On the Efficiency of Data Collection for Crowdsourced Classification

Edoardo Manino[1], **Long Tran-Thanh**[1] and **Nicholas R. Jennings**[2]
[1] University of Southampton
[2] Imperial College, London
em4e15@soton.ac.uk, l.tran-thanh@soton.ac.uk, n.jennings@imperial.ac.uk

Abstract

The quality of crowdsourced data is often highly variable. For this reason, it is common to collect redundant data and use statistical methods to aggregate it. Empirical studies show that the policies we use to collect such data have a strong impact on the accuracy of the system. However, there is little theoretical understanding of this phenomenon. In this paper we provide the first theoretical explanation of the accuracy gap between the most popular collection policies: the *non-adaptive* uniform allocation, and the *adaptive* uncertainty sampling and information gain maximisation. To do so, we propose a novel representation of the collection process in terms of random walks. Then, we use this tool to derive lower and upper bounds on the accuracy of the policies. With these bounds, we are able to quantify the advantage that the two adaptive policies have over the non-adaptive one for the first time.

1 Introduction

In the past decade crowdsourcing has emerged as an effective way to gather a temporary workforce, and execute large numbers of small and repetitive tasks at a competitive price. This allows an employer to complete large scale data-processing projects when alternative methods are either too expensive (e.g. hiring a team of full-time experts) or impractical (e.g. developing an ad hoc machine learning system of equivalent accuracy) [Lintott *et al.*, 2011]. The applications of this approach range from image and video annotation [Vondrick *et al.*, 2013], to speech recognition [Lasecki *et al.*, 2013], language processing [Snow *et al.*, 2008] and even research studies and surveys [Buhrmester *et al.*, 2011].

Currently, the most popular way of crowdsourcing data involves an online platform like Amazon Mechanical Turk[1] or Crowdflower[2], where workers from all around the world can login and execute the tasks submitted by the employers in exchange for a small payment [Ross *et al.*, 2010]. Despite the efforts to screen the workers and introduce qualifications and reputation systems, the crowdsourced data tends to contain a sizeable amount of errors or random answers [Downs *et al.*, 2010]. As cleaning the data by hand goes against the very reason behind using crowdsourcing in the first place, most of the literature focuses on automatically singling out the errors by collecting redundant data and aggregating it together [Whitehill *et al.*, 2009; Liu *et al.*, 2012; Augustin *et al.*, 2017]. On the theoretical side, the main concern is uncovering the relationship between the number of available data points and the accuracy of the aggregated estimates. Specifically, Berend and Kontorovich [2014] derive bounds on the accuracy of the weighted majority voting rule given a fixed set of workers, Gao et al. [2016] compare the asymptotic performance of several probabilistic inference methods and Bonald and Combes [2017] propose an optimal algorithm to estimate the reliability of the individual workers.

However, the aforementioned research focuses solely on the aggregation phase of a crowdsourcing project (see Figure 1), which happens after all the data has been collected. Due to the iterative nature of crowdsourcing, where the data is collected over several days or weeks, there is also an opportunity to improve the efficiency of the collection phase. In this regard, the simplest collection strategy is the *non-adaptive* uniform allocation policy, i.e. always collect a fixed number of data points on each task [Karger *et al.*, 2014]. Alternatively, some authors have proposed the use of *adaptive* policies, which use the data collected so far to inform their future decisions, in an attempt to optimise the subsequent aggregation phase. In particular, Barowy et al. [2012] suggest collecting more data on the tasks where a clear majority has not formed yet, Welinder and Perona [2010] propose retraining the aggregator on every new data point and collecting more data on the tasks with larger uncertainty, and Simpson and Roberts [2014] attempt to estimate the information gain of future data points.

Figure 1: High-level view of the crowdsourcing process. The data is collected from the crowd over a period of time and then aggregated in a final prediction over the classification of the tasks.

[1] mturk.com
[2] www.crowdflower.com

While these adaptive policies have been shown to have an empirical advantage over non-adaptive ones [Welinder and Perona, 2010; Barowy *et al.*, 2012; Simpson and Roberts, 2014], we are still missing a clear theoretical understanding of their performance. More specifically, a prominent result by Karger et al. [2014] states that every policy exhibits an exponential tradeoff between the number of data points R and the final accuracy in the form $\mathbb{P}(\text{error}) \leq \exp(-cR)$. However, the authors provide a value of the constant factor c only for non-adaptive policies. In contrast, other authors have proven the advantage of adaptive policies for some specific scenarios, but failed to address the general case above. In particular, Chen et al. [2013] assume that we can summon specific workers from the crowd at any time, while Ho et al. [2013] assume we can test the accuracy of each worker beforehand.

In this paper we provide the first theoretical explanation of the impact of the data collection process on the accuracy of a crowdsourcing system. Furthermore, we derive new bounds on the accuracy tradeoff of the most popular collection policies: uniform allocation, uncertainty sampling and information gain maximisation. More specifically, we make the following contributions to the state of the art. First, we propose a new way to represent the runtime behaviour of a collection policy in terms of a random walk in the log-odds domain. Second, we use this tool to analyse the tradeoff $\mathbb{P}(\text{error}) \leq \exp(-cR)$ of the existing collection policies under a weighted majority voting aggregator. In so doing, we are able to bound the performance of the uncertainty sampling policy proposed by Welinder and Perona [2010] from both sides, and show its equivalence with the information gain maximisation policy of Simpson et al. [2014]. Third, we repeat our analysis on the more challenging case of probabilistic inference aggregators, improve the bound of Karger et al. [2014] and derive new upper and lower bounds on the error rate of adaptive policies. Finally, with these bounds we are able to quantify the advantage that adaptive policies have over non-adaptive ones.

The paper is structured in the following way. In Section 2 we introduce all the relevant data aggregators and collection policies. In Section 3 we study the performance of the policies under weighted majority voting. In Section 4 we repeat our analysis under probabilistic inference. In Section 5 we conclude and outline possible future work. All the proofs of our theorems are collated in Appendix A.

2 Preliminaries

Among the models for crowdsourced classification, the one-coin Dawid-Skene model [Dawid and Skene, 1979] has received most attention from the theoretical literature [Liu *et al.*, 2012; Karger *et al.*, 2014; Bonald and Combes, 2017]. The reason for this lies in the simplicity of the model, coupled with its ability to capture all the major characteristics of the crowdsourcing scenario.

According to this model, we assume that the objective of the system is to recover the correct classification of M distinct tasks. We denote the underlying ground-truth vector as $\mathbf{y}$, with $y_i \in \{\pm 1\}$ being the true class of task i. Moreover, we assume the presence of a crowd of N workers who provide us with a set of labels $X = \{x_{ij}\}$ over the course of the crowdsourcing effort. These workers become available one by one in random order, get assigned to a task i and provide a label $x_{ij} \in \{\pm 1\}$ in exchange for a unitary payment. We assume that a maximum budget $B << MN$ can be spent on collecting new labels, and we set by convention $x_{ij} = 0$ for any missing task-worker pair. Furthermore, we assume that the probability of observing a correct label depends only on the accuracy of the individual worker, which we denote by $\mathbb{P}(x_{ij} = y_i) = p_j$. In other words, we assume that the workers act independently from each other, and their accuracy is not affected by the task they are assigned to (we plan to extend our analysis to more complicated models as future work, see Section 5). Finally, we assume that the population of workers is extracted from a common distribution $p_j \sim f_p$, and that the prior on the ground-truth labels is $\mathbb{P}(y_i = +1) = 1/2$ (it is trivial to extend our results to other priors).

2.1 Label Aggregation Methods

Given a set of labels X, we need a way to aggregate them into a vector $\hat{\mathbf{y}}$ of predictions over the task classes (where $\hat{y}_i$ is the prediction on task i). Here we consider two of the most common aggregators, one that assumes perfect knowledge over the workers' accuracy vector $\mathbf{p}$ and one that does not need this piece of information.

Weighted Majority Voting. The simple *majority voting* rule predicts the class of a task by $\hat{y}_i = \text{sign}\{\sum_{j=0}^{N} x_{ij}\}$, where ties are broken randomly. However, the presence of workers with varying degrees of accuracy, particularly when $p_j < 1/2$, may dramatically decrease the performance of this system. When the workers' accuracy $\mathbf{p}$ is known, it is possible to assign a larger weight to the more accurate workers, hence increasing the reliability of the system. This method is known as *weighted majority voting*. Nitzan and Paroush [1982] show that, by assigning each worker a weight $w_j = \log(p_j/(1-p_j))$, the resulting aggregation method $\hat{y}_i = \text{sign}\{\sum_{j=0}^{N} x_{ij} w_j\}$ achieves optimal accuracy. Notably, these weights stem from a probabilistic interpretation of the weighted majority rule. In particular, it is possible to show that the weighted sum $z_i = \sum_{j=0}^{N} x_{ij} w_j$ is equal to the posterior log-odds as follows:

$$z_i = \log\left[\prod_{j=0}^{N}\left(\frac{p_j}{1-p_j}\right)^{x_{ij}}\right] = \log\left[\frac{\mathbb{P}(y_i=+1|X,\mathbf{p})}{\mathbb{P}(y_i=-1|X,\mathbf{p})}\right] \quad (1)$$

Probabilistic Inference. When the workers' accuracy $\mathbf{p}$ is not known, we can resort to several unsupervised probabilistic methods [Liu *et al.*, 2012; Zhang *et al.*, 2014; Gao *et al.*, 2016; Bonald and Combes, 2017]. This family of methods infer an estimate $\hat{\mathbf{p}} \approx \mathbf{p}$ from the set of labels X itself, which can then be plugged into the weighted majority voting rule to compute the log-odds $\hat{z}_i = \sum_{j=0}^{N} x_{ij} \log(\hat{p}_j/(1-\hat{p}_j))$. In particular in Section 4 we present a result for the approximate variational inference method proposed by Liu et al. [2012] as it fits our theoretical framework well. This method computes $\hat{\mathbf{p}}$ in an expectation-maximisation fashion as follows:

$$\hat{p}_j = \frac{\sum_{i:x_{ij}\neq 0} \sigma(x_{ij}\hat{z}_i) + \alpha}{|\{x_{ij} \neq 0\}| + \alpha + \beta} \quad (2)$$

where the prior distribution of the workers' accuracy is $f_p \sim \text{Beta}(\alpha, \beta)$ and the sigmoid function $\sigma(z) = 1/(1+\exp(-z))$ is the inverse of the log-odds function.

2.2 Label Collection Policies

During the course of the crowdsourcing effort, we need to decide which task i we want to label next. The sequence of decisions is usually formalised in terms of a *collection policy*, a rule or heuristic that selects i given the incoming worker j^t and the set of labels X^t collected so far. The existing literature provides us with the following three main collection policies.

Uniform Allocation (`UNI`). This non-adaptive policy assigns the same number $R = B/M$ of labels to each task. Using a variant of this policy, Karger et al. [2014] are able to bound the probability of a classification error under probabilistic inference to:

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq 2\exp(-qR/32) \tag{3}$$

where $q = \mathbb{E}_{f_p}\{(2p_j - 1)^2\}$. We will improve upon this bound in Section 4.1.

Uncertainty Sampling (`US`). This adaptive policy maintains a measure of uncertainty over the current set of predictions $\hat{\mathbf{y}}$, and collects new labels over the most uncertain tasks. Proposed first in the active learning community [Lewis and Gale, 1994], uncertainty sampling has been empirically shown to be successful for crowdsourcing in a number of applications [Welinder and Perona, 2010; Barowy *et al.*, 2012]. We derive the first bounds on the performance of this policy in Sections 3.3 and 4.2.

Information Gain Maximisation (`IG`). This adaptive policy always chooses the action that maximises the expected information gain on the current posterior distribution [MacKay, 1992]. The empirical performance of this policy in crowdsourcing applications is described by Simpson and Roberts [2014]. We provide the first theoretical analysis on this policy in Section 3.4.

3 Data Collection under Weighted Majority

In this section, we analyse the performance of the collection policies when the workers' accuracy $\mathbf{p}$ is known and weighted majority voting is used to aggregate the data. More specifically, we take advantage of the properties of the weighted majority aggregator to model the data collection process as a random walk. In turn, this enables us to bound the performance of the policies.

3.1 Modeling as a Random Walk

From the point of view of a single task i, the collection policy selects a subset of workers $N_i \subseteq N$ to work on i throughout the crowdsourcing process. The subset of labels $X_i \subseteq X$ provided by these workers move the log-odds on task i from its starting point $z_i = 0$ to its final value $z_i = \sum_{j \in N_i} x_{ij} w_j$. Since the labels are collected one-by-one iteratively, we can model the evolution of z_i as a random walk that is made of a sequence of independent steps $s_{ij} = x_{ij} w_j$.

Key to our analysis is how these steps s_{ij} are distributed. First, let us derive the probability density function of the weights w_j from that of the workers' accuracy $p_j \sim f_p$:

$$f_w(w) = \sigma(w)\big[1 - \sigma(w)\big] f_p(\sigma(w)) \tag{4}$$

Now, assume by convention that the true class of task i is $y_i = +1$, and that the collection policy does not base its decisions on the accuracy of the incoming worker j^t. Then, if we consider that a step s_j can be taken either by workers with weight $w_j = s_j$ or $w_j = -s_j$, we can write the common probability density function of the steps as:

$$f_s(s) = \sigma(s)\big[f_w(s) + f_w(-s)\big] \tag{5}$$

In general, we have $\mathbb{E}\{f_s\} \geq 0$ for any f_p, with equality holding only if all the workers have $p_j = 1/2$ or, in other terms, when the whole crowd provides only random answers. This is due to the fact that workers with $p_j < 1/2$ get assigned a negative weight, so that even their labels move the weighted majority in the right direction. As a consequence, the random walk on the log-odds z_i of each task i will always drift towards the true class y_i. However, we show in the following sections how different policies capitalise on this drift at different rates.

3.2 Performance of the `UNI` Policy

The `UNI` policy always collects at least $R = \lfloor B/M \rfloor$ labels on each task i. We can thus interpret its behaviour as a sum of R independent random steps extracted from the same distribution $s_{ij} \sim f_s$. Keeping the convention that $y_i = +1$, we can compute the probability density function of the sum $z_i = \sum_{j \in N_i} s_{ij}$ as the convolution between R copies of f_s. The probability of a classification error is thus given by:

$$\mathbb{P}(\hat{y}_i \neq y_i) = \mathbb{P}(z_i \leq 0) = \int_{-\infty}^{0} \Big(\overset{R}{\underset{k=1}{\ast}} f_s \Big) ds \tag{6}$$

where $\ast$ is the convolution operator.

While Equation 6 provides us with the exact performance of the `UNI` policy, it suffers from two drawbacks. First, it might be difficult to compute it without access to the full probability density function f_s. Second, its lack of a simple closed form makes it difficult to compare it theoretically with the performance of other policies. We can solve both drawbacks by measuring the concentration of z_i around its expected value. This leads us to the following theorem:

Theorem 1. *Assume that f_s is subgaussian, with parameter γ such that $\mathbb{E}\{\exp(ts)\} \leq \exp(\gamma^2 t^2/2)$ for all $t \in \mathbb{R}$. Then, the probability of a classification error under the UNI policy is bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \exp\left(-\frac{R\mathbb{E}\{f_s\}^2}{2\gamma^2}\right) \tag{7}$$

Equation 7 clearly exposes the exponential tradeoff between number of labels and accuracy achieved by the `UNI` policy. We use this result to compare it with the other policies in Section 3.5. At the same time, note that the bound in Equation 7 is only tight in an asymptotical sense. For small values of R there exists a tighter bound that, however, quickly becomes inefficient as R increases in value (see Appendix A).

3.3 Performance of the US Policy

We move now to our analysis of the performance of the US policy. This policy always chooses the task with the largest uncertainty, i.e. the one whose log-odds are closest to zero:

$$i^* = \underset{i}{\operatorname{argmin}}\{|z_i|\} \tag{8}$$

Due to this specific behaviour, the magnitude of the log-odds $|z_i|$ tends to increase at the same pace on all M tasks during the collection process. This phenomenon allows us to prove the following two bounds:

Theorem 2. *The probability of a classification error under the US policy is upper bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \exp\big(-\mathbb{E}\{f_s\}(R-1)\big) \tag{9}$$

Theorem 3. *The probability of a classification error under the US policy is lower bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \geq \frac{1}{2}\exp\big(-\mathbb{E}\{f_s\}R - \mathbb{E}\{|f_s|\} - 0.56\big) \tag{10}$$

Notice that the bounds in Equations 9 and 10 match asymptotically and guarantee that the US policy has an exponential tradeoff with constant $c = \mathbb{E}\{f_s\}$. This means that the policy fully exploits the drift in the random walk over the log-odds z_i to reduce the probability of an error. We compare this result with the other policies in Section 3.5.

3.4 Performance of the IG Policy

The IG policy always selects the task that yields the largest information gain. Since we cannot predict the value of the next label x_{ij} in advance, we evaluate the impact of adding it to our current set X^t in expectation:

$$i^* = \operatorname{argmax}\Big\{\mathbb{E}_{x_{ij}}\big\{\mathcal{I}\big(X^t \cup x_{ij}, \mathbf{p}\big)\big|\big|X^t, \mathbf{p}\big)\big\}\Big\} \tag{11}$$

where the information gain $\mathcal{I}$ is defined as the Kullback-Leibler divergence between the future posterior and the current one. Given this definition, we can prove that the policies US and IG are in fact equivalent:

Theorem 4. *Given the current log-odds $\mathbf{z}^t$ and a worker with weight $w_j \neq 0$, the two policies US and IG select the same task i^*, except in case of a tie.*

3.5 Policy Comparison

We now have the tools to compare the performance of the non-adaptive UNI policy with the adaptive US and IG ones. On the one hand, all of them exhibit an asymptotic tradeoff between the number of labels R and the final accuracy in the form $\mathbb{P}(\text{error}) \leq \exp(-cR)$. However, the specific constant factor c varies from policy to policy. In fact, for the UNI policy we have $c_{uni} = \mathbb{E}\{f_s\}^2/2\gamma^2$ (see Equation 7), whereas for the two equivalent policies US and IG we have $c_{ada} = \mathbb{E}\{f_s\}$ (see Equation 9).

In general, a higher value of c means that the policy is more efficient in using additional labels to improve the accuracy. In this respect, we can prove that adaptive policies offer superior guarantees by examining the ratio $c_{ada}/c_{uni} = 2\gamma^2/\mathbb{E}\{f_s\}$:

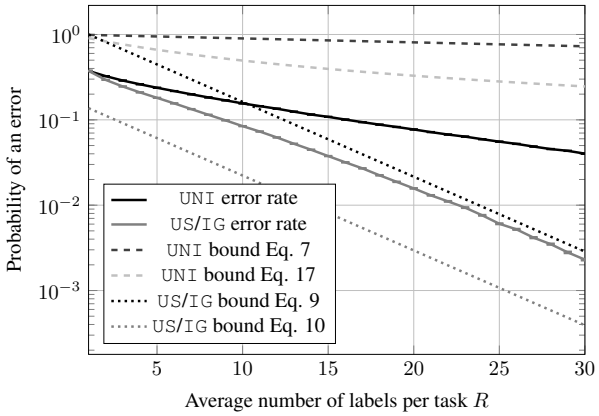

Figure 2: Comparison between the theoretical bounds and the empirical performance of the UNI, US and IG policies under the weighted majority voting aggregator.

Theorem 5. *For any distribution f_s bounded in $[-a, +a]$ with $0 < a < +\infty$, the efficiency ratio between adaptive (US, IG) and non-adaptive (UNI) policies is $c_{ada}/c_{uni} \geq 4$.*

On the other hand, the result in Theorem 5 is derived from upper bounds on the error rate of the policies, and thus it might not reflect the actual performance gap between the UNI, US and IG policies at runtime. For this reason, we run synthetic experiments with $M = 10000$ tasks and a uniform distribution of workers f_p over the interval $[0.4, 0.8]$ to simulate a mixed crowd (similar results can be obtained with different choices of f_p). We report the results in Figure 2, where each point is the average of 100 runs and has standard error below 5×10^{-4}. As the figure shows, the error rate of the US and IG policies quickly diverges from that of the UNI policy as the average budget per task R increases. At the same time, the observed c_{ada}/c_{uni} ratio has a value of 1.9. In order to overcome this discrepancy with the result in Theorem 5, tighter bounds on the UNI policy are needed.

4 Data Collection under Probabilistic Inference

We now remove the assumption that the workers' accuracy $\mathbf{p}$ is known, and repeat our analysis of the collection policies assuming that some unsupervised probabilistic method is employed to aggregate the crowdsourced data. In general, our results are valid for any state-of-the-art probabilistic method that is provably better than majority voting (see [Gao *et al.*, 2016]).

4.1 Performance of the UNI Policy

Similarly to our analysis in Section 3.2, we take advantage of Hoeffding's concentration inequality to bound the accuracy of the UNI policy:

Theorem 6. *The probability of a classification error under the UNI policy is bounded by*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \exp(-\frac{R}{2}(2\mathbb{E}\{f_p\} - 1)^2) \tag{12}$$

This bound is an improvement over the result of Karger et al. [2014] for two reasons. First it only depends on the number of labels R and the average accuracy of the workers $\bar{p}$, instead of the quantity $q = \mathbb{E}\{(2p_j - 1)^2\}$ which is more difficult to estimate in a practical scenario (see Equation 3). Second, for any distribution with the property $(2\bar{p} - 1)^2 > q/16$ our bound is tighter.

4.2 Performance of the US and IG Policies

Let us move on to the performance of adaptive collection policies. It is worth noting that the equivalence result between the US and IG policies shown in Section 3.4 is not valid here because the estimates $\mathbf{p}^t$ may change upon receiving a new label at time t. Nevertheless, the two following results still apply to both policies.

Theorem 7. *The performance of the US and IG policies is upper bounded by*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \exp(|\bar{w}|(1 - R|2\bar{p} - 1|)) \tag{13}$$

Theorem 8. *The performance of the US and IG policies is lower bounded by*

$$\mathbb{P}(\hat{y}_i \neq y_i) \geq \frac{1}{2}\exp\big(-\mathbb{E}\{f_{\hat{s}}\}R - \mathbb{E}\{|f_{\hat{s}}|\} - 0.56\big) \tag{14}$$

where $f_{\hat{s}}$ is the distribution of steps given the weight estimates $\hat{w}_j$ provided by the probabilistic inference method of choice at the end of the crowdsourcing process.

Theorem 8 can be adapted to any probabilistic method by providing a value for $\mathbb{E}\{f_{\hat{s}}\}$ and $\mathbb{E}\{|f_{\hat{s}}|\}$. In most cases this can be done only by numerical estimation. However, for the approximate variational inference algorithm proposed by Liu et al. [2012] we can derive an upper bound as follows:

Theorem 9. *Given a population of workers with accuracy $p_j \sim$ Beta(α, β), the expected value of a step under the approximate variational inference algorithm in [Liu* et al.*, 2012] is bounded by:*

$$\mathbb{E}\{f_{\hat{s}}\} \leq \sum_{Q=1}^{Q_{max}} \mathbb{P}(Q) \sum_{c=0}^{Q} \binom{Q}{c} \frac{B(c+\alpha, Q-c+\beta)}{B(\alpha,\beta)} \log\left(\frac{c+\alpha}{Q-c+\beta}\right)\frac{2c-Q}{Q} \tag{15}$$

where Q is the number of labels provided by a single worker.

Finally, the value of $\mathbb{E}\{|f_{\hat{s}}|\}$ can be computed from the result in Theorem 9 by taking the absolute value of the logarithm and discarding the $\frac{2c-Q}{Q}$ term.

4.3 Policy Comparison

We now have the tools to compare the performance of the three policies UNI, US and IG. As for the case with known workers' accuracies $\mathbf{p}$ in Section 3, all the policies have an asymptotic tradeoff between the number of labels R and the expected accuracy in the form $\mathbb{P}(\text{error}) \leq \exp(-cR)$. However, the constant factor c differs between the non-adaptive (UNI) and adaptive (US, IG) policies. For the UNI policy Equation 12 yields a factor $c_{uni} = \frac{1}{2}(2\bar{p} - 1)^2$, whereas for the UNI and IG policies Equation 13 yields a factor $c_{ada} = \bar{w}(2\bar{p} - 1)$ instead.

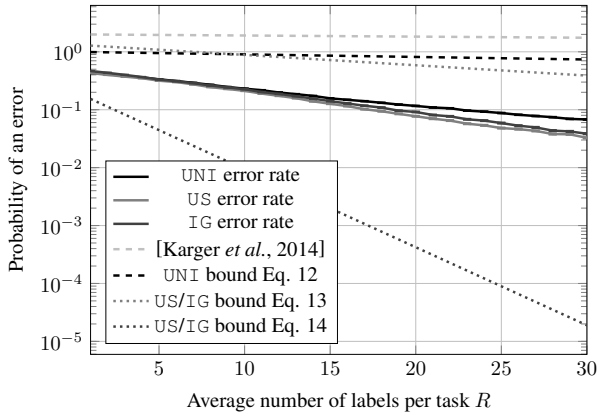

Figure 3: Comparison between the theoretical bounds and the empirical performance of the UNI, US and IG policies under the probabilistic inference aggregator in [Liu *et al.*, 2012].

Given this, we can prove that both US and IG offer superior guarantees over UNI when used in conjunction with a probabilistic inference method. In particular, by comparing the ratio $c_{ada}/c_{uni} = 2\bar{w}/(2\bar{p} - 1)$ we can state the following:

Theorem 10. *For any distribution of the workers' accuracy f_p with $\mathbb{E}\{f_p\} \neq 1/2$, the efficieny ratio between adaptive (US, IG) and non-adaptive (UNI) policies is $c_{ada}/c_{uni} \geq 4$.*

At the same time, the result in Theorem 10 is again derived from upper bounds on the error rate of the policies, and thus it might not reflect the actual performance gap between the UNI, US and IG policies at runtime. In order to rule out this eventuality, we run synthetic experiments with $M = 200$ tasks, $Q = 10$ labels per worker and $f_p \sim$ Beta$(\alpha=4, \beta=3)$ to simulate a mixed crowd. Moreover, we use the approximate variational inference method in [Liu *et al.*, 2012] to aggregate the labels, and average the results over 1000 runs, which yields a standard error below 2×10^{-3}. Note that similar results can be achieved with different values of Q, α and β, whereas the use of the Beta distribution is necessary to match the assumptions of Liu's algorithm. We report the results in Figure 3, which shows that the error rate of the US and IG policies becomes smaller than that of the UNI policy as the average budget per task R increases. On the other hand, the observed value of the c_{ada}/c_{uni} ratio is 1.3, which suggests that the empirical advantage of the adaptive US and IG policies is smaller than the theoretical prediction in Theorem 10.

5 Conclusions

We have analysed the performance of a number of the most common data collection policies for crowdsourced classification. By representing them as random walks in the log-odds domain, we derived new upper and lower bounds on their accuracy. Consequently, we were able to quantify the advantage that some adaptive policies can have over non-adaptive ones for the first time.

We believe that the techniques presented here can be extended to additional scenarios, both in terms of data aggregators and crowdsourcing models. Among them, the case of tasks with varying degrees of difficulty is of particular interest to us. Under this model, Khetan [2016] have already shown an exponential tradeoff in the form $\mathbb{P}(\text{error}) \leq \exp(-cR)$, but the values of c for different policies are still unknown. We aim to compute them using our methods as future work.

Acknowledgments

This research is funded by the UK Research Council project ORCHID, grant EP/I011587/1. The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton.

A Proofs

Proof of Theorem 1. Hoeffding's concentration inequality can be written as follows:

$$\mathbb{P}(z_i - \mathbb{E}\{z_i\} \leq t) \leq \exp\left(-\frac{t^2}{2R\gamma^2}\right) \tag{16}$$

where $R\mathbb{E}\{f_s\} = \mathbb{E}\{z_i\}$ since the R steps in the random walk are independent. Then, by substituting $t = -\mathbb{E}\{z_i\}$ and noting that $\mathbb{P}(\hat{y}_i \neq y_i) = \mathbb{P}(z_i \leq 0)$ we get the result in the theorem. □

As mentioned in Section 3.2, for small values of R there exist a tighter (but non-exponential) bound:

Theorem 1 bis. *Assume that f_s has finite variance. Then, the probability of a classification error under the UNI policy is bounded by:*

$$\mathbb{P}(\hat{y}_i \neq y_i) \leq \frac{Var\{f_s\}}{Var\{f_s\} + R\mathbb{E}\{f_s\}^2} \tag{17}$$

Proof. This result can be derived in the same way as Theorem 1, by using the Chebyshev-Cantelli inequality:

$$\mathbb{P}(z_i - \mathbb{E}\{z_i\} \leq t) \leq \frac{\text{Var}\{z_i\}}{\text{Var}\{z_i\} + t^2} \tag{18}$$

□

Proof of Theorem 2. From the perspective of a single task i, the US policy operates in short bursts of activity, as i keeps receiving new labels until it is no more the most uncertain one. We define $z_B = \min_i\{|z_i|\}$ as the threshold that all tasks have crossed at some point of the crowdsourcing effort. In this respect, we can model the evolution of the log-odds z_i as a bounded random walk, which starts in $z_i = 0$ and ends when z_i leaves the interval $(-z_B, +z_B)$.

Given this, let us assume that we can fix the threshold z_B and then collect as many labels as needed in order to cross it. We denote the log-odds after crossing the threshold as z_i^r, where $z_i^r \notin (-z_B, +z_B)$, and the log-odds at the step before as z_i^{r-1}. According to this definition, r is a stopping time since it is uniquely defined by the information collected before step r. Thus, we can use Wald's equation [Wald, 1944] to link the expected value of z_i^r and the stopping time r:

$$\mathbb{E}\{z_i^r\} = \mathbb{E}\{r\}\mathbb{E}\{f_s\} \tag{19}$$

Recall, however, that z_i^r is the sum of r i.i.d random variables, and that $z_i^{r-1} \in (-z_B, +z_B)$ by definition. As a consequence, we can further bound the expected value of z_i^r by:

$$\mathbb{E}\{z_i^r\} = \mathbb{E}\{z_i^{r-1}\} + \mathbb{E}\{f_s\} < z_B + \mathbb{E}\{f_s\} \tag{20}$$

Putting Equations 19 and 20 together, we can derive a bound for the threshold z_B:

$$z_B > \mathbb{E}\{f_s\}\big(\mathbb{E}\{r\} - 1\big) \tag{21}$$

At the same time, we also know that the random walks on the M tasks are independent, and that the variance of r for a bounded random walk with i.i.d. steps is finite. Therefore, as $M \to \infty$ the total number of steps required to cross the threshold will converge to its expected value, i.e. $B \to M\mathbb{E}\{r\}$. This property allows to substitute $\mathbb{E}\{r\} = R$.

However, we also know that the confidence in our prediction is $\mathbb{P}(\hat{y}_i = y_i) = \sigma(|z_i|)$. Thus, we can bound the final accuracy as:

$$\mathbb{P}(\hat{y}_i = y_i) \geq \sigma\big(\mathbb{E}\{f_s\}(R-1)\big) \tag{22}$$

which yields Equation 9 after some simple algebraic manipulations. □

Proof of Theorem 3. Similarly to the proof of Theorem 2, we can bound the value of the threshold x_B as follows:

$$z_B \leq \mathbb{E}\{|z_i^r|\} = \mathbb{E}\{z_i^r\} - 2\int_{-\infty}^{-z_B} z\mathbb{P}(z_i^r = z)dz \tag{23}$$

where

$$\begin{aligned}\int_{-\infty}^{-z_B} z\mathbb{P}(z_i^r{=}z)dz &= -z_B\mathbb{P}(z_i^r{\leq}{-}z_B) \\ &\quad + \int_{-\infty}^{0} t\mathbb{P}(z_i^r{=}t{-}z_B)dt \\ &\geq -z_B\sigma(-z_B) - \frac{1}{2}\mathbb{E}\{|f_s|\}\end{aligned} \tag{24}$$

Now, we can compute $\max_{z\geq 0}\{z\sigma(-z)\} \approx 0.28$. Additionally, we know that $\mathbb{E}\{z_i^r\} = R\mathbb{E}\{f_s\}$. Thus, the following is true:

$$z_B \leq R\mathbb{E}\{f_s\} + \mathbb{E}\{|f_s|\} + 0.56 \tag{25}$$

By noting that $\mathbb{P}(\hat{y}_i \neq y_i) = \sigma(-z_B) \geq \frac{1}{2}\exp(-z_B)$ we obtain the result of the theorem. □

Proof of Theorem 4. Let us write the information gain in closed form as follows:

$$\begin{aligned}&\mathcal{I}\big(X^t \cup x_{ij}, \mathbf{p}\big|\big|X^t, \mathbf{p}\big) \\ &\quad= \sigma(z_i^t + x_{ij}w_j)\log\left(\frac{\sigma(z_i^t + x_{ij}w_j)}{\sigma(z_i^t)}\right) \\ &\quad+ \sigma(-z_i^t - x_{ij}w_j)\log\left(\frac{\sigma(-z_i^t - x_{ij}w_j)}{\sigma(-z_i^t)}\right)\end{aligned} \tag{26}$$

since the log-odds vector $\mathbf{z}^t$ changes on task i only. The expected value of Equation 26 is the following:

$$\begin{aligned}&\mathbb{E}_{x_{ij}}\big\{\mathcal{I}\big(X^t \cup x_{ij}, \mathbf{p}\big|\big|X^t, \mathbf{p}\big)\big\} \\ &\quad= \sum_{x_{ij}\in\{\pm 1\}} \mathcal{I}\big(X^t \cup x_{ij}, \mathbf{p}\big|\big|X^t, \mathbf{p}\big)\mathbb{P}(x_{ij}|X^t, \mathbf{p})\end{aligned} \tag{27}$$

where

$$\mathbb{P}(x_{ij}|X^t,\mathbf{p}) = \sigma(z_i^t)\sigma(x_{ij}w_j) + \sigma(-z_i^t)\sigma(-x_{ij}w_j) \quad (28)$$

First, note that Equation 27 is even in z_i^t and w_j (this can be easily proven by substitution). Given this, we can focus on the positive intervals $z_i^t, w_j \in (0,\infty)$ and show that the function is monotonically decreasing. We can do so by taking the derivative of Equation 27 and proving that it is negative by contradiction:

$$\begin{aligned}\frac{d}{dz_i^t}\Big\{\mathbb{E}_{x_{ij}}\big\{\mathcal{I}\big(X^t \cup x_{ij},\mathbf{p}\big)\big|\big|X^t,\mathbf{p}\big)\big\}\Big\} \\ = \frac{\exp(z_i^t)(\exp(w_j)-1)}{(1+\exp(z_i^t))^2(1+\exp(w_j))} \\ \left[w_j + \log\left(\frac{1+\exp(z_i^t - w_j)}{1+\exp(z_i^t+w_j)}\right)\right] \geq 0\end{aligned} \quad (29)$$

which yields:

$$\frac{1+\exp(z_i^t - w_j)}{1+\exp(z_i^t+w_j)} \geq \exp(-w_j) \quad (30)$$

which is false for any $z_i^t, w_j \in (0,\infty)$. Finally, we can see that both the objective function $|z_i^t|$ for the US policy and Equation 27 for the IG policy are even in z_i^t, monotonically increasing (decreasing) for $z_i^t \to \infty$ and have a minimum (maximum) in $z_i^t = 0$ for any $w_j \neq 0$. Hence, if a task i_1 is preferred to i_2 under one policy, it is also preferred under the other. □

Proof of Theorem 5. The best-case scenario for the UNI policy is when f_s is such that $\mathbb{E}\{f_s\}$ has the largest possible value and γ has the smallest possible one. First, notice that the subgaussian parameter of f_s is $\gamma = a$ since f_s is bounded. Second, from Equation 5 we can derive that $f_s(s) = f_s(-s)\exp(s)$, thus the best-case is achieved when all the mass of f_s is concentrated in $\pm a$ and $\mathbb{E}\{f_s\} = a(2\sigma(a)-1)$. Under this scenario, the efficiency ratio becomes:

$$\frac{c_{ada}}{c_{uni}} = \frac{2a}{2\sigma(a)-1} \quad (31)$$

While numerator and denominator of Equation 31 are both zero for $a = 0$, the former has first-order derivative equal to 2 whereas the latter has derivative less than $1/2$. Hence, the ratio is larger than 4. □

Proof of Theorem 6. If we set our estimates of the workers' accuracy to $\hat{p}_j = \bar{p} = \mathbb{E}\{f_p\}$, $\forall j$, the distribution of steps f_s reduces to $s_{ij} = +\bar{w}$ with probability $\bar{p}$, and $s_{ij} = -\bar{w}$ with probability $1-\bar{p}$. Hence, f_s has mean $\mathbb{E}\{f_s\} = \bar{w}(2\bar{p}-1)$ and is bounded in $[-\bar{w},+\bar{w}]$, yielding the subgaussian parameter $\gamma = \bar{w}$. By plugging these values into Equation 7 we get the result of the theorem. Finally, note that by setting $\hat{p}_j = \bar{p}$, $\forall j$ the aggregation method becomes a simple majority voting rule. Therefore, any provably superior probabilistic aggregation method (see [Gao *et al.*, 2016]) must achieve a lower error rate and thus satisfy Equation 12. □

Proof of Theorem 7. As per the proof of Theorem 6, we set our estimates of the worker's accuracy to $\hat{p}_j = \bar{p} = \mathbb{E}\{f_p\}$, $\forall j$. In terms of the random walk interpretation, the presence of workers with the same weight $\bar{w} = \log(\bar{p}/(1-\bar{p}))$ means that all the steps have the same length, i.e. $s_{ij} \in \{\pm\bar{w}\}$. Thus, we can use a classic results on bounded discrete random walks [Feller, 1968] to derive the number of steps r required to reach one of the two boundaries $\pm K\bar{w}$ (with $K \in N$) from the initial starting point $z_i = 0$:

$$\mathbb{E}\{r\} = \frac{K\bar{w}}{\mathbb{E}\{f_s\}}(2\sigma(K\bar{w})-1) \leq \frac{K|\bar{w}|}{\mathbb{E}\{f_s\}} \quad (32)$$

By inverting this relationship, we can compute the minimum threshold $\pm K\bar{w}$ we can reach given a budget of R labels in expectation:

$$K \geq \left\lfloor \frac{R\mathbb{E}\{f_s\}}{|\bar{w}|} \right\rfloor \geq \frac{R\mathbb{E}\{f_s\}}{|\bar{w}|} - 1 \quad (33)$$

Thus, at the end of the crowdsourcing process the posterior on the majority class is, in expectation:

$$\mathbb{P}(\hat{y}_i = y_i) \geq \sigma(K|\bar{w}|) = \sigma(R\mathbb{E}\{f_s\} - |\bar{w}|) \quad (34)$$

which by substituting $\mathbb{E}\{f_s\} = \bar{w}(2\bar{p}-1)$ and considering the posterior probability of an error $\hat{y}_i \neq y_i$ yields the result in the theorem. □

Proof of Theorem 8. Define $\hat{z}_B = \min\{|\hat{z}_i| : |X| = B\}$ as the minimum estimated log-odds at the end of the collection process. Also, note that any non-trivial probabilistic inference method updates its estimates $\hat{\mathbf{p}}^t$ during the collection process as new information comes in. As a consequence, the US and IG policies may collect additional labels on a task i whose log-odds $|\hat{z}_i|$ are already above $\hat{z}_B$ (according to the final estimates $\hat{\mathbf{p}}^B$). If the policies had access to $\hat{z}_B$ from the beginning, they would avoid this inefficiency and achieve a larger $\hat{z}_B$ in expectation. Thus, the error rate in this ideal scenario is a lower bound on the real error rate and can be computed according to Theorem 3. □

Proof of Theorem 9. As the accuracy on the tasks increases to 1, i.e. $\hat{z}_i \to \infty$, the estimates in Equation 2 converge to $\hat{p}_j = (c_j+\alpha)/(Q_j+\alpha+\beta)$, where c_j is the number of the worker's correct answers. Moreover, the probability of c_j given that $p_j \sim \text{Beta}(\alpha,\beta)$ is $\binom{Q_j}{c_j}B(c_j+\alpha, Q_j-c_j+\beta)/B(\alpha,\beta)$. Finally, a worker with c_j correct answers has weight $\hat{w}_j = \log((c_j+\alpha)/(Q_j-c_j+\beta))$ and makes c_j/Q_j positive steps and $(Q_j-c_j)/c_j$ negative ones, which yields the equality in the theorem. For any $\hat{z}_i < \infty$ this result becomes an inequality. □

Proof of Theorem 10. The theorem can be proven by taking the first-order approximation of both the numerator and denominator of c_{ada}/c_{uni} centred in $\bar{p} = 1/2$. For the former, notice that $2\bar{w} = 2\log(\bar{p}/(1-\bar{p}))$, whose derivative in $\bar{p} = 1/2$ yields the following value:

$$\left.\frac{d(2\bar{w})}{d\bar{p}}\right|_{1/2} = \left.\frac{2}{\bar{p}(1-\bar{p})}\right|_{1/2} = 8 \quad (35)$$

Also note that $\bar{w}$ is monotonic in $\bar{p}$ and that $\bar{w} = +\infty$ for $\bar{p} = 1$ and $\bar{w} = -\infty$ for $\bar{p} = 0$. Thus in general the derivative of $2\bar{w}$ is always greater than the value in Equation 35. Finally, the first-order derivative of the denominator is $d(2\bar{p}-1)/d\bar{p} = 2$ and both c_{ada} and c_{uni} are zero with $\bar{p} = 1/2$. Hence, for any $\bar{p} \neq 1/2$ the ratio $c_{ada}/c_{uni} \geq 8/2$ which proves the theorem. □

References

[Augustin *et al.*, 2017] Alexandry Augustin, Matteo Venanzi, Alex Rogers, and Nicholas R. Jennings. Bayesian Aggregation of Categorical Distributions with Applications in Crowdsourcing. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1411–1417, 2017.

[Barowy *et al.*, 2012] Daniel W. Barowy, Charlie Curtsinger, Emery D. Berger, and Andrew McGregor. AutoMan: A Platform for Integrating Human-based and Digital Computation. In *Proceedings of the ACM International Conference on Object Oriented Programming Systems Languages and Applications*, pages 639–654, 2012.

[Berend and Kontorovich, 2014] Daniel Berend and Aryeh Kontorovich. Consistency of Weighted Majority Votes. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 3446–3454, 2014.

[Bonald and Combes, 2017] Thomas Bonald and Richard Combes. A Minimax Optimal Algorithm for Crowdsourcing. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4355–4363, 2017.

[Buhrmester *et al.*, 2011] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[Chen *et al.*, 2013] Xi Chen, Qihang Lin, and Dengyong Zhou. Optimistic Knowledge Gradient Policy for Optimal Budget Allocation in Crowdsourcing. In *Proceedings of the 30th International Conference on Machine Learning*, pages 64–72, 2013.

[Dawid and Skene, 1979] Alexander P. 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.

[Downs *et al.*, 2010] Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie F. Cranor. Are Your Participants Gaming the System? Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2399–2402, 2010.

[Feller, 1968] William Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, 1968.

[Gao *et al.*, 2016] Chao Gao, Yu Lu, and Dengyong Zhou. Exact Exponent in Optimal Rates for Crowdsourcing. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 603–611, 2016.

[Ho *et al.*, 2013] Chien-ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. Adaptive Task Assignment for Crowdsourced Classification. In *Proceedings of the 30th International Conference on Machine Learning*, pages 534–542, 2013.

[Karger *et al.*, 2014] David R. Karger, Sewoong Oh, and Devavrat Shah. Budget-Optimal Task Allocation for Reliable Crowdsourcing Systems. *Operations Research*, 62(1):1–24, 2014.

[Khetan and Oh, 2016] Ashish Khetan and Sewoong Oh. Achieving Budget-Optimality with Adaptive Schemes in Crowdsourcing. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, pages 4844–4852, 2016.

[Lasecki *et al.*, 2013] Walter S. Lasecki, Christopher D. Miller, and Jeffrey P. Bigham. Warping Time for More Effective Real-time Crowdsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2033–2036, 2013.

[Lewis and Gale, 1994] David D. Lewis and William A. Gale. A Sequential Algorithm for Training Text Classifiers. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12, 1994.

[Lintott *et al.*, 2011] Chris Lintott, Kevin Schawinski, Steven Bamford, Anže Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C. Nichol, M. Jordan Raddick, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166–178, 2011.

[Liu *et al.*, 2012] Qiang Liu, Jian Peng, and Alexander Ihler. Variational Inference for Crowdsourcing. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pages 692–700, 2012.

[MacKay, 1992] David J. C. MacKay. Information-Based Objective Functions for Active Data Selection. *Neural Computation*, 4(4):590–604, 1992.

[Nitzan and Paroush, 1982] Shmuel Nitzan and Jacob Paroush. Optimal Decision Rules in Uncertain Dichotomous Choice Situations. *International Economic Review*, 23(2):289–297, 1982.

[Ross *et al.*, 2010] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems – Extended Abstracts*, pages 2863–2872, 2010.

[Simpson and Roberts, 2014] Edwin Simpson and Stephen Roberts. Bayesian Methods for Intelligent Task Assignment in Crowdsourcing Systems. In *Scalable Decision Making: Uncertainty, Imperfection, Deliberation*, pages 1–32. Springer, 2014.

[Snow *et al.*, 2008] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and Fast – but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[Vondrick *et al.*, 2013] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently Scaling Up Crowdsourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013.

[Wald, 1944] Abraham Wald. On Cumulative Sums of Random Variables. *The Annals of Mathematical Statistics*, 15(3):283–296, 1944.

[Welinder and Perona, 2010] Peter Welinder and Pietro Perona. Online Crowdsourcing: Rating Annotators and Obtaining Cost-Effective Labels. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops*, pages 25–32, 2010.

[Whitehill *et al.*, 2009] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, pages 2035–2043, 2009.

[Zhang *et al.*, 2014] Yuchen Zhang, Xi Chen, Denny Zhou, and Michael I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 1260–1268, 2014.