{10, 1000\}\). The choice of 10 and 1000 is made our experimental experience on object annotations, as shown in Figure 3. This means that each image is assigned to three workers and each worker is assigned 10 images (\(\approx 24.2\) objects) to estimate all the corner points of the bounding boxes of objects in the set of images. We then gather the noisy estimations on the corner points and run BP, Average, Simple, and Oracle to produce four different crowdsourced training datasets, whose MSE values are presented in Table 1.

**Performance evaluation.** We train a CNN of single shot multibox detector (SSD) 300 \(\times\) 300 model developed in Liu et al. (2016) with the crowdsourced datasets from different estimators, separately. Then we compare the performance of SSD trained with different training datasets in terms of the mean average precision (mAP) which is a popular benchmarking metric for the datasets (see Table 1).

Comparing mAPs of Average and Simple, that of BP is 5% higher in the experiment with VOC-07+12 datasets. Note that achieving a similar amount of improvement is highly challenging, as evidenced in recent extensive research efforts on smarter machine learning algorithms. For example, Faster-RCNN in Ren et al. (2015) is proposed to improve

---

Table 1: Experiment results on the original VOC-07/12 datasets and the crowdsourced datasets by Average, Simple, BP, and Oracle in terms of mean squared errors (MSE) ; mean average precision (mAP) of SSD ; mean portion of overlap of the true positive bounding boxes of SSD with the ground truth bounding boxes.

| DATASET   | ESTIMATOR   | MSE    | MAP    | OVERLAP  |
|-----------|-------------|--------|--------|----------|
| VOC 07+12 | AVERAGE     | 355.6  | 71.80  | 0.741    |
|           | SIMPLE      | 171.1  | 75.17  | 0.763    |
|           | BP          | 109.8  | 75.94  | 0.772    |
|           | ORACLE      | 109.8  | 76.05  | 0.774    |
|           | GROUND TRUTH| 0      | 77.79  | 0.784    |
| VOC 07    | AVERAGE     | 353.1  | 62.70  | 0.711    |
|           | SIMPLE      | 170.6  | 66.383 | 0.726    |
|           | BP          | 111.2  | 67.52  | 0.736    |
|           | ORACLE      | 111.2  | 67.55  | 0.736    |
|           | GROUND TRUTH| 0      | 69.47  | 0.746    |

---

1 SSD shows the state-of-the-art performance on VOC-07/12 datasets and we use the hyper parameters suggested by Liu et al. (2016) (see Appendix A.1 for more details).
the mAP of Fast-RCNN in [Girshick (2015)] from 70.0% to 73.2%. Later, SSD in [Liu et al. (2016)] is proposed to achieve 4% mAP improvement over Faster-RCNN. In addition to the mAP improvement, more accurate training dataset with less MSE leads to more qualified detection with higher overlap ratio. We present how SSD detect objects in Figure 2, where we observe that the training dataset from BP or Oracle enables SSD to not only detect more objects but also draw tighter bounding boxes than Average or Simple.

6 Conclusion

We propose a new probabilistic model to address the problem of aggregating real-valued responses from a crowd of workers, when worker noise levels are heterogeneous. We pose this crowdsourced regression problem as an inference problem over a graphical model, naturally motivating the use of belief propagation. Typically, the performance of a BP algorithm is analytically intractable. However, we bring ideas from a long line of work in BP (e.g. [Mossel et al. (2014)]) to provide sharp analysis on the performance achieved by BP under our model and show its optimality for a broad range of parameters.

A promising research direction with significant practical interest is the question of how to adaptively assign tasks to make more efficient use of the budget. As workers typically arrive in an online fashion, such heuristics are used widely in practice with little theoretical understanding. Efficient and principled schemes have given significant gain in, for instance, voting in social media [Jun et al. (2016)]. There are recent advances for adaptive crowdsourced classification [Ho et al. (2013); Khetan & Oh (2016)], but these approaches rely on the discrete nature of the problem. For crowdsourced regression, it requires innovative ideas to characterize confidence intervals for non binary responses.
References

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

Bollobás, B. Random graphs. Springer, 1998.

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

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

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

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

Girshick, R. Fast r-cnn. In Proceedings of the IEEE ICCV, 2015.

Ho, C., Jabbari, S., and Vaughan, J. W. Adaptive task assignment for crowdsourced classification. In Proceedings of ICML, 2013.

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

Jun, K., Jamieson, K., Nowak, R., and Zhu, X. Top arm identification in multi-armed bandits with batch arm pulls. In Proceedings of AISTATS, 2016.

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.

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

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

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

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

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. SSD: Single shot multibox detector. In Proceedings of ECCV, 2016.

Marcus, A., Karger, D., Madden, S., Miller, R., and Oh, S. Counting with the crowd. In Proceedings of the VLDB Endowment. VLDB Endowment, 2012.

Minka, T. P. Expectation propagation for approximate bayesian inference. In Proceedings of UAI, 2001.

Moallemi, C. C. and Roy, B. Van. Convergence of min-sum message passing for quadratic optimization. IEEE Transactions on Information Theory, 55(5):2413–2423, 2009.
Mossel, E., Neeman, J., and Sly, A. Belief propagation, robust reconstruction and optimal recovery of block models. In *Proceedings of COLT*, 2014.

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

Noorshams, N. and Wainwright, M. J. Belief propagation for continuous state spaces: stochastic message-passing with quantitative guarantees. *Journal of Machine Learning Research*, 14(1):2799–2835, 2013.

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

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.

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

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of NIPS*, 2015.

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

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 KDD*, 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.

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

Wald, Y. and Globerson, A. Tightness results for local consistency relaxations in continuous mrfs. In *Proceedings of UAI*, 2014.

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

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

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.
A Supplementary Material of Theoretical Analysis

A.1 Proof of Lemma 1

Let \( s_\rho \in \{1, \ldots, S\} \) be the index of \( \hat{\sigma}_\rho^2 \), i.e., \( \hat{\sigma}_\rho^2 = \sigma_z^2 \). Consider the classification problem recovering given but latent \( s \) from \( A_{\rho,2k} \) in the following:

\[
\text{minimize} \quad \mathbb{P}[s_\rho \neq \hat{s}_\rho(A_{\rho,2k})]
\]

where the optimal estimator, denoted by \( \hat{s}_\rho^* \), minimizes the classification error rate. By standard Bayesian argument, it is not hard to check that the optimal estimator \( \hat{s}_\rho^* \) is given as follows:

\[
\hat{s}_\rho^*(A_{\rho,2k}) := \arg \max_{s_\rho' = 1, \ldots, S} \mathbb{P}[s_\rho' = s_\rho' | A_{\rho,2k}].
\]

From the above, it is not hard to check that

\[
\min_{s_\rho} \mathbb{P}[s_\rho \neq \hat{s}_\rho(A_{\rho,2k})] = \mathbb{P}[s_\rho \neq \hat{s}_\rho^*(A_{\rho,2k})] = \mathbb{E}_{\mathbb{P}_{\hat{\sigma}^2}} \left[ \mathbb{P}[\sigma_\rho^2 \neq \hat{\sigma}_\rho^2 | A_{\rho,2k}] \right].
\]

Thus an upper bound of the error rate of an arbitrary estimator for \( \hat{s}_\rho(A_{\rho,2k}) \) will provide an upper bound of \( \mathbb{E}_{\mathbb{P}_{\hat{\sigma}^2}} \left[ \mathbb{P}[\sigma_\rho^2 \neq \hat{\sigma}_\rho^2 | A_{\rho,2k}] \right] \). Consider a simple estimator for \( \hat{s}_\rho(A_{\rho,2k}) \), denoted by \( \hat{s}_\rho^\dagger \), which uses only \( A_{\rho,2} \subset A_{\rho,2k} \) as follows:

\[
\hat{s}_\rho^\dagger(A_{\rho,2}) = \arg \min_{s_\rho' = 1, \ldots, S} \left( \sigma_{s_\rho'}^2 + \sigma_{\text{avg}}^2(S) \right) - \hat{\sigma}_\rho^2(A_{\rho,2k})
\]

where we define

\[
\sigma_{\text{avg}}^2(S) := \frac{\sum_{i=1}^S \sigma_i^2}{S(\ell - 1)}, \quad \hat{\sigma}_\rho^2(A_{\rho,2}) := \frac{1}{\ell} \sum_{i \in N_\rho} \hat{\sigma}_\rho^2(A_i), \quad \text{and} \quad \hat{\sigma}_\rho^2(A_i) := \left\| \sum_{u \in M_i \setminus \{\rho\}} \frac{A_{iu}}{\ell - 1} - A_{i\rho} \right\|^2.
\]

From now on, we condition \( \sigma^2_{\hat{\sigma}_\rho^2} \), additionally to \( \sigma^2_\rho \) where \( \partial^2 \rho \) is the set of \( \rho \)'s grandchildren in \( G_{\rho,2} \). For every \( i \in N_\rho \), we define

\[
a_i := \sum_{u \in M_i \setminus \{\rho\}} \frac{\hat{\sigma}_u^2}{(\ell - 1)^2} + \hat{\sigma}_\rho^2, \quad \text{and} \quad Z_i := \sum_{u \in M_i \setminus \{\rho\}} \frac{A_{iu}}{\ell - 1} - A_{i\rho}.
\]

Since the conditional density of \( Z_i \) given \( \sigma^2 = \hat{\sigma}^2 \) is \( \phi(Z_i | 0, a_i) \), the conditional density of \( ||Z_i||^2 / a_i \) is \( \chi^2 \)-distribution with degree of freedom \( \ell \). In addition, it is not hard to check that \( ||Z_i||^2 \) is sub-exponential with parameters \( (2\alpha_i \sqrt{\ell})^2, 2\alpha_i \) such that for all \( |\lambda| < \frac{1}{2\alpha_i} \),

\[
\mathbb{E}_{\hat{\sigma}^2} \left[ \exp \left( \lambda \left( ||Z_i||^2 - da_i \right) \right) \right] = \left( \frac{e^{-a_i \lambda}}{\sqrt{1 - 2a_i \lambda}} \right)^\ell \leq \exp \left( \frac{(2a_i \sqrt{\ell})^2 \lambda^2}{2} \right).\]

Thus it follows that for all \( |\lambda| \leq \min_{i \in N_\rho} \frac{1}{2\alpha_i} \),

\[
\mathbb{E}_{\hat{\sigma}^2} \left[ \exp \left( \lambda \sum_{i \in N_\rho} \left( ||Z_i||^2 - da_i \right) \right) \right] = \prod_{i \in N_\rho} \mathbb{E}_{\hat{\sigma}^2} \left[ \exp \left( \lambda \left( ||Z_i||^2 - da_i \right) \right) \right] \leq \prod_{i \in N_\rho} \exp \left( \frac{(2a_i \sqrt{\ell})^2 \lambda^2}{2} \right).
\]
From this, it is straightforward to check that \( r \hat{\sigma}^2(A_{\rho, 2}) = \sum_{i \in N_{\rho}} \| Z_i \|^2 \) is sub-exponential with parameters \( (6\sigma_{\max}^2 \sqrt{d})^2, 6\sigma_{\max}^2 \) since

\[
0 \leq a_i \leq \sigma_{\max}^2 \left( \frac{\ell + 1}{\ell - 1} \right) \leq 3\sigma_{\max}^2.
\]

Using Bernstein bound, we have

\[
\Pr_{\hat{\sigma}} \left[ \left| \hat{\sigma}^2(A_{\rho, 2}) - \sum_{i \in N_{\rho}} a_i \right| \geq \frac{\varepsilon}{4} \right] \leq 2 \exp \left( -\frac{\varepsilon r}{48\sigma_{\max}^2} \right)
\]

where we let \( \Pr_{\hat{\sigma}} \) denote the conditional probability given \( \sigma^2 = \hat{\sigma}^2 \). Using Hoeffding bound with \( \Pr_{\hat{\sigma}} \), it follows that

\[
\Pr_{\hat{\sigma}} \left[ \left| \sum_{i \in N_{\rho}} a_i - (\sigma_{\text{avg}}^2(S) + \sigma_{\rho}^2) \right| \geq \frac{\varepsilon}{4} \right] \leq 2 \exp \left( -\frac{\varepsilon^2 r}{8\sigma_{\max}^2} \right)
\]

Combining (34) and (35) and using the union bound, it follows that

\[
\Pr_{\hat{\sigma}} \left[ s_{\rho} \neq \hat{s}_{\rho}(A_{\rho, 2}) \right] \geq \Pr_{\hat{\sigma}} \left[ \left| \hat{\sigma}^2(A_{\rho, 2}) - (\sigma_{\text{avg}}^2(S) + \sigma_{\rho}^2) \right| \leq \frac{\varepsilon}{2} \right]
\]

\[
\geq 1 - 2 \left( \exp \left( -\frac{\varepsilon r}{48\sigma_{\max}^2} \right) + \exp \left( -\frac{\varepsilon^2 r}{8\sigma_{\max}^2} \right) \right)
\]

\[
\geq 1 - 4 \exp \left( -\frac{\varepsilon^2 r}{8(8\varepsilon + 1)\sigma_{\max}^2} \right)
\]

where for the first inequality we use \( |\sigma_{s'}^2 - \sigma_{s''}^2| \geq \varepsilon \) for all \( 1 \leq s', s'' \leq S \) such that \( s' \neq s'' \). Hence, noting that \( \hat{s}_{\rho} \) cannot outperform the optimal one \( \hat{s}^* \) in (31), this performance guarantee on \( \hat{s}_{\rho} \) in (36) implies (24) and completes the proof of Lemma 1.

### A.2 Proof of Lemma 2

We start with several notations for convenience. For \( u \in W_{\rho, 2k} \), let \( T_u = (V_u, W_u, E_u) \) be the subtree rooted from \( u \) including all the offsprings of \( u \) in tree \( G_{\rho, 2k} \). Note that \( T_u = G_{\rho, 2k} \). We let \( \partial W_u \subset W_{\rho, 2k} \) denote the subset of worker on the leaves in \( T_u \) and let \( A_u := \{ A_{iv} : (i, v) \in E_u \} \). Since each worker \( u \)’s \( \sigma^2_u \) is a binary random variable, we define a function \( s_u : S \to \{ +1, -1 \} \) for the given \( \hat{\sigma}^2 \) as follows:

\[
s_u(\sigma^2_u) = \begin{cases} 
+1 & \text{if } \sigma^2_u = \hat{\sigma}^2_u \\
-1 & \text{if } \sigma^2_u \neq \hat{\sigma}^2_u
\end{cases}
\]

It is enough to show

\[
\mathbb{E}_{\hat{\sigma}} \left[ \mathbb{P}[s_{\rho}(\sigma^2_{\rho}) = +1 \mid A] - \mathbb{P}[s_{\rho}(\sigma^2_{\rho}) = +1 \mid A_{\rho, 2k}] \right] \leq 2^{-k}.
\]

since for each \( u \in W, \mathbb{P}[\sigma^2_u = \sigma^2_{+}] = \mathbb{P}[\sigma^2_u = \sigma^2_{-}] = \frac{1}{2} \).

To do so, we first define

\[
X_u := 2\mathbb{P}[s_u(\sigma^2_u) = +1 \mid A_u] - 1 , \text{ and} \\
Y_u := 2\mathbb{P}[s_u(\sigma^2_u) = +1 \mid A_u, A_{\rho}] - 1
\]

where we denote \( A_{\rho} := A \setminus A_{\rho} \) so that

\[
|\mathbb{P}[s_{\rho}(\sigma^2_{\rho}) = +1 \mid A] - \mathbb{P}[s_{\rho}(\sigma^2_{\rho}) = +1 \mid A_{\rho, 2k}]| = \frac{1}{2} |X_{\rho} - Y_{\rho}|.
\]

Then Using the above definitions of \( X_u \) and \( Y_u \) and noting \( |X_u - Y_u| \leq 2 \), it is enough to show that for given non-leaf worker \( u \in W_\rho \setminus \partial W_\rho \),

\[
\mathbb{E}_{\hat{\sigma}} \left[ |X_u - Y_u| \right] \leq \frac{1}{2|\partial^2 u|} \sum_{v \in \partial^2 u} \mathbb{E}_{\hat{\sigma}} \left[ |X_v - Y_v| \right]
\]

\[
(38)
\]
where we let $\partial^2 u$ denote the set of grandchildren of $u$ in $T_u$.

To do so, we study certain recursions describing relations among $X$ and $Y$. For notational convenience, we define $g_{iu}^+$ and $g_{iu}^-$ as follows:

$$
g_{iu}^+(X_{\partial_{ui}}; A_i) := \sum_{\sigma_i^2 \in \mathcal{S}_i: \sigma_i^2 = \sigma_u^2} C_i(A_i, \sigma_i^2) \prod_{v \in \partial_{ui}} \left( 1 + s_v(\sigma_v^2) \cdot X_v \right) / 2$$

$$
g_{iu}^-(X_{\partial_{ui}}; A_i) := \sum_{\sigma_i^2 \in \mathcal{S}_i: \sigma_i^2 \neq \sigma_u^2} C_i(A_i, \sigma_i^2) \prod_{v \in \partial_{ui}} \left( 1 + s_v(\sigma_v^2) \cdot X_v \right) / 2.
$$

where we may omit $A_i$ in the argument of $g_{iu}^+$ and $g_{iu}^-$ if $A_i$ is clear from the context. Recalling the factor form of the joint probability of $\sigma^2$ in (9) and using Bayes’ theorem with the fact that $\mathbb{P}[s_u(\sigma_u^2) = +1 \mid A_u] = \frac{1 + X_u}{2}$ and some calculus, it is not hard to check

$$
g_{iu}^+(X_{\partial_{ui}}; A_i) \propto \mathbb{P} \left[ s_u(\sigma_u^2) = +1 \mid A_i, X_{\partial_{ui}} \right]$$

$$
g_{iu}^-(X_{\partial_{ui}}; A_i) \propto \mathbb{P} \left[ s_u(\sigma_u^2) = -1 \mid A_i, X_{\partial_{ui}} \right].$$

From the above, it is straightforward to check that

$$
X_u = h_u(X_{\partial u}) := \frac{\prod_{i \in \partial u} g_{iu}^+(X_{\partial_{ui}}) - \prod_{i \in \partial u} g_{iu}^-(X_{\partial_{ui}})}{\prod_{i \in \partial u} g_{iu}^+(X_{\partial_{ui}}) + \prod_{i \in \partial u} g_{iu}^-(X_{\partial_{ui}})}
$$

where we let $\partial u$ be the task set of all the children of worker $u$ and $\partial_{ui}$ be the worker set of all the children of $i$ in tree $T_u$. Similarly, we also have

$$
Y_u = h_u(Y_{\partial u}).$$

For simplicity, we now pick an arbitrary worker $u \in \mathcal{W}$, which is neither the root nor a leaf, i.e., $u \notin \partial W_\rho$ and $u \neq \rho$, so that $|\partial^2 u| = (\ell - 1) \cdot (r - 1)$. It is easy to show (38) for only $u$. To do so, we will use the mean value theorem. We first obtain a bound on the gradient of $h_u(x)$ for $x \in [-1, 1]^{\partial^2 u}$. Define $g_u^+(x) := \prod_{i \in \partial u} g_{iu}^+(x_{\partial_{ui}})$ and $g_u^-(x) := \prod_{i \in \partial u} g_{iu}^-(x_{\partial_{ui}})$. Using basic calculus, we obtain that for $v \in \partial_{ui}$,

$$
\frac{\partial h_u}{\partial x_v} = \frac{\partial g_u^+ - g_u^-}{g_u^+ + g_u^-} \left( g_u^- \frac{\partial g_u^+}{\partial x_v} - g_u^+ \frac{\partial g_u^-}{\partial x_v} \right)
$$

$$
= \frac{2}{(g_u^+ + g_u^-)^2} \left( g_u^- \frac{\partial g_u^+}{\partial x_v} - g_u^+ \frac{\partial g_u^-}{\partial x_v} \right)
$$

$$
= \frac{2 g_u^-}{g_u^+ + g_u^-} \left( 1 \frac{\partial g_u^+}{\partial x_v} - 1 \frac{\partial g_u^-}{\partial x_v} \right).$$

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

$$
\frac{g_u^+}{g_u^+ + g_u^-} \leq \sqrt{\frac{g_u^-}{g_u^+}}.
$$

(42)

We note here that one can replace $g_u^- / g_u^+$ with $g_u^+ / g_u^-$ in the upper bound. However, in our analysis, we use (42) since we will take the conditional expectation $\mathbb{E}_{\sigma^2}$ which takes the randomness of $A$ generated by the condition $\sigma^2 = \bar{\sigma}^2$. Hence $X_u$ and $Y_u$ will be closer to 1 than $-1$ thus $g_u^- / g_u^+$ will be a tighter upper bound than $g_u^+ / g_u^-$. From (42), it follows that for $x \in [-1, 1]^{\partial^2 u}$ and $v \in \partial_{ui}$,

$$
\left| \frac{\partial h_u}{\partial x_v}(x) \right| \leq \left| g_u^+(x_{\partial_{ui}}) \right| \cdot \prod_{j \in \partial u: j \neq i} \sqrt{\frac{g_{ju}^+(x_{\partial_{uj}})}{g_{ju}^-(x_{\partial_{uj}})}}
$$

17
where we define

\[ g'_{uv}(x_{\partial v}) := 2 \sqrt{\frac{g_u^+(x_{\partial v})}{g_u^-(x_{\partial v})}} \left( \frac{1}{g_u^+(x_{\partial v})} \frac{\partial g_u^+(x_{\partial v})}{\partial x_v} - \frac{1}{g_u^-(x_{\partial v})} \frac{\partial g_u^-(x_{\partial v})}{\partial x_v} \right). \]

Further, we make the bound independent of \( x_{\partial v} \in [-1,1]^{\partial v} \) by taking the maximum of \(|g'_{uv}(x_{\partial v})|\), i.e.,

\[
\left| \frac{\partial h_u}{\partial x_v}(x) \right| \leq \eta_i(A_i) \cdot \prod_{j \in \partial u : j \neq i} \sqrt{\frac{g_{ju}^+(x_{\partial j}; A_j)}{g_{ju}^-(x_{\partial j}; A_j)}} \quad (43)
\]

where we define

\[
\eta_i(A_i) := \max_{x_{\partial v} \in [-1,1]^{\partial v}} g'_{uv}(x_{\partial v}; A_i).
\]

Now we apply the mean value theorem with (43) to bound \(|X_u - Y_u| = |h_u(X_{\partial v u}) - h_u(Y_{\partial v u})|\) by \(|X_v - Y_v|\) of \( v \in \partial^2 u \). It follows that for given \( X_{\partial v u} \) and \( Y_{\partial v u} \), there exists \( \lambda' \in [0,1] \) such that

\[
|X_u - Y_u| = |h_u(X_{\partial v u}) - h_u(Y_{\partial v u})| \leq \sum_{i \in \partial u} \sum_{v \in \partial u} |X_v - Y_v| \cdot \left| \frac{\partial h_u}{\partial x_v}(x) \right| \cdot \prod_{j \in \partial u : j \neq i} \max_{\lambda \in [0,1]} \left\{ \frac{g_{jx}^+(\lambda X_{\partial j} + (1-\lambda)Y_{\partial j}; A_j)}{g_{jx}^-(\lambda X_{\partial j} + (1-\lambda)Y_{\partial j}; A_j)} \right\}. \quad (44)
\]

where for the first and last inequalities, we use the mean value theorem and (43), respectively. We note that each term in an element of the summation in the RHS of (44) is independent to each other. Thus, it follows that

\[
E_{\tilde{\sigma}^2}\left[ |X_u - Y_u| \right] \leq \sum_{i \in \partial u} \sum_{v \in \partial u} E_{\tilde{\sigma}^2}\left[ |X_v - Y_v| \right] \cdot E_{\tilde{\sigma}^2}\left[ \eta_i(A_i) \right] \cdot \prod_{j \in \partial u : j \neq i} \max_{\lambda \in [0,1]} \Gamma_{ju}(\lambda X_{\partial j} + (1-\lambda)Y_{\partial j}) \quad (45)
\]

where we define function \( \Gamma_{iu}(x_{\partial j}; A_i) \) for given \( x_{\partial j} \in [-1,1]^{\partial j} \) as follows:

\[
\Gamma_{iu}(x_{\partial j}) := \sqrt{\frac{g_{iu}^+(x_{\partial j}; A_i)}{g_{iu}^-(x_{\partial j}; A_i)}}.
\]

Note that the assumption on \( \sigma_{min}^2 \) and \( \sigma_{max}^2 \), i.e., \( \sigma_{min}^2 + \varepsilon \leq \sigma_{max}^2 < \frac{5}{2} \sigma_{min}^2 \). This implies

\[
\left( \frac{-1}{\sigma_{max}^2} + \frac{1}{\sigma_{min}^2} \right) \frac{3}{2} - \frac{1}{\sigma_{max}^2} < 0.
\]

Hence, for constant \( \ell \) and \( \varepsilon > 0 \), it is not hard to check that there is a finite constant \( \eta \) with respect to \( r \) such that

\[
\max_{\tilde{\sigma}^2} E_{\tilde{\sigma}^2}\left[ \eta_i(A_i) \right] \leq \eta < \infty \quad (46)
\]

where \( \eta \) may depend on only \( \varepsilon, \sigma_{min}^2 \), and \( \sigma_{max}^2 \).

In addition, we also obtain a bound of the last term of (45), when \( r \) is sufficiently large, in the following lemma whose proof is presented in Section 3.3.

**Lemma 3.** For given \( \tilde{\sigma}^2_{M_i} \in S_{M_i} \) and \( u \in M_i \), let \( \tilde{\sigma}^2_{M_i u} \in S_{M_i} \) be the set of \( \tilde{\sigma}^2_{M_i} \) such that \( \tilde{\sigma}^2_{M_i} \neq \tilde{\sigma}^2_{u} \) and \( \tilde{\sigma}^2_{M_i} = \tilde{\sigma}^2_{v} \) for all \( v \in M_i \setminus \{u\} \). Then, there exists a constant \( C^{\ell,\varepsilon}_{t,\varepsilon} \) such that for any \( r \geq C^{\ell,\varepsilon}_{t,\varepsilon} \),

\[
E_{\tilde{\sigma}^2} \left[ \max_{\lambda \in [0,1]} \Gamma_{iu}(\lambda X_{\partial j} + (1-\lambda)Y_{\partial j}) \right] \leq 1 - \frac{\Delta_{min}^2}{2} < 1,
\]
where we let $\Delta_{\min}$ be the square of the minimum Hellinger distance between the conditional densities of $A_i$ given two different $\sigma_{M_i}^2$ and $\sigma_{M_i}^2$, i.e.,

$$\Delta_{\min} := \min_{\sigma_{M_i}^2, \sigma_{M_i}^2 \in \mathcal{M}_i : \sigma_{M_i}^2 \neq \sigma_{M_i}^2} H^2(f_{A_i|\sigma_{M_i}^2}, f_{A_i|\sigma_{M_i}^2}) > 0.$$ 

Using the above lemma, we can find a sufficiently large constant $C_{\ell,e} \geq C_{\ell,e}$ such that if $|\partial u| = r \geq C_{\ell,e}$,

$$\prod_{j \in \partial u : j \neq i} \mathbb{E}_{\sigma^2} \left[ \max_{\lambda \in [0,1]} \Gamma_{ju} (\lambda X_{\partial u} + (1 - \lambda) Y_{\partial u}) \right] \leq \eta \left( 1 - \psi_{\min} \right) \frac{C_{\ell,e} - 2}{2(\ell - 1)(r - 1)} \leq \frac{1}{2(\ell - 1)(C_{\ell,e} - 1)} \leq \frac{1}{2(\ell - 1)(r - 1)}$$

which implies (39) with (45) and completes the proof of Lemma [2].

### A.3 Proof of Lemma [3]

We first obtain a bound on $X_v$ and $Y_v$ for $v \in \partial_u i$. Noting that $v$ is a non-leaf node in $G_{\nu,2k}$ and $|\partial v| = r - 1$, Lemma [1] directly provides

$$\mathbb{E}_{\sigma^2}[\mathbb{P}[\sigma_v^2 \neq \tilde{\sigma}_v^2 | A_v, 2k]] = \mathbb{E}_{\sigma^2} \left[ \frac{1 - X_v}{2} \right] \leq 4 \exp \left( -\frac{\varepsilon^2}{8(8\varepsilon + 1)\sigma_{\max}^2} \cdot (r - 1) \right).$$

Using Markov inequality for $\frac{1 - X_v}{2} \geq 0$, it is easy to check that for any $\delta > 0$,

$$\mathbb{P}_{\sigma^2} \left[ X_v < 1 - \delta \right] \leq \frac{8}{\delta} \exp \left( -\frac{\varepsilon^2}{8(8\varepsilon + 1)\sigma_{\max}^2} \cdot (r - 1) \right). \tag{47}$$

Note that

$$4 \exp \left( -\frac{\varepsilon^2}{8(8\varepsilon + 1)\sigma_{\max}^2} \cdot (r - 1) \right) \geq \mathbb{E}_{\sigma^2} \left[ \mathbb{P}[\sigma_v^2 \neq \tilde{\sigma}_v^2 | A_v] \right] \geq \mathbb{E}_{\sigma^2} \left[ \mathbb{P}[\sigma_v^2 \neq \tilde{\sigma}_v^2 | A_v, A_{\partial v}] \right] = \mathbb{E}_{\sigma^2} \left[ \frac{1 - Y_v}{2} \right].$$

Hence, we have the same bound in (47) for $Y_v$, i.e.,

$$\mathbb{P}_{\sigma^2} \left[ Y_v < 1 - \delta \right] \leq \frac{8}{\delta} \exp \left( -\frac{\varepsilon^2}{8(8\varepsilon + 1)\sigma_{\max}^2} \cdot (r - 1) \right).$$

Using the assumption that $\sigma_{\min}^2 + \varepsilon \leq \sigma_{\max}^2 < \frac{5}{2} \sigma_{\min}^2$, similarly to (46), we can find finite constants $\eta'$ and $\eta''$ with respect to $r$ such that for all $x \in [0, 1]^{\partial_u i}$,

$$\max_{\sigma^2} \mathbb{E}_{\sigma^2}[|\Gamma_{iu}(x)|] \leq \eta', \quad \text{and} \quad \max_{\sigma^2} \mathbb{E}_{\sigma^2} \left[ \frac{\partial \Gamma_{iu}(x)}{\partial x_v} \right] \leq \eta''.$$  

Then, it follows that for given $\delta > 0$,

$$\mathbb{E}_{\sigma^2} \left[ \max_{\lambda \in [0,1]} \Gamma_{iu} (\lambda X_{\partial u} + (1 - \lambda) Y_{\partial u}) \right] \leq \left( 1 - \mathbb{P}_{\sigma^2} \left[ X_v > 1 - \delta \text{ and } Y_v > 1 - \delta \right] \right) \cdot \max_{x \in [1 - 1/\delta]^{\partial u i}} \mathbb{E}_{\sigma^2} \left[ \Gamma_{iu}(x) \right] + \max_{x \in [1 - 1/\delta]^{\partial u i}} \mathbb{E}_{\sigma^2} \left[ \Gamma_{iu}(x) \right]$$

$$\leq \left( \sum_{v \in \partial u i} \mathbb{P}_{\sigma^2} \left[ X_v \leq 1 - \delta \right] + \mathbb{P}_{\sigma^2} \left[ Y_v \leq 1 - \delta \right] \right) \cdot \max_{x \in [1 - 1/\delta]^{\partial u i}} \mathbb{E}_{\sigma^2} \left[ \Gamma_{iu}(x) \right] + \max_{x \in [1 - 1/\delta]^{\partial u i}} \mathbb{E}_{\sigma^2} \left[ \Gamma_{iu}(x) \right] \tag{48}$$

$$\leq r \cdot \eta' \cdot \frac{8}{\delta} \exp \left( -\frac{\varepsilon^2}{8(8\varepsilon + 1)\sigma_{\max}^2} \cdot (r - 1) \right) + \max_{x \in [1 - 1/\delta]^{\partial u i}} \mathbb{E}_{\sigma^2} \left[ \Gamma_{iu}(x) \right] \tag{49}$$
\[ \leq r \cdot \eta' \cdot \frac{8}{\delta} \exp \left( -\frac{\varepsilon^2}{8(8\varepsilon + 1)\sigma_{\text{max}}^2} \cdot (r - 1) \right) + \delta \eta'' + \mathbb{E}_{\hat{\sigma}^2} [\Gamma_{iu}(1_{\tilde{\sigma}, \tilde{\theta})}] \]  

(50)

where for (48), (49), and (50), we use the union bound, (47), and the mean value theorem, respectively. We will show there exists constant \( \Delta \) such that \( \mathbb{E}_{\hat{\sigma}^2} [\Gamma_{iu}(1_{\tilde{\sigma}, \tilde{\theta})}] \leq 1 - \Delta \), since the first term in (50) is exponentially decreasing with respect to \( r \) thus there exists a constant \( C_{r, \varepsilon} \) such that for \( r \geq C_{r, \varepsilon} \).

\[ \mathbb{E}_{\hat{\sigma}^2} \left[ \max_{\lambda \in [0,1]} \Gamma_{iu}(\lambda X_{\tilde{\theta}, \tilde{\sigma}} + (1 - \lambda) Y_{\tilde{\theta}, \tilde{\sigma}}) \right] \leq 1 - \frac{\Delta}{2} . \]

Recalling the property of \( g_{iu}^+ \) and \( g_{iu}^- \) in (39) and (40), it directly follows that
\[
\mathbb{E}_{\hat{\sigma}^2} [\Gamma_{iu}(1_{\tilde{\sigma}, \tilde{\theta})}] \\
= \int_{\mathbb{R}^{d \times M}} f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}}^2 | \left| \frac{g_{iu}^+(1_{\tilde{\sigma}, \tilde{\theta})} | A_i = x_i}{g_{iu}^+(1_{\tilde{\sigma}, \tilde{\theta})} | A_i = x_i} \right| dx_i \\
= \int_{\mathbb{R}^{d \times M}} f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}}^2 | \left| \frac{f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 \setminus (u) = \hat{\sigma}_{\hat{\sigma}}^2 | \sigma_u^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} {f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} \right| dx_i \\
= \int_{\mathbb{R}^{d \times M}} \sqrt{f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} \left| \frac{f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 \setminus (u) = \hat{\sigma}_{\hat{\sigma}}^2 | \sigma_u^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} {f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} \right| dx_i .
\]

For notational simplicity, we define

\[ \Delta(\hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2, \hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2) := \frac{1}{2} - \frac{1}{2} \int_{\mathbb{R}^{d \times M}} \sqrt{f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} \left| \frac{f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 \setminus (u) = \hat{\sigma}_{\hat{\sigma}}^2 | \sigma_u^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} {f_{A_i} | x_i, \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}}^2 |} \right| dx_i .
\]

Then \( 2 \cdot \Delta(\hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2, \hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2) \) is equal to the square of the Hellinger distance \( H \) between the conditional densities of \( A_i \) given \( \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2 \) and \( \sigma_{\hat{\sigma}, \tilde{M}}^2 = \hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2 \), i.e.,

\[ \Delta(\hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2, \hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2) = H^2(f_{A_i | \sigma_{\hat{\sigma}, \tilde{M}}^2}, f_{A_i | \sigma_{\hat{\sigma}, \tilde{M}}^2}) \geq 0 .
\]

This implies \( \Delta(\hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2, \hat{\sigma}_{\hat{\sigma}, \tilde{M}}^2) > 0 \) and taking the minimum \( \Delta \), we complete the proof of Lemma 3.

### A.4 Proof of Inequality (18)

Noting that \( \mu_i^{BP(k)}(A) \) is the weighted sum of \( \tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) \) as described in (13), we can rewrite \( \| \mu_i^{BP(k)}(A) - \mu_i \|^2 \) as follows:

\[
\| \mu_i^{BP(k)}(A) - \mu_i \|^2 = \sum_{\sigma_{\hat{\sigma}, \tilde{M}}^2, \sigma_{\hat{\sigma}, \tilde{M}}^2} (\tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) - \mu_i)^\top (\tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) - \mu_i) \cdot \prod_{u \in M_i} b_k(u)(\sigma_u^{(2)}) \cdot b_k(u)(\sigma_u^{(2)}) .
\]

Hence, using Cauchy-Schwarz inequality for random variables, it directly follows that

\[
\mathbb{E}_{\hat{\sigma}^2} \left[ \| \mu_i^{BP(k)}(A) - \mu_i \|^2 \right] \leq \mathbb{E}_{\hat{\sigma}^2} \left[ \left( \| \tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) - \mu_i \| \right)^2 \right] + \sum_{\sigma_{\hat{\sigma}, \tilde{M}}^2, \sigma_{\hat{\sigma}, \tilde{M}}^2} \sum_{\sigma_{\hat{\sigma}, \tilde{M}}^2, \sigma_{\hat{\sigma}, \tilde{M}}^2} \sqrt{\mathbb{E}_{\hat{\sigma}^2} \left[ \prod_{u \in M_i} (b_k(u)(\sigma_u^{(2)}) \cdot b_k(u)(\sigma_u^{(2)}))^2 \right]} \times \sqrt{\mathbb{E}_{\hat{\sigma}^2} \left[ (\tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) - \mu_i)^\top (\tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) - \mu_i) \right]}. 
\]

(51)

For any \( \sigma_{\hat{\sigma}, \tilde{M}}^2 \in S^d \), the conditional density of the random vector \( \tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) - \mu_i \) conditioned on \( \sigma^2 = \hat{\sigma}^2 \) is identical to

\[
f_{\tilde{\mu}_i(A_i, \sigma_{\hat{\sigma}, \tilde{M}}^2) - \mu_i | \sigma^2 = \hat{\sigma}^2} = \phi \left( x \mid \frac{(\hat{\sigma}^2(\sigma_{\hat{\sigma}, \tilde{M}}^2))^2}{\frac{1}{\hat{\sigma}^2} + \sum_{u \in M_i} \frac{\hat{\sigma}_{\hat{\sigma}}^2}{\sigma_u^2}} \right).
\]
Using this with some linear algebra and calculus, it is not hard to check that

$$\mathbb{E}_{\sigma^2} \left[ \| \hat{\mu}_i(A, \sigma^2_{M_i}) - \mu_i \|_2 \right] = \frac{d \cdot (\sigma_i^2(\sigma_{M_i}^2))}{\sigma_i^2 + \sum_{u \in M_i} \frac{\sigma_u^2}{\sigma_i^2}} \left( 2 + \frac{d \cdot (\sigma_i^2(\sigma_{M_i}^2))}{\sigma_i^2 + \sum_{u \in M_i} \frac{\sigma_u^2}{\sigma_i^2}} \right)$$

$$\leq \frac{d \sigma_{\text{max}}^2}{\ell^2 \sigma_{\text{min}}^2} \left( 2 + \frac{d \sigma_{\text{max}}^2}{\ell^2 \sigma_{\text{min}}^2} \right) \left( 2 + \frac{d \sigma_{\text{max}}^2}{\ell^2 \sigma_{\text{min}}^2} \right)$$

where for the last inequality, we use the fact that $\sigma_{\text{min}}^2 \leq \sigma_i^2 \leq \sigma_{\text{max}}^2$ for any $1 \leq s \leq S$ and $|M_i| = \ell$. Using the above calculation and Cauchy-Schwarz inequality, it is straightforward to check that for any $\sigma_{M_i}^2, \sigma_{M_i}^2 \in S^{M_i}$,

$$\mathbb{E} \left[ \left( \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right)^\top \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right) \right)^2 \right] \leq 8 \frac{d \sigma_{\text{max}}^2}{\ell^2 \sigma_{\text{min}}^2} \left( 2 + \frac{d \sigma_{\text{max}}^2}{\ell^2 \sigma_{\text{min}}^2} \right). \quad (52)$$

Applying this with the fact that $0 \leq b_k^k(\sigma_u^2) \leq 1$ to (51), it directly follows that

$$\mathbb{E}_{\sigma^2} \left[ \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right)^2 \right] \leq \mathbb{E}_{\sigma^2} \left[ \left( \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right)^\top \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right) \right)^2 \right] + 3S^4 \sqrt{\left( \frac{d \sigma_{\text{max}}^2}{\ell^2 \sigma_{\text{min}}^2} \right) \left( 2 + \frac{d \sigma_{\text{max}}^2}{\ell^2 \sigma_{\text{min}}^2} \right)} \sum_{u \in M_i, \sigma_u \neq \hat{\sigma}_u} \sqrt{\mathbb{E}_{\sigma^2} \left[ b_k^k(\sigma_u^2) \right]}. \quad (53)$$

Note that the conditional density of $X = (\hat{\mu}_i(A, \bar{\sigma}_{M_i}^2) - \mu_i)$ given $\sigma^2 = \hat{\sigma}^2$ is identical to $\phi(X \mid 0, \bar{\sigma}^2_{\bar{M}_i})$.

Thus, it follows that

$$\mathbb{E}_{\sigma^2} \left[ \left( \left( \hat{\mu}_i(A, \bar{\sigma}_{M_i}^2) - \mu_i \right)^\top \left( \hat{\mu}_i(A, \bar{\sigma}_{M_i}^2) - \mu_i \right) \right)^2 \right] = d \mathbb{E}[\sigma_i^2(\bar{\sigma}_{M_i}^2)]$$

which implies (18) with (53).

### A.5 Proof of Inequality (25)

We start with rewriting the difference between MSE’s of $\hat{\mu}_i^*(A)$ and $\hat{\mu}_i^{\text{BP}(k)}(A)$ for $i \in V$ as follows:

$$\| \hat{\mu}_i^*(A) - \mu_i \|_2^2 - \| \hat{\mu}_i^{\text{BP}(k)}(A) - \mu_i \|_2^2$$

$$= \sum_{\sigma_{M_i}^2, \sigma_{M_i}^2 \in S^4} \left( \prod_{u \in M_i} \mathbb{P}[\sigma_u^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} b_k^k(\sigma_u^2) \right) \left( \prod_{u \in M_i} \mathbb{P}[\sigma_u^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} b_k^k(\sigma_u^2) \right) \left( \prod_{u \in M_i} \mathbb{P}[\bar{\sigma}_{M_i}^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} \bar{b}_k^k(\sigma_u^2) \right) \left( \prod_{u \in M_i} \mathbb{P}[\bar{\sigma}_{M_i}^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} \bar{b}_k^k(\sigma_u^2) \right)$$

$$\times \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right)^\top \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right) \left( \hat{\mu}_i(A, \bar{\sigma}_{M_i}^2) - \mu_i \right)^\top \left( \hat{\mu}_i(A, \bar{\sigma}_{M_i}^2) - \mu_i \right)$$

$$\leq \sum_{\sigma_{M_i}^2, \sigma_{M_i}^2 \in S^4} 2 \left( \prod_{u \in M_i} \mathbb{P}[\sigma_u^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} b_k^k(\sigma_u^2) \right) \left( \prod_{u \in M_i} \mathbb{P}[\bar{\sigma}_{M_i}^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} \bar{b}_k^k(\sigma_u^2) \right) \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right)^\top \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right)$$

Then, using Cauchy-Schwarz inequality for random variables $X$ and $Y$, i.e., $|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2] \mathbb{E}[Y^2]}$, we have

$$\mathbb{E} \left[ \left| \text{MSE}(\hat{\mu}_i^*(A)) - \text{MSE}(\hat{\mu}_i^{\text{BP}(k)}(A)) \right| \right]$$

$$\leq \sum_{\sigma_{M_i}^2, \sigma_{M_i}^2 \in S^4} 2 \sqrt{\mathbb{E}\left[ \left( \prod_{u \in M_i} \mathbb{P}[\sigma_u^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} b_k^k(\sigma_u^2) \right)^2 \right]} \sqrt{\mathbb{E}\left[ \left( \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right)^\top \left( \hat{\mu}_i(A, \sigma_{M_i}^2) - \mu_i \right) \right)^2 \right]}.$$

(54)

Note that a simple calculus shows that

$$\mathbb{E}\left[ \left( \prod_{u \in M_i} \mathbb{P}[\sigma_u^2 = \sigma_{M_i}^2] - \prod_{u \in M_i} b_k^k(\sigma_u^2) \right)^2 \right] \leq 2 \sum_{u \in M_i} \mathbb{E}\left[ \left| \mathbb{P}[\sigma_u^2 = \sigma_{M_i}^2] - b_k^k(\sigma_u^2) \right| \right].$$

(55)

Using (52) and (55) to (54), we complete the proof of (25).
B Supplementary Material of Experiment Result

B.1 Worker’s Noisy Annotation

To give an intuition on the choice of $S = \{10, 1000\}$ in our experiment, we first note that the estimation of a worker $u$ with $\sigma_u^2 = 1000$ on a task is concentrated on the disk of radius 50 centered at the true position with probability more than 0.7, where the average size of images and bound boxes in VOC-07/12 are $359.5 \times 496.2$ and $113.5 \times 182.6$, respectively. We also provide few examples of the worker $u$’s noisy annotations with $\sigma_u^2 = 10$ and $\sigma_u^2 = 1000$ in Figures 4a and 4b respectively.

![Examples of object annotations by a worker $u$ with $\sigma_u^2 = 10$ or 1000.](image)

**Figure 4:** Examples of object annotations by a worker $u$ with $\sigma_u^2 = 10$ or 1000.

B.2 Hyper Parameter Settings of Single Shot Multibox Detector

We use the following hyper parameter settings. For VOC-07+12, we trained 120,000 iterations with initial learning rate $4 \times 10^{-5}$ and decrease it by factor of 0.1 at iteration 80,000 and 100,000 as [Liu et al.](2016) suggested. For VOC-07, we trained model 60,000 iterations with initial learning rate $10^{-5}$ and reduce it by factor of 0.1 at iteration 40,000. The other hyper parameter setting is equivalent to suggested in [Liu et al.](2016).