This is a fictitious manuscript for an experiment
(Version with error in optimization claim)

On Aggregating Evaluations with Expertise or Confidence Information

Anonymous submission

Abstract

In many contexts such as crowdsourced labeling and peer review, multiple evaluators assess a collection of items. These evaluations are often supplemented with additional information such as the evaluator's expertise or confidence, and must be aggregated into a final label for labeling tasks or an acceptance decision in peer review. In these settings, we delve into two primary questions: Can some known ground truths in crowdsourced labeling enhance aggregation for other items? Can we discern how meta-reviewers in peer review aggregate individual assessments to make their decisions? We formulate these problems within a unified modeling framework. We then propose estimators to address these questions, and prove theoretical results on identifiability and convexity. We finally perform empirical evaluations using data from crowdsourcing (Mechanical Turk) and peer review (ICLR conference) which reveal significant advantages of our methods.

Introduction

There are numerous application domains in which multiple evaluators assess a set of items, leading to potential variations in evaluations. These individual evaluations are then aggregated to produce a final overall evaluation for each item. We explore two primary applications of such scenarios, crowdsourced labeling and scientific peer review, while acknowledging that our methods have relevance in other domains, such as admissions or hiring processes where committees aggregate multiple evaluations for a decision.

Crowdsourced labeling serves as our first application of interest. Traditional supervised learning relies on ample labeled data, often obtained from crowdsourcing platforms like Amazon Mechanical Turk (mturk.com) or Prolific (prolific.co). Here, a requester puts up labeling tasks, compensating participating crowdsourcing workers. Multiple workers' labels are then aggregated into a single label per item, then used to train and test supervised learning algorithms. More recently, similar paradigms have emerged for the training of (large) language models via reinforcement learning with human feedback (RLHF). Here as well, we have annotators or workers who are shown prompts and model responses, and asked to score the responses based on various criteria such as factual correctness, toxicity, etc.

Our second emphasis is on scientific peer review, prevalent in computer science conferences such as AAAI. Submitted papers undergo evaluation by several reviewers, whose evaluations are aggregated by meta-reviewers into a final acceptance or rejection decision for each paper.

A facet of these applications that we focus on is the availability of additional data accompanying each evaluation. Evaluators are frequently asked to self-report their confidence or expertise for each evaluation. Sometimes, this additional data is evaluator-centric: crowdsourcing participants often possess "master" certifications, or peer reviewers might vary in seniority. Some other times, this additional data varies with each evaluation, for example, evaluators' self-reported confidence or expertise for each evaluation. We term this supplementary information as "levels", and these levels may encompass confidence, seniority, expertise, etc. In this paper, we delve into how these "levels" can be used to enhance the understanding and utility of evaluations.

In these settings, we consider the following questions. In application domains such as crowdsourced labeling, requesters frequently use questions whose ground truth answers are known to them (Le et al. 2010; Khan and Garcia-Molina 2017). These 'gold standard' questions typically serve either to train workers or to filter out workers providing low-quality answers or spamming the system. This raises the natural question: beyond these basic checks, how can we leverage gold standard questions to enhance the accuracy of labels for non-gold standard questions, in a principled manner? More accurate labels will positively influence downstream tasks of supervised learning or reinforcement learning with human feedback.

In scientific peer review, meta-reviewers aggregate individual reviewers' evaluations. Can we understand this aggregation process in a quantitative manner? Understanding this has multiple uses. First, it sheds light on the aggregation process, helping answer questions such as how meta-reviewers distribute emphasis based on varying confidence or seniority levels of reviewers. Second, it helps identify variations in the strategies of different meta-reviewer groups, for instance, whether there a distinction in aggregation methods between meta-reviewers handling theoretical papers and those handling applied papers. Lastly, many conferences equip meta-reviewers with preliminary aggregation guidelines such as displaying mean scores or using fixed weighted scores based on confidence levels. Our research yields a systematic approach to developing preliminary aggregation guidelines, by drawing on aggregation methods learnt from past confer-

This is a fictitious manuscript for an experiment (Nihar B. Shah)

ences.

With this background, we now list our main contributions.

- We identify and formulate these problems in a common mathematical framework.
- We provide both parametric and non-parametric estimators towards addressing these considerations.
- We provide theoretical results regarding identifiability of our proposed statistical model and convexity of our proposed estimators.
- We empirically evaluate our algorithms on a crowdsourced labeling dataset, revealing significant improvements over previous methods. We then perform a qualitative study on peer review using data from the ICLR 2022 conference, offering insights into the decision-making approaches employed by meta-reviewers.

Related literature

In crowdsourcing, aggregation of labels from multiple workers is a well-studied problem. The most commonly assumed model is that by Dawid and Skene (1979). This model is analyzed further by Karger, Oh, and Shah (2011); Sheng, Provost, and Ipeirotis (2008); Dalvi et al. (2013); Ghosh, Kale, and McAfee (2011); Gao and Zhou (2013); Zhou et al. (2015); Zhang et al. (2014); Khetan and Oh (2016); Shah, Balakrishnan, and Wainwright (2020), and generalized in various ways (Venanzi et al. 2014; Khetan and Oh 2016; Shah, Balakrishnan, and Wainwright 2020). This literature focuses on unsupervised aggregation, devoid of any gold standard questions. In contrast, our work targets using gold standard questions to improve the aggregation. A limitation in our study is the omission of specific worker identity-dependency, reserved for future exploration.

Oyama et al. (2013) present an extension of the Dawid-Skene model that also incorporates separate parameters for each expertise level. Rastogi et al. (2022) show that the maximum likelihood estimator under such extensions is inadmissible. Aydin et al. (2014) combine self-reported confidences with Dawid-Skene like models. However, neither of these papers exploit availability of gold standard questions.

Shah and Zhou (2016); Méndez Méndez et al. (2022) focus on elicitation of self-reported confidences. Ho, Jabbari, and Vaughan (2013) use gold standard questions to decide which workers to query for which task. Kühne and Böhm (2015) consider adaptively choosing the gold standard questions. These topics of eliciting answers or choosing gold standard questions are orthogonal to our focus of estimation using the answers and gold standards.

Our approach of understanding meta-reviewers' aggregation process is inspired by Noothigattu, Shah, and Procaccia (2021). Noothigattu, Shah, and Procaccia (2021) address the problem of "commensuration bias" wherein different reviewers combine their evaluations for individual criteria in different manners, causing arbitrariness in the review process (Lee 2015). They design an algorithm that learns the best fit mapping from individual criteria scores to overall scores, and use this mapping for all reviews, thereby standardizing the process of aggregating individual criteria. Like our model (5) to be introduced below, they also consider a non-parametric model with the only assumption being that of monotonicity of the mapping.

A potential application of our work is its use as a guideline for meta-reviewer decisions. Dycke et al. (2021) develop methods to process the text and scores in order to provide preliminary meta-reviewer decisions. Our approaches are markedly different, offer a multitude of other uses, and focus primarily on interpretability. Pearce and Erosheva (2022); Liu et al. (2022) elicit both ratings and rankings from reviewers based on the analysis by Shah et al. (2018), and develop methods to combine the two for meta-reviewer's use. In contrast, our setting or data does not have rankings.

In peer review and related domains, eliciting evaluations can introduce issues such as miscalibration and dishonesty (Shah 2022), and furthermore, phenomena like the Dunning-Krueger effect (Kruger and Dunning 1999) may arise in self-reported confidences. We leave the integration of these issues into our setting for future work.

Problem formulation

We begin with some general notation. For any positive integer κ, we let $[\kappa]$ denote the set $\{1, \ldots, \kappa\}$. We let $\mathbb{R}$ denote the set of real numbers. We let $\mathbf{I} : \{true, false\} \rightarrow \{0, 1\}$ represent the indicator function whose value is 1 if its argument is true, and is 0 if its argument is false. Finally, we adopt the convention that any vector is a column vector.

Moving on to our problem-specific notation and formulation, we let $n \geq 2$ be the number of items to be evaluated. These items may represent the questions in a crowdsourced labeling task or papers to be peer reviewed. For any item $i \in [n]$, we let d_i denote the number of evaluations it has received. We assume that each evaluation is a real-valued score. Let $x_i = [x_{i,1}, ..., x_{i,d_i}]^T \in \mathbb{R}^{d_i}$ be the scores given by the d_i evaluators for item $i \in [n]$. Let $y_i \in \mathbb{R}$ denote either a known ground truth for item i or the final decision made by a human aggregator on item i. We will simply term y_i as the overall score for item i.

We consider additional data available with each evaluation, such as the experience or expertise levels of evaluators or self-reported confidence or expertise in each evaluation, which we generically term as a "level." Let $\mathcal{Z}$ denote the set of possible levels associated with any response, e.g., $\mathcal{Z} = \{$Low (1), Medium (2), High (3)$\}$. We assume that $\mathcal{Z}$ is a finite set and that $|\mathcal{Z}| \geq 2$. In many applications, one may additionally have a total ordering among the entries of $\mathcal{Z}$, although that is not a necessity in our framework. For notational convenience, we index the elements of $\mathcal{Z}$ as $\{1, 2, \ldots, |\mathcal{Z}|\}$. For any item $i \in [n]$ we let $z_{i,1}, ..., z_{i,d_i} \in \mathcal{Z}$ respectively denote the levels associated to the d_i evaluations. As a shorthand, we define $z_i = [z_{i,1}, ..., z_{i,d_i}]^T$.

Next, we introduce our two models.

Convex combination generalized linear model

Under this model, for any item $i \in [n]$, we assume that the expected value of the overall score y_i is a convex combination of the scores given by its d_i evaluators, where the convex combination coefficients are associated to the level of each evaluation. Specifically, we assume existence of a latent vector $\beta^* \in [0,1]^{|\mathcal{Z}|}$ that comprises our parameters of interest, and will be the target of our subsequent estimation procedures. The model is also associated to a known function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, commonly termed as the link function in conventional generalized linear models (GLMs). Then under our model, for any item $i \in [n]$, the expected value of the answer y_i is assumed to take the following form:

$$\mathbb{E}[y_i] = \sigma\left(x_i^T \begin{bmatrix} \beta^*_{z_{i,1}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}} \\ \beta^*_{z_{i,2}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}} \\ \vdots \\ \beta^*_{z_{i,d_i}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}} \end{bmatrix} \right). \qquad (1)$$

With the expectation of y_i determined as above, we assume that its probability distribution does not depend on any of the parameters (e.g., the probability distribution may be obtained by adding i.i.d. standard Gaussian noise to the mean). This assumption is similar to that in conventional GLMs. The goal then is to estimate the parameters β^* given $\{x_i, z_i, y_i\}_{i \in [n]}$.

Let us interpret our model (1). For any item $i \in [n]$, the terms $\beta^*_{z_{i,1}}, \ldots, \beta^*_{z_{i,d_i}}$ represent latent parameters associated with the d_i evaluations for question i. Recall that the value of each of these parameters lies in the interval $[0,1]$. Then the terms $\beta^*_{z_{i,1}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}}, \ldots, \beta^*_{z_{i,d_i}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}}$ represent a normalized version of these parameters that sum to 1. These normalized values are then used to take a convex combination of the answers x_i given by the evaluators, and this convex combination is represented by the term $x_i^T \begin{bmatrix} \beta^*_{z_{i,1}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}} \\ \vdots \\ \beta^*_{z_{i,d_i}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}} \end{bmatrix}$. In this convex combination, any evaluator $\ell \in [d_i]$ is given weight $\beta^*_{z_{i,\ell}} / \sum_{j \in [d_i]} \beta^*_{z_{i,j}}$. Finally, the function σ is a link function that relates the latent convex combination parameters β^* and the known evaluations $\{x_i, z_i\}_{i \in [n]}$ to the expected values of $\{y_i\}_{i \in [n]}$. While we have specified the model in terms of a general link function σ, our subsequent results will focus on the linear link (that is, with σ as the identity function).

Connection to crowdsourced labeling and peer review: For peer review, we are concerned about how meta-reviewers aggregate individual reviewers' evaluations. One can interpret the model (1) as a generative model for this human-aggregation process. An estimate of the parameter vector β^* can yield insights into the process followed by meta-reviewers, and the program chairs may also use this estimate to form preliminary guidelines for combining reviews in subsequent editions of the conference.

For crowdsourcing on the other hand, we do not consider it as a generative model since in practice y_i will be the independent variable whereas x_i and z_i will be the dependent variables. Instead, through our estimators, we will aim to find the best function that maps $\{(x_i, z_i)\}_{i \in [n]}$ to the respective $\{y_i\}_{i \in [n]}$ based on the gold standard questions, and this mapping may subsequently be used to determine the overall scores for the non-gold standard questions.

Additional assumptions on β^*: We now describe few additional assumptions we make on the vector β^* in model (1).

(A1) The reader may be concerned that setting all or a subset of entries of β^* to be zero results in a 0/0 situation in the model (1). To avoid this situation, we assume that for a given $\epsilon > 0$, the parameters obey $\beta^*_\ell \geq \epsilon$ for every $\ell \in [|\mathcal{Z}|]$.[1] In practice, one may set ϵ to be a small positive number, and in our subsequent experiments we set $\epsilon = 10^{-4}$.

(A2) Observe that β^* cannot be recovered from the model (1). This is because the distribution of $\{y_i\}_{i \in [n]}$ is invariant to multiplying the entire vector β^* with any positive value. To resolve this, we assume that the entries of β^* sum to 1, that is, $\sum_{\ell \in [|\mathcal{Z}|]} \beta^*_\ell = 1$. We discuss such issues of "identifiability" further in Theorem 1.

(A3) Finally, one may also assume additional monotonicity conditions on the entries of β^* as appropriate for the problem setting. Specifically, in some applications, one may be able to assume that the accuracy of evaluations is monotonic in their levels (e.g., if the levels represent an ordered set of expertise levels). In this case, one may assume the parameters to be monotonic in the presumed accuracy level, that is, make the additional assumption $\beta^*_1 \leq \beta^*_2 \leq \ldots \leq \beta^*_{|\mathcal{Z}|}$.

Comparison to conventional generalized linear models (GLMs): We discuss the key difference between our model (1) and conventional GLMs. Conventional GLMs model the various quantities of interest as:

$$\mathbb{E}[y_i] = \sigma\left(\widetilde{x}_i^T \beta^*\right), \qquad (2)$$

for some known covariates $\widetilde{x}_i \in \mathbb{R}^{|\mathcal{Z}|}$. Instead, our model *normalizes the β^* vector differently for each question $i \in [n]$*, thereby introducing another non linearity. We emphasize that this is a very important distinction, rendering standard results in GLM inapplicable.

A natural GLM analogue of our model (1) is when $\widetilde{x}_i^T \beta^*$ represents a linear combination of the d_i evaluations for each question $i \in [n]$, where the coefficient for evaluation $j \in [d_i]$ is $\beta^*_{z_{i,j}}$. This is obtained by setting $\widetilde{x}_{i,\ell} = \sum_{j \in [d_i]} x_{i,j} \mathbf{I}\{z_{i,j} = \ell\}$, for every $i \in [n], \ell \in \mathcal{Z}$. Observe that as compared to our model (1), the normalization term $\sum_{j \in [d_i]} \beta^*_{z_{i,j}}$ is absent in this GLM analogue (2), which forms a key distinguisher from our model.

A non-parametric model

We generalize the model (1) to take a non-parametric form. For simplicity, we assume that the number of evaluators for every item is identical, that is, $d_1 = d_2 = \ldots = d_n = d$. In this non-parametric model, there is no parameter vector β^*, but instead, the model is associated with an unknown function $\Phi^* : (\mathbb{R} \times \mathcal{Z})^d \rightarrow \mathbb{R}$ as:

$$\mathbb{E}[y_i] = \Phi^*(x_{i,1}, z_{i,1}, x_{i,2}, z_{i,2}, \ldots, x_{i,d}, z_{i,d}), \qquad (3)$$

for every $i \in [n]$. It is this function Φ^* that we wish to estimate. We make two assumptions on Φ^*:

[1] Since all $|\mathcal{Z}|$ entries of β^* must have value at least ϵ, no choice of $\epsilon > \frac{1}{|\mathcal{Z}|}$ is feasible. Moreover, $\epsilon = \frac{1}{|\mathcal{Z}|}$ restricts the model to just one possible value of β^* where all its entries equal $\frac{1}{|\mathcal{Z}|}$. Hence we consider $\epsilon \in (0, \frac{1}{|\mathcal{Z}|})$.

(B1) The ordering of the evaluators does not matter, that is, for any $\ell_1, \ell_2 \in [d]$, swapping the arguments $(x_{i,\ell_1}, z_{i,\ell_1})$ and $(x_{i,\ell_2}, z_{i,\ell_2})$ does not change the output of Φ^*.
(B2) $\Phi^*(x_{i,1}, z_{i,1}, x_{i,2}, z_{i,2}, \ldots, x_{i,d}, z_{i,d})$ is coordinate-wise non-decreasing in each of $x_{i,1}, x_{i,2}, \ldots, x_{i,d}$. This implicitly assumes that evaluators are not adversarial, and hence an increase in the score given by an evaluator does not decrease the overall aggregate score, provided nothing else changes.

Loss function

We consider prediction losses for our estimators. For any estimates $\widehat{\beta}$ or $\widehat{\Phi}$ of β^* or Φ^* respectively, given any $(x, z) \in \mathbb{R} \times \mathcal{Z}$ and associated $y \in \mathbb{R}$, we measure the loss between this quantity y and the estimate $\widehat{y}$ obtained by applying $\widehat{\beta}$ under model (1) or $\widehat{\Phi}$ under model (3) to (x, z). In classification settings, we then consider the 0-1 loss $\mathbf{I}\{\widehat{y} \neq y\}$ and for regression settings, we consider the squared loss $(\widehat{y} - y)^2$.

Estimators

We now propose estimation procedures for our two models.

Parametric estimator

For model (1), for simplicity, we consider the linear link, that is, assume σ is the identity function. To estimate the parameters β^*, we consider the least squares estimator:

$$\boxed{\begin{aligned} &\widehat{\beta} \in \mathop{\arg\min}_{\beta \in \mathbb{R}^{|\mathcal{Z}|}} \sum_{i=1}^{n} \left(y_i - \frac{x_i^T [\beta_{z_{i,1}} \;\; \beta_{z_{i,2}} \;\; \ldots \;\; \beta_{z_{i,d_i}}]^T}{\sum_{j=1}^{d_i} \beta_{z_{i,j}}} \right)^2 \\ &\text{subject to } \; \epsilon \leq \beta_\ell \leq 1 \;\; \forall\, \ell \in [|\mathcal{Z}|], \\ &\qquad\qquad \sum_{\ell \in |\mathcal{Z}|} \beta_\ell = 1, \\ &\qquad \text{and } \; \beta_1 \leq \beta_2 \leq \ldots \leq \beta_{|\mathcal{Z}|}. \end{aligned} \qquad (4)}$$

The three constraints capture the three assumptions (A1), (A2) and (A3). One may remove the last constraint if (A3) is not suitable for the application at hand.

Non-parametric estimator

Next, we present an estimator for the non-parametric model (3). For this estimator, we assume that each evaluation $x_{i,j}$ lies in some finite-sized set $\mathcal{X} \subseteq \mathbb{R}$. Our estimator is associated with a hyperparameter $\lambda \geq 0$. It comprises an optimization problem whose solution $\widehat{\phi}$ yields an estimate of the evaluation of Φ^* for every input:

$$\boxed{\begin{aligned} &\widehat{\phi} \in \mathop{\arg\min}_{\phi \in \mathbb{R}^{(|\mathcal{X}| \times |\mathcal{Z}|)^d}} \left(\sum_{i \in [n]} \left(y_i - \phi_{x_{i,1} z_{i,1}, \ldots, x_{i,d} z_{i,d}} \right)^2 \right. \\ &\left. + \lambda \sum_{(a_1, b_1, \ldots, a_d, b_d) \in (|\mathcal{X}| \times |\mathcal{Z}|)^d} \left(\phi_{a_1, b_1, \ldots, a_d, b_d} - \frac{1}{d} \sum_{j \in [d]} a_j \right)^2 \right) \\ &\text{subject to } \phi_{a_1, b_1, \ldots, a_{\ell_1}, b_{\ell_1}, \ldots, a_{\ell_2}, b_{\ell_2}, \ldots, a_d, b_d} \\ &\qquad\qquad = \phi_{a_1, b_1, \ldots, a_{\ell_2}, b_{\ell_2}, \ldots, a_{\ell_1}, b_{\ell_1}, \ldots, a_d, b_d} \\ &\forall\, (a_1, b_1, \ldots, a_{\ell_1}, b_{\ell_1}, \ldots, a_{\ell_2}, b_{\ell_2} \ldots, a_d, b_d) \in (\mathcal{X} \times \mathcal{Z})^d, \\ &\forall\, \ell_1 \neq \ell_2 \in [d]; \\ &\text{and } \phi_{a_1, b_1, \ldots, a_\ell, b_\ell, \ldots, a_d, b_d} \geq \phi_{a_1, b_1, \ldots, a'_\ell, b_\ell, \ldots, a_d, b_d} \\ &\forall (a_1, b_1, \ldots, a_\ell, b_\ell, \ldots, a_d, b_d) \in (\mathcal{X} \times \mathcal{Z})^d, a'_\ell \in \mathcal{X}, a_\ell \geq a'_\ell. \end{aligned} \qquad (5)}$$

The first term in the objective of (5) captures the least squares difference between the observations and the variables. The second term biases the variables towards the mean of the d evaluations to which the respective variable is associated. The hyperparameter $\lambda \geq 0$ specifies the weight on the second term relative to the first. Finally, the two constraints capture assumptions (B1) and (B2).

Non-parametric estimator in a transductive setting

In machine learning, a transductive setting is one where the test points are available at training time. In our crowdsourcing application, the evaluations $\{(x_i, z_i)\}_{i \in [n]}$ by the workers on the non-gold standard questions are available when the estimator needs to be executed; similarly in peer review, all of the reviewer evaluations are available. We thus consider such a transductive setting, with the goal of drastically reducing the number of variables and inequalities in the optimization problem (5). Unlike the previous section, we do not need to restrict the $\{x_{i,j}\}$s to lie in a finite set.

Formally, assume that in addition to the data $\{x_i, z_i, y_i\}_{i \in [n]}$, we are also given $\{x_i, z_i\}_{i=n+1}^{n+m}$. Further, we are only concerned with the outputs of $\widehat{\Phi}^*$ (or its estimate $\widehat{\Phi}$) on $\{x_i, z_i\}_{i \in [n+m]}$ and not with its output on any other point. Such a setting indeed arises in both of our canonical applications. In crowdsourcing, the first n samples correspond to the gold standard data for which workers' evaluations and labels are available, and the last m samples correspond to non-gold standard data for which workers' evaluations are available but not the labels. We thus have access to outputs of Φ^* on $\{x_i, z_i\}_{i \in [n]}$ and are interested in its outputs on $\{x_i, z_i\}_{i=n+1}^{n+m}$. In peer review, we have access to all labels ($m = 0$) and wish to interpret the outputs of Φ^* on $\{x_i, z_i\}_{i \in [n]}$.

To present our estimator for this setting, we introduce some notation. Consider any arbitrary total ordering of the elements of set $\mathcal{Z}$. We say that a vector $(a_1, b_1, \ldots, a_d, b_d) \in (\mathbb{R} \times \mathcal{Z})^d$ is lexicographically sorted if it satisfies two conditions: (i) $a_j \geq a_{j+1}$ for every $j \in [d-1]$, and (ii) $a_j = a_{j+1}$ implies $b_j \geq b_{j+1}$, where this last relation is according to the ordering in $\mathcal{Z}$. For any $(a_1, b_1, \ldots, a_d, b_d)$, we let *sorted*$(a_1, b_1, \ldots, a_d, b_d) \in (\mathbb{R} \times \mathcal{Z})^d$ denote the vec-

This is a fictitious manuscript for an experiment (Nihar B. Shah)

tor obtained by reordering the d evaluators in a manner that the sequence becomes lexicographically sorted. Finally, define $\mathcal{W} = \{sorted(x_{i,1}, z_{i,1}, \ldots, x_{i,d}, z_{i,d}) | i \in [n+m]\}$. Our estimator is now presented in (6).

$$\boxed{\begin{aligned}
&\widehat{\phi} \in \mathop{\arg\min}_{\{\phi_w \in \mathbb{R}\}_{w\in\mathcal{W}}} \Bigg(\sum_{i\in[n]} \left(y_i - \phi_{sorted(x_{i,1} z_{i,1},\ldots,x_{i,d} z_{i,d})}\right)^2 \\
&\quad + \lambda \sum_{(a_1,b_1,\ldots,a_d,b_d)\in\mathcal{W}} \Big(\phi_{a_1,b_1,\ldots,a_d,b_d} - \frac{1}{d}\sum_{j\in[d]} a_j\Big)^2 \Bigg) \\
&\text{subject to } \phi_{a_1,b_1,\ldots,a_d,b_d} \geq \phi_{\widetilde{a}_1,b_1,\ldots,\widetilde{a}_d,b_d} \\
&\quad \forall (a_1,b_1,\ldots,a_d,b_d),\ (\widetilde{a}_1,b_1,\ldots,\widetilde{a}_d,b_d) \in \mathcal{W},\ \text{with} \\
&\qquad a_j \geq \widetilde{a}_j,\ \ \forall j \in [d].
\end{aligned}\qquad (6)}$$

The number of variables in (6) is $|\mathcal{W}| \leq (n+m)$, as opposed to $(|\mathcal{X}||\mathcal{Z}|)^d$ in (5). The number of inequalities is at most $(n+m)^2$ in (6), as opposed to at least $(|\mathcal{X}||\mathcal{Z}|)^d$ in (5).

Theoretical results

In this section, we present our main theoretical results.

Identifiability of the model

In any statistical model, it is natural to ask if one can recover the desired quantity given infinite samples. Identifiability is a fundamental property which goes one step beyond. Rather than pondering infinite samples from a probability distribution, it imagines a situation where we know the probability distribution itself. A model is identifiable when, by knowing the data's probability distribution, one can accurately recover the desired quantity. Formally, for any two valid parameter vectors β^* and $\widetilde{\beta}^*$, let $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ denote the probability distributions of $(y_1, \ldots, y_n)$ under β^* and $\widetilde{\beta}^*$ respectively. Then the model is said to be identifiable if $\beta^* \neq \widetilde{\beta}^* \implies \mathbb{P} \neq \widetilde{\mathbb{P}}$ for every pair of valid parameter vectors. In words, two distinct valid parameter vectors should result in different distributions of the data.

We prove a sufficient condition for identifiability of our model (1). We introduce some notation used in our result. For each item $i \in [n]$, we introduce the shorthand $\mu_i = \mathbb{E}[y_i]$. Define a matrix $\Sigma \in \mathbb{R}^{n\times|\mathcal{Z}|}$ whose $(i,\ell)^{\text{th}}$ entry equals $\Sigma_{i,\ell} = \sum_{j\in[d_i]} (\mu_i - x_{i,j})\mathbf{I}\{z_{i,j} = \ell\}$. In words, $\Sigma_{i,\ell}$ captures the difference between the expected overall score for item i and the provided scores with level ℓ.

Theorem 1. *Consider assumption (A1) with any $\epsilon \in (0, \frac{1}{|\mathcal{Z}|})$, assumption (A2), and and optionally also (A3). Then the model* (1) *with the linear link (i.e., σ as the identity function) is identifiable if the matrix $\begin{bmatrix}\Sigma \\ \mathbf{1}^T\end{bmatrix}$ is of rank $|\mathcal{Z}|$.*

It is important to note that this result is different from the well known identifiability condition for conventional linear models (model in (2) with linear link). The condition for identifiability in the conventional linear model is that the covariate matrix $[\widetilde{x}_1 \ldots \widetilde{x}_n]$ must have rank $|\mathcal{Z}|$. However, this condition is insufficient for our model (1). Here is a simple example to see this. Suppose that for each distinct item, all its evaluations have the same level (though the levels may differ between items). Further, assume that each level in $\mathcal{Z}$ is used for at least one item, and that all scores $\{x_{i,j}\}$ are strictly positive. Then the matrix of covariates in (2) is of full rank, which suffices for recoverability of the parameters in the conventional GLM. However, in this scenario, it is impossible to determine parameters β^* in model (1). As a sanity check, we can verify that our condition in Theorem 1 does call out this scenario. Based on our careful construction of the matrix Σ, all its entries are zero, and hence the matrix $\begin{bmatrix}\Sigma \\ \mathbf{1}^T\end{bmatrix}$ is of rank only $1 (< |\mathcal{Z}|)$.

Proof of Theorem 1. Let $(\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$ and $(\widetilde{\mu}_1, \ldots, \widetilde{\mu}_n) \in \mathbb{R}^n$ denote the mean of $(y_1, \ldots, y_n)$ under $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ respectively. Since $(\mu_1, \ldots, \mu_n) \neq (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_n) \implies \mathbb{P} \neq \widetilde{\mathbb{P}}$, the model is identifiable if $\beta^* \neq \widetilde{\beta}^* \implies (\mu_1, \ldots, \mu_n) \neq (\widetilde{\mu}_1, \ldots, \widetilde{\mu}_n)$. Under the linear link, we have $\mu_i = x_i^T \begin{bmatrix} \beta^*_{z_{i,1}} / \sum_{j\in[d_i]} \beta^*_{z_{i,j}} \\ \vdots \\ \beta^*_{z_{i,d_i}} / \sum_{j\in[d_i]} \beta^*_{z_{i,j}} \end{bmatrix}$ $\forall i \in [n]$. Rearranging the terms, we get $\sum_{j\in[d_i]} \Big((\mu_i - x_{i,j})\beta^*_{z_{i,j}}\Big) = 0$. This rearrangement is possible due to assumption (A1) that entries of β^* are lower bounded by $\epsilon > 0$. Taking together all n items, we get $\Sigma\beta^* = 0$. Finally, assumption (A2) implies $\mathbf{1}^T\beta^* = 1$. We thus have $n+1$ linear equations in terms of the $|\mathcal{Z}|$-length vector β^*, and hence β^* can be recovered exactly if the rank of matrix $\begin{bmatrix}\Sigma \\ \mathbf{1}^T\end{bmatrix}$ is $|\mathcal{Z}|$. □

Convexity of estimation

Theorem 2. *The three optimization problems defined in* (4), (5) *and* (6) *are all convex.*

Our result that each of our estimators amount to solving convex optimization problems is of vital importance, and form a key contribution of this paper. Numerous efficient algorithms have been designed specifically for convex problems, with guarantees on convergence to the optimal solution. Therefore, one can leverage a wealth of mathematical tools and computational techniques and have greater confidence in the solutions obtained.

Proof of Theorem 2. Consider first the estimator (5). It is easy to see that the objective function of the optimization problem is convex with respect to the optimization variables, since both terms in the objective are quadratic functions of these variables, and bear non-negative coefficients. Next, note that importantly, even though $\mathcal{X}$ and $\mathcal{Z}$ are discrete, the variables of the optimization problem do not take their values over $\mathcal{X}$ and $\mathcal{Z}$, but over $\mathbb{R}^{(|\mathcal{X}|\times|\mathcal{Z}|)^d}$ which is a convex set. Moving on to the constraints, the first set of constraints captures assumption (B1): $\phi_{a_1,b_1,\ldots,a_{\ell_1},b_{\ell_1},\ldots,a_{\ell_2},b_{\ell_2},\ldots,a_d,b_d} = \phi_{a_1,b_1,\ldots,a_{\ell_2},b_{\ell_2},\ldots,a_{\ell_1},b_{\ell_1},\ldots,a_d,b_d}$ for every $(a_1, b_1, \ldots, a_{\ell_1}, b_{\ell_1}, \ldots, a_{\ell_2}, b_{\ell_2} \ldots, a_d, b_d) \in (\mathcal{X} \times \mathcal{Z})^d$

and every $\ell_1 \neq \ell_2 \in [d]$. Observe that each constraint in this set is linear in the optimization variables. The second constraint captures assumption (B2): $\phi_{a_1,b_1,\ldots,a_\ell,b_\ell,\ldots,a_d,b_d} \geq \phi_{a_1,b_1,\ldots,a'_\ell,b_\ell,\ldots,a_d,b_d}$ for every $(a_1, b_1,\ldots, a_\ell, b_\ell,\ldots, a_d, b_d) \in (\mathcal{X} \times \mathcal{Z})^d$, and every $a'_\ell \in \mathbb{R}, a_\ell \geq a'_\ell$. Each constraint in this set is also linear in the optimization variables. We thus have a convex objective with linear constraints, thereby establishing that (5) is indeed a convex optimization problem. An identical argument holds for (6), where the variables take any real values, the objective is convex, and constraints are linear.

Turning our attention to (4), we first validate that the model (1) under the linear link meets assumptions (B1) and (B2). For (B1), by observing (1), we recognize that if we rearrange the entries in the vector x_i and correspondingly adjust the order of entries in the vector $\begin{bmatrix} \beta^*_{z_{i,1}} / \sum_{j\in[d_i]} \beta^*_{z_{i,j}} \\ \vdots \\ \beta^*_{z_{i,d_i}} / \sum_{j\in[d_i]} \beta^*_{z_{i,j}} \end{bmatrix}$, then $\mathbb{E}[y_i]$ does not change. Thus the ordering of evaluators does not matter, thereby satisfying assumption (B1). Regarding (B2), with every entry in the vector β^* being non-negative, an increase in any entry of x_i cannot decrease the outcome $\mathbb{E}[y_i]$, fulfilling assumption (B2). In summary, the model outlined in (1) is a special case of (3) with $\lambda = 0$. Since the least squares optimization in (5) for its overarching model (3) is convex as shown above, it directly implies the convexity of the least squares (4). □

Experimental results

In this section we present our experimental results in a crowdsourced labeling setting and a peer review setting. For all experiments, we set $\epsilon = 10^{-4}$.

Crowdsourced labeling

We consider the crowdsourced labeling dataset from Shah and Zhou (2016), who develop incentive mechanisms for crowdsourced labeling tasks. In their setting, workers can either skip questions they are unsure of or indicate their confidence in their answers. The mechanism ensures that workers maximize their expected payment if they choose appropriately, and they prove that this is the only mechanism to satisfy a so-called 'no free lunch' requirement. The paper then conducts experiments on the Amazon Mechanical Turk crowdsourcing platform, and we base our evaluations on the data from these experiments.

Methods The dataset encompasses nine tasks, with four centered on text transcription. Our emphasis is on the other five tasks, which are based on multiple-choice questions. These tasks are titled 'bridge,' 'dogs,' 'countries,' 'flag,' and 'texture'. They contain 21, 85, 20, 126, and 24 questions, respectively, and in the dataset, each question is answered by 35 workers. For every task, we report results averaged across 10,000 iterations of our experiment.

Recall that our framework assumes that both evaluations and true answers have real values. In the 'bridge' task, only two options exist, which we denote as 0 and 1. The remaining four experiments have multiple categorical options, prohibiting a direct translation to a real-valued scale. To address this, we transform the multi-class classification tasks to binary classification (Karger, Oh, and Shah 2013). This transformation involves randomly dividing the full set of options into two groups of equal size (or almost equal when dealing with an odd number of options), denoted by 0 and 1. Consequently, in each task, we have $\mathcal{X} = \{0, 1\}$.

We focus on the 'confidence-based setting' in Shah and Zhou (2016), where the worker could either say that they were "absolutely sure" or "moderately sure" in their answer, or otherwise "skip" the question if they were really unsure. Thus in our setting, we have $|\mathcal{Z}| = 3$. No response was elicited from the worker for the questions that they chose to skip. Since skipping meant they were not at all sure, we imputed these responses by choosing an option uniformly at random as their evaluation.

For each task, we randomly select 15 questions as the gold standard upon which our estimators operate. We underscore that although this represents a substantial portion of the dataset's total questions, in real-world scenarios, the volume of non-gold questions is much more extensive. Thus, practitioners can easily allocate 15 questions as gold standard. Additionally, these gold standard questions are heavily reused, being inserted into many workers' tasks. In our experiments, for every non-gold standard question, we pick evaluations from 3 random workers. For each gold standard question, we randomly partition the workers into 11 groups of 3 workers each (leaving out 2 workers), and treat them as 11 questions for the purposes of running the estimators. This results in $d = 3$ and $n = 11 \times 15 = 165$. Since our aggregate outcomes are binary, we quantize the output of any estimator to either 0 or 1 by thresholding at 0.5, and measure error in terms of the 0-1 loss. In terms of our notation, the workers' evaluations represent $\{x_{ij}\}$s, their self-reported confidences represent $\{z_{ij}\}$s, and the true answers represent $\{y_i\}$s.

We stress that our results are robust to the experimental design choices. Subsequent to these experiments, we assessed their robustness by exploring various alternative design elements. These included employing a softmax function as opposed to quantization for the output, considering the ℓ_2 loss of the non-quantized outputs, and opting for slightly varied counts of gold standard questions and workers. Across all variations, the outcomes qualitatively aligned with the key results presented below.

Among our proposed methods, we evaluate our proposed convex combination GLM model (1) with a linear link, estimated via (4). Recognizing that we are in a transductive setting, we employ the estimator (6) for evaluations within our non-parametric model (3). This estimator necessitates defining a parameter $\lambda \geq 0$. Initially, we test the estimator with a fixed choice $\lambda = 0$, executing it on the gold standard questions to derive an estimate $\widehat{\phi}$ which we subsequently apply to the non-gold standard questions. We also consider a version of (6) where we choose λ in a data-dependent fashion. Algorithm 1 details our procedure for determining λ and measuring the resulting performance; within this algorithm, the parameter ζ represents the error rate.

Algorithm ↓ \ Task →	Bridge	Dogs	Countries	Flag	Texture
Majority voting	13.5%	7.9%	6.5%	13.3%	22.5%
Dawid-Skene model and estimator (Dawid and Skene 1979)	12.9%	7.9%	6.3%	13.0%	16.5%
Dawid-Skene model, Spectral EM estimator (Zhang et al. 2014)	10.4%	7.7%	4.9%	11.1%	14.0%
Fixed weights (3, 2, 1) for the three confidence levels (Aydin et al. 2014)	13.4%	7.9%	6.1%	13.4%	18.2%
Confidence-Mine Voting (Aydin et al. 2014)	10.4%	6.8%	6.2%	10.1%	16.0%
Worker-dependent confidence model (Oyama et al. 2013)	9.9%	7.1%	5.8%	10.1%	16.3%
Worker ability as gold standard accuracy (Ho, Jabbari, and Vaughan 2013)	14.2%	6.9%	6.6%	12.8%	20.1%
Conventional GLM (2) with linear link, least squares estimator	8.2%	7.0%	4.7%	9.9%	14.0%
Conventional GLM (2) with logistic link, maximum likelihood estimator	8.1%	7.5%	4.3%	9.9%	13.9%
Our convex combination GLM estimator (4)	7.5%	**6.1%**	4.7%	9.8%	14.1%
Our non-parametric estimator (5) with $\lambda = 0$	11.0%	7.1%	6.1%	12.9%	23.3%
Our non-parametric estimator (error rate given by ζ from Algorithm 1)	**6.9%**	6.2%	**4.1%**	**9.1%**	**12.7%**

Table 1: Error rates in estimating the non-gold standard questions in our crowdsourced labeling experiment.

```
Algorithm 1: Implementation and evaluation of (6)
Let $G$ denote the set of all gold standard questions and
    $H$ denote the set of all non-gold standard questions
**for** $\lambda \in \{0, 2^{-16}, 2^{-15}, \ldots, 2^{15}, 2^{16}\}$ **do**:
    **for** each $g \in G$ **do**
        Solve (6) using $G \backslash \{g\}$ and denote the result as $\widehat{\phi}^{\lambda}$
        Let $\zeta_{\lambda,g} = \mathbf{I}\{\widehat{\phi}^{\lambda}_{x_{g,1},z_{g,1},x_{g,2},z_{g,2},x_{g,3},z_{g,3}} \neq y_g\}$
    **end for**
    Let $\zeta_{\lambda} = \text{Mean}\{\zeta_{\lambda,g}\}_{g \in G}$
**end for**
Let $\widehat{\lambda} \in \arg\min_{\lambda \in \{0,2^{-16},2^{-15},\ldots,2^{15},2^{16}\}} \zeta_{\lambda}$
    *//Computing error rate*
**for** each $h \in H$ **do**
    Compute $\zeta_h = \mathbf{I}\{\widehat{\phi}^{\widehat{\lambda}}_{x_{h,1},z_{h,1},x_{h,2},z_{h,2},x_{h,3},z_{h,3}} \neq y_h\}$
**end for**
Let $\zeta = \text{Mean}\{\zeta_h\}_{h \in H}$
```

Results We present the results in Table 1. (The standard error of the mean for every reported value is less than 0.02.) Algorithms which incorporate confidence data exhibit higher accuracy. Although conventional GLM models offer a slight improvement with linear and logistic links demonstrating similar efficacy, our convex combination GLM yields better results. This underscores the value of integrating gold standard questions in the aggregation procedure. Our non-parametric approach at $\lambda = 0$ performs worse, possibly due to overfitting. However, when λ is chosen in a data-dependent manner according to Algorithm 1, a balance between the non-parametric method's presumed overfitting and the underfitting of the "mean" term emerges, leading to the strongest performance in four of the five tasks.

Peer review

We conduct an exploratory study on how meta-reviewers aggregate reviewer scores with different confidences.

Methods We source our data from the 2022 edition of the International Conference on Learning Representations (ICLR), a premier machine learning conference, scraped from OpenReview.net. Of the total submissions, we focus on the 1095 accepted and the 1527 rejected papers. We exclude papers that were desk rejected or withdrawn, due to inconsistent availability of meta-reviewer decisions. This gives us a total of $n = 2622$ papers. Each paper was evaluated by three or more reviewers, and hence $d_i \geq 3 \; \forall \, i \in [n]$. Every review contained an overall recommendation score in $\{1, 2, \ldots, 10\}$ (10=best), coupled with a self-assessed confidence score in $\mathcal{Z} = \{1, 2, 3, 4, 5\}$ (5=highest confidence). The final accept/reject decisions form our labels $\{y_i\}_{i \in [n]}$, where accept is represented as 1 and reject as 0.

Since our primary objective is to interpret the model, we use our parametric model (1) with the linear link. Despite ongoing debates about linear parametric models (Stelmakh, Shah, and Singh 2019), their usage remains prevalent in analogous contexts (Tomkins, Zhang, and Heavlin 2017; Lane et al. 2022).

Results We execute our estimator (4) to compute an estimate $\widehat{\beta} \in [0, 1]^5$ of β^* as:

$$\widehat{\beta}_5 = 0.399, \; \widehat{\beta}_4 = 0.350, \; \widehat{\beta}_3 = 0.192, \; \widehat{\beta}_2 = 0.057, \; \widehat{\beta}_1 = 0.002.$$

The evaluations with the lowest confidence level have a negligible impact on the final decisions. The influence remains low at a confidence level of 2. There is a significant uptick in influence starting at level 3, and this influence doubles for levels 4 and 5. Notably, meta-reviewers attribute similar weights to both confidence levels 4 and 5. The R^2 value for the fitted model is 0.768, indicating a good fit.

Discussion

We propose a new methodology useful for applications where multiple evaluations are aggregated. Our work leads to several open problems. On the theoretical and algorithm-design front, our model presumes outcomes $\{y_i\}_{i \in [n]}$ to lie on the real line. We anticipate our approach to be versatile enough to encompass other outcome formats, like categorical data, which we aim to extend in subsequent studies. One should ensure computational efficiency of the estimators via convexity or other approaches, as done in this paper. On the empirical front, while our methods show efficacy in crowdsourced labeling tasks, a logical next step is to evaluate them within RLHF frameworks for large language models.

References

Aydin, B.; Yilmaz, Y. S. Y. Y. S.; Li, Y.; Li, Q.; Gao, J.; and Demirbas, M. 2014. Crowdsourcing for multiple-choice question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Dalvi, N.; Dasgupta, A.; Kumar, R.; and Rastogi, V. 2013. Aggregating Crowdsourced Binary Ratings. In *International Conference on World Wide Web*.

Dawid, A. P.; and Skene, A. 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of The Royal Statistical Society Series C-applied Statistics*.

Dycke, N.; Simpson, E.; Kuznetsov, I.; and Gurevych, I. 2021. Assisting decision making in scholarly peer review: A preference learning perspective. *arXiv preprint arXiv:2109.01190*.

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

Ghosh, A.; Kale, S.; and McAfee, P. 2011. Who Moderates the Moderators? Crowdsourcing Abuse Detection in User-Generated Content. In *Conference on Electronic Commerce*.

Ho, C.-J.; Jabbari, S.; and Vaughan, J. W. 2013. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning*.

Karger, D. R.; Oh, S.; and Shah, D. 2011. Iterative Learning for Reliable Crowdsourcing Systems. In *Conference on Neural Information Processing Systems*.

Karger, D. R.; Oh, S.; and Shah, D. 2013. Efficient crowdsourcing for multi-class labeling. In *Proceedings of the ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*.

Khan, A. R.; and Garcia-Molina, H. 2017. CrowdDQS: Dynamic question selection in crowdsourcing systems. In *International Conference on Management of Data*.

Khetan, A.; and Oh, S. 2016. Achieving budget-optimality with adaptive schemes in crowdsourcing. *Advances in Neural Information Processing Systems*.

Kruger, J.; and Dunning, D. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*.

Kühne, C.; and Böhm, K. 2015. Protecting the Dawid–Skene algorithm against low-competence raters and collusion attacks with gold-selection strategies. *Social Network Analysis and Mining*.

Lane, J. N.; Szajnfarber, Z.; Crusan, J.; Menietti, M.; and Lakhani, K. 2022. Are experts blinded by feasibility? Experimental evidence from a NASA robotics challenge. *SSRN*.

Le, J.; Edmonds, A.; Hester, V.; and Biewald, L. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*.

Lee, C. J. 2015. Commensuration bias in peer review. *Philosophy of Science*.

Liu, Y.; Xu, Y.; Shah, N.; and Singh, A. 2022. Integrating Rankings into Quantized Scores in Peer Review. *Transactions on Machine Learning Research*.

Méndez Méndez, A. E.; Cartwright, M.; Bello, J. P.; and Nov, O. 2022. Eliciting confidence for improving crowdsourced audio annotations. *Proceedings of the ACM on Human-Computer Interaction*.

Noothigattu, R.; Shah, N.; and Procaccia, A. 2021. Loss Functions, Axioms, and Peer Review. *Journal of Artificial Intelligence Research*.

Oyama, S.; Baba, Y.; Sakurai, Y.; and Kashima, H. 2013. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *International Joint Conference on Artificial Intelligence*.

Pearce, M.; and Erosheva, E. 2022. A Unified Statistical Learning Model for Rankings and Scores with Application to Grant Panel Review. *Journal of Machine Learning Research*.

Rastogi, C.; Stelmakh, I.; Shah, N.; and Balakrishnan, S. 2022. No Rose for MLE: Inadmissibility of MLE for Evaluation Aggregation Under Levels of Expertise. In *International Symposium on Information Theory*.

Shah, N. 2022. An Overview of Challenges, Experiments, and Computational Solutions in Peer Review. http://bit.ly/PeerReviewOverview (Abridged version published in the Communications of the ACM).

Shah, N.; Balakrishnan, S.; and Wainwright, M. 2020. A permutation-based model for crowd labeling: Optimal estimation and robustness. *IEEE Transactions on Information Theory*.

Shah, N.; Tabibian, B.; Muandet, K.; Guyon, I.; and Von Luxburg, U. 2018. Design and Analysis of the NIPS 2016 Review Process. *Journal of Machine Learning Research*.

Shah, N.; and Zhou, D. 2016. Double or Nothing: Multiplicative Incentive Mechanisms for Crowdsourcing. *Journal of Machine Learning Research*, 17: 1–52.

Sheng, V.; Provost, F.; and Ipeirotis, P. 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Stelmakh, I.; Shah, N.; and Singh, A. 2019. On Testing for Biases in Peer Review. In *NeurIPS*.

Tomkins, A.; Zhang, M.; and Heavlin, W. D. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*.

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*.

Zhang, Y.; Chen, X.; Zhou, D.; and Jordan, M. 2014. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Advances in neural information processing systems*.

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