Aggregating Binary Judgments Ranked by Accuracy

Daniel Halpern,[1] Gregory Kehne,[2] Dominik Peters,[1]
Ariel D. Procaccia,[1] Nisarg Shah,[3] Piotr Skowron[4]
[1] Harvard University, [2] Carnegie Mellon University, [3] University of Toronto, [4] University of Warsaw
dhalpern@g.harvard.edu,gkehne@andrew.cmu.edu,dpeters@seas.harvard.edu,
arielpro@seas.harvard.edu,nisarg@cs.toronto.edu,p.skowron@mimuw.edu.pl

Abstract

We revisit the fundamental problem of predicting a binary ground truth based on independent binary judgments provided by experts. When the accuracy levels of the experts are known, the problem can be solved easily through maximum likelihood estimation. We consider, however, a setting in which we are given only a ranking of the experts by their accuracy. Motivated by the worst-case approach to handle the missing information, we consider three objective functions and design efficient algorithms for optimizing them. In particular, the recently popular distortion objective leads to an intuitive new rule. We show that our algorithms perform well empirically using real and synthetic data in collaborative filtering and political prediction domains.

1 Introduction

Consider the task of predicting a binary ground truth $G \in \{0, 1\}$ by aggregating independent binary judgments provided by n experts. This models a wide range of real-world scenarios, where the judgments can be polls predicting the outcome of an upcoming political or sports event, a user's reviews of previously watched movies, weather forecasts, or juror opinions of a defendant's guilt.

The judgment of expert i, denoted X_i, is assumed to be a Bernoulli random variable, which coincides with the ground truth with probability p_i; this probability is referred to as the *accuracy* of the expert. If $\boldsymbol{p} = (p_1, \ldots, p_n)$ is known, then the classical *maximum likelihood estimation* approach chooses the ground truth estimate that maximizes the likelihood of inducing the vector of expert judgments $\boldsymbol{X} = (X_1, \ldots, X_n)$, i.e., the value of $y \in \{0, 1\}$ that maximizes $\mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}] = \prod_{i=1}^{n} p_i^{\mathbb{I}[X_i = y]} \cdot (1 - p_i)^{\mathbb{I}[X_i \neq y]}$, where $\mathbb{I}$ is the indicator variable.

However, sometimes we may not know the exact values of $p_1, \ldots, p_n$; instead, we may only know a *ranking* of the expert judgments by accuracy. This may be the case when there is metadata available about the judgments that is known to be correlated with accuracy, but the exact nature of the correlation is not known. For instance, if a pollster conducts multiple polls over time, polls conducted closer to the date of the event being predicted may be considered more accurate than the ones conducted earlier; the same reasoning applies to weather forecasts. Similarly, polls conducted concurrently may be ranked by their sample sizes. Sometimes, experts may participate in a judgment contest (such as the Good Judgment Project[1]), which may show their ranking by accuracy on the leaderboard.

Motivated by such settings, we address the following question in this work:

How should we aggregate n binary judgments ranked by accuracy in order to predict a binary ground truth?

Note that the n binary judgments ordered by accuracy can be represented as a bit-string of length n. Thus, we essentially study aggregation rules which take a bit-string as input and output a bit. Due to the fundamental nature of this setting, the rules designed in this work may have applications in other domains (see Section 6).

Our Contribution

Recall that the likelihood function $\mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$ depends on $p_1, \ldots, p_n$, i.e., on the accuracy of the experts. However, we are given only partial information about these values, namely, their ordering. To address this missing information, we take a worst-case viewpoint. Specifically, let $\mathcal{P}$ denote the set of all $\boldsymbol{p}$ which are consistent with the given ordering; we define the following three natural objectives that serve as proxies for the likelihood induced by a given estimate $y \in \{0, 1\}$, and design algorithms to compute the estimate optimizing these objectives.

1. *Distortion:* $\sup_{\boldsymbol{p} \in \mathcal{P}} \mathcal{L}[\boldsymbol{X}; G = 1-y, \boldsymbol{p}] / \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$. Note that this is worst-case ratio of the likelihood of the estimate not chosen $(1 - y)$ to the likelihood of the estimate chosen (y). Our aim is to minimize this objective.
2. *Optimistic likelihood:* $\sup_{\boldsymbol{p} \in \mathcal{P}} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$. Maximizing this objective can be thought of as a natural extension of the maximum likelihood approach, where we make an inference about $\boldsymbol{p}$ together with one about the estimate y.
3. *Pessimistic likelihood:* $\inf_{\boldsymbol{p} \in \mathcal{P}} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$. Maximizing this objective can be thought of as maximizing the worst-case likelihood.

[1] https://goodjudgment.com/

In Section 3, we characterize the rules which optimize these objectives, and show that they can be implemented in polynomial time. In particular, the rules optimizing the first two objectives are novel and elegant. In Section 4, we restrict our attention to a natural family of rules, which we refer to as *scoring rules*. These rules assign monotonic weights to judgments (i.e., judgments ranked higher by accuracy receive no less weight than those ranked lower), and return the estimate with the highest total weight. We characterize the scoring rules that optimize the three aforementioned objectives among all scoring rules. In the full version[2], we also consider three other approaches, namely, an axiomatic approach, a Bayesian approach, and a randomized approach.

Finally, in Section 5, we empirically evaluate the performance of the rules designed in this work against some baselines. The experiments use synthetic and real data in the domain of collaborative filtering, and real data in the domain of political predictions. Overall, given their low information requirements, our rules do remarkably well.

Related Work

Our paper contributes to a large body of work in computational social choice (Brandt et al. 2016). A central feature that separates our setting from the vast majority of papers in the area is that the judgments (or opinions, or preferences) that are being aggregated are typically assumed to be anonymous, in the sense that individuals are indistinguishable. However, it has been noted that there are important contexts where anonymity leads to bad outcomes (Lang 2019).

Our setting is related to judgment aggregation (Endriss 2016), an area that also aggregates binary judgments. Typically, this literature does not assume the existence of a ground truth, except for some work on epistemic judgment aggregation (Hartmann and Sprenger 2012; Bozbay, Dietrich, and Peters 2014; Terzopoulou and Endriss 2019). While our model deals with only a single issue, judgment aggregation focuses on problems arising from the aggregation of several logically related issues simultaneously.

In statistics there is influential work on the problem of estimating the common mean of multiple normal distributions (Cohen and Sackrowitz 1974; Jordan and Krishnamoorthy 1996), where the unknown variance of each distribution can be seen as a measure of (in)accuracy. Our setting is more closely related to the work of Ghosh, Kale, and McAfee (2011), who, like us, consider a binary ground truth (for each "item"), and binary judgments, each of which is correct with some probability that depends on the expert's unknown accuracy. The central idea that distinguishes our work from these papers is that we assume a known ranking of the experts by accuracy. This assumption also guides our choice of (worst-case) optimization objectives, which are different from the statistical estimation problems considered in previous work.

Some of our main results pertain to the distortion objective. This objective was conceived in the context of social-welfare maximization in voting settings (Procaccia and Rosenschein 2006; Boutilier et al. 2015; Caragiannis et al. 2017; Anshelevich et al. 2018), but several papers have applied the idea to other problems such as matching, facility location, and even traveling salesperson (Anshelevich and Sekar 2016; Abramowitz and Anshelevich 2018; Anshelevich and Zhu 2018).

Our aggregation rules can be viewed as *simple games* (Taylor and Zwicker 1999) where the experts are players, and winning coalitions correspond to sets of experts such that when all these experts report 1, then so does the aggregation rule. The simple games literature has also studied *linear* simple games, which correspond to games with ranked players. This literature includes characterization results for *weighted* simple games (Taylor and Zwicker 1992), which correspond to what we call scoring rules.

2 Model

For $k \in \mathbb{N}$, let us denote $[k] = \{1, \dots, k\}$. Let $G \in \{0, 1\}$ denote an unknown binary ground truth. Let $N = [n]$ denote a set of experts. Each expert $i \in N$ provides a binary judgment $X_i \in \{0, 1\}$, which is a Bernoulli random variable that is correct with probability p_i, i.e., $\Pr[X_i = G] = p_i$. We refer to $\boldsymbol{X} = (X_1, \dots, X_n)$ as the *judgment profile* and $\boldsymbol{p} = (p_1, \dots, p_n)$ as the *accuracy profile*.

In this work, we make two crucial assumptions regarding $\boldsymbol{X}$ and $\boldsymbol{p}$. First, we assume that the expert judgments (i.e. $X_1, \dots, X_n$) are independent. Second, we assume that each expert is at least as accurate as a coin toss, i.e., $p_i \geqslant 1/2$ for each $i \in N$. For a discussion about relaxing these assumptions, see Section 6.

For $y \in \{0, 1\}$, the likelihood of observing $\boldsymbol{X}$ when the ground truth is $G = y$ can now be written as

$$\mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}] = \textstyle\prod_{i=1}^{n} p_i^{\mathbb{I}[X_i = y]} \cdot (1 - p_i)^{\mathbb{I}[X_i \neq y]},$$

where $\mathbb{I}$ denotes the indicator variable. If the accuracy profile $\boldsymbol{p}$ is known, then a classical approach to aggregating the expert judgments would be to return the *maximum likelihood estimate* (MLE) of the ground truth given by $\operatorname{argmax}_{y \in \{0,1\}} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$.

In this work, we assume that we do not know $\boldsymbol{p}$. Instead, we are given a ranking of the experts by their accuracy, and we are interested in aggregating the expert judgments subject to this ordinal information. Without loss of generality, assume that $p_1 \geqslant p_2 \geqslant \dots \geqslant p_n$. Thus, expert 1 is the most accurate, while expert n is the least accurate. Let $\mathcal{P}_n = \{\boldsymbol{p} : 1 \geqslant p_1 \geqslant \dots \geqslant p_n \geqslant 1/2\}$ denote the set of feasible accuracy profiles. Note that $\mathcal{P}_n$ contains the accuracy profile $\boldsymbol{p} = (1, \dots, 1)$, under which the likelihood of any non-unanimous judgment profile $\boldsymbol{X}$ is zero, regardless of the estimate y. This makes some of our objectives not well-defined or uninteresting. Of course, in practice, no judgment is perfectly accurate. To circumvent this hypothetical inconsistency, we define $\mathcal{P}_n^\epsilon = \{\boldsymbol{p} : 1 - \epsilon \geqslant p_1 \geqslant \dots \geqslant p_n \geqslant 1/2\}$, analyze the aggregation rules optimizing our objectives defined with respect to $\mathcal{P}_n^\epsilon$, and then take the limit $\epsilon \to 0$. In the limit, these rules "converge", in the sense that they become fixed once ϵ is small enough. When the objective is well-defined directly with respect to $\mathcal{P}_n$, we avoid taking this longer route.

[2] Full version: www.cs.toronto.edu/ nisarg/papers/ranked-binary-judgments.pdf

Formally, our input is the bit-string $\boldsymbol{X} \in \{0,1\}^n$, where we refer to X_1 as the *most accurate bit* and X_n as the *least accurate*. An *aggregation function* is denoted $f : \{0,1\}^n \to \{0,1,\perp\}$, where $\perp$ denotes a tie.[3] We will alternatively represent a tie as the function returning $\{0,1\}$ instead of $\perp$.

We are also interested in a natural family of aggregation functions that we refer to as *scoring rules*. A scoring rule $f_{\boldsymbol{w}}$ is parametrized by a weight vector $\boldsymbol{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n_{\geqslant 0}$, where w_i is the weight associated with the i-th most accurate bit. Given input $\boldsymbol{X}$, $f_{\boldsymbol{w}}$ returns the bit with the highest total weight, i.e., $\operatorname{argmax}_{y\in\{0,1\}} \sum_{i=1}^n w_i \cdot \mathbb{I}[X_i = y]$. This definition is inspired by that of a prominent family of voting rules called positional scoring rules, which includes well-known rules such as plurality and Borda count.

3 Worst-Case Optimal Aggregation Rules

Given incomplete information about the accuracy profile $\boldsymbol{p}$, we cannot compute the MLE, since different accuracy profiles $\boldsymbol{p}$ consistent with the given ordinal information may induce different likelihoods. Our approach is to define an objective function that summarizes the likelihoods induced by all feasible $\boldsymbol{p}$ and optimize it; we consider three proposals.

Distortion

Informally, given an objective function and ordinal information about cardinal inputs to the function, the distortion approach selects an outcome minimizing the ratio between the optimal objective value and the objective value under the selected outcome, in the worst case over all cardinal inputs consistent with the given ordinal information. The objective we are interested in is the likelihood function $\mathcal{L}$, and we are given ordinal information about $\boldsymbol{p}$ (specifically, that $\boldsymbol{p} \in \mathcal{P}_n$). Given a judgment profile $\boldsymbol{X}$, the *distortion* of ground truth estimate $y \in \{0,1\}$ is then defined as

$$\begin{aligned}\operatorname{dist}(y;\boldsymbol{X}) &= \sup_{\boldsymbol{p}\in\mathcal{P}_n} \frac{\max\left(\mathcal{L}[\boldsymbol{X};G=0,\boldsymbol{p}],\mathcal{L}[\boldsymbol{X};G=1,\boldsymbol{p}]\right)}{\mathcal{L}[\boldsymbol{X};G=y,\boldsymbol{p}]} \\ &= \sup_{\boldsymbol{p}\in\mathcal{P}_n} \frac{\mathcal{L}[\boldsymbol{X};G=1-y,\boldsymbol{p}]}{\mathcal{L}[\boldsymbol{X};G=y,\boldsymbol{p}]}.\end{aligned}$$

Here, the second equality is due to the fact that with $G = y$, the ratio is always 1, which can also be achieved with $G = 1 - y$ at $p_1 = \ldots = p_n = 1/2$ (which makes the likelihoods given both possible ground truths equal). Hence, the worst case is achieved with $G = 1 - y$ in the numerator. Given this definition, the *distortion-optimal* estimate is $y^* \in \operatorname{argmax}_{y\in\{0,1\}} \operatorname{dist}(y;\boldsymbol{X})$.

We borrow the idea of distortion from the voting literature, where the goal is to select a candidate maximizing total voter happiness, but only voters' ranked preferences (and not the exact intensities) are known. Distortion offers a prior-independent evaluation of candidates; a candidate minimizing distortion is the best choice given only the information available.

This objective requires attention to the technicality mentioned in Section 2. Let $\boldsymbol{p} = (1,\ldots,1) \in \mathcal{P}_n$, and consider a profile $\boldsymbol{X}$ in which not all judgments agree. Then $\mathcal{L}[\boldsymbol{X};G=0,\boldsymbol{p}] = \mathcal{L}[\boldsymbol{X};G=1,\boldsymbol{p}] = 0$, making distortion undefined. Hence, we use $\mathcal{P}^\epsilon_n = \{\boldsymbol{p} : 1-\epsilon \geqslant p_1 \geqslant \ldots \geqslant p_n \geqslant 1/2\}$ to redefine the distortion as

$$\operatorname{dist}^\epsilon(y;\boldsymbol{X}) = \sup_{\boldsymbol{p}\in\mathcal{P}^\epsilon_n} \frac{\mathcal{L}[\boldsymbol{X};G=1-y,\boldsymbol{p}]}{\mathcal{L}[\boldsymbol{X};G=y,\boldsymbol{p}]}.$$

The distortion-optimal rule f_{dist} is defined as $f_{\text{dist}}(\boldsymbol{X}) = \lim_{\epsilon\to 0} \operatorname{argmin}_{y\in\{0,1\}} \operatorname{dist}^\epsilon(y;\boldsymbol{X})$. Interestingly, we show that the estimate y minimizing $\operatorname{dist}^\epsilon(y;\boldsymbol{X})$ is independent of ϵ, making the limit unnecessary. First, we define a quantity that we will later show to be closely related to distortion.

Definition 1. Given $\boldsymbol{X} \in \{0,1\}^n$, the *strength* $s_{\boldsymbol{X}}(y)$ of estimate y is the maximum difference between the number of occurrences of y and that of $1-y$ in any prefix of $\boldsymbol{X}$, i.e.,

$$s_{\boldsymbol{X}}(y) = \max_{k\in[n]\cup\{0\}} \textstyle\sum_{i=1}^k \left\{\mathbb{I}[X_i = y] - \mathbb{I}[X_i = 1-y]\right\}.$$

Lemma 1. *For $\epsilon \in (0, 1/2)$, $n \in \mathbb{N}$, $\boldsymbol{X} \in \{0,1\}^n$, and $y \in \{0,1\}$, we have $\operatorname{dist}^\epsilon(y;\boldsymbol{X}) = \left(\frac{1-\epsilon}{\epsilon}\right)^{s_{\boldsymbol{X}}(1-y)}$.*

Proof. Fix $y \in \{0,1\}$. Given a sequence $\boldsymbol{p}$, we say that it has a *jump* at $i \in [n-1]$ if $p_i > p_{i+1}$.

We first show that in the definition of $\operatorname{dist}^\epsilon(y;\boldsymbol{X})$, the supremum over $\boldsymbol{p}$ is achieved at an accuracy profile with at most one jump. Let $\boldsymbol{p}$ be a vector with the minimum jumps at which the supremum is achieved. Suppose for contradiction that it has at least two jumps, and let k and j be indices such that $p_k > p_{k+1} = \ldots = p_j > p_{j+1}$.

Define $\boldsymbol{p}^1$ and $\boldsymbol{p}^2$ such that $p^1_i = p^2_i = p_i$ for $i \in [n] \setminus \{k+1,\ldots,j\}$, $p^1_i = p_k$ for $i \in \{k+1,\ldots,j\}$, and $p^2_i = p_{j+1}$ for $i \in \{k+1,\ldots,j\}$. That is, in $\boldsymbol{p}^1$, we shift the block $(p_{k+1},\ldots,p_j)$ up and make it equal to p_k, and in $\boldsymbol{p}^2$, we shift it down and make it equal to p_{j+1}.

We show that at least one of these two vectors must yield an approximation ratio no better than that at $\boldsymbol{p}$, and is therefore also a point where the supremum is achieved; this is a contradiction because they both have one fewer jump than $\boldsymbol{p}$. To see why the claim is true, let $a = p_k$, $b = p_{k+1} = \ldots = p_j$, and $c = p_{j+1}$. Thus, $a > b > c \geqslant 1/2$. Denoting $S = \{k+1,\ldots,j\}$, we have that

$$\begin{aligned}\frac{\mathcal{L}[\boldsymbol{X};G=1-y,\boldsymbol{p}]}{\mathcal{L}[\boldsymbol{X};G=y,\boldsymbol{p}]} &= \prod_{i\in[n]\setminus S} \frac{p_i^{\mathbb{I}[X_i=1-y]}\cdot(1-p_i)^{\mathbb{I}[X_i=y]}}{p_i^{\mathbb{I}[X_i=y]}\cdot(1-p_i)^{\mathbb{I}[X_i=1-y]}} \\ &\quad\times \prod_{i\in S} \frac{b^{\mathbb{I}[X_i=1-y]}\cdot(1-b)^{\mathbb{I}[X_i=y]}}{b^{\mathbb{I}[X_i=y]}\cdot(1-b)^{\mathbb{I}[X_i=1-y]}} \\ &= \prod_{i\in[n]\setminus S} \frac{p_i^{\mathbb{I}[X_i=1-y]}\cdot(1-p_i)^{\mathbb{I}[X_i=y]}}{p_i^{\mathbb{I}[X_i=y]}\cdot(1-p_i)^{\mathbb{I}[X_i=1-y]}} \\ &\quad\times \left(\tfrac{b}{1-b}\right)^{\sum_{i\in S}(\mathbb{I}[X_i=1-y]-\mathbb{I}[X_i=y])}.\end{aligned}$$

In the last expression, as $b > 1/2$, we have $b/(1-b) > 1$. Thus, if the exponent of $b/(1-b)$ is non-positive, then decreasing b to c does not decrease the expression, and the expression changes from the approximation ratio at $\boldsymbol{p}$ to that

[3] Allowing ties does not significantly alter most of our results; we discuss some of the implications of ties in later sections.

at $\boldsymbol{p}^2$. Similarly, if the exponent is non-negative, then increasing b to a does not decrease the expression, and the expression changes from the approximation ratio at $\boldsymbol{p}$ to that at $\boldsymbol{p}^1$. Hence, at least one of $\boldsymbol{p}^1$ and $\boldsymbol{p}^2$ achieves a ratio at least as high as $\boldsymbol{p}$, as desired.

We have established that the supremum is achieved at some $\boldsymbol{p}$ which has at most one jump. Then, there exist $a, b \in [1/2, 1-\epsilon]$ with $a > b$ and an index $k \in [n] \cup \{0\}$ such that $p_i = a$ for all $i \leqslant k$ and $p_i = b$ for all $i > k$. Note that allowing $k = 0$ and $k = n$ permits zero jumps. We show that we can let $a = 1-\epsilon$ and $b = 1/2$ without loss of generality. The approximation ratio at this $\boldsymbol{p}$ is given by

$$\left(\tfrac{a}{1-a}\right)^{\sum_{i=1}^{k}(\mathbb{I}[X_i=1-y]-\mathbb{I}[X_i=y])} \times \left(\tfrac{b}{1-b}\right)^{\sum_{i=k+1}^{n}(\mathbb{I}[X_i=1-y]-\mathbb{I}[X_i=y])}.$$

The exponent of $a/(1-a)$ must be non-negative (otherwise decreasing a to b would strictly increase the approximation ratio). Hence, increasing a to $1-\epsilon$ does not decrease the approximation ratio. Similarly, we can let $b = 1/2$.

We have thus established that the supremum is achieved at $\boldsymbol{p}$ such that for some $k \in [n] \cup \{0\}$, $p_i = 1-\epsilon$ for $i \leqslant k$ and $p_i = 1/2$ for $i > k$. Thus, the distortion of y given X is

$$\mathrm{dist}^{\epsilon}(y; X) = \max_{k \in [n] \cup \{0\}} \left(\tfrac{1-\epsilon}{\epsilon}\right)^{\sum_{i=1}^{k}(\mathbb{I}[X_i=1-y]-\mathbb{I}[X_i=y])},$$

which is $\left(\frac{1-\epsilon}{\epsilon}\right)^{s_{\boldsymbol{X}}(1-y)}$, as desired. □

We can immediately obtain a characterization of the distortion-optimal estimate $y^* \in \{0,1\}$ by observing that $\operatorname{argmin}_{y\in\{0,1\}} s_{\boldsymbol{X}}(1-y) = \operatorname{argmax}_{y\in\{0,1\}} s_{\boldsymbol{X}}(y)$ and applying Lemma 1: The distortion-optimal estimate is the estimate with the greatest strength in $\boldsymbol{X}$.

Theorem 1. *For any $\epsilon \in (0, 1/2)$, $n \in \mathbb{N}$, and $\boldsymbol{X} \in \{0,1\}^n$,*

$$f_{\mathrm{dist}}(\boldsymbol{X}) = \operatorname*{argmin}_{y\in\{0,1\}} \mathrm{dist}^{\epsilon}(y; \boldsymbol{X}) = \operatorname*{argmax}_{y\in\{0,1\}} s_{\boldsymbol{X}}(y),$$

where f_{dist} is the distortion-optimal rule. Further, this can be computed in linear time.

Note that in case of both estimates having equal strength, the result also implies that their distortion will be equal.

A notable property of f_{dist} is that if more than $n/3$ most accurate judgments or more than $2n/3$ least accurate judgments are identical, then that will be the output of f_{dist}, regardless of the remaining judgments.

Other Objectives

We now turn our attention to two other objectives, namely maximization of optimistic and pessimistic likelihoods. Recall that the reason we cannot directly compute the MLE $\operatorname{argmax}_{y\in\{0,1\}} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$ is because we do not know the exact accuracy profile $\boldsymbol{p}$. Instead, we know that $\boldsymbol{p} \in \mathcal{P}_n$. Given this, we define the optimistic and pessimistic likelihoods by taking the best case and the worst case over the choice of $\boldsymbol{p}$, respectively.

The *optimistic likelihood* $\mathcal{L}_{\uparrow}$ of observing $\boldsymbol{X}$ when the ground truth is $G = y$ is $\mathcal{L}_{\uparrow}[\boldsymbol{X}; G = y] = \sup_{\boldsymbol{p}\in\mathcal{P}_n} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$. The *optimistic MLE* rule which maximizes this objective, denoted $f_{\mathrm{MLE}\uparrow}$, is given by $f_{\mathrm{MLE}\uparrow}(\boldsymbol{X}) = \operatorname{argmax}_{y\in\{0,1\}} \mathcal{L}_{\uparrow}[\boldsymbol{X}; G = y]$. We can view $f_{\mathrm{MLE}\uparrow}$ as simply performing a joint maximum likelihood estimation over $(y, \boldsymbol{p}) \in \{0,1\} \times \mathcal{P}_n$, and returning the y component of the resulting estimate.[4]

Algorithm 1: OPT-LIKELIHOOD

Input: Judgment profile $\boldsymbol{X} \in \{0,1\}^n$, $y \in \{0,1\}$
Output: Optimistic likelihood $\mathcal{L}_{\uparrow}[\boldsymbol{X}; G = y]$
if $n = 1$ **then**
  return $1^{\mathbb{I}[X_1=y]} \cdot (1/2)^{\mathbb{I}[X_1\neq y]}$
end
[Find the prefix of $\boldsymbol{X}$ with the highest density of y]
$i \leftarrow$ index maximizing $(1/i) \cdot \sum_{j=1}^{i} \mathbb{I}[X_j = y]$, breaking ties in favor of larger indices
$d \leftarrow (1/i) \cdot \sum_{j=1}^{i} \mathbb{I}[X_j = y]$
$r \leftarrow \max\{d, 1/2\}$
$L \leftarrow$ OPT-LIKELIHOOD$((X_{i+1}, \ldots, X_n), y)$
return $\left(r^d(1-r)^{1-d}\right)^i \cdot L$

We begin by presenting an algorithm that calculates the optimistic likelihood of an estimate $y \in \{0,1\}$ given a judgment profile $\boldsymbol{X}$. The algorithm repeatedly identifies a prefix of $\boldsymbol{X}$ with the highest density of y, and imputes that the accuracies of judgments in that prefix are equal to this density. The following is proved in the full version.

Theorem 2. *Algorithm 1 calculates the optimistic likelihood $\mathcal{L}_{\uparrow}[\boldsymbol{X}; G = y]$ in polynomial time. Thus, the aggregation rule $f_{\mathrm{MLE}\uparrow}$ can be implemented in polynomial time.*

We illustrate Algorithm 1 by an example.

Example 1. Let us consider running OPT-LIKELIHOOD with $\boldsymbol{X} = (0,1,1,1,0,1,1,0,0,1)$ and $y = 1$.

The first iteration selects $i = 4$ (i.e. prefix $(0,1,1,1)$) because the density of $y = 1$ in this prefix is $3/4$, and this is the highest density in any prefix. This leads to $d = r = 3/4$. The second iteration selects $i = 3$ (i.e. prefix $(0,1,1)$), leading to $d = r = 2/3$. The final iteration selects the remaining string, and sets $d = 1/3$ but $r = 1/2$.

Thus, $\boldsymbol{p} = (3/4, 3/4, 3/4, 3/4, 2/3, 2/3, 2/3, 1/3, 1/3, 1/3)$ is the accuracy profile leading to the optimistic likelihood of

$$\left(\frac{3}{4}^{\frac{3}{4}} \cdot \frac{1}{4}^{\frac{1}{4}}\right)^4 \cdot \left(\frac{2}{3}^{\frac{2}{3}} \cdot \frac{1}{3}^{\frac{1}{3}}\right)^3 \cdot \left(\frac{1}{2}^{\frac{1}{3}} \cdot \frac{1}{2}^{\frac{2}{3}}\right)^3.$$ □

We now turn our attention to maximizing the pessimistic likelihood $\mathcal{L}_{\downarrow}$. If we define it as $\mathcal{L}_{\downarrow}[\boldsymbol{X}; G = y] = \inf_{\boldsymbol{p}\in\mathcal{P}_n} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$, then we run into the issue discussed in Section 2: the pessimistic likelihood of any non-unanimous $\boldsymbol{X}$ becomes 0 under both values of y due to the accuracy profile $\boldsymbol{p} = (1, \ldots, 1) \in \mathcal{P}_n$, leading to unnecessary ties. Hence, we again consider $\mathcal{P}_n^{\epsilon}$ instead of $\mathcal{P}_n$, define $\mathcal{L}_{\downarrow}^{\epsilon}[\boldsymbol{X}; G = y] = \inf_{\boldsymbol{p}\in\mathcal{P}_n^{\epsilon}} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$, and define

[4] This is also equivalent to computing the maximum a posteriori estimate (MAP) when we are given a uniform prior over $\boldsymbol{p}$. For computing MAP given other priors, see the full version.

the *pessimistic MLE* rule, denoted $f_{\mathrm{MLE}\downarrow}$, as $f_{\mathrm{MLE}\downarrow}(\boldsymbol{X}) = \lim_{\epsilon\to 0}\operatorname{argmax}_{y\in\{0,1\}} \mathcal{L}^{\epsilon}_{\downarrow}[\boldsymbol{X}; G = y]$. When compared to $f_{\mathrm{MLE}\uparrow}$, $f_{\mathrm{MLE}\downarrow}$ is more in line with a worst-case approach. Unlike in the case of distortion, the choice of y does not turn out to be independent of ϵ, but as we see in the proof of Theorem 3, the rule converges once $\epsilon < 2^{-n}$.

The next result identifies $f_{\mathrm{MLE}\downarrow}$ analytically. This is possible because the accuracy profile resulting in the pessimistic likelihood always consists of only $1-\epsilon$ and $1/2$. This is in contrast to the one leading to the optimistic likelihood, which, as Example 1 demonstrates, can be more complex. The following is proved in the full version.

Theorem 3. *The pessimistic MLE rule $f_{\mathrm{MLE}\downarrow}$, given a judgment profile $\boldsymbol{X}$, outputs the majority judgment; if tied, it outputs the opposite of the least accurate judgment (i.e. $1-X_n$).*

4 Optimal Scoring Rules

We now turn our attention to a natural class of aggregation rules, scoring rules. Specifically, we are interested in how well we can optimize certain objectives when we are restricted to this class of functions. There are two clear ambiguities about scoring rules that we must take into account. First, for many choices of objectives, there is no scoring rule which is *instance optimal*, i.e., at least as good as every other scoring rule on *all* possible judgment profiles. To handle this, we modify our goal slightly and instead choose scoring rules that are optimal in the worst case. Next, is the issue of ties. In this case we'll take a pessimistic view (in line with our worst case objective goals) and say that when a scoring rule outputs a tie, the value of the objective in this instance will be the worse of the two outcomes.

Theorem 4. *For any $\epsilon \in (0, 1/2)$ and $n \in \mathbb{N}$, the scoring rule given by $\boldsymbol{w}^* = (1,\dots,1,0,\dots,0)$ with exactly $2\lfloor n/3\rfloor + 1$ ones minimizes the worst case distortion $\max_{\boldsymbol{X}\in\{0,1\}^n} \mathrm{dist}^{\epsilon}(f_{\boldsymbol{w}}(\boldsymbol{X}); \boldsymbol{X})$ over all possible scoring rules parametrized by $\boldsymbol{w} \in \mathbb{R}^n_{\geqslant 0}$.*

Proof. Fix $\epsilon \in (0, 1/2)$ and $n \in \mathbb{N}$. Recall that minimizing the distortion, $\mathrm{dist}^{\epsilon}(y; \boldsymbol{X})$ is equivalent to minimizing the strength of the unchosen judgment, $s_{\boldsymbol{X}}(1-y)$.

First, we'll show that no rule f (scoring or otherwise) can guarantee $s_{\boldsymbol{X}}(1-f(\boldsymbol{X})) < \lfloor n/3\rfloor$ for all $\boldsymbol{X} \in \{0,1\}^n$. This will imply $\max_{\boldsymbol{X}\in\{0,1\}^n} s_{\boldsymbol{X}}(1-f_{\boldsymbol{w}}(\boldsymbol{X})) \geqslant \lfloor n/3\rfloor$ for all $\boldsymbol{w} \in \mathbb{R}^n_{\geqslant 0}$. To see this, we construct $\boldsymbol{X} \in \{0,1\}^n$ such that both $s_{\boldsymbol{X}}(0)$ and $s_{\boldsymbol{X}}(1)$ are at least $\lfloor n/3\rfloor$. Consider the judgment profile $\boldsymbol{X}^s = (1,\dots,1,0,\dots,0)$ with $\lfloor n/3\rfloor$ ones and $n - \lfloor n/3\rfloor$ zeros. On the prefix of the first $\lfloor n/3\rfloor$ judgments, there are $\lfloor n/3\rfloor$ more 1s than there are 0s. Thus, the strength of 1 is at least $\lfloor n/3\rfloor$. On the other hand, on the entire profile, there are $(n - \lfloor n/3\rfloor) - \lfloor n/3\rfloor \geqslant n - \frac{2n}{3} \geqslant \lfloor n/3\rfloor$ more 0s then 1s. Hence, the strength of 0 is at least $\lfloor n/3\rfloor$, as desired.

Next, we show that $s_{\boldsymbol{X}}(1 - f_{\boldsymbol{w}^*}(\boldsymbol{X})) \leqslant \lfloor n/3\rfloor$ for all judgment profiles $\boldsymbol{X} \in \{0,1\}^n$. Qualitatively, $f_{\boldsymbol{w}^*}$ simply picks the majority bit of the first $2\lfloor n/3\rfloor + 1$ bits. Note that since $2\lfloor n/3\rfloor + 1$ is odd, there is always a majority bit and thus $f_{\boldsymbol{w}^*}$ will never output a tie.

Let $\boldsymbol{X} \in \{0,1\}^n$ and, without loss of generality, suppose $f_{\boldsymbol{w}^*}(\boldsymbol{X}) = 1$. We show that $s_{\boldsymbol{X}}(0) \leqslant \lfloor n/3\rfloor$. Since $f_{\boldsymbol{w}^*}$ chose 1, there cannot be a majority of 0s in the first $2\lfloor n/3\rfloor + 1$ bits. Hence, 0 occurs at most $\lfloor n/3\rfloor$ times in this prefix. This implies that for $k \leqslant 2\lfloor n/3\rfloor + 1$, $\sum_{i=1}^{k}\{\mathbb{I}[X_i = 0] - \mathbb{I}[X_i = 1]\} \leqslant \lfloor n/3\rfloor$. Next, since 1 has a majority among the first $2\lfloor n/3\rfloor + 1$ bits, $\sum_{i=1}^{2\lfloor n/3\rfloor+1}\{\mathbb{I}[X_i = 0] - \mathbb{I}[X_i = 1]\} \leqslant -1$. or $k > 2\lfloor n/3\rfloor + 1$,

$$\begin{aligned}
&\textstyle\sum_{i=1}^{k}\{\mathbb{I}[X_i = 0] - \mathbb{I}[X_i = 1]\}\\
&\leqslant \textstyle\sum_{i=2\lfloor n/3\rfloor+2}^{k}\{\mathbb{I}[X_i = 0] - \mathbb{I}[X_i = 1]\} - 1\\
&\leqslant k - (2\lfloor n/3\rfloor + 1) - 1\\
&\leqslant n - (2\lfloor n/3\rfloor + 1) - 1\\
&= n - (3\lfloor n/3\rfloor + 2) + \lfloor n/3\rfloor\\
&\leqslant n - n + \lfloor n/3\rfloor = \lfloor n/3\rfloor.
\end{aligned}$$

So, for all $k \in \{0\} \cup [n]$, $\sum_{i=1}^{k}\{\mathbb{I}[X_i = 0] - \mathbb{I}[X_i = 1]\} \leqslant \lfloor n/3\rfloor$, and hence $s_{\boldsymbol{X}}(0) \leqslant \lfloor n/3\rfloor$ as desired. □

Other Objectives

We also investigate optimal scoring rules with respect to the optimistic and pessimistic MLE rules $f_{\mathrm{MLE}\uparrow}$ and $f_{\mathrm{MLE}\downarrow}$. It is easy to see that maximizing the optimistic (resp. pessimistic) likelihood of the chosen estimate is equivalent to minimizing the optimistic (resp. pessimistic) likelihood of the unchosen estimate; thus, we can also view $f_{\mathrm{MLE}\uparrow}$ and $f_{\mathrm{MLE}\downarrow}$ as minimizing the optimistic and pessimistic likelihoods of the unchosen estimate, respectively. This connection holds only because we are looking for the optimal rule within the family of all possible rules; however, when we look for the optimal rule within the family of scoring rules, we have to consider four — not two — objectives.

Definition 2. We define optimal scores $\boldsymbol{w}^{\uparrow}_{\circ}, \boldsymbol{w}^{\uparrow}_{\times}, \boldsymbol{w}^{\downarrow}_{\circ}, \boldsymbol{w}^{\downarrow}_{\times}$ as

- $\boldsymbol{w}^{\uparrow}_{\circ} \in \arg\max_{\boldsymbol{w}\in\mathbb{R}^n_{\geqslant 0}} \min_{\boldsymbol{X}} \mathcal{L}_{\uparrow}[\boldsymbol{X}; G = f_{\boldsymbol{w}}(\boldsymbol{X})]$
- $\boldsymbol{w}^{\uparrow}_{\times} \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^n_{\geqslant 0}} \max_{\boldsymbol{X}} \mathcal{L}_{\uparrow}[\boldsymbol{X}; G = 1 - f_{\boldsymbol{w}}(\boldsymbol{X})]$
- $\boldsymbol{w}^{\downarrow}_{\circ} \in \arg\max_{\boldsymbol{w}\in\mathbb{R}^n_{\geqslant 0}} \min_{\boldsymbol{X}} \mathcal{L}_{\downarrow}[\boldsymbol{X}; G = f_{\boldsymbol{w}}(\boldsymbol{X})]$
- $\boldsymbol{w}^{\downarrow}_{\times} \in \arg\min_{\boldsymbol{w}\in\mathbb{R}^n_{\geqslant 0}} \max_{\boldsymbol{X}} \mathcal{L}_{\downarrow}[\boldsymbol{X}; G = 1 - f_{\boldsymbol{w}}(\boldsymbol{X})]$.

For example, $\boldsymbol{w}^{\uparrow}_{\circ}$ maximizes the optimistic likelihood of its chosen answer in the worst case. For this rule it suffices to always choose the most accurate expert's judgment:

Theorem 5. *The score $\boldsymbol{w}^{\uparrow}_{\circ} = (1, 0, \dots, 0)$ is optimal.*

For the cases based on pessimistic likelihood (both maximizing it for the chosen answer and minimizing it for the unchosen answer), characterizing the optimum scoring rule is easy, since the optimum rule we identified in Theorem 3 can be represented as a scoring rule.

Theorem 6. *Scores $\boldsymbol{w}^{\downarrow}_{\circ} = \boldsymbol{w}^{\downarrow}_{\times} = (1,\dots,1,1/2)$ are optimal for $\epsilon \leqslant 2^{-n}$, and coincide with the rule of Theorem 3.*

Note that in the scoring vectors of Theorem 6, the $1/2$ component could be replaced with any value strictly between 0 and 1 without changing the aggregation rule. In general, throughout this section, the scoring rules we identify are optimal but not uniquely optimal (cf. Definition 2).

The remaining case, $\boldsymbol{w}^{\uparrow}_{\times}$, that is minimizing the optimistic likelihood of the unchosen answer, is less straightforward. Using a linear program, we have obtained optimal scores $\boldsymbol{w}^{\uparrow}_{\times}$ for $n \leqslant 20$; these are cataloged in the full version. For general n, $\boldsymbol{w}^{\uparrow}_{\times}$ is unknown, but in the full version we also show that there exists an optimal scoring rule that is nonincreasing, for all n.

5 Experiments

In this section we assess through computer simulations the quality of decisions made by our aggregation functions in the context of two example applications.

Collaborative Filtering

Consider a set of agents N, a set of issues I, and a partially observed binary matrix $(x_{ij})_{i\in N, j\in I}$. We interpret an entry $x_{ij} \in \{0, 1\}$ as the decision of agent i on issue j (for example, reviewer i bids on paper j). In each run of the experiment, we randomly select an entry of the matrix, hide it, and use several algorithms to guess its value. An algorithm is successful if it guesses correctly. We repeat the experiment 1,000 times to assess the average accuracy of the algorithms.

We use our aggregation functions to predict hidden values in a matrix as follows. For an agent $i \in N$, let $R(i)$ be the set of issues j such that the entry x_{ij} is observed. Given a hidden entry x_{ij^*} we first identify the set of agents $k \in N$ for which the value x_{kj^*} is observed. Second, we rank those agents by their similarity to i. Formally, we define the similarity score of two agents i and k as $\mathrm{sim}(i, k) = |\{j \in R(k) \cap R(i) \colon x_{ij} = x_{kj}\}|/|R(k) \cap R(i)|$, that is the fraction of issues on which i and k agreed among all issues for which we have data from both agents. We rank the agents k in the descending order of their similarity score with i. Thus, we assume that more similar agents are better predictors of the hidden decision x_{ij^*} of agent i. We truncate the list of agents to the first half (we use this heuristic since our algorithms where designed for the case where $p > 1/2$). The ranked decisions by agents on issue j^* then form the input to the aggregation functions from Sections 3 and 4 and to the Bayesian algorithm from full version (with a prior estimated from the data).

We compare our algorithms with three standard Recommender Systems algorithms for matrix completion, implemented in the `fancyimpute` python library[5]: MatrixFactorization (MF), Iterative SVD (ISVD), and Soft Impute (SI). We evaluate the rules on two datasets from the PrefLib library (Mattei and Walsh 2013), and on a synthetic dataset.

Sushi. This dataset contains information about individuals' preferences on various types of sushi. There are 100 types of sushi, and each individual assigns scores from $\{1, \ldots, 4\}$ to 10 randomly selected sushi sets. We filter only those individuals who assigned 4 different scores to the sets (there are 2737 such agents), and convert their preferences to binary judgments as follows. For a fixed value of $d \in \{3, 4\}$ we set the decision of an agent i for sushi j to 1 if i assigns to j the score at least equal to d; the decision is 0 if the corresponding score is lower than d. Note that only 10% of entries of this matrix are observed.

Conference Bidding (CONF). This dataset contains reviewers' bids on papers at a major computer science conference. We convert the reviewers' bids to their binary judgments over papers by setting the decision to one if they bid "yes" for a paper ($d = \mathrm{Y}$) or by setting it to one if they bid "yes" or "maybe" for a paper ($d = \mathrm{M}$). Additionally, we hide an h fraction of randomly selected entries in the matrix ($h \in \{0.5, 0.8, 0.9\}$).

Synthetic Model (SYNT) Each agent and each issue is represented by a d-dimensional vector of attributes ($d \in \{5, 10\}$). For each agent and each issue we sample the value of each attribute independently and uniformly from $[-1, 1]$. An agent i decides 1 on an issue j if the dot product of their corresponding attribute vectors is positive. Otherwise i decides 0. We hide an h fraction of randomly selected entries in the matrix ($h \in \{0.5, 0.8, 0.9\}$).

Political Predictions

We use a dataset from FiveThirtyEight of polling data from the 2016 US Presidential Election. We convert this data into a binary format by choosing a threshold, the mean of the number of votes the candidate received over all polls in that state, then reporting 1 if the poll was above this threshold and 0 if it was below. In addition, we assume that polls' accuracies are sorted by their recency, that is later polls are more accurate than earlier ones.

We run an experiment for each US state and each candidate. The ground truth is taken to be whether the true number of votes the candidate received in the general election was above or below the threshold. We then analyze our algorithms: given the sorted binary polling data, do they correctly predict the ground truth? For each state and candidate, algorithms get a score of 1 for getting the ground truth correctly, 0 for being incorrect, and $1/2$ for a tie. The algorithms' overall scores are their average over all states. Note that for a few states, there is no polling data for a certain candidate in which case the state was not included in the score. This is why the scores are not all multiples of $1/50$. Finally, since older data may be inaccurate and could even hurt accuracy, we compare two settings: using all available polls and restricting the algorithms to polls conducted on or after October 1, 2016. The election took place on November 8, 2016.

Results

Representative results of our experiments for selected values of the parameters are summarized in Table 1.

1. The scoring rules using vectors $\boldsymbol{w}^{\uparrow}_{\times}$ and $\boldsymbol{w}^{\uparrow}_{\circ}$ are suboptimal: for most datasets the distortion-optimal rule achieved better accuracies than these rules (the only exception is JOHNSON-ALL, where f_{dist} performed worse than $\boldsymbol{w}^{\uparrow}_{\times}$). Similarly, $f_{\mathrm{MLE}\downarrow}$ performed (slightly) better than f_{dist} on only two datasets (CLINTON-OCT and CONF ($d = \mathrm{M}, h = 0.9$)), and for several other datasets it produced significantly worse results.
2. f_{dist}, $f_{\mathrm{MLE}\uparrow}$, and the scoring rule using $\boldsymbol{w}^*$ perform comparably well, though each excelled in different datasets.

[5] https://github.com/iskandr/fancyimpute

	f_{dist}	$f_{\text{MLE}\uparrow}$	$f_{\text{MLE}\downarrow}$	sc ($\boldsymbol{w}^*$)	sc ($\boldsymbol{w}^{\uparrow}_{\times}$)	sc ($\boldsymbol{w}^{\uparrow}_{\circ}$)	Bayesian	MT	ISVD	SI
SUSHI ($d = 3$)	65.3	**66.6**	65.2	65.5	62.4	57.0	48.8	50.1	57.5	49.5
SUSHI ($d = 4$)	68.6	69.4	67.2	**70.1**	67.9	63.3	57.6	60.4	63.3	66.7
CONF ($d = \text{M}, h = 0.5$)	94.8	94.8	94.8	94.8	90.3	94.8	94.8	96.5	94.5	**96.8**
CONF ($d = \text{M}, h = 0.8$)	**95.3**	95.2	**95.3**	**95.3**	92.0	**95.3**	95.2	93.0	91.0	93.5
CONF ($d = \text{M}, h = 0.9$)	95.1	94.6	95.2	95.2	92.3	91.9	90.5	92.0	94.6	**95.6**
SYNT ($h = 0.8, d = 10$)	76.5	73.4	73.7	68.1	64.6	74.2	77.1	46.0	**87.0**	73.5
SYNT ($h = 0.8, d = 5$)	85.4	84.0	83.4	81.5	78.8	85.5	90.1	49.0	**91.5**	89.0
SYNT ($h = 0.5, d = 5$)	89.9	91.2	88.9	87.7	86.7	89.9	92.4	94.0	**94.1**	92.0
CLINTON-ALL	83.3	82.4	74.5	**86.3**	52.9	78.4	84.3	–	–	–
JOHNSON-ALL	68.4	79.6	51.0	79.6	73.5	55.1	**85.7**	–	–	–
TRUMP-ALL	90.2	92.2	82.4	90.2	53.9	90.2	**94.1**	–	–	–
CLINTON-OCT	86.3	82.4	**88.2**	86.3	52.9	78.4	72.5	–	–	–
JOHNSON-OCT	80.6	**81.6**	77.6	71.4	73.5	55.1	71.4	–	–	–
TRUMP-OCT	92.2	92.2	92.2	92.2	53.9	90.2	**98.0**	–	–	–

Table 1: Summary of the experiments comparing accuracies (given as percentages) of aggregation functions. In each row, the best performing algorithms are bolded; those that perform within 1 and 2 percentage points of the best algorithm are shaded dark grey and light grey, respectively. Simulations for parameter values omitted in the table led to qualitatively similar conclusions. We use $sc(\boldsymbol{w})$ to denote the scoring rule parametrized by the vector $\boldsymbol{w}$.

3. For many datasets the Bayesian algorithm outperforms the rules with worst-case guarantees, yet there are instances (such as SUSHI) where the Bayesian algorithm is much worse. If we could pick the best response out of those produced by the Bayesian algorithm and $f_{\text{MLE}\uparrow}$ (or f_{dist}), we would always obtain high-quality results.
4. For some datasets, notably SUSHI, our algorithms outperform standard algorithms for matrix completion. For other datasets, the Bayesian algorithm is comparable to the matrix-completion algorithms. This is promising since our algorithms use less information.
5. In the political domain our best rules produced considerably more accurate predictions than simply trusting the most accurate (most recent) predictions ($\boldsymbol{w}^{\uparrow}_{\circ}$).

We can compare these experimental results to a setting in which the true accuracies of the experts are known. For example, if we take 10 experts and sample their accuracies uniformly from $(0.5, 0.7)$, then the maximum likelihood estimate using both the judgments along with the expert accuracies recovers the ground truth in 76.7% of cases. On the other hand, knowing simply the judgments ordered by accuracies, f_{dist} predicts correctly in 76.1% of cases, $f_{\text{MLE}\uparrow}$ in 74.1%, and $f_{\text{MLE}\downarrow}$ in 74.6%.

6 Discussion

Our setting boils down to the design of Boolean functions that take a string of bits as input and output a single bit — with the twist that the order of bits matters, in that earlier bits are given greater importance. We view this as a fundamental problem, and there are many ways to approach it. In addition to the objectives and algorithms described in Sections 3 and 4, we present three additional approaches in the full version: axiomatic, Bayesian, and randomized .

One might ask whether the assumption that $p_i \geqslant 1/2$ for all $i \in N$ can be relaxed. If the identities of experts with $p_i < 1/2$ are known, we can simply flip their judgments and reverse their order (as the flipped judgment of the least accurate expert is now the most accurate). Interestingly, our problem now becomes that of aggregating *two* strings of judgments, ordered by accuracy, into a single bit. This problem is potentially richer than ours because there is no information on the relative accuracy of experts associated with two different strings. An even more general setup simply provides a *partial* order of the experts by accuracy.

Another natural variant of our setting is one where, instead of binary judgments, experts provide real-valued judgments in, say, $[0, 1]$, and the goal is to aggregate them to return a single real number in $[0, 1]$. Interestingly, given a binary aggregation rule f from our work, one can compute the greatest value $x \in [0, 1]$ such that converting expert judgments to binary depending on whether they are at least x and feeding them to f gives output 1; this is well-defined when f satisfies a natural monotonicity condition. We leave such directions for future work.

Acknowledgements

Halpern, Kehne, Peters and Procaccia were partially supported by the National Science Foundation under grants CCF-2007080, IIS-2024287 and CCF-1733556; and by the Office of Naval Research under grant N00014-20-1-2488. Skowron was supported by Poland's National Science Center grant UMO-2019/35/B/ST6/02215. Shah was partially supported by an NSERC Discovery Grant.

References

Abramowitz, B.; and Anshelevich, E. 2018. Utilitarians without utilities: Maximizing social welfare for graph prob-

lems using only ordinal preferences. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 894–901.

Anshelevich, E.; Bhardwaj, O.; Elkind, E.; Postl, J.; and Skowron, P. 2018. Approximating Optimal Social Choice under Metric Preferences. *Artificial Intelligence* 264: 27–51.

Anshelevich, E.; and Sekar, S. 2016. Blind, Greedy, and Random: Algorithms for Matching and Clustering using only Ordinal Information. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 390–396.

Anshelevich, E.; and Zhu, W. 2018. Ordinal approximation for social choice, matching, and facility location problems given candidate positions. In *Proceedings of the 14th Conference on Web and Internet Economics (WINE)*, 3–20. Springer.

Boutilier, C.; Caragiannis, I.; Haber, S.; Lu, T.; Procaccia, A. D.; and Sheffet, O. 2015. Optimal Social Choice Functions: A Utilitarian View. *Artificial Intelligence* 227: 190–213.

Bozbay, I.; Dietrich, F.; and Peters, H. 2014. Judgment aggregation in search for the truth. *Games and Economic Behavior* 87: 571–590.

Brandt, F.; Conitzer, V.; Endress, U.; Lang, J.; and Procaccia, A. D., eds. 2016. *Handbook of Computational Social Choice*. Cambridge University Press.

Caragiannis, I.; Nath, S.; Procaccia, A. D.; and Shah, N. 2017. Subset Selection Via Implicit Utilitarian Voting. *Journal of Artificial Intelligence Research* 58: 123–152.

Cohen, A.; and Sackrowitz, H. B. 1974. On Estimating the Common Mean of Two Normal Distributions. *Annals of Statistics* 2(6): 1274–1282.

Endriss, U. 2016. Judgment Aggregation. In Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D., eds., *Handbook of Computational Social Choice*, chapter 17. Cambridge University Press.

Ghosh, A.; Kale, S.; and McAfee, P. 2011. Who Moderates the Moderators? Crowdsourcing Abuse Detection in User-Generated Content. In *Proceedings of the 12th ACM Conference on Economics and Computation (EC)*, 167–176.

Hartmann, S.; and Sprenger, J. 2012. Judgment aggregation and the problem of tracking the truth. *Synthese* 187(1): 209–221.

Jordan, S. M.; and Krishnamoorthy, K. 1996. Exact Confidence Intervals for the Common Mean of Several Normal Populations. *Biometrics* 52(1): 77–86.

Lang, J. 2019. In Praise of Nonanomymity: Nonsubstitutable Voting. In Laslier, J.-F.; Moulin, H.; Sanver, R.; and Zwicker, W. S., eds., *The Future of Economic Design*, 97–102. Springer.

Mattei, N.; and Walsh, T. 2013. PrefLib: A Library of Preference Data. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT)*, 259—270.

Procaccia, A. D.; and Rosenschein, J. S. 2006. The Distortion of Cardinal Preferences in Voting. In *Proceedings of the 10th International Workshop on Cooperative Information Agents (CIA)*, 317–331.

Taylor, A. D.; and Zwicker, W. S. 1992. A Characterization of Weighted Voting. *Proceedings of the American Mathematical Society* 115(4): 1089–1094.

Taylor, A. D.; and Zwicker, W. S. 1999. *Simple Games: Desirability Relations, Trading, Pseudoweightings*. Princeton University Press.

Terzopoulou, Z.; and Endriss, U. 2019. Optimal Truth-Tracking Rules for the Aggregation of Incomplete Judgments. In *International Symposium on Algorithmic Game Theory*, 298–311. Springer.