2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834403

A Generalized Worker-Task Specialization Model for Crowdsourcing: Optimal Limits and Algorithm

Doyeon Kim*, Jeonghwan Lee† and Hye Won Chung*
*School of Electrical Engineering, †Department of Mathematical Sciences
KAIST
E-mail: {highlowzz, sa8seung, hwchung}@kaist.ac.kr

***Abstract*—Crowdsourcing has emerged as an effective platform to label data with low cost by using non-expert workers. However, inferring correct labels from multiple noisy answers on data has been a challenging problem, since the quality of answers varies widely across tasks and workers. Many existing prior works have assumed a simple model where the order of workers in terms of their reliabilities is fixed across tasks, and focused on estimating the worker reliabilities to aggregate responses with different weights. We propose a highly general crowdsourcing model in which the reliability of each worker can vary depending on the type of a given task, where the number of types d can scale in the number of tasks. In this model, we characterize the optimal sample complexity to correctly infer the unknown labels within any given accuracy, and propose an algorithm achieving the order-wise optimal result. We conduct experiments on synthetic and real datasets, and show that our algorithm outperforms the existing ones developed based on strict model assumptions.** [1]

I. Introduction

Crowdsourcing systems have allowed us to collect a large amount of useful data by assigning tasks to human workers, requesting them to provide responses to these tasks, and offering them compensations in monetary terms. The main goal of tasks in crowdsourcing lies in the reliable estimation of the unknown ground-truth labels, so-called the *crowdsourced labeling*. Low-cost human workers are often non-experts and this issue may lead to the necessity to ask redundant questions and to collect multiple answers for each task with a heterogeneity in the quality of answers across workers and tasks. Thus, it has been a challenging problem to infer the ground-truth labels from multiple noisy responses while minimizing total queries.

Many existing works have adopted simple yet meaningful model assumptions to analyze and improve the sample efficiency. In the Dawid-Skene model [6], which is the most extensively studied model in this line of work, the worker reliability is assumed to be fixed across all tasks, and various inference algorithms have been proposed to better estimate the worker reliabilities and to infer the true labels by combining the responses with proper weights via statistical aggregation rules, based on the expectation-maximization (EM) algorithm [6], [8], message-passing [10], [16], [19], [17], [12], spectral methods [22], [5], [9], and gradient descent methods [18]. In some recent works [11], [21], [4], task difficulties are additionally considered in modeling the fidelity of responses. However, all these works are based on strict assumptions that each worker is either associated with its own reliability parameter, fixed across all tasks, or the order of workers in terms of their reliabilities does not change depending on tasks.

In this paper, we propose a general model that better represents real-world data, especially when the tasks are heterogeneous and the worker reliability can vary with a given task type. Specifically, we assume that each worker and task has its own type among $[d] := \{1, \ldots, d\}$, and the reliability of a worker may change by the task type and worker type. Under this general model, the worker reliabilities can be completely changed for each task, and the main questions are how to estimate the types of tasks and workers, and how to choose proper weights for answers from each worker depending on the task type and worker type, where neither the task types nor the worker types are known. We consider a high-dimensional regime where the number d of types can scale in the number of tasks, and the framework we develop is non-asymptotic.

We first fully characterize the optimal sample complexity to infer the correct labels with any target accuracy, and then design an inference algorithm achieving the order-wise optimal sample complexity. To further demonstrate the benefits of our model and the proposed algorithm in real applications, we present experimental results both on synthetic and real datasets and show that our algorithm outperforms the existing baselines that are mainly developed based on the strict model assumptions on consistent worker reliabilities across all tasks.

The proofs of our results are available at full version [14].

II. Model and problem formulation

Let m and n be the number of tasks and workers, respectively. Let $\mathbf{a} \in \{\pm 1\}^m$ denote the vector of unknown binary labels associated with these tasks, and $\mathcal{A} \subseteq [m] \times [n]$ be the *worker-task assignment set*, *i.e.*, $(i, j) \in \mathcal{A}$ if and only if the i-th task is assigned to the j-th worker.

The *crowdsourcing system with a fidelity matrix* $\mathbf{F} \in [0,1]^{m \times n}$ is an observation model, which samples a data $(M_{ij} : (i,j) \in [m] \times [n]) \in \{-1, 0, +1\}^{m \times n}$ according to the following rule: $M_{ij} = 0$ if $(i,j) \in ([m] \times [n]) \setminus \mathcal{A}$, and

$$M_{ij} = \begin{cases} a_i & \text{with probability } F_{ij}; \\ -a_i & \text{with probability } 1 - F_{ij}. \end{cases} \tag{1}$$

[1] This research was supported by the National Research Foundation of Korea under grant 2021R1C1C11008539, and by the Ministry of Science and ICT, South Korea, under the ITRC support program grant IITP-2022-2018-0-01402, and under the Institute of Information and Communications Technology Planning & Evaluation (IITP) grant No.2020-0-00626.

978-1-6654-2159-1/22/$31.00 ©2022 IEEE

We further assume the independence of the aggregation of noisy answers $\{M_{ij} : (i,j) \in \mathcal{A}\}$.

In previous models, it is often assumed that the worker reliability is fixed across tasks. In single-coin Dawid-Skene (DS) model [6], each worker is associated with a reliability parameter r_j and $F_{ij} = r_j$ for $i \in [m]$. In some recent works, task difficulties are additionally considered in modeling $\mathbf{F}$. In [11], the task difficulty is quantified by $c_i \in [1/2, 1]$, which is the probability that a task is perceived correctly, and the fidelity matrix is modeled by $F_{ij} = c_i r_j + (1 - c_i)(1 - r_j)$. In [21], a permutation-based model is considered, where there is a fixed order of workers in terms of their reliabilities that does not change for tasks, and a fixed order of task difficulties, perceived equally by all workers. For all such models, however, the order of workers in terms of their reliabilities is still assumed to be fixed for all tasks.

We introduce a highly general model, termed by the d-type specialization model, where each worker and task is associated with a certain type in $[d]$ and the value of F_{ij} is determined by the type of i-th task and the type of j-th worker. Since it is natural to assume that worker types and task types are unknown at the crowdsourcing system, we assume that those types are independently and uniformly distributed over $[d]$. For the d-type specialization model with a reliability matrix $\mathcal{Q}(\cdot,\cdot) : [d] \times [d] \to \left[\frac{1}{2}, 1\right]$, denoted by $\mathsf{SM}(d; \mathcal{Q})$, the fidelity matrix $\mathbf{F}$ is not deterministic but stochastic with the following prior distribution of $\mathbf{F}$ over $\left[\frac{1}{2}, 1\right]^{m \times n}$:

1) A *task-type vector* $\mathbf{t} = (t_i : i \in [m])$ and a *worker-type vector* $\mathbf{w} = (w_j : j \in [n])$ are drawn independently and uniformly over $[d]^m$ and $[d]^n$, resp.;
2) The value of F_{ij} is completely determined by the pair of the i-th task type and the j-th worker type (t_i, w_j): for each $(i,j) \in [m] \times [n]$, $F_{ij} = \mathcal{Q}(t_i, w_j)$.

In this model, the order of workers in terms of their reliabilities may vary depending on the task type. The d-type specialization model was first studied in [20], but with a restrictive assumption: $\mathcal{Q}(t, w) = p > 1/2$ if $t = w$; $\mathcal{Q}(t, w) = 1/2$ otherwise, *i.e.*, the workers give answers with fidelity better than random guess only when the worker type and the task type match. We extend this model by allowing any $\mathcal{Q}$ satisfying only two assumptions below.

Assumption 1 (Weak assortativity of $\mathcal{Q}$)**.** *Let* $p^*(t) := \mathcal{Q}(t,t)$ *and* $q^*(t) := \max_{w \in [d] \setminus \{t\}} \mathcal{Q}(t, w)$ *be the matched reliability and the maximum mismatched reliability for the task type* $t \in [d]$. *Then, we have* $p^*(t) > q^*(t),\ \forall t \in [d]$.

Assumption 2 (The strong assortativity of $\Phi(\mathcal{Q})$)**.** *We define* $\Phi(\mathcal{Q})(a,b) := \frac{1}{d}\sum_{t=1}^{d} \{2\mathcal{Q}(t,a) - 1\}\{2\mathcal{Q}(t,b) - 1\}$ *for* $(a,b) \in [d] \times [d]$, *and call* $\Phi(\mathcal{Q}) : [d] \times [d] \to [0,1]$ *the collective quality correlation matrix. Also, we define* $p_m := \min_{a \in [d]} \Phi(\mathcal{Q})(a,a)$ *and* $p_u := \max_{a \neq b} \Phi(\mathcal{Q})(a,b) : a \neq b$ *denote the minimum intra-cluster collective quality correlation and the maximum inter-cluster collective quality correlation, respectively. Then, we have* $p_m > p_u$.

Assumption 1 implies that the workers whose types match the type of a given task respond more reliably than the workers of other types. In Assumption 2, the diagonal entry $\Phi(\mathcal{Q})(a,a)$ represents the average quality of the type-a worker cluster in answering over all task types. The off-diagonal entry $\Phi(\mathcal{Q})(a,b)$, where $a \neq b$, represents the quality correlation between the type-a and the type-b clusters of workers over all task types. Assumption 2 asserts that the collective quality correlation between any two workers of the same type is higher than that of any two workers of different types.

We measure the quality of an estimator $\hat{\mathbf{a}}(\cdot) : \{\pm 1\}^{\mathcal{A}} \to \{\pm 1\}^m$ by the expected fraction of labels that do not match with the ground-truth: $\mathcal{R}(\mathbf{a}, \hat{\mathbf{a}}) := \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\{\hat{a}_i(\mathbf{M}) \neq a_i\}$. The main question is to find the minimal number of queries per task, $|\mathcal{A}|/m$, required to obtain the recovery performance

$$\mathcal{R}(\mathbf{a}, \hat{\mathbf{a}}) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\{\hat{a}_i(\mathbf{M}) \neq a_i\} \leq \alpha, \tag{2}$$

for an arbitrarily given target accuracy $\alpha \in (0,1)$.

III. Performance Baselines

A. Baseline estimators

Throughout this section, we review some baseline estimators and analyze their performance under the proposed model.

1) Weighted majority voting rule: A weighted majority voting infers the ground-truth label by aggregating the responses for the i-th task with weights $\{\theta_{ij} : j \in \mathcal{A}(i)\}$: $\hat{a}_i^{\mathrm{WMV}} := \operatorname{sign}\left(\sum_{j \in \mathcal{A}(i)} \theta_{ij} M_{ij}\right)$, where $\mathcal{A}(i) := \{j \in [n] : (i,j) \in \mathcal{A}\}$ denotes the set of workers assigned to the i-th task.

2) Maximum likelihood (ML) estimator: The ML estimator takes the weight $\theta_{ij} = \log\left(\frac{F_{ij}}{1 - F_{ij}}\right)$ on M_{ij}: for each $i \in [m]$,

$$\hat{a}_i^{\mathrm{ML}} = \operatorname{sign}\left(\sum_{j \in \mathcal{A}(i)} \log\left(\frac{F_{ij}}{1 - F_{ij}}\right) M_{ij}\right). \tag{3}$$

The ML estimator (3) requires the knowledge of the fidelity matrix $\mathbf{F}$ a priori, which is impossible in practice.

3) Standard majority voting (MV) rule: The majority voting rule takes weight $\theta_{ij} = 1$ for all $j \in \mathcal{A}(i)$:

$$\hat{a}_i^{\mathrm{MV}} := \operatorname{sign}\left(\sum_{j \in \mathcal{A}(i)} M_{ij}\right), \ \forall i \in [m]. \tag{4}$$

Proposition III.1 (Statistical analysis of the majority voting)**.** *In the d-type worker-task specialization model* $\mathsf{SM}(d; \mathcal{Q})$, *it is possible to achieve the recovery accuracy* (2) *via the majority voting rule* (4) *with the average number of queries per task*

$$\frac{|\mathcal{A}|}{m} \geq \frac{1}{\min_{t \in [d]} \theta_1(t; \mathcal{Q})} \log\left(\frac{1}{\alpha}\right) \tag{5}$$

for any given target accuracy $\alpha \in \left(0, \frac{1}{2}\right]$ *(*α *may depend on* m*), where* $\theta_1(t; \mathcal{Q}) := \frac{1}{2}\left[\frac{1}{d}\sum_{w=1}^{d}\{2\mathcal{Q}(t,w) - 1\}\right]^2$.

4) Type-dependent subset-selection (SS) scheme: The last baseline is the type-dependent subset-selection scheme [20]. The idea is to exploit the answers from the workers whose type matches the given task only. Since neither task types nor worker types are known, the main task is to estimate the task type $\hat{t}_i$ and infer the set of workers among $\mathcal{A}(i)$ whose type matches $\hat{t}_i$, denoted by $\mathcal{A}_{\hat{t}_i}(i)$. Then, a_i is estimated by MV using responses from the *workers of the matched type* only:

$$\hat{a}_i^{\mathrm{SS}} := \operatorname{sign}\left(\sum_{j \in \mathcal{A}_{\hat{t}_i}(i)} M_{ij}\right), \ \forall i \in [m]. \tag{6}$$

The basic idea to infer $\mathcal{A}_{\hat{t}_i}(i)$ in [20] is to first cluster workers by sequentially comparing the similarity on responses between every pair of workers, and then find a cluster whose answers for the given task are the most biased.

Proposition III.2 (Statistical analysis of the SS scheme)**.** *In the d-type specialization model* $\mathrm{SM}(d;\mathcal{Q})$*, with* $\mathcal{Q}$ *satisfying Assumption 1 and 2, the SS scheme achieves* (2) *provided that*

$$\frac{|\mathcal{A}|}{m} \geq \min\left\{ \frac{4d \cdot \log\left(\frac{6d+3}{\alpha}\right)}{\min_{t\in[d]}\left\{\left(p^*(t)-q^*(t)\right)^2+\theta_2(t;\mathcal{Q})\right\}}, \right. \\ \left. \frac{4d \cdot \log\left(\frac{3}{\alpha}\right)}{\min_{t\in[d]} \theta_2(t;\mathcal{Q})} \right\} \tag{7}$$

for every sufficiently large d, where $m \geq C_1 \cdot \frac{n^{1+\epsilon}}{(p_m - p_u)^2}$ *for some* $C_1, \epsilon > 0$*, and* $\theta_2(t;\mathcal{Q}) := \left[2\min_{w\in[d]} \mathcal{Q}(t,w) - 1\right]^2$.

Note that $\theta_2(t;\mathcal{Q})$ is the worst-case error exponent for the task type t. This exponent appears when the task-type matching fails, and thus the aggregated responses might come from the mismatched worker cluster with the worst reliability.

B. Baseline comparison for a special model

We next discuss a specific model where the MV and the SS algorithm can strictly perform better than the other depending on a model parameter. Consider a special d-type specialization model

$$\mathcal{Q} = q\mathbf{1}_{d\times d} + (p-q)\mathbf{I}_d, \tag{8}$$

where $\frac{1}{2} \leq q < p < 1$ are universal constants [13], [20], i.e., each worker provides an answer with fidelity $p > q$ if the task type matches the worker type, and with $q \geq 1/2$ otherwise.

For the standard majority voting estimator (4), Proposition III.1 implies that the sufficient condition for (2) is

$$\frac{|\mathcal{A}|}{m} = \begin{cases} \Omega\left(\log\left(\frac{1}{\alpha}\right)\right) & \text{if } q > \frac{1}{2}; \\ \Omega\left(d^2 \log\left(\frac{1}{\alpha}\right)\right) & \text{otherwise.} \end{cases} \tag{9}$$

For the subset-selection scheme [20], Proposition III.2 implies that the subset-selection algorithm succeeds if

$$\frac{|\mathcal{A}|}{m} = \begin{cases} \Omega\left(d\log\left(\frac{1}{\alpha}\right)\right) & \text{if } q > \frac{1}{2}; \\ \Omega\left(d\log\left(\frac{d}{\alpha}\right)\right) & \text{otherwise.} \end{cases} \tag{10}$$

By (9) and (10), the majority voting rule (4) and the subset-selection algorithm (6) do not consistently beat each other. In order to understand the reason, consider the spammer/hammer model [10]: the j-th worker is referred to as a *hammer* for the i-th task if $F_{ij} = 1$; a *spammer* if $F_{ij} = \frac{1}{2}$. If all workers are nearly hammers, *i.e.*, $\mathcal{Q}(t,w) \approx 1$ for all $(t,w) \in [d]\times[d]$, the majority voting using all responses outperforms the subset-selection scheme since the subset-selection scheme abandons $\left(\frac{d-1}{d}\right)$-fraction of answers. On the other hand, if we consider the regime where $q^*(t) \approx \frac{1}{2}$ and $p^*(t) - q^*(t) = \Theta(1)$ for all $t \in [d]$, then all workers with types different from a given task type are nearly spammers. For this case, the subset-selection scheme is far better than the majority voting, since the majority voting does not rule out the dominant random noisy answers. Indeed, as shown in (9) and (10), the subset-selection scheme requires d times more queries than the standard majority voting if $q > 1/2$, while it requires only $1/d$ times queries if $q = 1/2$.

The main question is how to design an algorithm achieving the superior performance in both parameter regimes when the model parameters are unknown, which is common in practice.

IV. Main Results

A. Fundamental limits

We establish the fundamental limits on the required number of queries. The optimality result will be characterized in terms of the *minimax risk*: $\mathcal{R}^*(\mathcal{A}) := \inf_{\hat{\mathbf{a}}}\left(\sup_{\mathbf{a}\in\{\pm1\}^m} \mathcal{R}\left(\mathbf{a},\hat{\mathbf{a}}\right)\right)$. We first present a sufficient condition from ML estimator (3).

Theorem IV.1 (Information-theoretic achievability)**.** *For any target accuracy* $\alpha \in \left(0, \frac{1}{2}\right]$*, the ML estimator* (3) *achieves the desired recovery accuracy* (2)*,* $\mathcal{R}^*(\mathcal{A}) \leq \mathcal{R}\left(\mathbf{a}, \hat{\mathbf{a}}^{\mathrm{ML}}\right) \leq \alpha$*, under* $\mathrm{SM}\left(d;\mathcal{Q}\right)$*, provided that the worker-task assignment set* $\mathcal{A} \subseteq [m]\times[n]$ *satisfies*

$$\min_{i\in[m]} |\mathcal{A}(i)| \geq \frac{1}{\gamma_1\left(d;\mathcal{Q}\right)} \log\left(\frac{1}{\alpha}\right), \tag{11}$$

where $\gamma_1\left(d;\mathcal{Q}\right) := \log\left(\frac{d}{2\max_{t\in[d]}\left(\sum_{w=1}^{d}\sqrt{\mathcal{Q}(t,w)(1-\mathcal{Q}(t,w))}\right)}\right)$.

Next, the corresponding impossibility result is summarized.

Theorem IV.2 (Statistical impossibility)**.** *For any* $\alpha \in \left(0, \frac{1}{8}\right]$ *and worker-task assignment set* $\mathcal{A} \subseteq [m]\times[n]$ *satisfying*

$$\gamma_2\left(d;\mathcal{Q}\right)\left(\frac{|\mathcal{A}|}{m}\right) + \Gamma\left(d;\mathcal{Q}\right)\sqrt{\frac{|\mathcal{A}|}{m}} < \log\left(\frac{1}{4\alpha}\right), \tag{12}$$

no inference methods based on the worker-task assignment set $\mathcal{A}$ *can achieve* (2) *in the model* $\mathrm{SM}(d;\mathcal{Q})$*. Here,* $\gamma_2\left(d;\mathcal{Q}\right) := \log\left(\frac{d^2}{2\sum_{(t,w)\in[d]\times[d]}\sqrt{\mathcal{Q}(t,w)(1-\mathcal{Q}(t,w))}}\right)$*, and* $\Gamma(d;\mathcal{Q})$ *denotes the log-odds of the maximum reliability, that is,* $\Gamma\left(d;\mathcal{Q}\right) := \log\left(\frac{\max_{(t,w)\in[d]\times[d]}\mathcal{Q}(t,w)}{1-\max_{(t,w)\in[d]\times[d]}\mathcal{Q}(t,w)}\right)$.

Note that the error exponents for the information-theoretic upper bound $\gamma_1\left(d;\mathcal{Q}\right)$ and the lower bound $\gamma_2\left(d;\mathcal{Q}\right)$ coincide when $\frac{1}{d}\sum_{w=1}^{d}\sqrt{\mathcal{Q}(t,w)\left(1-\mathcal{Q}(t,w)\right)}$ are equal for all $t \in [d]$, *i.e.*, when all task types have the same overall difficulty, when averaged over all worker types.

Remark 1 (Fundamental limits under a special model)**.** Under the special model (8), by Theorem IV.1, the recovery accuracy (2) is achievable via the ML estimator (3) if

$$\frac{|\mathcal{A}|}{m} = \begin{cases} \Omega\left(\log\left(\frac{1}{\alpha}\right)\right) & \text{if } q > \frac{1}{2}; \\ \Omega\left(d\log\left(\frac{1}{\alpha}\right)\right) & \text{otherwise}, \end{cases} \tag{13}$$

while it is statistically impossible by Theorem IV.2 whenever

$$\frac{|\mathcal{A}|}{m} = \begin{cases} o\left(\log\left(\frac{1}{\alpha}\right)\right) & \text{if } q > \frac{1}{2}; \\ o\left(d\log\left(\frac{1}{\alpha}\right)\right) & \text{if } q = \frac{1}{2} \text{ and } \log\left(\frac{1}{\alpha}\right) = \Omega(d); \\ o\left(\left(\log\left(\frac{1}{\alpha}\right)\right)^2\right) & \text{if } q = \frac{1}{2} \text{ and } \log\left(\frac{1}{\alpha}\right) = o(d). \end{cases} \tag{14}$$

The order analyses (13) and (14) match up to a constant factor when either $q > \frac{1}{2}$ or $q = \frac{1}{2}$ and $\log\left(\frac{1}{\alpha}\right) = \Omega(d)$. From (9) and (10), the order-wise optimal result is achievable by the majority voting if $q > 1/2$ and by the subset-selection scheme if $q = 1/2$ and $\log\left(\frac{1}{\alpha}\right) = \Omega(d)$. We will develop an algorithm achieving the order-wise optimal result for both cases.

B. Proposed algorithm

Our main algorithm consists of two stages, where *Stage #1* recovers the hidden clusters of workers, and *Stage #2* uses the recovered cluster structure to choose the cluster of the matched type for each task and then use this information to infer true labels via weighted majority voting. Our algorithm takes the advantages of both the MV and the subset-selection algorithm.

Algorithm 1 (The proposed inference algorithm)**.**

1) *Stage #1*: (Worker clustering via convex optimization).
   (a) Let $\mathcal{S} \subseteq [m]$ be a set of randomly chosen r tasks and assign each task in $\mathcal{S}$ to all n workers. Based on the responses $\mathbf{M}_{i*} = (M_{ij} : j \in [n])$ for $i \in \mathcal{S}$, we define the *similarity matrix* $\mathbf{A} := \mathcal{P}_{\text{off-diag}}\left(\sum_{i\in\mathcal{S}} \mathbf{M}_{i*}^\top \mathbf{M}_{i*}\right)$, where $\mathcal{P}_{\text{off-diag}}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ zeroes out all diagonal entries of an input $n \times n$ matrix;
   (b) Solve the following semi-definite program:

$$\begin{aligned} &\max_{\mathbf{X}\in\mathbb{R}^{n\times n}} \quad \langle \mathbf{A} - \eta \mathbf{1}_{n\times n}, \mathbf{X} \rangle \\ &\text{subject to } \mathbf{X} \succeq \mathbf{0}; \quad \langle \mathbf{I}_n, \mathbf{X} \rangle = n; \\ &\qquad\qquad 0 \le X_{ij} \le 1, \ \forall (i,j) \in [n] \times [n], \end{aligned} \tag{15}$$

   where $\eta > 0$ is a tuning parameter. Perform the approximate k-medoids clustering (*Algorithm 1* in [7]) on the optimal solution $\hat{\mathbf{X}}_{\text{SDP}}$ to the SDP (15) to extract d worker clusters $\left\{\hat{\mathcal{W}}_1, \cdots, \hat{\mathcal{W}}_d\right\}$, when d is known;
   (c) For each task $i \in [m] \setminus \mathcal{S}$ and cluster $w \in [d]$, assign task i to randomly selected l workers from each $\hat{\mathcal{W}}_w$.
2) *Stage #2*: (Task-type matching and label inference).
   (a) For every $i \in [m]$, we first select $\mathcal{A}_w(i) \in \binom{\mathcal{A}(i)\cap\hat{\mathcal{W}}_w}{l}$[2] for every $w \in [d]$ and define $\mathcal{A}'(i) := \bigcup_{w=1}^{d} \mathcal{A}_w(i) \subseteq \mathcal{A}(i)$. Then, we estimate the task type of $i \in [m]$ by computing $\hat{t}_i := \text{argmax}_{w\in[d]} \left|\sum_{j\in\mathcal{A}_w(i)} M_{ij}\right|$;
   (b) Designate weights $\boldsymbol{\theta}_{i*} = (\theta_{ij} : j \in \mathcal{A}'(i))$ for each $i \in [m]$ as per the following rule:

$$\theta_{ij} := \begin{cases} 1 & \text{if } j \in \mathcal{A}_{\hat{t}_i}(i); \\ \frac{1}{\sqrt{d-1}} & \text{otherwise}, \end{cases} \tag{16}$$

   and infer the label a_i via the weighted majority voting using weights (16): $\hat{a}_i := \text{sign}\left(\sum_{j\in\mathcal{A}'(i)} \theta_{ij} M_{ij}\right)$.

Theorem IV.3 (Statistical analysis of Alg.1)**.** *We consider the same setting with Proposition III.2. Then, the same result holds when we replace the error exponent $\theta_2(d; \mathcal{Q})$ by*

$$\theta_3(t;\mathcal{Q}) := \frac{1}{2}\left[\frac{1}{\sqrt{d-1}}\sum_{w=1}^{d}\{2\mathcal{Q}(t,w) - 1\} + \left(1 - \frac{1}{\sqrt{d-1}}\right)\left\{2\min_{w\in[d]}\mathcal{Q}(t,w) - 1\right\}\right]^2, \tag{17}$$

when d is sufficiently large and $m = \omega\left(\frac{n^3}{(p_m - p_u)^2}\right)$.

Remark 2 (Comparison of the sample complexity)**.** We first compare the sample complexity of Alg.1 against that of the subset-selection scheme. As $\theta_3(t;\mathcal{Q}) \ge \frac{\left(1+\sqrt{d-1}\right)^2}{2}\theta_2(t;\mathcal{Q})$, the error exponent $\theta_3(t;\mathcal{Q})$ of Alg.1 is strictly larger than the error exponent $\theta_2(t;\mathcal{Q})$ of the SS scheme. Thus, Alg.1 can be viewed as a strict improvement over the SS algorithm.

We then compare the sample complexity of Alg.1 and the standard majority voting rule. Since $\theta_3(t;\mathcal{Q}) \gtrsim d\cdot\theta_1(t;\mathcal{Q})$ for every $t \in [d]$, it follows that $\frac{4d\cdot\log\left(\frac{3}{\alpha}\right)}{\min_{t\in[d]}\theta_3(t;\mathcal{Q})} \lesssim \frac{\log\left(\frac{1}{\alpha}\right)}{\min_{t\in[d]}\theta_1(t;\mathcal{Q})}$ as $d \to \infty$. This implies that the sample complexity of Alg.1 is either smaller than or equal to that of the standard majority voting in an order-wise sense.

Remark 3 (Order-wise optimality of Alg.1 under the special model)**.** Let us revisit the special model (8). By Theorem IV.3, the recovery accuracy (2) is achievable by Alg.1 provided that

$$\frac{|\mathcal{A}|}{m} = \begin{cases} \Omega\left(\log\left(\frac{1}{\alpha}\right)\right) & \text{if } q > \frac{1}{2}; \\ \Omega\left(d\log\left(\frac{d}{\alpha}\right)\right) & \text{otherwise}, \end{cases} \tag{18}$$

which meets the bound (13) of the sample complexity per task required for the ML estimator (3) in both regimes $q > \frac{1}{2}$ and $q = \frac{1}{2}$ (up to logarithmic factors when $\alpha = \omega(1/d)$).

Remark 4 (Main differences from SS algorithm)**.** Alg.1 has two remarkable differences from the subset-selection (SS) algorithm [20]. First, the SS algorithm recovers the hidden type structure of workers by counting the same responses between every pair of workers sequentially, while Alg.1 unveils the type structure by solving the SDP (15). The SDP relaxation approach has been extensively used in the community detection literature [1], [2], [3], [15]. This approach makes the clustering stage more robust against the model variants and allows an easier parameter tuning for η in (15). Second, the SS algorithm estimates the ground-truth labels via the MV using answers from matched workers only. Alg.1, on the other hand, infers the labels via the weighted MV by utilizing all responses with proper weights based on the result from task-type matching.

[2] $\binom{\mathcal{X}}{l}$ denotes the set of all size-l subsets of the set $\mathcal{X}$.

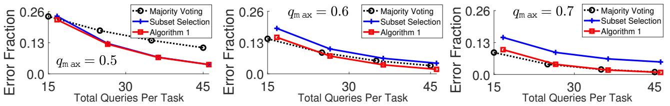

Fig. 1: Synthetic data experiment for $q_{\max} \in \{0.5, 0.6, 0.7\}$.

Remark 5 (Weights on responses)**.** We discuss the reason for the choice of specific weights (16). Suppose that we choose weights $\boldsymbol{\theta}_{i*} = (\theta_{ij} : j \in \mathcal{A}'(i))$, where $\theta_{ij} := 1$ if $j \in \mathcal{A}_{\hat{t}_i}(i)$ and $\theta_{ij} := \delta(d)$ otherwise, for some function $\delta(\cdot) : \mathbb{N} \to \mathbb{R}_+$. Alg.1 with weights $\boldsymbol{\theta}_{i*}$, $i \in [m]$, achieves the target accuracy (2) in the model (8) if

$$\frac{|\mathcal{A}|}{m} \geq \min\left\{ \frac{4d \cdot \log\left(\frac{6d+3}{\alpha}\right)}{\min\{\pi_m(d;\mathcal{Q}), (p-q)^2 + \pi_u(d;\mathcal{Q})\}}, \frac{4d \cdot \log\left(\frac{3}{\alpha}\right)}{\min\{\pi_m(d;\mathcal{Q}), \pi_u(d;\mathcal{Q})\}} \right\}, \tag{19}$$

where $\pi_m(d;\mathcal{Q})$ and $\pi_u(d;\mathcal{Q})$ denote the error exponents of matched type and mismatched type, respectively, such that

$$\pi_m(d;\mathcal{Q}) = \begin{cases} \Theta\left(\frac{1+d^2\{\delta(d)\}^2}{1+d\{\delta(d)\}^2}\right) & \text{if } q > \frac{1}{2}; \\ \Theta\left(\frac{1}{1+d\{\delta(d)\}^2}\right) & \text{otherwise,} \end{cases}$$
$$\pi_u(d;\mathcal{Q}) = \begin{cases} \Theta\left(\frac{1+d^2\{\delta(d)\}^2}{1+d\{\delta(d)\}^2}\right) & \text{if } q > \frac{1}{2}; \\ \Theta\left(\frac{\{\delta(d)\}^2}{1+d\{\delta(d)\}^2}\right) & \text{otherwise.} \end{cases} \tag{20}$$

To make (19) meet the desired order (13), we need to choose $\delta(\cdot)$ to satisfy $\delta(d) \asymp 1/\sqrt{d}$. For the sake of simplicity, we choose $\delta(d) := 1/\sqrt{d-1}$ as (16).

V. Empirical results

We provide the empirical performance comparison. The inference quality is measured by the fraction of labels that do not match with the ground-truth, *i.e.*, $\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left(\hat{a}_i(\mathbf{M}) \neq a_i\right)$.

A. Experiments with synthetic data

We compare the empirical performance of the proposed algorithm with two main baselines, the standard majority voting (MV) rule and the subset-selection (SS) algorithm in Fig.1, when $(m, n, r, d) = (25000, 100, 300, 5)$ with varying $(p_{\min}, q_{\max})$, where $p_{\min} := \min_{a\in[d]} \mathcal{Q}(a,a)$ and $q_{\max} := \max_{a\neq b} \mathcal{Q}(a,b)$. For the fixed value $p_{\min} = 0.9$, as $q_{\max}$ increases, the quality difference between the answers from the matched workers and the mismatched workers decreases.

As shown in Fig.1, the performance of the SS is better than that of the MV for a smaller $q_{\max}$, while that of the MV gets improved for a larger $q_{\max}$. Our algorithm attains consistently the best performances across all considered parameters as our theory implies (Remark 3).

B. Experiments with real-world data

We conduct experiments on real data collected from MTurk. Binary tasks using 600 images of athletes are designed where each a quarter is from one of four sports types ($d = 4$): football, baseball, soccer and basketball. Each human intelligent task (HIT) is designed to contain 80 images, where four types are evenly covered with 20 randomly sampled images from each type. We ask whether the athlete in each image is over 30 years old. Total 60 HITS are assigned to 60 workers.

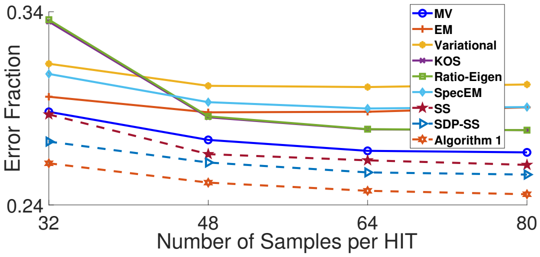

Fig. 2: Experiment with real data from the Amazon MTurk

We first check whether the collected real data indeed follows a *type structure*. Since only the task types are known, we infer the ground-truth worker types based on the correct answer rate of each worker on each task type, calculated using the ground-truth label information. Then, the reliability matrix $\mathcal{Q}$ can be computed by averaging the empirical correct answer rate for each task-worker type pair $(t,w) \in [d] \times [d]$:

$$\mathcal{Q} = \begin{bmatrix} 0.8647 & 0.5467 & 0.4962 & 0.5700 \\ 0.5765 & 0.9000 & 0.4846 & 0.5833 \\ 0.5573 & 0.5344 & 0.7825 & 0.7025 \\ 0.6131 & 0.5611 & 0.4542 & 0.9379 \end{bmatrix}.$$

The diagonal entries are larger than the off-diagonal entries, showing the existence of type structure in this real-world data.

In Fig.2, we compare our algorithm with other algorithms, including EM [6], Variational [17], KOS [10], Ratio-Eigen [5], and specEM [22], all of which are developed under the Dawid-Skene model. The performances of MV and SS are also plotted. For ablation study of our algorithm, which has two prominent differences from the SS, we also consider the SS scheme with only clustering stage replaced by our SDP clustering (SDP-SS). From each HIT of 80 answers, $[32, 48, 64]$-answers are randomly sampled total 100 times and used to calculate the empirical average for the fraction of errors $\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left(\hat{a}_i(\mathbf{M}) \neq a_i\right)$, plotted in Fig.2. From this plot, we observe that Alg.1 outperforms all the other algorithms developed based on strict model assumptions, and the benefits come from both the improved clustering (*Stage #1*) and the weighted majority voting with properly chosen weights (*Stage #2*).

VI. Conclusion

We explored the crowdsourced labeling problem in a highly generalized d-type specialization model. Our algorithm infers the types of workers and tasks, and use this information to utilize all the responses from workers with a proper weighting scheme. It achieves the order-wise optimal result across general parameter regimes, and also empirically performs better than the existing algorithms for real-world datasets.

References

[1] Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

[2] T Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015.

[3] Yudong Chen, Sujay Sanghavi, and Huan Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, 2014.

[4] Hye Won Chung, Ji Oon Lee, and Alfred O Hero. Fundamental limits on data acquisition: Trade-offs between sample complexity and query difficulty. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 681–685. IEEE, 2018.

[5] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd international conference on World Wide Web*, pages 285–294, 2013.

[6] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

[7] Yingjie Fei and Yudong Chen. Exponential error rates of sdp for block models: Beyond grothendieck's inequality. *IEEE Transactions on Information Theory*, 65(1):551–571, 2018.

[8] Chao Gao and Dengyong Zhou. Minimax optimal convergence rates for estimating ground truth from crowdsourced labels. *arXiv preprint arXiv:1310.5764*, 2013.

[9] Arpita Ghosh, Satyen Kale, and Preston McAfee. Who moderates the moderators?: crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 167–176. ACM, 2011.

[10] David R Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62(1):1–24, 2014.

[11] Ashish Khetan and Sewoong Oh. Achieving budget-optimality with adaptive schemes in crowdsourcing. *Advances in Neural Information Processing Systems*, 29:4844–4852, 2016.

[12] Daesung Kim and Hye Won Chung. Binary classification with XOR queries: Fundamental limits and an efficient algorithm. *IEEE Transactions on Information Theory*, 2021.

[13] Doyeon Kim and Hye Won Chung. Crowdsourced labeling for worker-task specialization model. In *IEEE International Symposium on Information Theory, ISIT*, pages 3191–3195. IEEE, 2021.

[14] Doyeon Kim, Jeonghwan Lee, and Hye Won Chung. A worker-task specialization model for crowdsourcing: Efficient inference and fundamental limits. *arXiv preprint arXiv:2111.12550*, 2021.

[15] Jeonghwan Lee, Daesung Kim, and Hye Won Chung. Robust hypergraph clustering via convex relaxation of truncated MLE. *IEEE Journal on Selected Areas in Information Theory*, 1(3):613–631, 2020.

[16] Hongwei Li and Bin Yu. Error rate bounds and iterative weighted majority voting for crowdsourcing. *arXiv preprint arXiv:1411.4086*, 2014.

[17] Qiang Liu, Jian Peng, and Alexander T Ihler. Variational inference for crowdsourcing. In *Advances in neural information processing systems*, pages 692–700, 2012.

[18] Yao Ma, Alexander Olshevsky, Csaba Szepesvari, and Venkatesh Saligrama. Gradient descent for sparse rank-one matrix completion for crowd-sourced aggregation of sparsely interacting workers. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3335–3344, 2018.

[19] Jungseul Ok, Sewoong Oh, Jinwoo Shin, and Yung Yi. Optimality of belief propagation for crowdsourced classification. In *International Conference on Machine Learning*, pages 535–544, 2016.

[20] Devavrat Shah and Christina Lee. Reducing crowdsourcing to graphon estimation, statistically. In *International Conference on Artificial Intelligence and Statistics*, pages 1741–1750, 2018.

[21] Nihar B Shah, Sivaraman Balakrishnan, and Martin J Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*, 67(6):4162–4184, 2020.

[22] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Advances in neural information processing systems*, pages 1260–1268, 2014.