2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834340

No Rose for MLE: Inadmissibility of MLE for Evaluation Aggregation Under Levels of Expertise

Charvi Rastogi[1], Ivan Stelmakh[1], Nihar Shah[1,2], Sivaraman Balakrishnan[1,3]
Machine Learning Department[1], Computer Science Department[2], Department of Statistics and Data Science[3]
Carnegie Mellon University

***Abstract*—A number of applications including crowdsourced labeling and peer review require aggregation of labels or evaluations sourced from multiple evaluators. There is often additional information available pertaining to the evaluators' expertise. A natural approach for aggregation is to consider the widely studied Dawid-Skene model (or its extensions incorporating evaluators' expertise), and employ the standard maximum likelihood estimator (MLE). While MLE is in general widely used in practice and enjoys a number of appealing theoretical guarantees, in this work we provide a negative result for the MLE. Specifically, we prove that the MLE is asymptotically inadmissible for a special case of evaluation aggregation with expertise level information. We show this by constructing an alternative estimator that we show is significantly better than the MLE in certain parameter regimes and at least as good elsewhere. Finally, simulations reveal that our findings may hold in more general conditions than what we theoretically analyze.**

***Index Terms*—MLE, admissibility, crowdsourcing**

I. Introduction

A number of applications involve evaluations from multiple people with varying levels of expertise, and an eventual objective of aggregating the different evaluations to obtain a final decision. For instance, in peer-review, multiple reviewers provide their evaluation regarding the acceptance of the submission and their expertise on the submission matter. Another instance is found in crowdlabeling, where multiple crowdworkers provide labels for the same question. Additionally, one often has access to the evaluators' level of expertise, for instance, from their known expertise [15], self-reported confidence [17] or their prior approval rating [20]. Other such applications include decision-making in university admissions, grant allotments etc., where the quality of individual decisions obtained generally varies across individuals because of varying levels of expertise. Each of the aforementioned problems involves aggregation of multiple evaluations with varying expertise.

Moreover, in such settings, it is frequently the case that the set of evaluators are *deliberately* chosen in a certain manner based on their expertise levels. As an example, in the peer-review process of the AAAI conference on Artificial Intelligence in 2022 [3], due to lack of sufficient senior reviewers, each paper was assigned one senior and one junior reviewer. Similarly, in crowdlabeling, budget constraints impose the need for balancing out high expertise (but more expensive) and low expertise (but cheaper) crowdworkers.

There is a vast literature on the problem of aggregation of multiple evaluations in crowdsourcing [1], [5], [6], [9], [10], [16], [18], [23], [25]. The bulk of this past work is based on the classical Dawid-Skene model [2], in which each evaluator is associated with a single scalar parameter corresponding to their probability of correctness. While the Dawid-Skene model does not incorporate expertise levels, a natural extension [13] incorporates them with separate parameters for each expertise level.

Dawid and Skene [2] propose the maximum likelihood estimator (MLE) for estimation. They use the Expectation-Maximization algorithm to approximately compute the MLE. The correct answers and the evaluator's parameter for correctness are jointly estimated by maximizing the likelihood of the observed evaluations. This MLE-based approach has had huge empirical success [14], [19], [24]. Moreover, theoretical analyses [5] have shown that global optimal solutions of the MLE can achieve minimax rates of convergence in simplified scenarios such as "one-coin" Dawid-Skene. Paralelly, computationally efficient approximations of the MLE have proven to be useful for crowdlabelling, with many of the desired properties of the MLE [23]. Prior work on crowdlabeling with multiple expertise levels [13] also uses MLE for label estimation. With this motivation, we focus on analyzing the MLE in our problem of aggregating evaluations with multiple expertise levels. Our work contributes to the body of

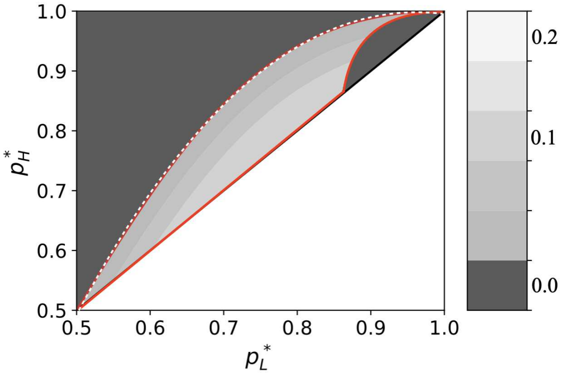


Fig. 1: Pictorial illustration of the main theoretical results: risk of MLE minus risk of our proposed estimator for different parameter values. The two axes represent the two latent (nuisance) parameters. The MLE performs significantly worse than our constructed estimator in the light gray region enclosed within the red lines, whereas everywhere else above the diagonal our estimator is asymptotically as good as the MLE.

literature on Neyman-Scott problems [7], [12] that focus on the behavior of MLE where the number of nuisance parameters grows with the size of the problem.

We focus on objective tasks involving binary choices, meaning that each question or task is associated with a single correct binary answer. We consider the extension of the Dawid-Skene model from prior work [13] which incorporates multiple expertise levels, in a simplified form. Our main contribution is a surprising negative result of asymptotic inadmissibility of the MLE in this context. Specifically, we consider a setting wherein each question is evaluated by exactly two low-level experts (or non-experts) and one (high-level) expert. We prove that MLE is asymptotically inadmissible even in this simplified setting. To prove this result, we construct an alternative polynomial-time-computable estimator and show that for all possible parameter values, the alternative estimator is as good as or better than the MLE. Importantly, for some parameter values, we show that the risk of MLE is higher than that of our estimator by a positive constant.

We pictorially illustrate this in Figure 1. For parameter values in the light gray region our estimator is significantly better than MLE, and for parameter values lying in the dark gray region our estimator is as good as MLE. We subsequently provide simulations to qualitatively show that this finding extends to other combinations of expertise evaluations in a finite sample setting.

II. Problem Setup

Let m denote the number of questions. We assume every question has two possible answers, denoted by 0 and 1, of which exactly one is correct. Each question is answered by multiple evaluators, and for each question-answer pair, we have access to the evaluator's level of expertise in the corresponding question; examples of such expertise level include the evaluator's self-reported confidence or their seniority in the application domain. We assume there are two expertise levels, which we refer to as *low*, denoted by L, and *high*, denoted by H. Under this expertise-level information, we model the question-answering as follows.

We will show that MLE is asymptotically inadmissible even in a simplified setting: we consider the setting where each question has exactly two evaluators with a low level of expertise and one evaluator with a high level of expertise, and that the probability of correctness is governed only by the expertise level. For each question without loss of generality, we assume that the first two evaluators have low expertise level and the third evaluator has a high expertise level. For any question $i \in [m]$, we let x_i^* denote the correct answer. The evaluation of the j^{th} evaluator ($j \in \{1,2,3\}$) for the i^{th} question is denoted by y_{ij}. The probability of correctness, $\mathbb{P}(y_{ij} = x_i^*)$ depends on the associated expertise level of the evaluator, and is independent of all else. Specifically, we assume existence of two *unknown* values $p_L^*, p_H^* \in [0,1]$ that govern the correctness probabilities of low and high expertise evaluators respectively. We assume that

$$y_{ij} = \begin{cases} x_i^* & \text{wp } p, \\ 1 - x_i^* & \text{wp } 1-p, \end{cases} \tag{1}$$

where $p = p_L^*$ for $j \in \{1,2\}$ and $p = p_H^*$ for $j = 3$. We further assume that $0.5 \leq p_L^* \leq p_H^* \leq 1$, which indicates that the evaluators are not adversarial [4], [16], [22], and that the high-expertise evaluator answers correctly with a probability at least as high as that for a low-expertise evaluator [8], [11], [21]. We make the standard assumption that for all $i \in [m]$ and $j \in [3]$, given the values of x_i^* and p_L^*, p_H^*, the evaluations y_{ij} are mutually independent.

For ease of exposition subsequently in the paper, for all $i \in [m]$ we introduce the notation $y_{L_i} := y_{i1} + y_{i2} \in \{0,1,2\}$ and $y_{H_i} := y_{i3} \in \{0,1\}$.

Evaluation metric: Consider any estimator $\widehat{x} : \{0,1\}^{3\times m} \rightarrow \{0,1\}^m$ as a function that maps the received evaluations to answers for all questions. We let $\widehat{x}_i$ denote the output of the estimator for question $i \in [m]$ wherein we drop the dependence on y from the notation for brevity. We then evaluate any estimator $\widehat{x}$ in terms of the 0-1 loss, and focus on the risk:

$$R(\widehat{x}) = \mathbb{E}_{\{y_{ij}\}_{(i,j)\in[m]\times[3]}}\left[\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(\widehat{x}_i \neq x_i^*\right)\right]. \qquad (2)$$

Note that the risk for any estimator lies in the interval $[0,1]$.

The goal of any estimator is to minimize the risk (2). In this setting, a widely studied and used estimator is the MLE. In this work, we provide a negative result for the MLE. We first formally describe the MLE for our problem.

Maximum likelihood estimator (MLE): The values p_L^*, p_H^* are unknown, and thus MLE simultaneously estimates the correct answers x^* and the values p_L^*, p_H^*. Given answers $\vec{y}_L \in \{0,1,2\}^m$ and $\vec{y}_H \in \{0,1\}^m$, under our model (1), the negative log-likelihood $W(\vec{x}, p_L, p_H, \vec{y}_L, \vec{y}_H)$ is given by

$$W(\vec{x}, p_L, p_H, \vec{y}_L, \vec{y}_H) = \sum_{i=1}^{m}\Big((y_{H\,i} - x_i)^2 \log\frac{p_H}{1-p_H} - \log p_H + (y_{L\,i} + 2(1-y_{L\,i})x_i)\log\frac{p_L}{1-p_L} - 2\log p_L\Big). \qquad (3)$$

The MLE minimizes the negative log-likelihood function (3) to obtain an estimate of the probability values, denoted by $\widehat{p}_L, \widehat{p}_H$ and estimator of the correct answers denoted by $\widehat{x}_{\text{MLE}} : \{0,1,2\}^m \times \{0,1\}^m \to \{0,1\}^m$, where $\widehat{x}_{\text{MLE}_i}$ denotes the estimate for the i^{th} question. Thus, we have

$$\widehat{x}_{\text{MLE}}, \widehat{p}_L, \widehat{p}_H \in \underset{\substack{\vec{x}\in\{0,1\}^m;\\ p_L, p_H \in [0.5,1]^2;\\ p_L \leq p_H}}{\arg\min}\; W(\vec{x}, p_L, p_H, \vec{y}_L, \vec{y}_H), \qquad (4)$$

where for concreteness we assume that for all $i \in [m]$ the estimator $\widehat{x}_{\text{MLE}_i}$ breaks ties in favour of $y_{H\,i}$.

III. Main Result

In this section, we provide our main result that the MLE is asymptotically inadmissible. In order to prove this result, we construct another estimator which we call the plug-in estimator.

A. Proposed estimator

As an intermediary in constructing the plug-in estimator, we first introduce and analyze an estimator we call the oracle MLE.

Oracle MLE: The oracle MLE is an estimator that is assumed to have access to the true values p_L^* and p_H^* (and is hence not realizable in our problem setting). It computes the maximum likelihood estimate $\widehat{x}_{\text{OMLE}}$ given p_L^* and p_H^* as:

$$\widehat{x}_{\text{OMLE}} \in \underset{\vec{x}\in\{0,1\}^m}{\arg\min}\; W(\vec{x}, p_L^*, p_H^*, \vec{y}_L, \vec{y}_H). \qquad (5)$$

Observe that with the true p_L^*, p_H^*, the objective function for each question can be treated separately. In the following lemma, we characterise the estimation by oracle MLE. We will see that, for all questions, it either goes with the high expertise evaluation or goes with the majority vote of the three evaluators.

Lemma 1. *For any given value of $p_L^*, p_H^* \in [0.5,1]^2$ with $p_L^* \leq p_H^*$ the solution of (5), for all $i \in [m]$ is given by $\widehat{x}_{OMLE_i} = f_{t^*}(y_{L\,i}, y_{H\,i})$, defined as follows. For any question i, let $a_i \in \{0,1,2\}$ denote the number of low expertise evaluations that agree with the high expertise evaluation, that is, $a_i = \sum_{j=1}^{2}\mathbb{I}(y_{ij} = y_{H\,i})$. Let $t^* \in \{1,2\}$ be defined for $p_L^*, p_H^* \in (0.5,1)^2$ as*

$$t^* = \max\left(\left\lceil \frac{1}{2}\left(2 - \frac{\log\frac{p_H^*}{1-p_H^*}}{\log\frac{p_L^*}{1-p_L^*}}\right)\right\rceil, 0\right) + 1, \qquad (6)$$

and, if $p_L^* = 0.5$ or $p_H^* = 1$ we set $t^* = 1$. Now, we have

$$f_{t^*}(y_{L\,i}, y_{H\,i}) = \begin{cases} 1 - y_{H\,i} & \textit{if } a_i + 1 < t^* \\ y_{H\,i} & \textit{otherwise.} \end{cases} \qquad (7)$$

We pictorially illustrate the operation of the oracle MLE in Figure 1, where for (p_L^*, p_H^*) to the left of the red dashed line it picks $t^* = 1$ and to the right of this line it picks $t^* = 2$.

Next, we present our constructed estimator, the plug-in estimator using the functional form derived in Lemma 1.

Plug-in estimator: This is a two-stage polynomial-time-computable estimator and is described in Algorithm 1. In the first stage (steps 1, 2 and 3 of Algorithm 1), the probability values p_L^* and p_H^* are estimated (with estimates denoted as $\widetilde{p}_L$ and $\widetilde{p}_H$) by measuring the agreement between the two low expertise evaluations, and one low and one high expertise evaluation respectively, for $\sqrt{m}$ questions. In the second stage (step 4 and output of Algorithm 1), $\widetilde{p}_L$ and $\widetilde{p}_H$ are plugged-in to the MLE objective function (3) to get the estimator $\widehat{x}_{\text{PI}}$. The functional form of the output of Algorithm 1 — specifically, (10) and $\widehat{x}_{\text{PI}}$ — is based on the form of the oracle MLE derived in Lemma 1. We note that purpose of sample-splitting in Algorithm 1 is for showing theoretical

```
Input: $m$ and $\{y_{ij}\}_{i\in[m],j\in[3]}$, where recall that
  $y_{Li} = y_{i1} + y_{i2}, y_{Hi} = y_{i3}$ for all $i \in [m]$.
(1) Define $\mu_L = \frac{2}{\sqrt{m}} \sum_{i=1}^{\sqrt{m}/2} \mathbb{I}(y_{i1} = y_{i2})$.
  Compute $\widetilde{p}_L$ as
      $\widetilde{p}_L = 0.5\left(1 + \sqrt{\max\{2\mu_L - 1, 0\}}\right).$    (8)
(2) Define $\mu_H = \frac{2}{\sqrt{m}} \sum_{i=\sqrt{m}/2+1}^{\sqrt{m}} \mathbb{I}(y_{i1} = y_{i3})$.
  Compute $\widetilde{p}_H$ as
      $\widetilde{p}_H = \min\left\{1, \frac{\widetilde{p}_L + \mu_H - 1}{2\widetilde{p}_L - 1}\right\}.$    (9)
(3) If $\widetilde{p}_L > \widetilde{p}_H$, then reset
  $\widetilde{p}_L = \widetilde{p}_H = (\widetilde{p}_L + \widetilde{p}_H)/2$.
(4) Define $t_{\text{PI}}$ as follows. For $\widetilde{p}_L, \widetilde{p}_H \in (0.5, 1)^2$
  set
      $t_{\text{PI}} = \max\left(\left\lceil \frac{1}{2}\left(2 - \frac{\log \frac{\widetilde{p}_H}{1-\widetilde{p}_H}}{\log \frac{\widetilde{p}_L}{1-\widetilde{p}_L}}\right)\right\rceil, 0\right) + 1.$
                                                   (10)
  For $\widetilde{p}_L = 0.5$ or $\widetilde{p}_H = 1$ set $t_{\text{PI}} = 1$.
Output: For each question $i \in [m]$, output
  $\widehat{x}_{\text{PI}_i} = f_{t_{\text{PI}}}(y_{Li}, y_{Hi})$ with $f_{t_{\text{PI}}}$ as defined in (7).
```

Algorithm 1: The proposed plug-in estimator.

results. In practice, one may use all m questions for estimating $\widetilde{p}_L, \widetilde{p}_H$ in step 1 and 2.

B. Asymptotic inadmissibility of MLE

Let $R_m(\widehat{x}_{\text{MLE}})$ and $R_m(\widehat{x}_{\text{PI}})$ denote the risk of the MLE and the plug-in estimator respectively, as defined in (2). To prove that the MLE is asymptotically inadmissible in our setting, we show that there exist no values of p_L, p_H such that the MLE has a lower risk than the constructed plug-in estimator, described in Algorithm 1. We do this in two steps. First we show that there exist p_L^*, p_H^* such that the risk of MLE is higher than the risk of plug-in estimator, by more than a positive constant. Second, we show that asymptotically the risk of the plug-in estimator is as good as or better than that of MLE for all p_L^*, p_H^*.

a) Negative result: Through the following theorem, we show that for some p_L^*, p_H^* the risk of MLE is worse than that of the plug-in estimator by a constant.

Theorem 1. *There exist $p_L^*, p_H^* \in [0.5, 1]^2$ with $p_L^* \le p_H^*$ and m_0 such that for all $m \ge m_0$, we have $R_m(\widehat{x}_{MLE}) > R_m(\widehat{x}_{PI}) + c$, where $c > 0$ is a universal constant.*

We provide a sketch of the proof of Theorem 1 in Section III-C(a).

Remark 1. *Theorem 1 holds true for a set of p_L^*, p_H^*, in the light gray region in Figure 1, enclosed by a red boundary. This set has a non-zero measure.*

Thus, there are many p_L^*, p_H^* for which the risk of MLE is worse than the risk of plug-in by a constant.

b) Positive result: We now present a positive result for the plug-in estimator.

Theorem 2. *For any $p_L^*, p_H^* \in [0.5, 1]^2$ such that $p_L^* \le p_H^*$, there exists m_0 such that for all $m \ge m_0$, we have*

$$R_m(\widehat{x}_{PI}) \le R_m(\widehat{x}_{MLE}) + \frac{c'}{\sqrt{m}}, \tag{11}$$

where c' is a universal constant. Thus, we have

$$\liminf_{m\to\infty} \left[R_m(\widehat{x}_{MLE}) - R_m(\widehat{x}_{PI})\right] \ge 0. \tag{12}$$

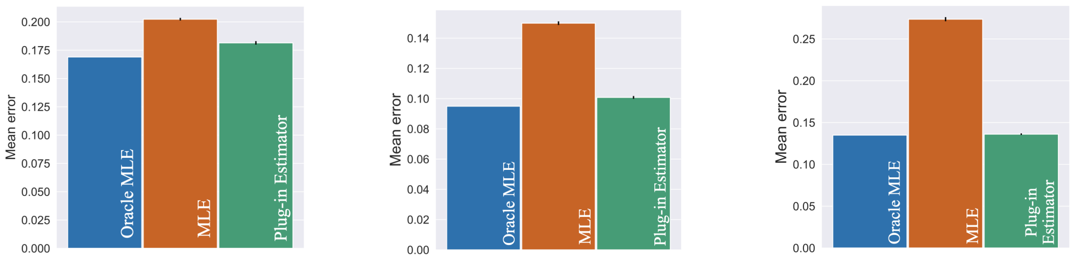
We provide a sketch of the proof of Theorem 2 in Section III-C(b). Theorem 2 provides a positive result for the plug-in estimator by stating that asymptotically it is as good as the MLE or better, pointwise, for all p_L^*, p_H^*. Finally, by combining Theorem 1 and Theorem 2, we see that our constructed plug-in estimator deems the MLE asymptotically inadmissible for our setting.

C. Proof sketch for Theorem 1 and Theorem 2

Our proofs rely on the certain structure of both MLE and plug-in estimators. Specifically, we show that both algorithms operate by picking one of the decision rules defined in (7) (i.e., $t = 1$ for high-level expert-based or $t = 2$ for majority vote-based) and applying it to all the questions $i \in [m]$ to obtain $\widehat{x}_i$. The choice of the decision rule (7) is fully determined by the estimates of true probabilities p_L^*, p_H^* obtained in the inner-workings of the estimators. With these preliminaries, we separately show negative and positive results.

a) Negative result: The crux of the proof is to find p_L^*, p_H^* such that with high probability (i) MLE picks $t = 1$, (ii) the plug-in estimator picks $t = 2$, and (iii) the choice of $t = 2$ leads to a smaller risk than $t = 1$. We approach the proof in three steps and the key challenge is to get a handle on the sample-level behavior of estimators (steps 1 and 2).

Step 1. Starting from MLE, we use a subgaussian argument to show that in the region of interest, the value of the MLE objective (3) uniformly concentrates around its expectation. We then study the corresponding expected value to derive closed-form minimizers and describe the



(a) 2L1H. $p_L^* = 0.7, p_H^* = 0.8$. (b) 3L1H. $p_L^* = 0.75, p_H^* = 0.85$. (c) 5L1H. $p_L^* = 0.72, p_H^* = 0.72$.

Fig. 2: Mean 0-1 error of the three estimators described in this work: Oracle MLE, MLE, and plug-in estimator under three settings with $m = 1000$ questions, computed over 100 trials, with error bars to represent the standard error. Here, xLyH indicates that each question is evaluated by x low-level experts and y high-level experts.

behavior of MLE in terms of the mapping between $\widehat{p}_L, \widehat{p}_H$ and the choice of decision rule (7) it makes.

Step 2. We show that Algorithm 1 obtains unbiased estimates of the true values p_L^*, p_H^*. We then establish convergence rates, thereby characterizing the choice of the decision rule made by the plug-in estimator.

Step 3. With these results, we carefully choose p_L^*, p_H^* that results in requested conditions (i) — (iii), leading to a significant difference in the risks of the two estimators.

b) Positive result: To prove the positive result, we introduce an auxiliary estimator that picks the best decision rule (7) for each instance of $\vec{y}_L, \vec{y}_H$. First, we observe that this auxiliary estimator is as good as or better than both plug-in and MLE. Hence, to prove our result, it remains to show that the risk of plug-in asymptotically converges to that of the auxiliary estimator.

Step 1. We study the behavior of the auxiliary estimator which we illustrate in Figure 1. For all p_L^*, p_H^* to the left of the red dashed line, with high probability, it chooses the high expertise-based decision rule ($t = 1$). To the right of the red dashed line, with high probability, it chooses the majority vote-based decision rule ($t = 2$).

Step 2. To conclude the proof, we establish a convergence result which confirms that with high probability plug-in picks the same decision rule ($t = 1$ or $t = 2$) as the auxiliary estimator.

IV. Simulations

In this section, we simulate settings that relax assumptions in our theoretical analysis, investigating settings where the number of questions m is finite, and under different combinations of evaluators' expertise. We find that our plug-in estimator continues to outperform or perform at least as well as the MLE.

We consider $m = 1000$ questions. In each of our experiments and for each estimator, we compute the average error over 100 trials, where in each trial we generate $\vec{x}^* \in \{0,1\}^m$ uniformly at random and then generate $\vec{y}_L, \vec{y}_H$ based on (1). We consider three settings in our simulations. In Figure 2a each question is evaluated by 2 low-level experts and 1 high-level expert, same as the setting for our theoretical results in Section III, with $p_L^* = 0.7, p_H^* = 0.8$. In Figure 2b each question is evaluated by 3 low-level experts and 1 high-level expert, with $p_L^* = 0.75, p_H^* = 0.85$. In Figure 2c each question is evaluated by 5 low-level experts and 1 high-level expert, with $p_L^* = 0.72, p_H^* = 0.72$. In each setting, we simulate the oracle MLE, MLE and plug-in estimator as described in (4), (5) and Algorithm 1 respectively. Note that for our simulations of the plug-in estimator, we use all the questions for estimation of $\widetilde{p}_L, \widetilde{p}_H$ defined in (8), (9). Observe in Figure 2 that in each setting, the mean 0-1 error of MLE is higher than that of our plug-in estimator. This suggests that our result on the asymptotic inadmissibility of MLE may be true more generally.

V. Conclusion

In this work, we show that the widely used estimator MLE is asymptotically inadmissible in a simplified setting of the Dawid-Skene model with expertise information. For this we construct an alternative estimator, the plug-in estimator. In the future, it will be interesting to investigate the optimality of the plug-in estimator for this setting. More generally, finding the optimal estimator for evaluation aggregation with expertise-level information is an open question of interest.

Acknowledgments: This work was supported in part by NSF grants DMS 1713003, CCF 1763734, 1755656 and CAREER 1942124. CR was partly supported by a J.P. Morgan AI research fellowship.

References

[1] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. Aggregating crowdsourced binary ratings. In *Proceedings of the 22nd International Conference on World Wide Web*, page 285–294. Association for Computing Machinery, 2013.

[2] A. P. Dawid and A. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*, 28:20–28, 1979.

[3] A. for the Advancement of Artificial Intelligence. 36th AAAI conference on artificial intelligence 2022. https://aaai.org/Conferences/AAAI-22/, 2022. [Online; accessed 05-November-2021].

[4] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, page 1631–1640, 2015.

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

[6] A. Ghosh, S. Kale, and P. McAfee. Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, page 167–176, New York, NY, USA, 2011. Association for Computing Machinery.

[7] J. K. Ghosh. The new likelihoods and the Neyman-Scott problems. In *Higher Order Asymptotics*, pages 99–105. Institute of Mathematical Statistics, 1994.

[8] R. Hertwig. Tapping into the wisdom of the crowd—with confidence. *Science*, 336(6079):303–304, 2012.

[9] D. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. *2011 49th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2011*, 09 2011.

[10] D. R. Karger, S. Oh, and D. Shah. Iterative learning for reliable crowdsourcing systems. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1953–1961, Red Hook, NY, USA, 2011. Curran Associates Inc.

[11] A. Koriat. When are two heads better than one and why? *Science*, 336(6079):360–362, 2012.

[12] J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16:1–32, 1948.

[13] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. pages 2554–2560, 08 2013.

[14] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322, 2010.

[15] N. B. Shah. An overview of challenges, experiments, and computational solutions in peer review. Communications of the ACM (to appear). Preprint available at http://bit.ly/PeerReviewOverview, July 2021.

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

[17] N. B. Shah and D. Zhou. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. *Advances in neural information processing systems*, 28, 2015.

[18] V. Sheng, F. Provost, and P. Ipeirotis. Get another label? Improving data quality and data mining using multiple, noisy labelers. pages 614–622, 08 2008.

[19] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. 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. Association for Computational Linguistics, 2008.

[20] M. Staffelbach, P. Sempolinski, D. Hachen, A. Kareem, T. Kijewski-Correa, D. Thain, D. Wei, and G. Madey. Lessons learned from an experiment in crowdsourcing complex citizen engineering tasks with amazon mechanical turk. *arXiv preprint arXiv:1406.7588*, 2014.

[21] I. Stelmakh, N. Shah, and A. Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *JMLR*, 2021.

[22] M.-C. Yuen, I. King, and K. Leung. A survey of crowdsourcing systems. pages 766–773, 10 2011.

[23] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Advances in neural information processing systems*, 27:1260–1268, 2014.

[24] D. Zhou, Q. Liu, J. Platt, and C. Meek. Aggregating ordinal labels from crowds by minimax conditional entropy. volume 2, 06 2014.

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