Adaptive Polling for Information Aggregation

Thomas Pfeiffer
Harvard University
pfeiffer@fas.harvard.edu

Xi Alice Gao
Harvard University
xagao@seas.harvard.edu

Andrew Mao
Harvard University
mao@seas.harvard.edu

Yiling Chen
Harvard University
yiling@seas.harvard.edu

David G. Rand
Harvard University
drand@fas.harvard.edu

Abstract

The flourishing of online labor markets such as Amazon Mechanical Turk (MTurk) makes it easy to recruit many workers for solving small tasks. We study whether information elicitation and aggregation over a combinatorial space can be achieved by integrating small pieces of potentially imprecise information, gathered from a large number of workers through simple, one-shot interactions in an online labor market. We consider the setting of predicting the ranking of n competing candidates, each having a hidden underlying strength parameter. At each step, our method estimates the strength parameters from the collected pairwise comparison data and adaptively chooses another pairwise comparison question for the next recruited worker. Through an MTurk experiment, we show that the adaptive method effectively elicits and aggregates information, outperforming a naïve method using a random pairwise comparison question at each step.

1 Introduction

Decision making often relies on collecting small pieces of relevant information from many individuals and aggregating such information into a consensus that forecasts some event of interest. Such information elicitation and aggregation is especially challenging when the outcome space of the event is large, due to the inherent difficulties in reasoning over and propagating information through the large outcome space in a consistent and efficient manner.

In recent years, online labor markets, such as Amazon Mechanical Turk (MTurk), have become a burgeoning platform for human computation (Law and von Ahn 2011). MTurk provides easy access to an ever-growing workforce that is readily available to solve complex problems such as image labeling, translation, and speech-to-text transcriptions. One salient feature of MTurk is that the tasks typically offer small monetary rewards (e.g. 10 cents) and involve simple, one-shot interactions. This leads to a natural problem solving approach where a complex problem is decomposed into many simple, manageable subtasks, such that each worker can make a small, relatively independent contribution towards the overall solution. The algorithm then takes care of integrating the solutions to the subtasks into a coherent final solution to the entire problem.

In this paper, we examine whether we can leverage online labor markets' easy access to participants to effectively solve the information elicitation and aggregation problem for an event with an exponentially large outcome space. Our proposed algorithm, through simple, one-shot interactions, adaptively collects many small pieces of potentially imprecise information from a large number of participants recruited through an online labor market, and integrates these information together into an accurate solution.

We consider a setting with n competing candidates, each characterized by a hidden strength parameter. Our goal is to predict a ranking of these candidates by producing accurate estimates of their strength parameters. Participants have noisy information about the strengths of the candidates. We design an adaptive algorithm that at each step estimates the strength parameters based on collected pairwise comparison data and presents another pairwise comparison question that myopically maximizes the expected information gain to a recruited participant. We then evaluate our algorithm through an MTurk experiment for a set of candidates for which we know the underlying true ranking. Our experimental results show that the adaptive method can gradually incorporate small pieces of collected information and improve the estimates of the strength parameters over time. Compared with presenting a random pairwise comparison question at each step, adaptive questioning has the advantage of reducing the uncertainty of the estimates and increasing the accuracy of the prediction more quickly. Interestingly, this is achieved by asking more pairwise comparison questions that are less likely to be answered correctly.

2 Related Work

Many elaborate approaches have been developed for event forecasting. For example, prediction markets (Wolfers and Zitzewitz 2004) allow participants to wager on the outcomes of uncertain events and make profits by improving market predictions. There have been several attempts to design expressive prediction markets (Chen et al. 2008; Abernethy, Chen, and Wortman Vaughan 2011; Xia and Pennock 2011; Pennock and Xia 2011), especially for forecasting an event with a combinatorial outcome space (e.g. permutation of n candidates). However, these combinatorial prediction markets can be computationally intractable to operate, and it is more complicated for humans to interact with the markets than participate in simpler elicitation mechanisms such as surveys. A study by Goel et al. (2010) showed that, for predicting outcomes of binary sports events, the relative advantage of using prediction markets instead of polls was very small. This suggests that methods requiring simple interactions with participants may still provide accurate results for the purpose of eliciting and aggregating information.

There is a rapidly evolving human computation literature on designing workflows for solving complex problems using crowdsourcing platforms. The simpler approaches either allow for participants to iteratively improve the solution, or to work on the same problems in parallel (Little et al. 2009; 2010). More complex workflows attempt to break a problem down into small chunks so that the participants can make relatively independent contributions to the final solution (Kittur et al. 2011; Liem, Zhang, and Chen 2011; Noronha et al. 2011). Our method can be seen as a workflow that aggregates pairwise comparison results from many participants using an adaptive algorithm, and integrates these results into an accurate total ordering of the candidates.

Our adaptive algorithm characterizes the participants' noisy information on the strength parameters using the Thurstone-Mosteller model (Thurstone 1927; Mosteller 1951), which is a special case of the well known random utility model (RUM) (McFadden 1974) in economics with Gaussian noise. The Thurstone-Mosteller model has a long history in psychology, econometrics, and statistics, and has been used in preference learning (Brochu, de Freitas, and Ghosh 2007; Houlsby et al. 2011) and rating chess players (Elo 1978). Carterette et al. (2008) demonstrate from an information retrieval perspective that pairwise comparisons such as used in the Thurstone-Mosteller model are more natural and effective for human preference elicitation than absolute judgments. When the noise follows a Gumbel distribution, the RUM model becomes the Plackett-Luce model (Plackett 1975; Luce 2005). For pairwise comparison, the Plackett-Luce model reduces to the well known Bradley-Terry model (Bradley and Terry 1952). We choose to use the Thurstone-Mosteller model because of the tractability in model estimation when using Gaussian noise.

The way our algorithm selects the next pair of candidates takes an active learning approach. Interested readers can refer to Settles (2009) for a comprehensive survey on active learning. Our approach is in the same spirit as those that maximize information gain according to an information-theoretic metric. The metric we use is the expected Kullback-Leibler divergence between the current and updated estimated parameter distributions. Glickman and Jensen (2005) also used this metric to optimally find pairs for tournaments using the Bradley-Terry model. There exists some work on predicting rankings using active learning and pairwise comparison data (Long et al. 2010; Ailon 2011); however, these assume that the labeled data are accurate, whereas our method allows for erroneous answers from the participants.

3 Method

We are interested in predicting the ranking of n competing candidates, where the true ranking is determined by hidden strength parameters s_i for each candidate. If $s_i > s_j$, candidate i is ranked higher than candidate j. Participants have noisy information on the strength parameters.

Our method presents simple pairwise comparison questions to participants and elicits information only on the presented pair of candidates. Based on the data collected, we estimate the strength parameters of all the candidates. As it is costly to poll the participants, we adaptively choose (in each iteration) the next pair of candidates that can provide the largest expected (myopic) improvement to the current estimation.

Let M be a $n \times n$ nonnegative matrix used to record the pairwise comparison results. $\mathrm{M}_{i,j}$ denotes the number of times candidate i has been ranked higher than candidate j. Let $\mathrm{M}_{i,i} = 0, \forall i$. Then, a high-level summary of our method with T iterations is presented in Algorithm 1 below.

Algorithm 1 Adaptive Information Polling and Aggregation

```
1. Initialize M to a nonnegative, invertible matrix, with value 0 on the diagonal.
2. $t = 1$.
while $t \leq T$ do
    3. Estimate the strength parameters based on M. We use the Thurstone-Mosteller model to capture the noisiness of participants' information and obtain the maximum likelihood estimates of the strength parameters. See Sections 3.1 and 3.2 for details.
    4. Select a pair of candidates that maximizes the expected information gain of the parameter estimation. See Section 3.3 for details.
    5. Obtain the answer to the pairwise comparison question from an participant and update the matrix M.
    6. $t = t + 1$.
end while
```

In Section 3.1, we introduce the Thurstone-Mosteller model adopted for modeling the noisiness of the participants' information. We discuss the method for estimating the strength parameters of candidates in Section 3.2. Together, these two parts detail how step 3 of Algorithm 1 is carried out. Finally, we explain step 4 of Algorithm 1 in Section 3.3.

3.1 Noisy Information Model

To model the noisiness of the participants' information, we adopt the Thurstone-Mosteller model or the Probit model with Gaussian noise. One may also adopt the Bradley-Terry model, also called the Logit model, by setting $\mathrm{P}(r_i > r_j)$ to $\frac{1}{1+e^{s_j - s_i}}$, the cdf of the logistic distribution. The difference between the two models is very slight, but the Gaussian distribution of the Thurstone-Mosteller model is more tractable for the adaptive approach in our algorithm.

Let $\boldsymbol{s'} = (s'_1, s'_2, \ldots, s'_n)$ represent the absolute strength of the n candidates. We model a random participant's perceived absolute strength of candidate i as a random variable: $r'_i = s'_i + \epsilon'_i$, where the noise term is Gaussian,

$\epsilon_i' \sim \mathcal{N}(0, \sigma^2)$ with unknown σ^2. Thus, the probability for the participant to rank candidate i higher than candidate j is

$$\mathrm{P}(r_i' > r_j') = \mathrm{P}(\epsilon_j' - \epsilon_i' < s_i' - s_j') = \Phi\left(\frac{s_i' - s_j'}{\sqrt{2}\sigma}\right) \tag{1}$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard Gaussian distribution $\mathcal{N}(0, 1)$.

We note that the σ^2 term only affects scaling. Furthermore, with a fixed number of parameters n, only their differences affect the probabilities. Without loss of generality, let

$$s_i = \frac{1}{\sqrt{2}\sigma}(s_i' - s_k'), \tag{2}$$

and

$$r_i = \frac{1}{\sqrt{2}\sigma}(r_i' - s_k'), \tag{3}$$

where k is an arbitrary reference candidate. We then have $s_k = 0$, and $r_i = s_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1/2)$. Effectively we only have $n - 1$ unknown parameters. The probability that a participant ranks candidate i higher than candidate j can be written as

$$\mathrm{P}(r_i > r_j) = \Phi(s_i - s_j). \tag{4}$$

From now on, for simplicity, we will call $\boldsymbol{s}$ the strength parameters of the candidates and $\boldsymbol{r}$ the perceived strength of the candidates.

3.2 Maximum Likelihood Estimation

Given the pairwise comparison results M, we will obtain the maximum likelihood estimates of the strength parameters for the Thurstone-Mosteller model introduced above.

The log likelihood given M is

$$L(\boldsymbol{s}|\mathrm{M}) = \sum_{i,j} \mathrm{M}_{i,j} \log(\Phi(s_i - s_j)). \tag{5}$$

The maximum likelihood estimators, $\hat{\boldsymbol{s}}$, are the strength parameters that maximize the log likelihood, i.e. $\hat{\boldsymbol{s}} \in \arg\max_{\boldsymbol{s}} L(\boldsymbol{s}|\mathrm{M})$.

Let $\phi(x)$ be the probability density function (pdf) of the standard Gaussian distribution: $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Note that $\phi(x)$ is log-concave, that is, $\log \phi(x)$ is concave. According to Bagnoli and Bergstrom (1989), the cdf of a log-concave and differentiable pdf is also log-concave. This means that $\log \Phi(x)$ is concave in x. Thus, the log likelihood function $L(\boldsymbol{s}|\mathrm{M})$ in (5) is a concave function of $\boldsymbol{s}$ and we only need to consider the first order conditions to solve the optimization problem.

The derivatives of $L(\boldsymbol{s}|\mathrm{M})$ are

$$\frac{\partial L(\mathrm{M}|\boldsymbol{s})}{\partial s_i} = \sum_j \mathrm{M}_{i,j} \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)} - \sum_j \mathrm{M}_{j,i} \frac{\phi(s_j - s_i)}{\Phi(s_j - s_i)}$$

for all i. Hence, $\hat{\boldsymbol{s}}$ is the solution to the equation system $\frac{\partial L(\mathrm{M}|\boldsymbol{s})}{\partial s_i} = 0, \forall i$. This does not have a closed-form solution, but can be solved using numerical methods.

The maximum likelihood estimators $\hat{\boldsymbol{s}}$ asymptotically follow a multivariate Gaussian distribution. The variance and covariance of $\hat{\boldsymbol{s}}$ can be estimated using the Hessian matrix of the log likelihood evaluated at $\hat{\boldsymbol{s}}$. The Hessian matrix has elements

$$\begin{aligned}\frac{\partial^2 L}{\partial s_j \partial s_i} &= \mathrm{M}_{i,j} \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)} \left(s_i - s_j + \frac{\phi(s_i - s_j)}{\Phi(s_i - s_j)}\right) \\ &\quad + \mathrm{M}_{j,i} \frac{\phi(s_j - s_i)}{\Phi(s_j - s_i)} \left(s_j - s_i + \frac{\phi(s_j - s_i)}{\Phi(s_j - s_i)}\right)\end{aligned}$$

for $i \neq j$, and

$$\frac{\partial^2 L}{\partial s_i^2} = - \sum_{j: j \neq i} \frac{\partial^2 L}{\partial s_j \partial s_i}$$

for all i. Let $\mathrm{H}(\hat{s})$ be the Hessian matrix at $\boldsymbol{s} = \hat{\boldsymbol{s}}$. Then, the estimated covariance matrix of $\hat{\boldsymbol{s}}$ is the inverse of negative $\mathrm{H}(\hat{\boldsymbol{s}})$, i.e.

$$\hat{\Sigma} = (-\mathrm{H}(\hat{\boldsymbol{s}}))^{-1}.$$

Therefore, given M, our knowledge on $\boldsymbol{s}$ can be approximated by the multivariate Gaussian distribution $\mathcal{N}(\hat{\boldsymbol{s}}, \hat{\Sigma})$.

3.3 Adaptive Approach

At each iteration, the most valuable poll to present to a participant is on a pair of candidates that can best improve our current knowledge of the strength parameters.

Let M^c be the matrix of observations, $\hat{\boldsymbol{s}}^{\boldsymbol{c}}$ be the estimation of $\boldsymbol{s}$, and $\hat{\Sigma}^c$ be the estimated covariance matrix of $\hat{\boldsymbol{s}}^{\boldsymbol{c}}$ in the current round. Because $r_i = s_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1/2)$ is independent Gaussian noise, the predicted perceived strength of candidates by a random participant follows a multivariate Gaussian distribution: $\hat{\boldsymbol{r}}^{\boldsymbol{c}} \sim \mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{c}}, \hat{\Sigma}^c + \Sigma^{\boldsymbol{\epsilon}})$, where $\Sigma^{\boldsymbol{\epsilon}}$ is the covariance matrix of the ϵ_i and has value $1/2$ on the diagonal and 0 everywhere else. Hence, given a pair of candidates i and j, the predicted probability that a random participant will rank candidate i higher than candidate j is

$$\hat{p}_{i,j}^c = \mathrm{P}(\hat{r}_i^c > \hat{r}_j^c) = \Phi\left(\frac{\hat{s}_i^c - \hat{s}_j^c}{1 + \hat{\Sigma}^c(i,i) + \hat{\Sigma}^c(j,j) - 2\hat{\Sigma}^c(i,j)}\right)$$

where $\hat{\Sigma}^c(i, j)$ is the element of $\hat{\Sigma}^c$ at row i and column j. This means that at each iteration, for each pair of candidates i and j, we can predict how likely a random participant will rank i higher than j and similarly will rank j higher than i.

Suppose we present the pair of candidates i and j to a participant. If the participant ranks i higher than j, our matrix of observations will become M^{ij}, which is identical to M^c everywhere except $\mathrm{M}^{ij}(i, j) = \mathrm{M}^c(i, j) + 1$. We denote the approximate distribution obtained from the maximum likelihood estimation given M^{ij} as $\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{ij}}, \hat{\Sigma}^{ij})$. Intuitively, if $\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{ij}}, \hat{\Sigma}^{ij})$ is very different from our current estimation $\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{c}}, \hat{\Sigma}^c)$, the extra observation has a large information value. Thus, we use the Kullback-Leibler divergence, also called relative entropy, to measure the information value. The Kullback-Leibler divergence between the two multivariate normal distributions is

$$\begin{aligned}D_{\mathrm{KL}}(\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{ij}}, \hat{\Sigma}^{ij}) \| \mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{c}}, \hat{\Sigma}^c)) &= \frac{1}{2}\Big[\mathrm{tr}\left((\hat{\Sigma}^c)^{-1} \widehat{\Sigma^{ij}}\right) \qquad (6) \\ &+ \left(\hat{\boldsymbol{s}}^{\boldsymbol{c}} - \hat{\boldsymbol{s}}^{\boldsymbol{ij}}\right)^\top (\hat{\Sigma}^c)^{-1} (\hat{\boldsymbol{s}}^{\boldsymbol{c}} - \hat{\boldsymbol{s}}^{\boldsymbol{ij}}) - \log\left(\frac{|\hat{\Sigma}^{ij}|}{|\hat{\Sigma}^c|}\right) - n\Big],\end{aligned}$$

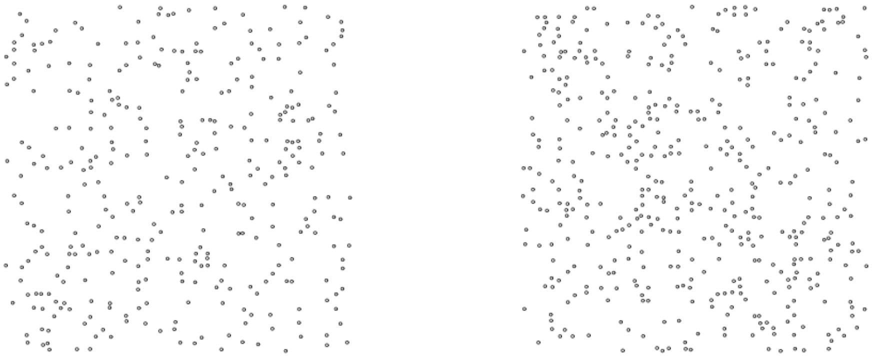

Figure 1: Two example pictures. The left picture has 342 dots, and the right one has 447 dots.

where n is the dimension of the random vectors, which equals the number of candidates, and $|\hat{\Sigma}^{ij}|$ is the determinant of $\hat{\Sigma}^{ij}$. Similarly, if the participant ranks j higher than i, our matrix of observations will become M^{ji}, which is identical to M^{c} everywhere except $\mathrm{M}^{ji}(j,i) = \mathrm{M}^{c}(j,i) + 1$. The new approximate distribution becomes $\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{ji}}, \hat{\Sigma}^{ji})$. The Kullback-Leibler divergence $D_{\mathrm{KL}}(\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{ji}}, \hat{\Sigma}^{ji}) \| \mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{c}}, \hat{\Sigma}^{c}))$ can be calculated analogously to (6).

Putting all pieces together, for each pair of candidates i and j, we can calculate the expected information gain of polling an participant on the pair as

$$\begin{aligned} g(i,j) =& \hat{p}^{c}_{i,j} D_{\mathrm{KL}}(\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{ij}}, \hat{\Sigma}^{ij}) \| \mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{c}}, \hat{\Sigma}^{c})) \qquad (7) \\ &+ \hat{p}^{c}_{j,i} D_{\mathrm{KL}}(\mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{ji}}, \hat{\Sigma}^{ji}) \| \mathcal{N}(\hat{\boldsymbol{s}}^{\boldsymbol{c}}, \hat{\Sigma}^{c})). \end{aligned}$$

At each iteration, we pick the pair with the maximum expected information gain and present it to another participant.

4 Experiment

We experimentally evaluate the effectiveness of our method in polling and aggregating information through many simple, one-shot interactions with participants recruited from MTurk. In our experiment, each candidate was a picture containing a relatively large number of dots (Horton 2010). We generated 12 different pictures, each having 318, 335, 342, 344, 355, 381, 383, 399, 422, 447, 460, and 469 non-overlapping dots respectively. The number of dots x in each picture was independently drawn according to a distribution such that $\mathrm{P}(x) \propto 1/x$ for $x \in [300, 500]$. Figure 1 presents two example pictures used in the experiment. The goal was to use the method introduced in the previous section to estimate the relative number of dots in these 12 pictures in order to correctly rank these pictures in decreasing number of dots. There are several reasons that we chose pictures with dots as the candidates for our experiment: (1) we know the correct ranking and can more objectively evaluate the proposed method; (2) the number of dots in each picture is large enough that counting is not an option for participants, introducing uncertainty; (3) the differences in number of dots across pictures vary and some pairs are more difficult to compare than others; for example, pictures in some adjacent pairs differ by only 2 dots, while those in some other adjacent pair are separated by 26 dots.

We ran our experiment on MTurk. For each HIT (Human Intelligence Task in MTurk's terminology), we presented a pair of pictures, randomly placing one on the left and the other on the right, and asked a MTurk user (Turker) to choose the picture that contained more dots. The base reward for completing a HIT was \$0.05. If the Turker correctly selected the picture with more dots, we provided another \$0.05 as a bonus. Using the adaptive method described in the previous section, we compute an estimate of the strength parameters which reflect the relative differences between the number of dots in the pictures, and decide which pair of pictures to present to the next Turker.

The matrix M was initialized to have value 0 on the diagonal and 0.08 everywhere else. The effect of this was that our initial estimate of the strength parameter was $\mathcal{N}(\mathbf{0}, \Sigma^{0})$, where Σ^{0} had value 1.64 on the diagonal and value 0.82 everywhere else. This can be interpreted as our prior belief of the strength parameters without any information.

For adaptive polling, we ran 6 trials. For each trial, we recruited 100 participants assuming that the budget is only enough for collecting 100 correct answers. To evaluate the advantage of our adaptive approach, we ran another 6 trials (with 100 HITs in each trial) of the random polling method, where the pair in each HIT was randomly selected. In our experiment, each HIT was completed by a Turker with a unique ID. In other words, we interacted with each participating Turker only once.

5 Results

In our setting, we consider the number of dots in each picture as its absolute strength parameter s'_i, the value of which we know as the experimenter. However, in order to evaluate our method, we need to establish a "gold standard" for the strength parameters, which are relative to the strength $s_k = 0$ of a reference candidate k, as defined in equation (2). Thus, we need to transform the number of dots in each picture into their strength parameters, which means that we need a good estimate of $\frac{1}{\sqrt{2}\sigma}$ according to equation (2), $s_i = \frac{1}{\sqrt{2}\sigma}(s'_i - s'_k)$. We run a Probit regression (McCullagh and Nelder 1989) on the 1200 pairwise comparison results collected from all 12 trials. Specifically, let Y be 1 if the left picture is selected and 0 if the right picture is selected. Let

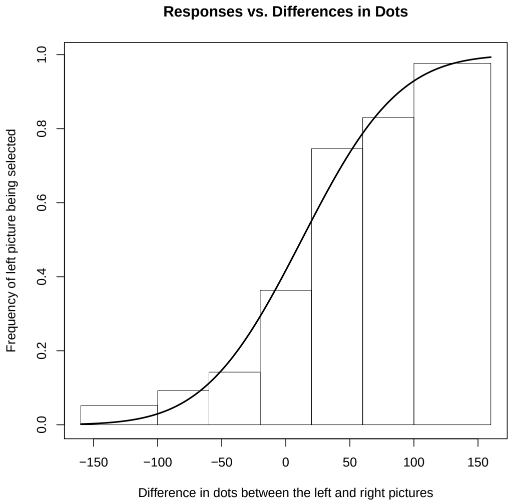

Figure 2: Frequency of the left picture being selected in the 1200 pairwise comparisons of all 12 trials. The x-axis represents the difference in number of dots between the left and right pictures (left $-$ right). The observations are grouped into 7 buckets according to the difference in dots. Each bar represents the empirical frequency for the corresponding bucket.The curve is $\Phi(0.017x)$.

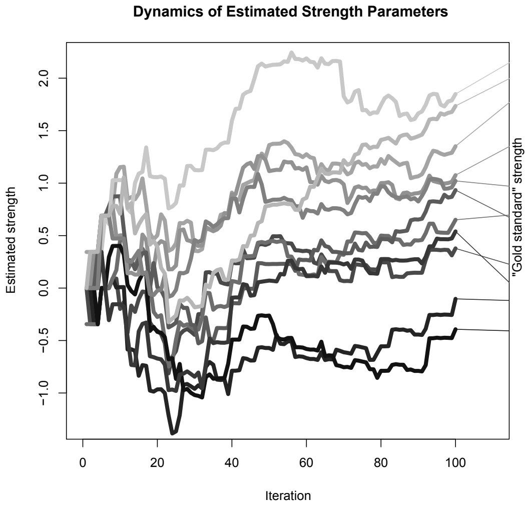

Figure 3: The dynamics of the estimated strength parameters for an adaptive polling trial. The x-axis is the number of iterations. The y-axis is the value of the estimated strength parameters. The rightmost part of the figure labels the value of the "gold standard" strength parameter for each picture.

X be the number of dots in the left picture minus the number of dots in the right picture. Then, $\mathrm{P}(Y = 1|X) = \Phi(X\beta)$, where $\beta = \frac{1}{\sqrt{2}\sigma}$, and we have 1200 observations for (X, Y). The Probit regression gives us an estimate $\hat{\beta} = 0.017$. Multiplying $(s'_i - s'_k)$ by $\hat{\beta}$, we obtain the "gold standard" strength parameters -0.41, -0.12, 0, 0.03, 0.22, 0.66, 0.7, 0.97, 1.36, 1.79, 2.01, and 2.16 for the 12 pictures. The picture with 342 dots (the third lowest) is used as the reference picture and hence has a strength parameter of 0. Since we only perform a linear transformation, a picture with more dots has a larger "gold standard" strength parameter.

A fair concern with our model is whether the Thurstone-Mosteller model accurately characterizes the participants' information in our setting. To evaluate this assumption, we compare the empirical frequencies of the Turkers' responses with those predicted by the Thurstone-Mosteller model. By equation (1), the probability for a participant to select picture i in a pairwise comparison between pictures i and j is $\Phi\left(\frac{s'_i - s'_j}{\sqrt{2}\sigma}\right)$, and we estimated $\frac{1}{\sqrt{2}\sigma} = 0.017$ using all the collected data. Thus, the Thurstone-Mosteller model predicts that the empirical frequencies of the Turkers' responses should closely follow the distribution $\Phi(0.017(s'_i - s'_j))$. Figure 2 plots the empirical frequency of the left picture being selected in our experiment for seven brackets of differences in dots between the left and right pictures. The empirical frequency matches the cdf well, indicating that our setting does not significantly deviate from the Thurstone-Mosteller model. We notice that Turkers have a slight bias toward selecting the picture on the right, because when the difference in number of dots is around 0, the frequency of the left picture being selected is about 40%, in contrast to the 50% predicted by the model.

Next, we look into whether our method effectively incorporates information over time. Figure 3 shows the dynamics of the estimated strength parameters for one of the adaptive polling trials [1]. Since the strength parameter for the picture with 342 dots is set to 0, the estimates are for the other 11 pictures. The lines are colored in grayscale such that the lightest color corresponds to the picture with the most dots and the darkest line corresponds to the picture with fewest dots. We can see that all pictures start with an estimated strength parameter of 0. As more pairwise comparisons are polled, the estimated strength parameters diverge. The overall trend is that the estimated strength parameters of pictures with more dots increase and those of pictures with less dots decrease, showing that information is aggregated into the estimates over time. The right side of Figure 3 labels the value of the "gold standard" strength parameter for each picture. At the end of 100 iterations, the estimated strength parameters are close to the gold standard strength parameters. The produced ranking is generally correct, except that two adjacent pairs are flipped. A closer look reveals that these two flipped pairs have the smallest difference in dots among all adjacent pairs of the 11 pictures, with 381 and 383 dots and 344 and 355 dots respectively.

Finally, we compare the performance of adaptive polling with that of random polling. In addition to our collected data, we also run 100 trials of simulation for each method using the "gold standard" strength parameters to understand what we should expect to see if our model perfectly captures the noisiness of the setting and we know the strength parame-

[1] The other 5 adaptive polling trials exhibit similar dynamics.

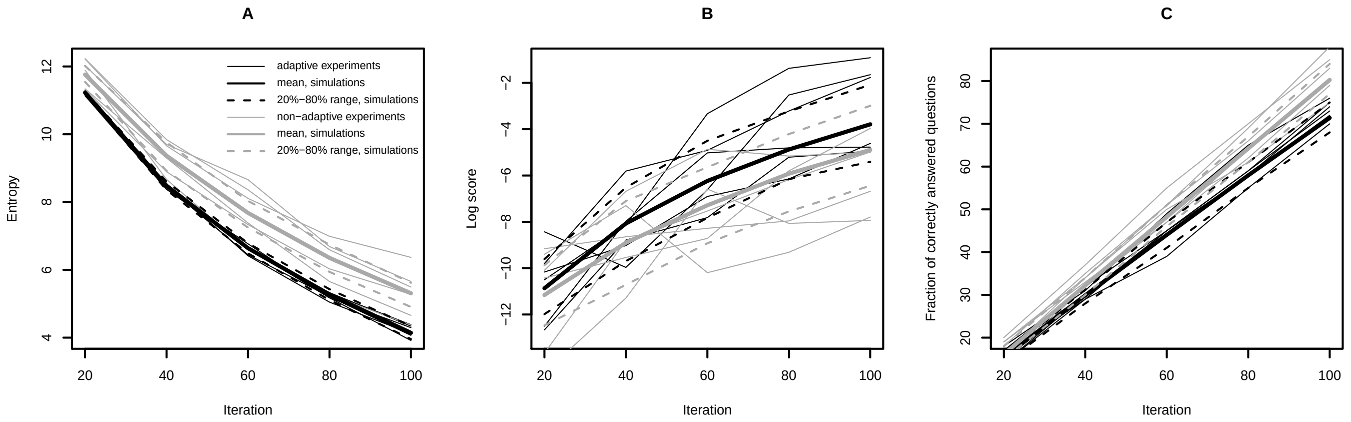

Figure 4: Performance comparison between adaptive polling and random polling. Black lines are for adaptive polling while grey lines are for random polling. Each thin line corresponds to an experimental trial. The thick lines are the average value of 100 simulations using the "gold standard" strength parameters. The two dashed lines of the same color give the 20%-80% range of the simulated values. The x-axes are the number of iterations. Figure A plots the entropy of the estimated distribution $\mathcal{N}(\hat{s}, \hat{\Sigma})$. Figure B shows the log score — the logarithm of the pdf of $\mathcal{N}(\hat{s}, \hat{\Sigma})$ evaluated at the "gold standard" strength parameters. Figure C presents the fraction of the pairwise comparison questions that are correctly answered.

ters. Figure 4 presents the results of the MTurk experiments and our simulations.

Intuitively, we expect the adaptive method to reduce the entropy of the estimated distribution more quickly than the random method, since the adaptive method is optimized for quickly reducing the uncertainty of the probabilistic estimates of the strength parameters. In Figure 4.A, we show a plot of the entropy of the estimated distribution $\mathcal{N}(\hat{s}, \hat{\Sigma})$, which is calculated as $\log \sqrt{(2\pi e)^n |\hat{\Sigma}|}$ where $|\hat{\Sigma}|$ is the determinant of $\hat{\Sigma}$. This figure confirms that the entropy of the estimated distribution indeed decreases faster for the adaptive polling than for the random polling. The difference between the two methods in terms of entropy is statistically significant by two-tailed t-test ($p = 0.01$).

Next, Figure 4.B presents a comparison of the log score of the estimated distributions for the two methods. The log score is often used to measure the accuracy of a probabilistic prediction, so it is a good indicator for how well our method performs in estimating the strength parameters. Having a high log score means that our method produces accurate estimates of the strength parameters. Given an estimated distribution $\mathcal{N}(\hat{s}, \hat{\Sigma})$ and the "gold standard" strength parameters $\boldsymbol{s}$, the log score is the logarithm of the pdf of $\mathcal{N}(\hat{s}, \hat{\Sigma})$ evaluated at $\boldsymbol{s}$. Figure 4.B shows that the log scores for both adaptive and random polling increase over time. The log scores for adaptive polling are higher but the variation is large. The differences between the two methods in terms of log score is statistically significant by two-tailed t-test ($p = 0.016$).

Interestingly, according to Figure 4.C, the fraction of pairwise comparison questions that are answered correctly is lower for adaptive polling than for the random polling, and the difference is statistically significant by two-tailed t-test ($p = 0.005$). This observation suggests that adaptive polling tends to ask relatively difficult comparison questions. The answers to these questions are more valuable for improving the estimates of the strength parameters, even though the participants are less likely to answer them correctly. Moreover, since we pay Turkers a bonus only for correct answers, this implies that the cost of adaptive polling is lower than that of random polling. For our experiment, an average of 10% in bonus payment is saved per trial by using adaptive polling instead of the random method.

6 Discussion and Conclusion

Although the Thurstone-Mosteller model suitably captures the noisiness of participants' information in our experiments, it has some limitations. The model implicitly assumes that participants are ex-ante equally informed and their mistakes are independent. These may not hold in some settings where some participants are better informed than others and mistakes of participants are correlated. In future work, we are interested in studying how our approach performs in such settings and developing suitable methods for them.

Even though we only evaluated our method for a setting with a known underlying ranking of the candidates, our method can be easily adapted for settings when the underlying ranking is unknown. In this case, it is crucial to decide on a suitable termination condition for our algorithm. Since our model produces probabilistic estimates of the strength parameters, we could, for instance, choose to stop the algorithm once a desired entropy of the estimated distribution is reached. It is an interesting future direction to explore different termination conditions for applying our algorithm to such settings.

In conclusion, we demonstrate that eliciting and aggregating information about the ranking of n competing candidates can be effectively achieved by adaptively polling participants recruited from an online labor market on simple pairwise comparison questions and gradually incorporating the collected information into an overall prediction. Our experiments demonstrate that this method is robust against the unpredictable noise in the participants' information and it is effective in eliciting and aggregating information while requiring only simple interactions with the participants.

7 Acknowledgments

This material is based upon work supported by NSF Grant No. CCF-0915016 and Xerox Foundation. Pfeiffer and Rand gratefully acknowledge support from the John Templeton Foundation. Gao is partially supported by an NSERC Postgraduate Scholarship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors alone.

References

Abernethy, J.; Chen, Y.; and Wortman Vaughan, J. 2011. An optimization-based framework for automated market-making. In *Proceedings of the 12th ACM conference on Electronic commerce*, EC '11, 297–306. New York, NY, USA: ACM.

Ailon, N. 2011. Active learning ranking from pairwise preferences with almost optimal query complexity. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems 24*. 810–818.

Bagnoli, M., and Bergstrom. 1989. Log-concave probability and its applications. *Economic Theory* 26:445–469.

Bradley, R. A., and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* 39(3/4):pp. 324–345.

Brochu, E.; de Freitas, N.; and Ghosh, A. 2007. Active preference learning with discrete choice data. In *Advances in Neural Information Processing Systems*.

Carterette, B.; Bennett, P. N.; Chickering, D. M.; and Dumais, S. T. 2008. Here or there: Preference judgments for relevance.

Chen, Y.; Fortnow, L.; Lambert, N.; Pennock, D. M.; and Wortman, J. 2008. Complexity of combinatorial market makers. In *EC '08: Proceedings of the 9th ACM conference on Electronic commerce*, 190–199. New York, NY, USA: ACM.

Elo, A. 1978. *The Rating of Chess Players: Past and Present*. New York: Acro Publishing.

Glickman, M. E., and Jensen, S. T. 2005. Adaptive paired comparison design. *Journal of Statistical Planning and Inference* 127:279–293.

Goel, S.; Reeves, D. M.; Watts, D. J.; and Pennock, D. M. 2010. Prediction without markets. In *EC '10: Proceedings of the 11th ACM Conference on Electronic Commerce*, 357–366. New York, NY, USA: ACM.

Horton, J. J. 2010. The dot-guessing game: A fruit fly for human computation research. *SSRN eLibrary*.

Houlsby, N.; Huszár, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian active learning for classification and preference learning. arXiv:1112.5745.

Kittur, A.; Smus, B.; Khamkar, S.; and Kraut, R. E. 2011. Crowdforge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, 43–52. New York, NY, USA: ACM.

Law, E., and von Ahn, L. 2011. *Human Computation*. Morgan & Claypool Publishers.

Liem, B.; Zhang, H.; and Chen, Y. 2011. An iterative dual pathway structure for speech-to-text transcription. In *HCOMP '11: The 3rd Human Computation Workshop*.

Little, G.; Chilton, L. B.; Goldman, M.; and Miller, R. C. 2009. Turkit: tools for iterative tasks on mechanical turk. In *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation*, 29–30. New York, NY, USA: ACM.

Little, G.; Chilton, L.; Goldman, M.; and Miller, R. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*, 68–76.

Long, B.; Chapelle, O.; Zhang, Y.; Chang, Y.; Zheng, Z.; and Tseng, B. 2010. Active learning for ranking through expected loss optimization. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, 267–274. New York, NY, USA: ACM.

Luce, R. 2005. Individual choice behavior: A theoretical analysis. *books.google.com*.

McCullagh, P., and Nelder, J. A., eds. 1989. *Generalized Linear Models*. Boca Raton, FL, USA: Chapman and Hall/CRC.

McFadden, D. 1974. *Conditional logit analysis of qualitative choice behavior*, volume 1. Academic Press. 105–142.

Mosteller, F. 1951. Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* 16(1):3–9.

Noronha, J.; Hysen, E.; Zhang, H.; and Gajos, K. Z. 2011. Platemate: Crowdsourcing nutrition analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, 1–12. New York, NY, USA: ACM.

Pennock, D., and Xia, L. 2011. Price updating in combinatorial prediction markets with bayesian networks. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, 581–588. Corvallis, Oregon: AUAI Press.

Plackett, R. 1975. The analysis of permutations. *Applied Statistics*.

Settles, B. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Thurstone, L. L. 1927. A law of comparative judgement. *Psychological Review* 34:273–286.

Wolfers, J., and Zitzewitz, E. 2004. Prediction markets. *Journal of Economic Perspective* 18(2):107–126.

Xia, L., and Pennock, D. 2011. An efficient monte-carlo algorithm for pricing combinatorial prediction markets for tournaments. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI), Barcelona, Catalonia, Spain*.