

---

# Robust Deep Learning from Crowds with Belief Propagation

---

Hoyoung Kim\*

Seunghyuk Cho\*

Dongwoo Kim

Jungseul Ok<sup>†</sup>

Pohang University of Science and Technology

## Abstract

Crowdsourcing systems enable us to collect large-scale dataset, but inherently suffer from noisy labels of low-paid workers. We address the inference and learning problems using such a crowdsourced dataset with noise. Due to the nature of sparsity in crowdsourcing, it is critical to exploit both probabilistic model to capture worker prior and neural network to extract task feature despite risks from wrong prior and overfitted feature in practice. We hence establish a neural-powered Bayesian framework, from which we devise deepMF and deepBP with different choice of variational approximation methods, mean field (MF) and belief propagation (BP), respectively. This provides a unified view of existing methods, which are special cases of deepMF with different priors. In addition, our empirical study suggests that deepBP is a new approach, which is more robust against wrong prior, feature overfitting and extreme workers thanks to the more sophisticated BP than MF.

## 1 Introduction

Crowdsourcing systems, such as Amazon Mechanical Turk, enable us to collect huge labeled datasets at low budget and in short time by distributing labeling tasks over the crowd workers. However, low-paid workers are liable to provide noisy labels, and even trustworthy workers, called hammers, have non-zero probability to make mistakes. In addition, there often exist spammers randomly labeling and adversaries incorrectly labeling due to misinterpretation of task

description or malicious intention (Jagabathula et al., 2017). Therefore, in order to fully utilize the crowd-sourced dataset from such various workers for either inferring true labels, i.e., *inference from crowds* (Dawid and Skene, 1979), or training a model to perform the same labeling task, i.e., *learning from crowds* (Raykar et al., 2010), it is a fundamental problem to jointly estimate worker abilities and true labels. Indeed, with the presence of diverse workers, majority voting (MV) giving the same weight to each worker often fails at recovering true labels due to spammers or adversaries, whereas a weighted MV giving higher weight to more reliable workers would not if the worker estimation is accurate.

A principled approach for this problem is to establish a probabilistic generative model on the behaviors of worker labeling and apply a standard method for inference or learning. Dawid and Skene (1979) propose a pioneering generative model where each worker is associated with confusion matrix, and an expectation-maximization (EM) algorithm to recursively infer true labels and estimate confusion matrices. Since then, early works (Whitehill et al., 2009; Raykar et al., 2010; Welinder et al., 2010; Liu et al., 2012) have developed sophisticated methods for inference and learning methods to be capable of exploiting additional information such as a task feature or worker prior. Raykar et al. (2010) propose a framework to conduct inference and learning from crowds in an iterative manner, where the learning part enables us to utilize task features. Liu et al. (2012) establish a flexible Bayesian approach that allows us to plug-in any worker prior distribution.

Meanwhile, with the recent advances in deep learning, there have been proposed a number of methods to train neural network model directly from crowd-sourced dataset (Rodrigues and Pereira, 2018; Tanno et al., 2019). However, it is well known that such methods often suffer from unstable learning mainly since it is inevitable to use local search methods such as a gradient-based optimizer in deep learning. The instability issue becomes severe in canonical scenarios that the number of workers per task is limited by budget

---

Proceedings of the 25<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s). \*Equal contribution. <sup>†</sup>Corresponding author.

constraints, and each worker has a limited capacity to perform trustworthy labeling, i.e., the information in crowdsourced dataset is *large but sparse*.<sup>1</sup> To mitigate this issue, Tanno et al. (2019) introduces a specific regularization term based on the model of worker labeling. However, we empirically found that Tanno et al. (2019)’s approach is sensitive to the initialization or hyperparameter of deep learning<sup>2</sup>, in particular, when their worker model is mismatched to the real perhaps with outliers.

In this paper, we hence aim at *robust* deep learning from crowds. To do so, we first establish a generative model containing a neural network and worker prior to which we easily insert any worker prior distribution. This provides not only a flexibility but also plausible interpretation to the choice of worker prior. Using our model, we then devise an EM framework alternating variational inference, which is used to infer true label on the probabilistic model, and deep learning, which is used to update the probabilistic model. In Section 3.1, adopting mean-field (MF) approximation for the variational inference, we devise *deepMF* as a simple procedure of minimizing a loss function. We also show that the existing methods can be interpreted as an instance of deepMF with a specific choice of worker prior in Section 3.2.

Although MF-based approaches are popular thanks to its simplicity (Rodrigues and Pereira, 2018; Tanno et al., 2019), belief-propagation (BP) is known to be a better approximation method in general (Weiss, 2001; Murphy, 2013). To be specific, we note that the crowdsourced dataset can be interpreted as a bipartite weighted graph consisting of edges between tasks and workers with the corresponding worker labels as weights. If the graph is tree, the inference based on BP is exact on the graph (Pearl, 1982). Furthermore, Ok et al. (2016) prove the asymptotic optimality of BP in the canonical scenario where the assignment graph is a random bipartite graph with limited degrees so that locally tree-like with high probability. Intuitively, to infer a task label, BP is able to fully utilize the information spread over the local tree rooted from the task in a non-backtracking manner (Yedidia et al., 2003), while MF approximation is not precise at this level. To take the advantages of BP, in Section 3.3, we propose *deepBP* using BP for the variational inference.

As intended in the design of deepBP, it empirically outperforms deepMF and the other algorithms. In particular, the advantage of using deepBP is robustness

in the set of canonical crowdsourcing scenarios. We consider the following three specific scenarios, where any workers or practitioners would face in the real world. First, workers are commonly faced with non-informative or even irrelevant task features in a given task. Consequently, the features become obstacles in learning. Second, practitioners are likely to choose a wrong worker prior for inference and learning, or the model would not work properly with even true prior. Third, malicious workers are everywhere, especially the ones who submit a lot of random answers. All the scenarios are simulated with synthetic data. In all cases, deepBP performs robustly than the other algorithms, although some perform comparably in a particular case but not for all. In an additional experiment on a real-world dataset augmented with spammers and non-informative features, deepBP still maintains its robustness.

**Contribution.** In this work, to robustify deep learning from crowds, we propose a principled framework alternating variational inference and deep learning to utilize both the prior of worker behavior and the features of tasks. Our contributions are summarized in three folds. First, we devise deepMF in a simple form of deep learning and reveal that the previous methods (Rodrigues and Pereira, 2018; Tanno et al., 2019) are special cases of deepMF with specific choices of worker prior. This provides a useful guideline to select algorithms with a plausible interpretation. Second, inspired by the strong theoretical guarantee on BP for inference from crowds (Pearl, 1982; Ok et al., 2016), we propose deepBP, to our best knowledge, which is the first attempt to use BP for deep learning from crowds. Third, we empirically show that deepBP overall outperforms deepMF and the other existing methods. The advantages of deepBP in terms of robustness is clearer particularly in the set of canonical crowdsourcing scenarios with i) non-informative features; ii) multi-modal or mismatched worker prior; or iii) extreme spammers who submit noises to several tasks.

**Related Work.** As mentioned earlier, sparse dataset is common in crowdsourcing system due to limited budget, but it is challenging by the risk of overfitting. As a part of robustifying deep learning on the sparse regime, our approach is along with the Bayesian view (Liu et al., 2012) in the sense that the knowledge on workers is given as a prior distribution of workers’ confusion matrices. Meanwhile, there are diverse ways to express and exploit the worker prior. Assuming that workers can be clustered into few different types with similar confusion matrix, Venanzi et al. (2014) propose to involve a worker clustering mechanism. In (Nguyen et al., 2017), the confusion matrix is converted into

<sup>1</sup>In Appendix A, we empirically show that given robust algorithms and budget limit, it is better to obtain large but sparse dataset than small but dense one in terms of the performance in learning.

<sup>2</sup>More details are in Appendix D.

a confusion vector to stabilize learning by dimensionality reduction. Interestingly, Chu and Wang (2021) introduce the generative adversarial networks to learn worker behavior. We believe that our approach provides ease of adopting different worker prior through the lens of the probabilistic perspective.

As deepBP is a deep learning method equipped with BP, it is worth to mention studies to take advantages of BP and deep learning for other learning problems. To capture high order dependencies on factor graphs, Zhang et al. (2020) propose a factor graph neural network which can mimic the max-product BP. Kuck et al. (2020) generalize BP with neural network to find better fixed posteriors faster than loopy BP. Recently, Satorras and Welling (2021) propose a hybrid model which runs conjointly a graph neural network with BP. The above methods are designed for flexible or fast inference by training neural network from multiple instances of inference on a *fixed* factor graph. However, in crowdsourcing systems, the factor graph varies easily per each instance. We hence need to devise a new framework alternating inference and learning.

## 2 Model

We consider a classification model to predict latent label  $z \in [K] := \{1, 2, \dots, K\}$  from feature  $x$  such as an image or audio track. We model the probability of class  $k$  given feature  $x$  using an arbitrary function  $f_\phi(k; x)$  parameterized by  $\phi$ , i.e.,

$$p(z = k \mid x, \phi) = f_\phi(k; x), \quad (1)$$

which can be logistic regression model as in (Raykar et al., 2010), or deep neural network as in (Rodrigues and Pereira, 2018).

To collect training dataset, we consider a crowdsourcing model consisting of  $N$  classification tasks and  $M$  workers. Each task  $i \in [N]$  is associated with true label  $z_i \in [K]$  and feature  $x_i$ . We assume each task is sampled independently, i.e.,  $p(\mathbf{z} \mid \mathbf{x}, \phi) = \prod_{i \in [N]} p(z_i \mid x_i, \phi)$ , where  $\mathbf{z}$  and  $\mathbf{x}$  are the sets of all  $z_i$ 's and  $x_i$ 's, respectively. Each worker  $u \in [M]$  is associated with confusion matrix  $\theta^{(u)} \in [0, 1]^{K \times K}$ , parameterizing worker  $u$ 's average ability, such that  $\theta_{kk'}^{(u)}$  is the probability of answering  $k'$  for tasks with true label  $k$ :

$$p(y_i^{(u)} = k' \mid z_i = k, \theta^{(u)}) = \theta_{kk'}^{(u)}, \quad (2)$$

where  $y_i^{(u)} \in [K]$  is worker  $u$ 's answer for task  $i$ , and  $\sum_{k' \in [K]} \theta_{kk'}^{(u)} = 1$ .

We assume that each confusion matrix  $\theta^{(u)}$  is drawn independently from prior distribution  $p(\theta^{(u)} \mid \alpha)$  with parameter  $\alpha$ , i.e.,  $p(\theta \mid \alpha) = \prod_{u \in [M]} p(\theta^{(u)} \mid \alpha)$ ,

where  $\theta$  is the set of all  $\theta^{(u)}$ 's. The worker prior  $p(\theta^{(u)} \mid \alpha)$  is often equipped with Dirichlet distribution  $\text{Dir}(\alpha)$  since it provides an analytical tractability from the fact that it is a conjugate prior of the categorical distribution, and also represents a wide set of distributions by simply manipulating  $\alpha$ . For instance, in (Liu et al., 2012), the one-coin model for binary classification uses the worker prior such that for  $\alpha_1, \alpha_2 > 0$ ,

$$(\theta_{11}^{(u)}, \theta_{12}^{(u)}) = (\theta_{22}^{(u)}, \theta_{21}^{(u)}) \sim_{\text{i.i.d.}} \text{Dir}(\alpha_1, \alpha_2), \quad (3)$$

where the expected probability of being correct is  $\frac{\alpha_1}{\alpha_1 + \alpha_2}$ . The two coin model in (Liu et al., 2012) independently draws  $(\theta_{11}^{(u)}, \theta_{12}^{(u)})$  and  $(\theta_{22}^{(u)}, \theta_{21}^{(u)})$  from the same distribution.

We assume that  $y_i^{(u)}$ 's are conditionally independent to each other given  $\mathbf{z}$  and  $\theta$ . Let  $\mathbf{y}$  be the set of all  $y_i^{(u)}$ 's, and  $\mathbb{N}_u$  be the set of tasks labeled by worker  $u$ . The model described above can be summarized as:

$$\begin{aligned} p(\mathbf{y}, \mathbf{z}, \theta \mid \mathbf{x}, \alpha, \phi) &= p(\mathbf{z} \mid \mathbf{x}, \phi) p(\mathbf{y} \mid \mathbf{z}, \theta) p(\theta \mid \alpha) \\ &= \prod_{i \in [N]} f_\phi(z_i; x_i) \left( \prod_{u \in [M]} p(\theta^{(u)} \mid \alpha) \prod_{j \in \mathbb{N}_u} \theta_{z_j, y_j^{(u)}}^{(u)} \right). \end{aligned} \quad (4)$$

Given this model equipped with task feature  $f_\phi(z_i; x_i)$  and worker prior  $p(\theta^{(u)} \mid \alpha)$ , in what follows, we describe two fundamental problems: inference and learning from crowds.

**Inference from crowds.** Given crowdsourced dataset  $(\mathbf{x}, \mathbf{y})$ , worker prior  $\alpha$ , and classifier  $f_\phi$  based on task feature, the Bayes decision rule is given as:

$$\hat{z}_i = \arg \max_{z_i \in [K]} p(z_i \mid \mathbf{x}, \mathbf{y}, \alpha, \phi), \quad (5)$$

which is optimal in the sense of minimizing the expected bit-wise error rate. Noting  $p(\mathbf{z}, \theta \mid \mathbf{x}, \mathbf{y}, \alpha, \phi) \propto p(\mathbf{z} \mid \mathbf{x}, \phi) p(\mathbf{y} \mid \mathbf{z}, \theta) p(\theta \mid \alpha)$ , the marginal probability in (5) can be computed as follows:

$$\begin{aligned} p(z_i \mid \mathbf{x}, \mathbf{y}, \alpha, \phi) &= \sum_{\mathbf{z}_{-i}} \int p(\mathbf{z}, \theta \mid \mathbf{x}, \mathbf{y}, \alpha, \phi) d\theta \\ &\propto \sum_{\mathbf{z}_{-i}} p(\mathbf{z} \mid \mathbf{x}, \phi) \int p(\mathbf{y} \mid \mathbf{z}, \theta) p(\theta \mid \alpha) d\theta, \end{aligned} \quad (6)$$

where  $\mathbf{z}_{-i} := (z_j : i \neq j \in [N])$ . In general, the marginalization is computationally intractable since the summation in (6) takes over exponentially many  $\mathbf{z}_{-i} \in [K]^{N-1}$ . To approximate such an intractable marginalization, one can apply a variational method such as MF or BP.

---

**Algorithm 1** deepMF( $\mathbf{x}, \mathbf{y}, \phi, \alpha, \theta, \beta, c$ )
 

---

1: **while** not converged **do**  
 2:    $\hat{f}_\phi(\mathbf{x}) \leftarrow \text{Clip}(f_\phi(\mathbf{x}), c)$   
 3:    $q(\mathbf{z}) \leftarrow \text{MF}(\hat{f}_\phi(\mathbf{x}), \mathbf{y}, \alpha, \theta, \beta)$   
 4:    $\phi, \beta \leftarrow \arg \max_{\phi, \beta} \mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \theta, \beta)$   
 5: **return**  $q(\mathbf{z}), \phi$

---

**Learning from crowds.** Given crowdsourced dataset  $(\mathbf{x}, \mathbf{y})$  and worker prior  $\alpha$ , we can formulate the problem of learning classifier  $f_\phi$  as the maximization of the posterior:

$$\hat{\phi} = \arg \max_{\phi} p(\phi | \mathbf{x}, \mathbf{y}, \alpha). \quad (7)$$

The maximum a posterior (MAP) estimate  $\hat{\phi}$  can be used for not only the optimal inference in (5), but also for predicting label  $z_*$  of unseen  $x_*$  as follows:

$$\hat{z}_* = \arg \max_{k \in [K]} f_\phi(z_* = k; x_*). \quad (8)$$

Using the Bayes rule, the posterior is written as:

$$p(\phi | \mathbf{x}, \mathbf{y}, \alpha) \propto p(\mathbf{y} | \mathbf{x}, \alpha, \phi) p(\phi), \quad (9)$$

where if the prior  $p(\phi)$  is assumed to be zero-mean Gaussian, then it will be translated into the negative of L2-norm regularizer of  $\phi$  in (7). Hence, to obtain the MAP estimate in (7), we write the likelihood  $p(\mathbf{y} | \mathbf{x}, \alpha, \phi)$  as follows:

$$p(\mathbf{y} | \mathbf{x}, \alpha, \phi) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} | \mathbf{x}, \alpha, \phi) \quad (10)$$

$$= \sum_{\mathbf{z}} \int p(\mathbf{y}, \mathbf{z}, \theta | \mathbf{x}, \alpha, \phi) d\theta, \quad (11)$$

where analog to (6), in general, it is intractable to marginalizing out  $\mathbf{z}$  due to the exponentially many summations over  $\mathbf{z} \in [K]^N$ .

### 3 Method

We propose deepMF and deepBP, each of which is essentially an expectation-maximization (EM) algorithm consisting of iterations of E- and M-steps addressing the intractable marginalization issues for the inference (5) and the learning (7), respectively, with different variational Bayesian approaches. To be specific, E-step estimates  $q_i(z_i)$  to approximate  $p(z_i | \mathbf{x}, \mathbf{y}, \alpha, \phi)$  in (6), where deepMF and deepBP use MF approximation and BP algorithm, respectively. In M-step,

---

**Algorithm 2** deepBP( $\mathbf{x}, \mathbf{y}, \phi, \alpha, c$ )
 

---

1: **while** not converged **do**  
 2:    $\hat{f}_\phi(\mathbf{x}) \leftarrow \text{Clip}(f_\phi(\mathbf{x}), c)$     $\triangleright$  Clipping  
 3:    $q(\mathbf{z}) \leftarrow \text{BP}(\hat{f}_\phi(\mathbf{x}), \mathbf{y}, \alpha)$     $\triangleright$  Inference  
 4:    $\phi \leftarrow \arg \max_{\phi} \mathcal{L}_{\text{BP}}(q(\mathbf{z}); \phi)$     $\triangleright$  Learning  
 5: **return**  $q(\mathbf{z}), \phi$

---

given  $q(\mathbf{z})$  from E-step, classifier  $f_\phi$  is trained to maximize an evidence lower bound (ELBO) of the likelihood  $p(\mathbf{y} | \mathbf{x}, \alpha, \phi)$ , derived from (10) and (11) for deepMF and deepBP, correspondingly. In what follows, we first present formal descriptions of deepMF (Section 3.1) and deepBP (Section 3.3), and then in Section 3.4, we provide useful remarks.

#### 3.1 Deep Mean-Field

We derive deepMF from an ELBO of the marginalization in (11) by introducing a variational distribution  $q(\mathbf{z}, \theta)$  in the followings:

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{x}, \alpha, \phi) &= \log \mathbb{E}_{q(\mathbf{z}, \theta)} \left[ \frac{p(\mathbf{y}, \mathbf{z}, \theta | \mathbf{x}, \alpha, \phi)}{q(\mathbf{z}, \theta)} \right] \\ &\geq \mathbb{E}_{q(\mathbf{z}, \theta)} \left[ \log \frac{p(\mathbf{y}, \mathbf{z}, \theta | \mathbf{x}, \alpha, \phi)}{q(\mathbf{z}, \theta)} \right]. \end{aligned} \quad (12)$$

Note that the lower bound is maximized and achieves the equality if  $q(\mathbf{z}, \theta) = p(\mathbf{z}, \theta | \mathbf{x}, \mathbf{y}, \alpha, \phi)$ . We apply MF approximation, assuming a conditional independence of  $z_i$ 's and  $\theta^{(u)}$ 's given  $(\mathbf{x}, \mathbf{y}, \alpha, \phi)$ , such that:

$$q(\mathbf{z}, \theta) = q(\mathbf{z}) q(\theta; \beta) = \prod_{i \in [N]} q_i(z_i) \prod_{u \in [M]} q_u(\theta^{(u)}; \beta^{(u)}), \quad (13)$$

where we use a parametric estimation  $q_u(\theta^{(u)}; \beta^{(u)}) = \text{Dir}(\theta^{(u)}; \beta^{(u)})$  for each worker  $u$ , and approximating  $q_i(z_i) \approx p(z_i | \mathbf{x}, \mathbf{y}, \alpha, \phi)$  and  $q_u(\theta^{(u)}; \beta^{(u)}) \approx p(\theta^{(u)} | \mathbf{x}, \mathbf{y}, \alpha, \phi)$ , we let  $q(\mathbf{z}) = \prod_{i \in [N]} q_i(z_i)$  and  $q(\theta) = \prod_{u \in [M]} q_u(\theta^{(u)})$ . Then, using (4) and (13), the ELBO in (12) is written as follows:

$$\begin{aligned} \mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \theta, \beta) &:= \mathbb{E}_{q(\mathbf{z})q(\theta; \beta)} [\log p(\mathbf{y} | \mathbf{z}, \theta)] \\ &\quad - D_{\text{KL}}(q(\mathbf{z}) || p(\mathbf{z} | \mathbf{x}, \phi)) - D_{\text{KL}}(q(\theta; \beta) || p(\theta | \alpha)). \end{aligned} \quad (14)$$

In each iteration of deepMF, we seek  $q(\mathbf{z})$ ,  $\beta$  and  $\phi$ , sequentially, to maximize  $\mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \theta, \beta)$  by fixing the others. Recalling  $p(\mathbf{z} | \mathbf{x}, \phi) := \prod_{i \in [N]} f_\phi(z_i; x_i)$ , given the classifier  $f_\phi$  and the other variational distributions, the ELBO is maximized at  $q_i(z_i)$  such that:

$$\begin{aligned} &\log q_i(z_i) \quad (15) \\ &= \sum_{u \in \mathbb{M}_i} \mathbb{E}_{q_u(\theta^{(u)}; \beta^{(u)})} [\log \theta_{z_i, y_i^{(u)}}^{(u)}] + \log f_\phi(z_i; x_i) - 1, \end{aligned}$$

Table 1: *Comparison between deepMF and the previous methods: CL (Rodrigues and Pereira, 2018), Trace (Tanno et al., 2019), BayesDGC (Li et al., 2021).* **Prior** indicates the use of worker specific prior in each model. **Variational** indicates the use of variational methods in the parameter estimation. **Analytic  $\beta$**  shows the presence of an analytic solution for the point-wise estimation of  $\beta$ .

	Prior	Variational	Analytic $\beta$
CL	×	×	×
Trace	△	×	×
BayesDGC	○	○	×
deepMF (ours)	○	○	○

where  $\mathbb{M}_i$  denotes the workers who label task  $i$ . The collection is denoted by  $\text{MF}(f_\phi(\mathbf{x}), \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta})$ . The inference of deepMF is based on  $q_i(z_i)$  from  $\text{MF}(\cdot)$ .

Then, provided  $q(\mathbf{z})$  from  $\text{MF}(\cdot)$ , the ELBO is maximized at  $q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta})$  such that:

$$\begin{aligned} \log q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}) & \quad (16) \\ &= \sum_{i \in \mathbb{N}_u} \mathbb{E}_{q(z_i)} \left[ \log p(y_i^{(u)} | z_i, \boldsymbol{\theta}^{(u)}) \right] + \log p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) - 1. \end{aligned}$$

When Dirichlet distribution is used for worker prior, i.e.,  $p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) := \text{Dir}(\boldsymbol{\theta}^{(u)}; \boldsymbol{\alpha})$ , the parameter  $\boldsymbol{\beta}$  is updated as:

$$\beta_{k_1 k_2}^{(u)} = \alpha_{k_1 k_2} + \sum_{\{i | i \in \mathbb{N}_u, y_i^{(u)} = k_2\}} q_i(z_i = k_1), \quad (17)$$

for all  $k_1, k_2 \in [K]$ . Finally, deepMF finds  $\phi$  independently of  $\boldsymbol{\beta}$  by maximizing the ELBO in (14):

$$\hat{\phi} = \arg \max_{\phi} \mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \boldsymbol{\theta}, \boldsymbol{\beta}), \quad (18)$$

which corresponds to training classifier  $f_\phi$ .

We note that deepMF is easily implementable as an iteration of the inference (15) and learning (18) is just a sequence of maximizations for the same objective (14) but different variables. Hence, it is a popular framework despite the coarse approximation in (13). The overall procedure of deepMF is summarized in Algorithm 1, and the detailed equations of deepMF can be found in Appendix B.

### 3.2 Connection to existing methods.

Rodrigues and Pereira (2018) and Tanno et al. (2019) are specific instances of deepMF with different choices of variational distributions and worker prior  $\boldsymbol{\alpha}$  in (14). Both Rodrigues and Pereira (2018) and Tanno et al.

(2019) use a neural network parameterized by  $\phi$  as a variational distribution to approximate the posterior of the true label and the Dirac delta function to confusion matrix  $q(\boldsymbol{\theta})$ , i.e., point-wise estimation on the confusion matrix. With the variational distributions, the ELBO in (14) is simplified as:

$$\mathcal{L}(\mathbf{z}; \phi, \boldsymbol{\theta}) := \mathbb{E}_{p(\mathbf{z} | \mathbf{x}, \phi)} [\log p(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta})] + \log p(\boldsymbol{\theta} | \boldsymbol{\alpha}). \quad (19)$$

Rodrigues and Pereira (2018) further assumes that the ability of workers follows the uniform distribution by fixing all  $\alpha_i$  to 1. Separately, Tanno et al. (2019) assumes the presence of more adversaries than hammers using an adversarial prior where  $\alpha_{k_1 k_2}$  less than one when  $k_1 = k_2$  and one otherwise. In this work, we do not predetermine the hyperparameter, and instead, we show the robustness of the proposed framework to the choice of hyperparameters. Furthermore, deepMF approximates the distribution over the confusion matrix by placing a Dirichlet variational distribution.

BayesDGC (Li et al., 2021) combines deep neural networks and probabilistic graphical model and is the most similar to deepMF. BayesDGC uses a natural-gradient stochastic variational inference combining variational message passing and stochastic gradient descent. Unlike BayesDGC, we derive a closed form update for parameter  $\boldsymbol{\beta}$ . The comparison between our methods and previous approaches is summarized in Table 1 and Appendix C

### 3.3 Deep Belief-Propagation

To derive deepBP, we start with an ELBO for the alternative marginalization in (10) using a variational distribution  $q(\mathbf{z})$  as follows:

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{x}, \boldsymbol{\alpha}, \phi) &= \log \mathbb{E}_{q(\mathbf{z})} \left[ \frac{p(\mathbf{y}, \mathbf{z} | \mathbf{x}, \boldsymbol{\alpha}, \phi)}{q(\mathbf{z})} \right] \\ &\geq \mathbb{E}_{q(\mathbf{z})} \left[ \log \frac{p(\mathbf{y}, \mathbf{z} | \mathbf{x}, \boldsymbol{\alpha}, \phi)}{q(\mathbf{z})} \right], \end{aligned} \quad (20)$$

where the lower bound is maximized with the equality when  $q(\mathbf{z}) = p(\mathbf{z} | \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \phi)$ . In deepBP, we again recursively update  $q(\mathbf{z})$  and  $\phi$ , but employ BP to compute  $q(\mathbf{z})$ . Unlike the MF approximation, it is known that BP computes the exact posterior when tree structure of the assignment graph is given (Pearl, 1982). Furthermore, without the features  $\mathbf{x}$ , the inference with BP is exactly optimal under some mild assumptions (Ok et al., 2016).

We first describe the use of BP for  $q(\mathbf{z})$ . To do so, using (4), we correspond the complete-data likelihood

in (20) to the following factor graph form:

$$\begin{aligned}
 p(\mathbf{y}, \mathbf{z} \mid \mathbf{x}, \boldsymbol{\alpha}, \phi) &= \int p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} \mid \mathbf{x}, \boldsymbol{\alpha}, \phi) d\boldsymbol{\theta} \quad (21) \\
 &\propto \prod_{i \in [N]} \underbrace{f_\phi(z_i; x_i)}_{=: h_i(z_i; \phi)} \prod_{u \in [M]} \underbrace{\int p(\boldsymbol{\theta}^{(u)} \mid \boldsymbol{\alpha}) \prod_{j \in \mathbb{N}_u} \theta_{z_j, y_j}^{(u)} d\boldsymbol{\theta}^{(u)}}_{=: g_u(\mathbf{z}_{\mathbb{N}_u}; \boldsymbol{\alpha})},
 \end{aligned}$$

where we let  $h_i(z_i; \phi)$  be the task  $i$ 's factor containing feature information and  $g_u(\mathbf{z}_{\mathbb{N}_u}; \boldsymbol{\alpha})$  be the worker  $u$ 's factor including worker prior. Given  $f_\phi$ , the sum-product BP on the factor graph can be written as the following iterative updates of three types of messages:

$$\begin{aligned}
 m_{i \rightarrow g_u}^{t+1}(z_i) &\propto m_{h_i \rightarrow i}(z_i) \prod_{v \in \mathbb{M}_i \setminus \{u\}} m_{g_v \rightarrow i}^t(z_i), \\
 m_{g_u \rightarrow i}^{t+1}(z_i) &\propto \sum_{\mathbf{z}_{\mathbb{N}_u \setminus \{i\}}} g_u(\mathbf{z}_{\mathbb{N}_u}) \prod_{j \in \mathbb{N}_u} m_{j \rightarrow g_u}^{t+1}(z_j), \quad (22) \\
 m_{h_i \rightarrow i}(z_i) &= f_\phi(z_i; x_i),
 \end{aligned}$$

where the superscription  $t$  denotes the index of BP iteration and is omitted for  $m_{h_i \rightarrow i}$  as it has no change unless  $f_\phi$  is updated. We take the standard initialization of the worker-to-task message at uniform distribution, i.e.,  $m_{g_u \rightarrow i}^0 = [1/K]^K$ . After updating the messages  $T$  iterations, we use

$$q_i(z_i) \propto m_{h_i \rightarrow i}(z_i) \prod_{u \in \mathbb{M}_i} m_{g_u \rightarrow i}^T(z_i), \quad (23)$$

where  $\text{BP}(f_\phi(\mathbf{x}), \mathbf{y}, \boldsymbol{\alpha})$  denotes the entire collections of  $q_i(z_i)$  from BP. The inference of deepBP is directly based on  $q_i(z_i)$  from  $\text{BP}(\cdot)$ .

We now describe the update of classifier  $f_\phi$ . Using (21), the ELBO in (20) can be simplified as:

$$\begin{aligned}
 \mathcal{L}_{\text{BP}}(q(\mathbf{z}); \phi) &:= \mathbb{E}_{q(\mathbf{z})} \left[ \log \frac{h(\mathbf{z}; \phi) g(\mathbf{z}; \boldsymbol{\alpha})}{q(\mathbf{z})} \right] \\
 &= \mathbb{E}_{q(\mathbf{z})} [\log g(\mathbf{z}; \boldsymbol{\alpha})] - D_{\text{KL}}(q(\mathbf{z}) \parallel h(\mathbf{z}; \phi)), \quad (24)
 \end{aligned}$$

where  $h(\mathbf{z}; \phi) := \prod_{i \in [N]} h_i(z_i; \phi)$  and  $g(\mathbf{z}; \boldsymbol{\alpha}) := \prod_{u \in [M]} g_u(\mathbf{z}_{\mathbb{N}_u}; \boldsymbol{\alpha})$ . Then, given  $q(\mathbf{z})$  from  $\text{BP}(\cdot)$ , deepBP obtains  $\phi$  to maximize the ELBO in (24):

$$\hat{\phi} = \arg \max_{\phi} \mathcal{L}_{\text{BP}}(q(\mathbf{z}); \phi). \quad (25)$$

The overall process of deepBP is described at Algorithm 2.

### 3.4 Implementation Remarks

**Clipping trick.** In both deepMF and deepBP, the deep learning step can be unstable in early phase in particular when we train neural network from scratch.

In addition, we often find overfitting in deep learning. These observation motivates us to regulate the influence of deep classifier  $f_\phi$  in the inference step. To do so, to update  $q(\mathbf{z})$  in both algorithms, we clip the output of  $f_\phi(\cdot)$  at  $c$  and then evenly distribute the clipped amount to the other classes for normalization, i.e., for each class  $k$ , we have the clipped output:

$$\text{Clip}(f_\phi(k; \mathbf{x}), c) \leq c \quad \forall k \in [K]. \quad (26)$$

Then, the clipping parameter  $c$  can be adjusted accordingly to the certainty on features. For instance of classification from severely blurred images, we may set  $c$  close to  $1/K$  to cope overfitting.

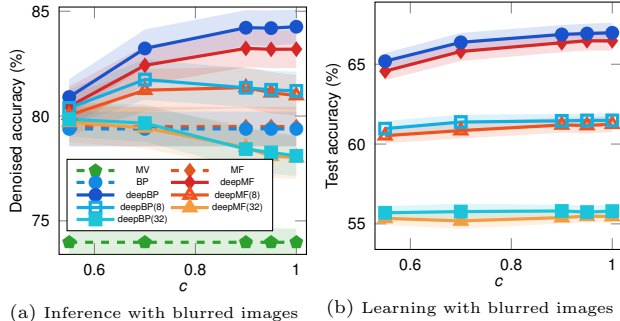
**Fast BP message update.** We remark that the update of  $m_{g_u \rightarrow i}^{t+1}$  in (22) seems intractable as it requires exponentially many summations in the number of tasks labeled by worker, i.e.,  $O(2^{|\mathbb{N}_u|})$ . To bypass the intractable computation, we use a Monte-Carlo method using  $S$  samples, of which computational cost is bounded by  $O(|\mathbb{N}_u| \cdot K \cdot S)$ . The detailed derivation and the performance with respect to the sample size can be found in Appendix E. We note that for a wide family of worker prior, Liu et al. (2012) propose another method with divide & conquer and fast Fourier transformation of  $O(|\mathbb{N}_u| \log^2 |\mathbb{N}_u| \cdot K^2)$  complexity. Liu et al. (2012)'s method is exact for the prior family. However, our method is universal to any prior distribution, but also is simply implementable with the utilization of GPU's parallel computing.

## 4 Numerical Analysis

We evaluate our algorithms, deepMF and deepBP, on inference and learning tasks for a image classification on both synthetic and real-world datasets. We compare the proposed algorithms with learning algorithms: CL (Rodrigues et al., 2014), BayesDGC (Li et al., 2021) and classic inference algorithms: MV, MF and BP (Liu et al., 2012). In the inference task, we report the accuracy of the estimated true labels obtained from (5). In the learning task, we report the accuracy of the classifier on unseen data. For classifier  $f_\phi$ , we use a four-layer convolutional network and train with an Adam optimizer. An average outcome with 50 different random seeds and the 99% confidence interval are reported for all experiments. More details on settings can be found in Appendix F.

### 4.1 Robustness Analysis

Several inference and learning algorithms have been proposed, but their robustness on various environments has not been analyzed thoroughly yet. In this



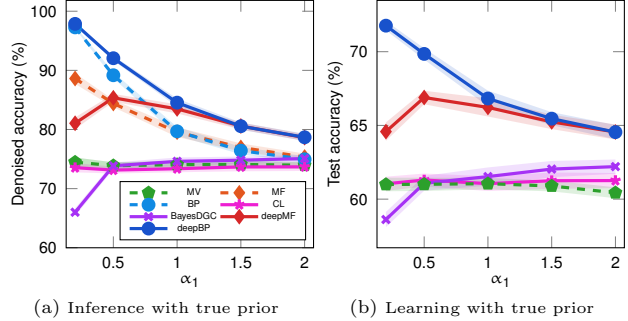
(a) Inference with blurred images (b) Learning with blurred images

Figure 1: *Robustness to blurred feature.* Inference (a) and learning (b) performance on blurred images with clipping parameter  $c$ . deepMF( $r$ ) and deepBP( $r$ ) indicate the performance of algorithms on the dataset with  $r$  radius of Gaussian blur. In all settings, the performance of deepMF and deepBP decrease as we blur images more. By clipping the classifier output with low values, we can bound the performance of deepMF and deepBP over the non-feature algorithms. We observe that deepBP outperforms the other existing methods.

experiment, we test algorithmic robustness from different perspectives. We use synthetic dataset for controlled experiments. The synthetic dataset is generated by assuming 1,000 tasks and 750 workers. The assignment structure between the tasks and workers is generated from a randomly sampled  $(l, r)$ -regular bipartite graph, where  $l$  indicates the number of workers assigned to each task and  $r$  indicates the number of tasks per worker. Each task is a binary classification associated with an image feature sampled from the Dogs vs. Cats dataset (Kaggle, 2013). Workers' confusion matrices are randomly drawn from the one-coin model as (3), where we set the parameter of the Dirichlet distribution differently for each experiment. We set  $l = 3$ ,  $r = 4$  for our synthetic data similar to the statistics of known datasets (Snow et al., 2008; Welinder et al., 2010; Han et al., 2015). Note that the assignment graph satisfies the sparse regime reported in the earlier work (Ok et al., 2016). To evaluate the classifiers obtained from learning methods, we use the test set from the Dogs vs. Cats dataset (Kaggle, 2013).

**Robustness to Feature.** We first test robustness of inference and learning algorithms against non-informative features. To generate a non-informative feature, we add Gaussian blur to the images<sup>3</sup> and vary the blur radius from 0 to 32. When features become more non-informative, i.e., as the radius increases, the message coming from task features becomes more irrelevant. To prevent the influence of features, we employ the clipping technique introduced in Section 3.4. We vary the clipping parameter  $c$  from 0.55 to 1. The

<sup>3</sup>The examples of blurred images are in Appendix G



(a) Inference with true prior (b) Learning with true prior

Figure 2: *Robustness to true prior.* Inference (a) and learning (b) performance with varying  $\alpha_1$  of the true prior:  $\text{Dir}(\alpha_1, \alpha_1/2)$ . As the prior becomes sparser, i.e.,  $\alpha_1 < 1$ , the gap between MF-based and BP-based algorithms increases. DeepBP outperforms the other methods across all settings.

dataset is generated with worker prior  $\text{Dir}(1, 0.5)$ , and we use the true prior for experiments.

With informative features, Figure 1a shows the inference accuracy of deepMF and deepBP can be improved to compare with their non-feature counter-parts. The learning accuracy is also increased with informative features as shown in Figure 1b. The performance of feature-based algorithms decrease as the images get more corrupted. However, we can still bound the performance of the proposed algorithms to those of non-feature algorithms by clipping the influence of the classifier when the features are highly non-informative.

**Robustness to Prior.** A prior distribution over the confusion matrices needs to be set a priori. The exact prior is unknown in general, but in the following experiments, we investigate how the algorithms work when the models know the true prior. This study reveals a failure of algorithms even with the exact knowledge on the true prior. We use  $\text{Dir}(\alpha_1, \alpha_1/2)$  with varying  $\alpha_1$  from 0.2 to 2 as a prior over the confusion matrices. As  $\alpha_1$  goes to zero, the prior becomes sparse, i.e., the worker is either an adversary or a hamper. The population of spammer increases as  $\alpha_1$  goes to two.

The results on the inference and learning tasks in Figure 2 show the MF-based algorithms perform worse than BP-based when the prior distribution is sparse, i.e.,  $\alpha_1 < 1$ . Although deepMF performs better than MF except the sparse prior case, both algorithms perform worse than their BP counter-parts.

**Robustness to Extreme-spammer.** In a typical crowdsourcing system, workers get a fixed amount of reward by solving a single task, and therefore some workers attempt as many tasks as possible to maximize the reward (Gadiraju et al., 2015). Among those,

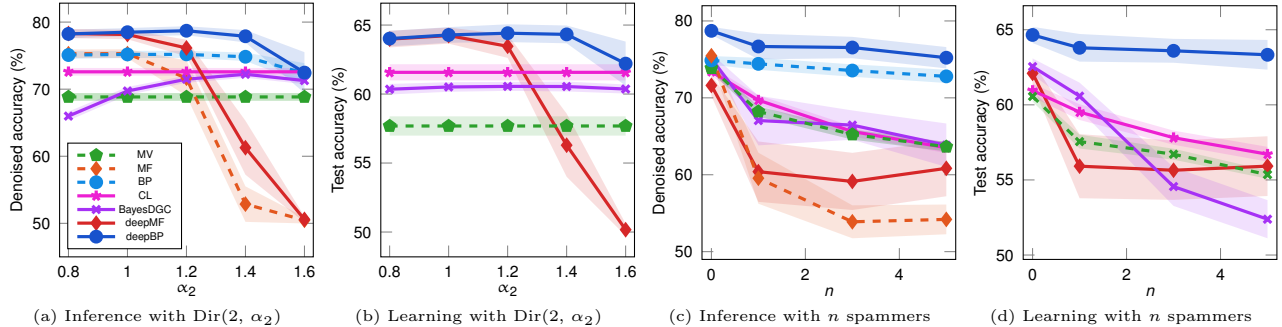


Figure 3: *Robustness to extreme-spammers with mismatched prior.* (a, b) The performance of MF-based methods decrease as the model prior deviates from the true prior:  $\text{Dir}(2, 1)$ . BP-based methods perform robustly than the MF-based methods in both inference and learning. (c, d) BP-based methods are influenced less from the spammers than the other methods regardless of the number of spammers. We use  $\text{Dir}(2, 1.4)$  for the model prior.

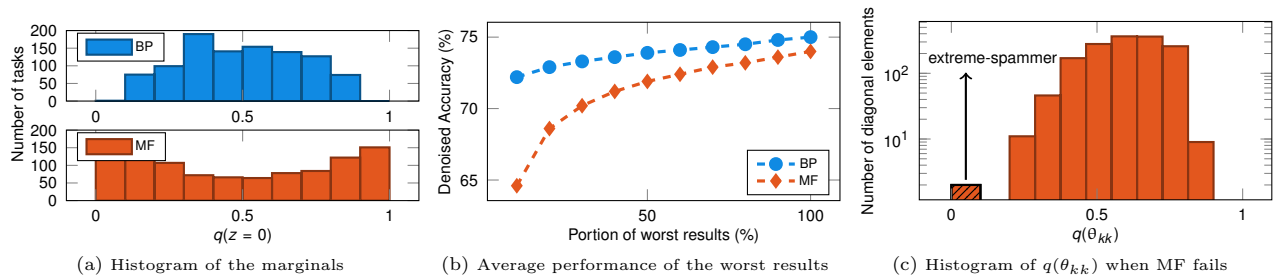


Figure 4: *Overconfidence of MF-based methods.* (a) The distribution of marginals on  $\mathbf{z}$  obtained from the variational distribution  $q(\mathbf{z})$ . (b) The inference accuracy sorted in increasing order from multiple experiments with different seeds. In the worst case, the performance of MF significantly worse than that of BP, although their performances are not on average. (c) Histogram of diagonal elements in confusion matrices when MF fails. MF is likely to categorize the extreme-spammer as an adversary or a hamper.

we focus on *extreme-spammer*, who labels uniformly across all tasks, i.e.  $p(y_i^{(u)} = z_i) \approx 1/K$ . Indeed, we observe the extreme-spammer appears frequently in real-world datasets. Check Appendix H for the existence of the extreme-spammer. We add the extreme-spammer to a synthetic dataset with worker prior  $\text{Dir}(2, 1)$  to see the effect of an extreme-spammer on each algorithm. We further assume that the true prior is unknown, making the simulated environment more realistic. All algorithms are evaluated with  $\text{Dir}(2, \alpha_2)$ , where  $\alpha_2$  varies from 0.8 to 1.6.

Figure 3a and Figure 3b show the inference and learning accuracy with an extreme-spammer, respectively. When the true prior is given, i.e.  $\alpha_2 = 1$ , both MF-based and BP-based algorithms perform well. However, the accuracy of MF-based algorithms starts to decrease as the mismatch between true and model prior becomes bigger. BP-based algorithms are robust to extreme-spammer than MF-based algorithms regardless of the given prior in this setting. When we add more extreme-spammers, the accuracy of BP-based algorithms decreases less than those of MF-based algorithms. The results are shown in Figure 3c and Fig-

ure 3d. This result suggests that MF-based algorithms are more vulnerable to a few workers who labels a lot.

**Overconfidence Issues.** We experimentally show that the BP-based algorithms are more robust than the MF-based algorithms on both inference and learning task under various scenarios. To understand the difference, we focus on the phenomenons which are known to appear in mean-field approximation: overconfidence and local minima (Weiss, 2001; Murphy, 2013). In toy problems, Weiss (2001) experimentally shows that the variational distributions of MF are likely to be overconfident and fall into local minima easily.

We investigate the problems in crowdsourcing systems. To understand overconfidence, we analyze the results used in Figure 3a. Figure 4a shows the distribution over marginal of true labels from the results. The marginals of MF skew to either zero or one, whereas those of BP distributed evenly around a half. To understand how easily MF falls into a local minima, we run 100 experiments with different seeds and sort the inference accuracy in ascending order. Figure 4b shows that there is a significant gap between MF and BP in



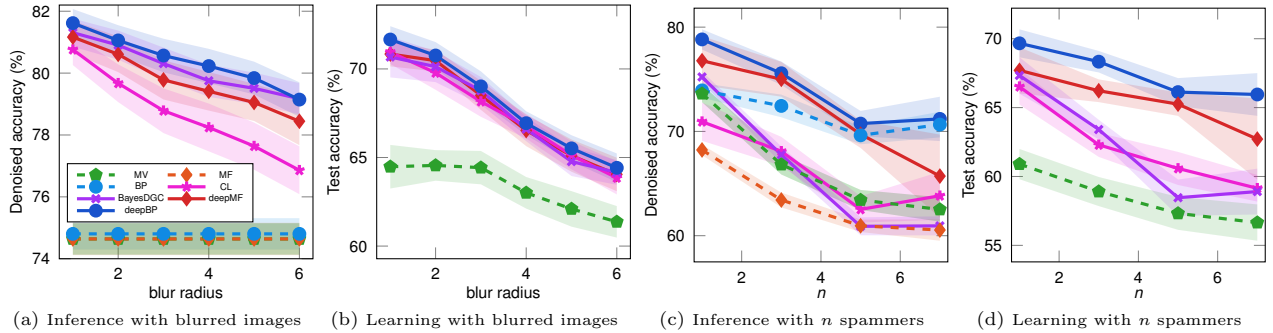


Figure 5: *Face-age dataset experiments*. (a) The inference accuracy decreases as we blur the images more. (b) DeepBP outperforms the others in learning task when the images are relatively clear. (c, d) DeepBP is more robust than the other models against extreme-spammers added to the face-age dataset.

Table 2: *LabelMe experiments*. The performance is measured by training from scratch with 20 different seeds. DeepMF and deepBP outperform the other baseline models.

Models	MV	CL	BayesDGC	deepMF	deepBP
Denoised accuracy (%)	77.22 ± 0.10	73.55 ± 4.53	77.91 ± 0.43	80.86 ± 0.23	<b>81.03 ± 0.22</b>
Test accuracy (%)	61.28 ± 0.86	63.38 ± 3.81	53.93 ± 3.05	<b>65.47 ± 0.80</b>	65.45 ± 0.74

the worst case, although their performance is not on average. These two analyses show the overconfidence and local minima problems still exist in the crowdsourcing systems. In addition, we investigate the distribution over the diagonal elements of the confusion matrices when MF trapped into a local minima. Figure 4c shows that MF estimates the extreme-spammer as an adversary, and therefore the labels of spammers misguide the inference steps in MF. We conjecture that the overconfident makes MF-based algorithms sensitive but leave a thorough investigation for future work.

## 4.2 Experiments on Real-world Datasets

For the real-world datasets, we use the face-age dataset from Han et al. (2015) and LabelMe (Rodrigues and Pereira, 2018).

**Face-age.** The face-age dataset is consisted of 1,002 human face images and their predicted ages answered by 165 workers from Amazon Mechanical Turk. Since 20% of the facial dataset are biased to infant images, we use 800 facial images over five-years-old. We then transform the remaining dataset into a binary image classification that classifies whether a face is over 16-years-old to make a balanced classification problem. The dataset is divided into 500 training and 300 test sets. On average, each worker labels 30 images in the original dataset. We generate a new assignment graph by randomly subsampling four tasks for each worker, i.e.,  $r = 4$  to make the assignment structure sparse. For the model prior, we use  $\text{Dir}(1, 0.5)$  in figures 5a

and 5b, and  $\text{Dir}(2, 1.3)$  in figures 5c and 5d. Figure 5 shows the performance of the algorithms with blurred images and extreme-spammers. The results show deepBP is robust against corrupted features and extreme-spammers in the real-world dataset as well.

**LabelMe.** LabelMe is a multi-class image classification task with eight classes consisted of 2,688 images. We use 1,000 training images labeled by 59 workers from the Amazon Mechanical Turk. The rest 1,688 images are used for testing. We set the hyperparameter of prior  $\alpha_{kk'}$  to 8 if  $k = k'$  and 1 otherwise. Table 2 shows that both deepMF and deepBP outperform the existing methods.

## 5 Conclusion

In this work, we propose principled frameworks, named deepMF and deepBP, alternating variational inference and deep learning to utilize both the prior of worker behavior and the features of tasks. We reveal that the previous deep crowdsourcing methods are special cases of deepMF with specific choices of worker prior. We propose deepBP inspired by the strong theoretical guarantee on BP for inference from crowds. The experimental results show that the deepBP algorithm is more robust than the deepMF algorithm in canonical scenarios with non-informative features, multi-modal or mismatched worker prior, or the presence of extreme spammers.

## Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No.2021R1C1C1011375)

## References

- Cao, P., Xu, Y., Kong, Y., and Wang, Y. (2019). MaxMIG: an information theoretic approach for joint learning from crowds. In *International Conference on Learning Representations*.
- Chu, Z. and Wang, H. (2021). Improve learning from crowds via generative augmentation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 167–175, New York, NY, USA. Association for Computing Machinery.
- Dawid, A. and Skene, A. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, pages 20–28.
- Gadiraju, U., Demartini, G., Kawase, R., and Dietze, S. (2015). Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems*, 30(4):81–85.
- Han, H., Otto, C., Liu, X., and Jain, A. K. (2015). Demographic estimation from face images: Human vs. machine performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1148–1161.
- Jagabathula, S., Subramanian, L., and Venkataraman, A. (2017). Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *The Journal of Machine Learning Research*, 18(1):3233–3299.
- Kaggle (2013). Dogs vs. cats. <https://www.kaggle.com/c/dogs-vs-cats/overview>.
- Kuck, J., Chakraborty, S., Tang, H., Luo, R., Song, J., Sabharwal, A., and Ermon, S. (2020). Belief propagation neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 667–678. Curran Associates, Inc.
- Li, S.-Y., Huang, S.-J., and Chen, S. (2021). Crowdsourcing aggregation with deep bayesian learning. *SCIENCE CHINA-INFORMATION SCIENCES*, 64(3).
- Liu, Q., Peng, J., and Ihler, A. (2012). Variational inference for crowdsourcing. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1*, pages 692–700.
- Murphy, K. P. (2013). *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.].
- Nguyen, A. T., Wallace, B., Li, J. J., Nenkova, A., and Lease, M. (2017). Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 299–309, Vancouver, Canada. Association for Computational Linguistics.
- Ok, J., Oh, S., Shin, J., and Yi, Y. (2016). Optimality of belief propagation for crowdsourced classification. In *International Conference on Machine Learning*, pages 535–544. PMLR.
- Pearl, J. (1982). Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the Second AAAI Conference on Artificial Intelligence*, AAAI'82, page 133–136. AAAI Press.
- Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, 11(4).
- Rodrigues, F. and Pereira, F. (2018). Deep learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Rodrigues, F., Pereira, F., and Ribeiro, B. (2014). Gaussian process classification and active learning with multiple annotators. In *International conference on machine learning*, pages 433–441. PMLR.
- Russell, B. C., Torralba, A., Murphy, K. P., and Freeman, W. T. (2008). Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173.
- Satorras, V. G. and Welling, M. (2021). Neural enhanced belief propagation on factor graphs. In *International Conference on Artificial Intelligence and Statistics*, pages 685–693. PMLR.
- Snow, R., O’connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.
- Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D. C., and Silberman, N. (2019). Learning from noisy labels by regularized estimation of annotator

confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11244–11253.

Venanzi, M., Guiver, J., Kazai, G., Kohli, P., and Shokouhi, M. (2014). Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web, WWW '14*, page 155–164, New York, NY, USA. Association for Computing Machinery.

Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in mrfs. *Advanced mean field methods: theory and practice*, pages 229–240.

Welinder, P., Branson, S., Perona, P., and Belongie, S. (2010). The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23:2424–2432.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J., and Ruvolo, P. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, 22:2035–2043.

Yedidia, J. S., Freeman, W. T., Weiss, Y., et al. (2003). Understanding belief propagation and its generalizations.

Zhang, Z., Wu, F., and Lee, W. S. (2020). Factor graph neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8577–8587. Curran Associates, Inc.

---

## Supplementary Material: Robust Deep Learning from Crowds with Belief Propagation

---

### A Large but Sparse Dataset given Fixed Budget

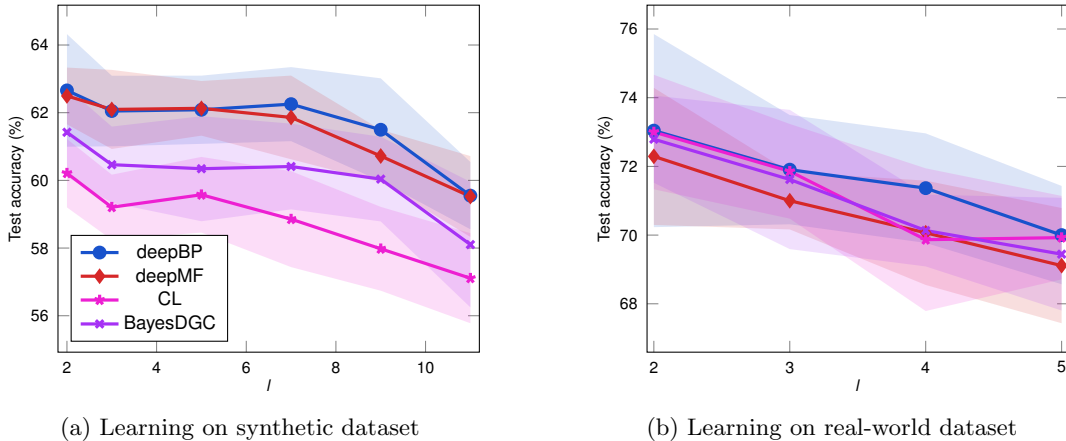


Figure 6: *Fixed budget experiments.* Learning performance with varying  $l$ : (a) given  $B = 2,000$ , synthetic dataset generated with worker prior  $\text{Dir}(2, 1)$ ; and (b) given  $B = 1,000$ , the same real-world dataset used in Figure 5 for which the algorithms use  $\text{Dir}(2, 1)$  as worker prior.

In our experiment, we mainly focus on the large-but-sparse dataset, where the number of tasks  $N$  is large but the number of workers per task is limited by a small constant  $l$  due to budget constraint. We often find such a dataset in practice, e.g., (Russell et al., 2008; Rodrigues et al., 2014), although it is possible to use the same budget for small-but-dense dataset with small  $N$  and large  $l$ . In this section, we formally study the trade-off between large-but-sparse and small-but-dense datasets, where, in turn, the crowdsourced dataset is better to be large-but-sparse than small-but-dense. In Figure 6, we plot the performance of the set of learning algorithms varying the choice of  $N$  and  $l$  given budget  $B$  such that  $N = \lfloor \frac{B}{l} \rfloor$ . We observe that regardless of which algorithm is employed, the learning performance decreases as  $l$  increases. This suggests the large-but-sparse regime rather than the small-but-dense one.

## B Details of Deep Mean-Field

We derive a detailed equation of deepMF alternating variational inference and learning described in Section 3.1.

DeepMF is derived from an ELBO of the marginalization in (11) by introducing a variational distribution  $q(\mathbf{z}, \boldsymbol{\theta})$  in the followings:

$$\log p(\mathbf{y} | \mathbf{x}, \boldsymbol{\alpha}, \phi) = \log \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\theta})} \left[ \frac{p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \mathbf{x}, \boldsymbol{\alpha}, \phi)}{q(\mathbf{z}, \boldsymbol{\theta})} \right] \geq \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\theta})} \left[ \log \frac{p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \mathbf{x}, \boldsymbol{\alpha}, \phi)}{q(\mathbf{z}, \boldsymbol{\theta})} \right]. \quad (27)$$

Note that the lower bound is maximized and achieves the equality if  $q(\mathbf{z}, \boldsymbol{\theta}) = p(\mathbf{z}, \boldsymbol{\theta} | \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \phi)$ . We apply MF approximation, assuming a conditional independence of  $z_i$ 's and  $\boldsymbol{\theta}^{(u)}$ 's given  $(\mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \phi)$ , such that:

$$q(\mathbf{z}, \boldsymbol{\theta}) = q(\mathbf{z})q(\boldsymbol{\theta}; \boldsymbol{\beta}) = \prod_{i \in [N]} q_i(z_i) \prod_{u \in [M]} q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)}), \quad (28)$$

where we use a parametric estimation  $q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)}) = \text{Dir}(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)})$  for each worker  $u$ , and approximating  $q_i(z_i) \approx p(z_i | \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \phi)$  and  $q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)}) \approx p(\boldsymbol{\theta}^{(u)} | \mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \phi)$ , we let  $q(\mathbf{z}) = \prod_{i \in [N]} q_i(z_i)$  and  $q(\boldsymbol{\theta}; \boldsymbol{\beta}) = \prod_{u \in [M]} q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)})$ .

Then, using (4) and (28), the ELBO in (27) is written as follows:

$$\begin{aligned} \mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \boldsymbol{\theta}, \boldsymbol{\beta}) &:= \mathbb{E}_{q(\mathbf{z})q(\boldsymbol{\theta}; \boldsymbol{\beta})} \left[ \log \frac{p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \mathbf{x}, \boldsymbol{\alpha}, \phi)}{q(\mathbf{z})q(\boldsymbol{\theta}; \boldsymbol{\beta})} \right] = \mathbb{E}_{q(\mathbf{z})q(\boldsymbol{\theta}; \boldsymbol{\beta})} \left[ \log \frac{p(\mathbf{z} | \mathbf{x}, \phi)p(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta})p(\boldsymbol{\theta} | \boldsymbol{\alpha})}{q(\mathbf{z})q(\boldsymbol{\theta}; \boldsymbol{\beta})} \right] \\ &= \mathbb{E}_{q(\mathbf{z})q(\boldsymbol{\theta}; \boldsymbol{\beta})} [\log p(\mathbf{y} | \mathbf{z}, \boldsymbol{\theta})] - D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x}, \phi)) - D_{\text{KL}}(q(\boldsymbol{\theta}; \boldsymbol{\beta}) \| p(\boldsymbol{\theta} | \boldsymbol{\alpha})). \end{aligned} \quad (29)$$

In each iteration of deepMF, we seek  $q(\mathbf{z})$ ,  $\boldsymbol{\beta}$  and  $\phi$ , sequentially, to maximize  $\mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \boldsymbol{\theta}, \boldsymbol{\beta})$  by fixing the others. To be specific, recalling  $p(\mathbf{z} | \mathbf{x}, \phi) := \prod_{i \in [N]} f_\phi(z_i; x_i)$ , given the classifier  $f_\phi$  and the other variational distributions, the ELBO is maximized at  $q_i(z_i)$  such that:

$$\log q_i(z_i) = \sum_{u \in \mathbb{M}_i} \mathbb{E}_{q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)})} [\log \theta_{z_i, y_i^{(u)}}^{(u)}] + \log f_\phi(z_i; x_i) - 1, \quad (30)$$

where we let  $\mathbb{M}_i$  be the set of workers labeling task  $i$  and collection is denoted by  $\text{MF}(f_\phi(\mathbf{x}), \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta})$ . The inference of deepMF is based on  $q_i(z_i)$  from  $\text{MF}(\cdot)$ .

Then, provided  $q(\mathbf{z})$  from  $\text{MF}(\cdot)$ , deepMF finds  $\phi$  and  $\boldsymbol{\beta}$  independently to maximize the ELBO in (29):

$$\hat{\phi}, \hat{\boldsymbol{\beta}} = \arg \max_{\phi, \boldsymbol{\beta}} \mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \boldsymbol{\theta}, \boldsymbol{\beta}). \quad (31)$$

First, the ELBO is maximized at  $q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)})$  such that:

$$\log q_u(\boldsymbol{\theta}^{(u)}; \boldsymbol{\beta}^{(u)}) = \sum_{i \in \mathbb{N}_u} \mathbb{E}_{q(z_i)} [\log p(y_i^{(u)} | z_i, \boldsymbol{\theta}^{(u)})] + \log p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) - 1, \quad (32)$$

where we use a worker prior  $p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}^{(u)}; \boldsymbol{\alpha})$  for each worker  $u$ . Each element of  $\boldsymbol{\beta}^{(u)}$  is updated as:

$$\beta_{k_1 k_2}^{(u)} = \alpha_{k_1 k_2} + \sum_{\{i | i \in \mathbb{N}_u, y_i^{(u)} = k_2\}} q_i(z_i = k_1), \quad (33)$$

for all  $k_1, k_2 \in [K]$ .

Second, the ELBO is maximized at  $\phi$  such that:

$$\hat{\phi} = \arg \max_{\phi} \mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \boldsymbol{\theta}, \boldsymbol{\beta}) = \arg \min_{\phi} D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x}, \phi)), \quad (34)$$

which corresponds to training classifier  $f_\phi$ .

We note that deepMF is easily implementable as an iteration of the inference in (30) and learning in (34) is just a sequence of maximizations for the same objective in (29) but different variables. Hence, it is a popular framework despite the coarse approximation in (28).

## C Correspondence of Deep Mean-Field to Existing Algorithms

Table 3: *Choices of variational distributions and hyperparameter.* All the methods use fully-factorized variational distribution which follows MF approximation and Dirichlet distribution for worker prior  $\alpha$ . To approximate the posterior of true label  $q_i(z_i)$ , CL and Trace use the output of a neural network  $f_\phi(z_i; x_i)$ . BayesDGC and deepMF estimate the Categorical distribution of  $\gamma$ . For the posterior of confusion matrix  $q_u(\theta^{(u)})$ , BayesDGC and deepMF estimate the Dirichlet distribution of  $\beta$  to capture uncertainty. For the worker prior, CL set all  $\alpha$  to 1 which indicates no prior distribution on  $\theta$ . Trace posits an adversary prior that assumes more adversaries than hammers.

Model	$q_i(z_i)$	$q_u(\theta^{(u)})$	$\alpha$
CL (Rodrigues and Pereira, 2018)	$f_\phi(z_i; x_i)$	$\delta(\theta = \theta^{(u)})$	$\alpha = \mathbf{1}$
Trace (Tanno et al., 2019)	$f_\phi(z_i; x_i)$	$\delta(\theta = \theta^{(u)})$	$\alpha_{kk} < 1, \alpha_{kk'} = 1 (k' \neq k)$
BayesDGC (Li et al., 2021)	$\text{Cat}_i(z_i   \gamma)$	$\text{Dir}(\theta^{(u)}   \beta^{(u)})$	-
deepMF	$\text{Cat}_i(z_i   \gamma)$	$\text{Dir}(\theta^{(u)}   \beta^{(u)})$	-

In Section 3.2, we show that deepMF is a generalization of some previous studies based on choices of variational distributions and worker prior (Rodrigues and Pereira, 2018; Tanno et al., 2019; Li et al., 2021). In this section, we show that how these studies can be seen as a special case of deepMF. The choices of variational distributions and hyperparameters are summarized in Table 3.

First, the ELBO of deepMF in (14) is written as follows:

$$\mathcal{L}_{\text{MF}}(q(\mathbf{z}); \phi, \theta, \beta) := \mathbb{E}_{q(\mathbf{z})q(\theta; \beta)}[\log p(\mathbf{y} | \mathbf{z}, \theta)] - D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x}, \phi)) - D_{\text{KL}}(q(\theta; \beta) \| p(\theta | \alpha)). \quad (35)$$

Each Rodrigues and Pereira (2018) and Tanno et al. (2019) use a neural network parameterized by  $\phi$  to model the posterior of true label  $q(\mathbf{z})$  and use the Dirac delta function as a variational distribution of confusion matrices  $q(\theta)$  to simplify the training process. Then, the new ELBO is given by:

$$\mathcal{L}_{\text{new}}(\mathbf{z}; \phi, \theta) := \mathbb{E}_{p(\mathbf{z} | \mathbf{x}, \phi)}[\log p(\mathbf{y} | \mathbf{z}, \theta)] + \log p(\theta | \alpha). \quad (36)$$

The two algorithms choose different concentration parameters  $\alpha$  of the Dirichlet distribution. Assuming each confusion matrix  $\theta^{(u)}$  is conditionally independent given  $\alpha$ , the log-likelihood is written as:

$$\log p(\theta | \alpha) = \sum_{u \in [M]} \log p(\theta^{(u)} | \alpha) = \sum_{u \in [M]} \sum_{k_1 \in [K]} \sum_{k_2 \in [K]} \log p(\theta_{k_1 k_2}^{(u)} | \alpha_{k_1 k_2}) \propto \sum_{u \in [M]} \sum_{k_1 \in [K]} \sum_{k_2 \in [K]} (\alpha_{k_1 k_2} - 1) \log \theta_{k_1 k_2}^{(u)}. \quad (37)$$

Rodrigues and Pereira (2018) set all  $\alpha$  to 1, which assumes that the workers are from a uniform distribution. Then, the ELBO of CL (Rodrigues and Pereira, 2018) is given by:

$$\mathcal{L}_{\text{CL}}(\mathbf{z}; \phi, \theta) := \mathbb{E}_{p(\mathbf{z} | \mathbf{x}, \phi)}[\log p(\mathbf{y} | \mathbf{z}, \theta)], \quad (38)$$

where its negative represents the cross-entropy of the distribution  $p(\mathbf{y} | \mathbf{z}, \theta)$  relative to the distribution  $p(\mathbf{z} | \mathbf{x}, \phi)$ . Tanno et al. (2019) set  $\alpha_{k_1 k_2}$  less than 1 when  $k_1 = k_2$  and 1 otherwise, which assumes that adversaries exist more than hammers, i.e.,  $\text{Dir}(0.8, 1)$ . Then, the ELBO of log version of Trace (Tanno et al., 2019) is given by:

$$\mathcal{L}_{\text{Trace}}(\mathbf{z}; \phi, \theta) := \mathbb{E}_{p(\mathbf{z} | \mathbf{x}, \phi)}[\log p(\mathbf{y} | \mathbf{z}, \theta)] - \lambda \text{tr}(\log \theta), \quad (39)$$

where  $\lambda$  is hyperparameter, i.e.  $\lambda \in \mathbb{R}^{K \times K}$  and  $\lambda_{k_1 k_2} = 1 - \alpha_{k_1 k_2}$ .  $\text{tr}(\cdot)$  denotes trace and the difference from Tanno et al. (2019) is the existence of log in the trace. The ELBO in (39) simply add a regularization term of worker prior to CL (Rodrigues and Pereira, 2018).

However, the above uniform and adversary prior seem unreasonable in the real world. We find that the difference in prior is correlated with the initialization of the confusion matrix. Rodrigues and Pereira (2018) initialize the confusion matrix with identities, however, Tanno et al. (2019) initialize the confusion matrix to totally experts, i.e., the diagonal component of the confusion matrix is greater than 1, for adversary prior, which can lower the value of the confusion matrix.

## D Sensitivity of Adversary Prior

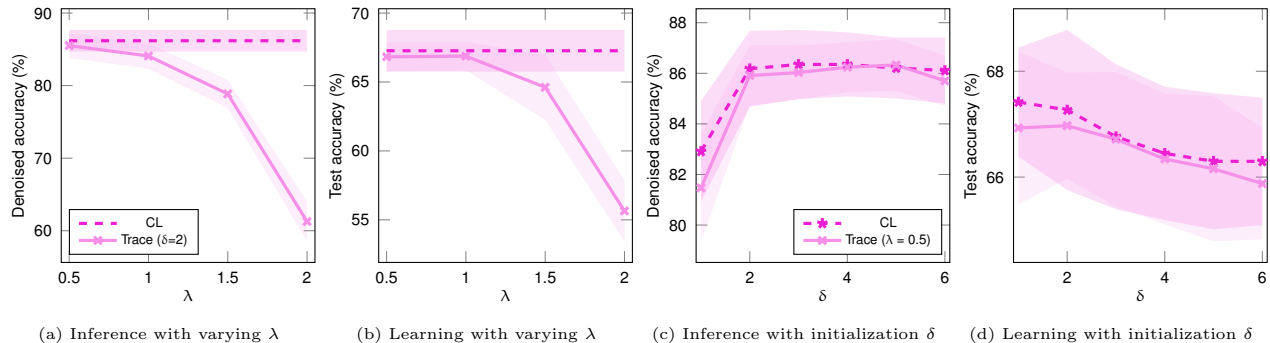


Figure 7: *Sensitivity of Adversary Prior*. (a, b) The inference and learning performance of Trace decrease as  $\lambda$  increases. We use confusion matrices initialized with  $\delta = 2$ . (c, d) The inference and learning performance of both CL and Trace are sensitive to the initialization of confusion matrices. We set the hyperparameter  $\lambda$  to 0.5.

Trace (Tanno et al., 2019) proposes the adversary prior which assumes more adversaries than hammers. However, we empirically found the adversarial prior is sensitive to the choice of different hyperparameters. Specifically, we focus on  $\lambda$  and  $\delta$ , which indicates the degree of adversary and the initialization value of confusion matrices, respectively. For the experiments, the confusion matrix is initialized as an diagonal matrix with the same diagonal entry  $\delta$ , e.g.,  $[[\delta, 0], [0, \delta]]$  for a binary classification. We use a synthetic dataset generated with worker prior  $\text{Dir}(3, 1)$ . Different learning rates are set for classifier parameters and confusion matrix parameters:  $1e-3$  for training the classifier and  $1e-2$  for the confusion matrices, respectively. We conduct experiments from 7 different seeds. As Trace constrains  $\lambda$  to be positive, we vary  $\lambda$  from 0.5 to 2. To initialize confusion matrices of workers to totally experts, we vary  $\delta$  from 1 to 6. Figure 7 shows that in both inference and learning task, the performance of Trace seems to be sensitive to  $\lambda$  and  $\delta$ . Even in the above situations, Trace performs worse than CL which assumes uniform prior.

## E BP Optimization

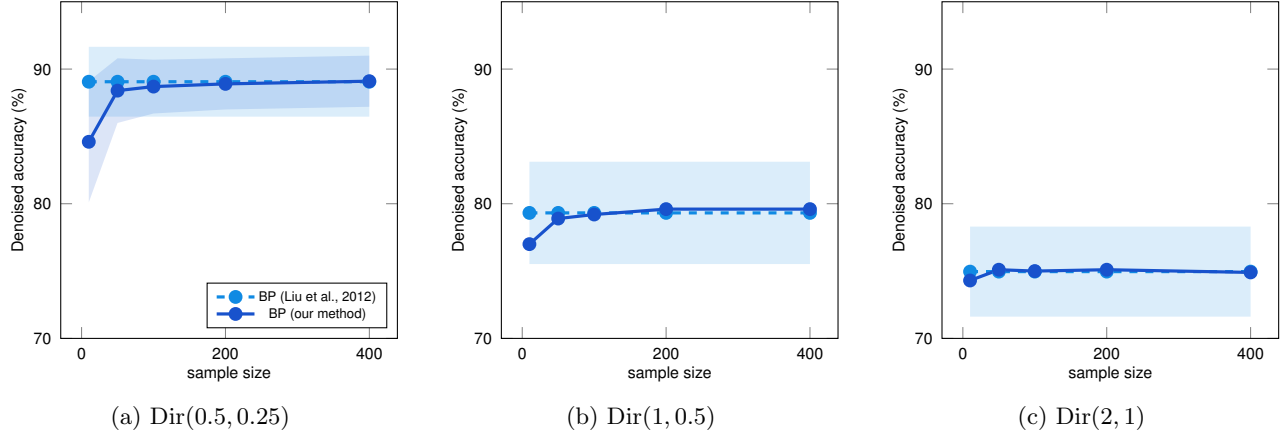


Figure 8: *Inference performance of our BP and Liu et al. (2012)'s BP.* We compare the inference performance of our BP algorithm and Liu et al. (2012)'s BP algorithm in three different worker priors. The inference performance is averaged from 100 different seeds. (a, b, c) In all the settings, as the sample size increases, our BP's inference performance is improved and shows similar performance to Liu et al. (2012). We use sample size 400.

### E.1 Factor-to-Variable Message

We remark that the update of  $m_{g_u \rightarrow i}^{t+1}$  in (22) seems intractable as it requires exponentially many summations in the number of tasks labeled by worker, i.e.,  $O(2^{|\mathbb{N}_u|})$ . To bypass the intractable computation, we use a Monte-Carlo method using  $S$  samples, of which computational cost is bounded by  $O(|\mathbb{N}_u| \cdot K \cdot S)$ .

**Expansion.** The update of factor-to-variable message in (22) can be written as:

$$m_{g_u \rightarrow i}(z_i) \propto \sum_{\mathbf{z}_{\mathbb{N}_u \setminus \{i\}}} g_u(\mathbf{z}_{\mathbb{N}_u}) \prod_{j \in \mathbb{N}_u \setminus i} m_{j \rightarrow g_u}(z_j) \quad (40)$$

$$= \sum_{\mathbf{z}_{\mathbb{N}_u \setminus \{i\}}} \left( \int_{\boldsymbol{\theta}^{(u)}} p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) \prod_{(k_1, k_2) \in [K] \times [K]} (\theta_{k_1, k_2}^{(u)})^{\gamma_{k_1 k_2}^{(u)}} d\boldsymbol{\theta}^{(u)} \right) \prod_{j \in \mathbb{N}_u \setminus i} m_{j \rightarrow g_u}(z_j) \quad (41)$$

$$= \int_{\boldsymbol{\theta}^{(u)}} p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) \cdot \sum_{\mathbf{z}_{\mathbb{N}_u \setminus \{i\}}} \left( \prod_{(k_1, k_2) \in [K] \times [K]} (\theta_{k_1, k_2}^{(u)})^{\gamma_{k_1 k_2}^{(u)}} \prod_{j \in \mathbb{N}_u \setminus i} m_{j \rightarrow g_u}(z_j) \right) d\boldsymbol{\theta}^{(u)} \quad (42)$$

$$= \int_{\boldsymbol{\theta}^{(u)}} p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) \cdot \theta_{z_i, y_i^{(u)}}^{(u)} \cdot \sum_{\mathbf{z}_{\mathbb{N}_u \setminus i}} \prod_{j \in \mathbb{N}_u \setminus i} \theta_{z_j, y_j^{(u)}}^{(u)} \cdot m_{j \rightarrow g_u}(z_j) d\boldsymbol{\theta}^{(u)} \quad (43)$$

$$= \int_{\boldsymbol{\theta}^{(u)}} p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) \cdot \theta_{z_i, y_i^{(u)}}^{(u)} \cdot \prod_{j \in \mathbb{N}_u \setminus \{i\}} \sum_{z \in [K]} \theta_{z, y_j^{(u)}}^{(u)} \cdot m_{j \rightarrow g_u}(z) d\boldsymbol{\theta}^{(u)} \quad (44)$$

$$= \int_{\boldsymbol{\theta}^{(u)}} p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha}) \cdot \theta_{z_i, y_i^{(u)}}^{(u)} \cdot \prod_{j \in \mathbb{N}_u \setminus i} \langle \theta_{\cdot, y_j^{(u)}}^{(u)}, \mathbf{m}_{j \rightarrow g_u} \rangle d\boldsymbol{\theta}^{(u)} \quad (45)$$

$$= \mathbb{E}_{p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha})} \left[ \theta_{z_i, y_i^{(u)}}^{(u)} \cdot \prod_{j \in \mathbb{N}_u \setminus i} \langle \theta_{\cdot, y_j^{(u)}}^{(u)}, \mathbf{m}_{j \rightarrow g_u} \rangle \right], \quad (46)$$

where  $\gamma_{k_1 k_2}^{(u)}$  is the number tasks that the true label is  $k_1$  and the worker  $u$ 's label is  $k_2$ , and  $\mathbf{m}_{j \rightarrow g_u}$  is a vector of the messages from task  $j$  to worker  $u$  where  $\mathbf{m}_{j \rightarrow g_u} \in \mathbb{R}^K$ . By approximating  $p(\boldsymbol{\theta}^{(u)} | \boldsymbol{\alpha})$  with  $S$  Monte-Carlo samples, we can compute the expectation (46) in  $O(|\mathbb{N}_u| \cdot K \cdot S)$ .



## E.2 Effect of the Sample Size

As our BP algorithm in (46) uses Monte-Carlo sampling, its performance is inconsistent when using continuous worker prior, for example, Beta and Dirichlet. To provide a performance boundary, we test our BP algorithm in synthetic datasets, where the true prior is continuous, with varying sample size from 10 to 400 and compare it with Liu et al. (2012)'s BP algorithm when the true prior is given to the algorithms. Each of the synthetic dataset is generated with worker prior  $\text{Dir}(2, 1)$ ,  $\text{Dir}(1, 0.5)$ , and  $\text{Dir}(0.5, 0.25)$ . Figure 8 shows that in all the settings, the inference performance of our BP algorithm is improved as the sample size increases. Furthermore, with many samples, it shows comparable performance to Liu et al. (2012).

## F Experimental Details

Table 4: *Learning rates of the experiments.* CL denotes Rodrigues and Pereira (2018) and BayesDGC denotes Li et al. (2021).

Models	MV	CL	BayesDGC	deepMF	deepBP
<b>Figure 1</b>	-	-	-	1e-4	1e-4
<b>Figure 2</b>	1e-4	1e-4	1e-4	1e-4	1e-4
<b>Figures 3a, 3b</b>	1e-4	1e-4	1e-5	1e-4	1e-4
<b>Figures 3c, 3d</b>	1e-4	1e-4	1e-4	1e-4	1e-4
<b>Figures 5a, 5b</b>	1e-3	1e-3	1e-3	1e-3	1e-3
<b>Figures 5c, 5d</b>	5e-5	5e-5	5e-5	5e-5	5e-5
<b>Table 2</b>	1e-4	1e-3	1e-4	1e-4	1e-4
<b>Figures 6</b>	-	1e-4	1e-4	1e-4	1e-4

### F.1 Model Architecture

Due to the complexity between the input features and the true labels, the classifier in (1) is often modeled with neural networks. We follow the standard architecture of neural network in crowdsourcing literature for fair comparison, e.g., (Rodrigues and Pereira, 2018) and (Cao et al., 2019). For the classifier, we use a four-layer convolutional network with kernel size 3 and stride 1 without padding, and the number of output channels for each convolutional layer is 32, 32, 64, and 64. Each of the convolutional layer’s output passes through a max-pooling layer with kernel size 2 and stride 2. Batch normalization is used at the output of the first convolutional layer. By passing the last convolutional layer’s output to two fully-connected layers with hidden size 128 and ReLU activation, we finally retrieve the prediction of the input image.

### F.2 Hyperparameters

We train the classifier with full-batch to maintain the whole structure of the given assignment graph in all the experiments. For approximating the factor-to-variable messages in (46), we use sample size 400. We clip the classifier’s output with value 0.9 for all the experiments, except Figure 1. Table 4 shows the learning rates for each experiment. For the computing infrastructures, we use RTX 3090 and RTX A5000.

## G Examples of Blurred Images

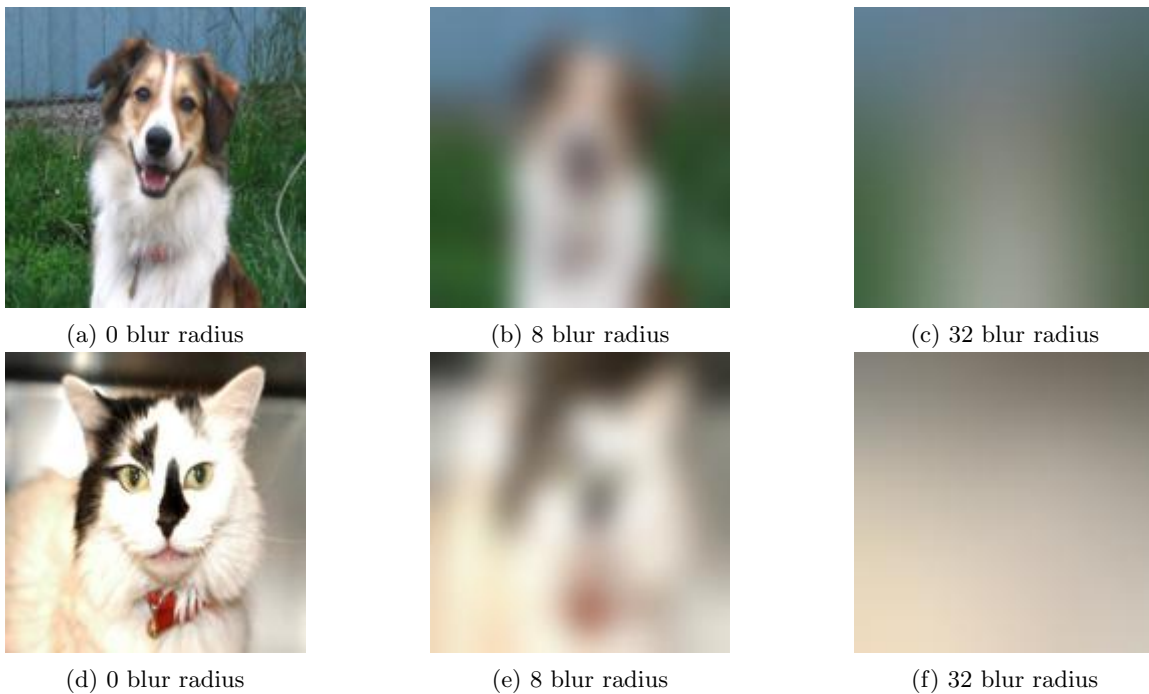


Figure 9: *Blurred images.* (a, b, c) Dog images with blur radius from 0 to 32. (d, e, f) Cat images with blur radius from 0 to 32. When blur radius reaches 32, it is tricky to distinguish between dogs and cats with the naked eyes.

In Section 4.1, we show the performance of deepMF and deepBP when the task features are corrupted by the Gaussian blur. To exemplify the effect of Gaussian blur, we demonstrate the changes in images with varying radius of Gaussian blur in Figure 9. When the radius of Gaussian blur reaches 32, it is almost impossible to identify the original class of the image by naked eyes as shown in Figure 9c and Figure 9f.

## H Existence of Extreme-spammers

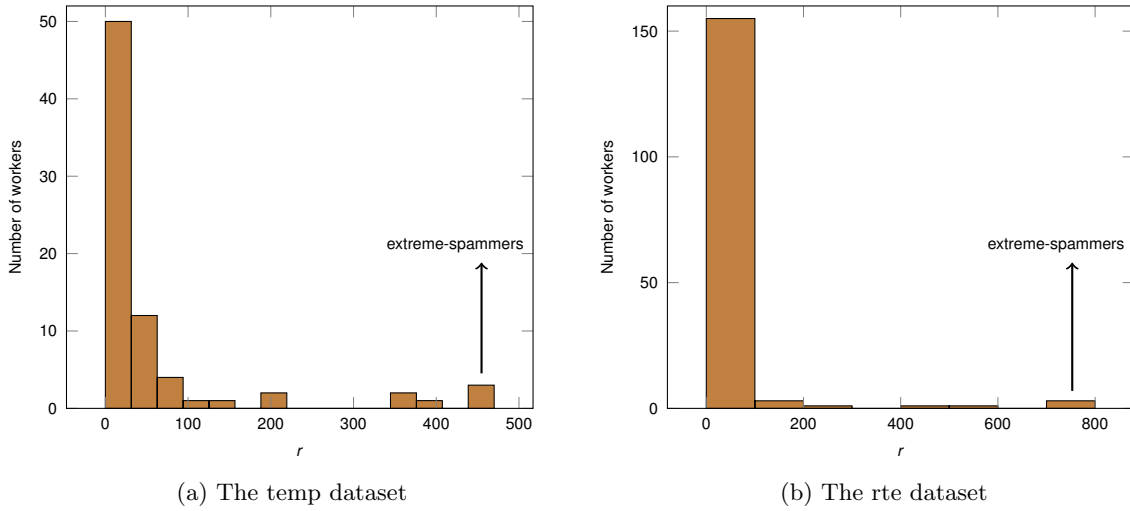


Figure 10: *Existence of extreme-spammers.* (a) The temp dataset (Snow et al., 2008) consists of 462 tasks and 76 workers. Three extreme-spammers solve 442, 452, and 462 tasks with correct answer rate 0.50, 0.44, and 0.44, respectively. (b) The rte dataset (Snow et al., 2008) consists of 800 tasks and 164 workers. Four extreme-spammers solve 540, 700, 760, and 800 tasks with correct answer rate 0.50, 0.58, 0.49, and 0.50, respectively. We observe that extreme-spammers exists in the real-world datasets.

BP-based algorithms, such as BP and deepBP, are robust to extreme-spammer who labels uniformly across all tasks than MF-based algorithms as show in Section 4.1. To show that these extreme-spammers not only exist in the imaginary scenarios, we investigate the existence of extreme-spammers in real-world datasets. Figure 10 describes histograms of  $r$ , the number of tasks per worker, from the temp dataset (Snow et al., 2008) and the rte dataset (Snow et al., 2008). In this figure, the workers who attempt the most of the tasks are identified as an extreme-spammers. Specifically, we can identify three and four extreme-spammers in the temp and the rte dataset, respectively. As we have shown in the main experiments, a few number of extreme-spammer can mis-guide the entire inference and learning algorithms. Therefore, it is important to have robust algorithms working with these extreme-spammers.