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

[1] https://goodjudgment.com/

1.1 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.

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 appendix, we also consider three other approaches, namely, an axiomatic approach (Appendix A), a Bayesian approach (Appendix B), and a randomized approach (Appendix C).

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.

1.2 Related Work

Our paper contributes to a large body of work in computational social choice [6]. 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 [12].

Our setting is related to judgment aggregation [9], an area that also aggregates binary judgments. However, that literature focuses on problems arising from the aggregation of several logically related issues simultaneously, and does not typically assume a ground truth.

In statistics there is influential work on the problem of estimating the common mean of multiple normal distributions [8, 11], 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 [10], 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 [14, 5, 7, 2], but several paper have applied the idea to other problems such as matching, facility location, and even traveling salesperson [3, 1, 4].

Our aggregation rules can be viewed as *simple games* [16] 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 [15], 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

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

where $\mathbb{1}$ 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.

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.[2] 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 $\boldsymbol{w} = (w_1, \dots, w_n) \in \mathbb{R}^n_{\geqslant 0}$, 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., $\operatorname{argmax}_{y \in \{0,1\}} \sum_{i=1}^{n} w_i \cdot \mathbb{1}[X_i = y]$. This definition is novel in our setting of binary judgments, but it 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.

3.1 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,

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

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

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

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 $y^* \in \operatorname{argmax}_{y\in\{0,1\}} \operatorname{dist}(y;\boldsymbol{X})$.

This objective requires attention to the technicality mentioned in Section 2. Consider $\boldsymbol{X}$ in which some two judgments disagree. Then, under $\boldsymbol{p} = (1,\ldots,1) \in \mathcal{P}_n$, we have $\operatorname{dist}(0;\boldsymbol{X}) = \operatorname{dist}(1;\boldsymbol{X}) = 0$, making distortion undefined. Hence, we use $\mathcal{P}_n^\epsilon = \{\boldsymbol{p} : 1-\epsilon \geqslant p_1 \geqslant \ldots \geqslant p_n \geqslant 1/2\}$ to redefine the distortion as

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

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.,

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

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 $p_i^1 = p_i^2 = p_i$ for $i \in [n] \setminus \{k+1,\ldots,j\}$, $p_i^1 = p_k$ for $i \in \{k+1,\ldots,j\}$, and $p_i^2 = p_{j+1}$ 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 yields 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, $a > b > c$. Denoting $S = \{k+1,\ldots,j\}$, we have that

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

In the last expression, 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 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

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

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

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

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 $\mathrm{argmin}_{y \in \{0,1\}}\, s_{\boldsymbol{X}}(1 - y) = \mathrm{argmax}_{y \in \{0,1\}}\, s_{\boldsymbol{X}}(y)$ 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.

3.2 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 $\mathrm{argmax}_{y \in \{0,1\}}\, \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$ 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.[3] On the other hand, we can view $f_{\mathrm{MLE}\downarrow}$ as inspired by worst-case analysis. Further, 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; in

[3] 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 Appendix B.

Section 4.1, we will see that when we look for the optimal rule within the family of scoring rules, we have to consider four — not two — objectives.

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.

```
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 $1^{\mathbb{1}[X_1 = y]} \cdot (1/2)^{\mathbb{1}[X_1 \neq y]}$
  end
  [Find the prefix of $\boldsymbol{X}$ with the highest density of $y$]
  $i \leftarrow$ index maximizing $(1/i) \cdot \sum_{j=1}^{i} \mathbb{1}[X_j = y]$, breaking ties in favor of larger indices
  $d \leftarrow (1/i) \cdot \sum_{j=1}^{i} \mathbb{1}[X_j = y]$
  $r \leftarrow \max\{d, 1/2\}$
  $L \leftarrow$ OPT-LIKELIHOOD$((X_{i+1}, \ldots, X_n), y)$
  return $\left(r^d (1-r)^{1-d}\right)^i \cdot L$
```

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.*

It is clear that Algorithm 1 runs in polynomial time. Here we prove the claim that for any $\boldsymbol{X}$ and answer y, the probability vector

$$\boldsymbol{p}^* := \arg\max_{\boldsymbol{p} \in \mathcal{P}_n} \mathcal{L}[\boldsymbol{X} | G = y, \boldsymbol{p}]$$

can be found iteratively by identifying the index i which maximizes the density $r(\boldsymbol{X}^i)$ of y over all prefixes $\boldsymbol{X}^i := (X_1, \ldots, X_i)$, taking $p_1^*, \ldots p_i^* = \max\{r(\boldsymbol{X}^i), 1/2\}$, and recursing on the suffix $\overline{\boldsymbol{X}}^i := (X_{i+1}, \ldots, X_n)$. We show that $\boldsymbol{p}^*$ takes this form for $\boldsymbol{p} \in \mathcal{P}_n$ and only remark that the analogous structure emerges in maximizing over the larger set $\mathcal{Q}_n := \{\boldsymbol{p} : 1 \geqslant p_1 \geqslant \ldots \geqslant p_n \geqslant 0\}$.

We begin with two observations. First, that for a given run of experts which share a p_i, the maximizing value of these p_i is the density of the answer y in this run:

Lemma 2. *If $p_1 = p_2 = \ldots p_n$ then $\boldsymbol{p} = \arg\max_{\boldsymbol{p} \in \mathcal{Q}_n} \mathcal{L}_\uparrow[\boldsymbol{X} | G = y]$ is given by $r(\boldsymbol{X}) = \frac{||\mathbb{1}[X_i = y]||_1}{n}$.*

Proof. Given that $p := p_1 = p_2 = \ldots p_n$, the likelihood is

$$\mathcal{L}[\boldsymbol{X} | G = y, \boldsymbol{p}] = p^{||\mathbb{1}[X_i = y]||_1} (1-p)^{||\mathbb{1}[X_i \neq y]||_1}.$$

Taking derivatives shows that this is concave with respect to p, with unique maximum over $[0, 1]$ at $p = r(\boldsymbol{X})$. □

Next, if multiple vectors of experts have likelihoods maximized by compatible probability vectors, then the concatenation of these probability vectors maximizes the likelihood of the concatenated expert vectors:

Lemma 3. *If $\boldsymbol{p}_1 = \arg\max_{\boldsymbol{p} \in \mathcal{P}_{n_1}} \mathcal{L}[\boldsymbol{X}_1 | G = y, \boldsymbol{p}], \ldots, \boldsymbol{p}_k = \arg\max_{\boldsymbol{p} \in \mathcal{P}_{n_k}} \mathcal{L}[\boldsymbol{X}_k | G = y, \boldsymbol{p}]$ and $p \geqslant p'$ for all $p \in \boldsymbol{p}_j$, $p' \in \boldsymbol{p}_{j+1}$ and for all j, then*

$$\boldsymbol{p} = \arg\max_{\boldsymbol{p} \in \mathcal{P}_n} \mathcal{L}[\boldsymbol{X} | G = y, \boldsymbol{p}],$$

where $\boldsymbol{p} := (\boldsymbol{p}_1 | \ldots | \boldsymbol{p}_k)$ and $\boldsymbol{X} := (\boldsymbol{X}_1 | \ldots | \boldsymbol{X}_k)$ and $n := \sum_{i=1}^{i} n_i$.

Proof. In short, this is because $\mathcal{P}_n \subset \mathcal{P}_{n_1} \times \ldots \times \mathcal{P}_{n_k}$ and because

$$\mathcal{L}[(\boldsymbol{X}_1 | \ldots | \boldsymbol{X}_k) | G = y, (\boldsymbol{p}_1 | \ldots | \boldsymbol{p}_k)] = \mathcal{L}[\boldsymbol{X}_1 | G = y, \boldsymbol{p}_1] \times \ldots \times \mathcal{L}[\boldsymbol{X}_k | G = y, \boldsymbol{p}_k]. \tag{1}$$

Since the $\boldsymbol{p}_i$ maximize the (nonnegative) terms $\mathcal{L}[\boldsymbol{X}_i|G = y, \boldsymbol{p}_i]$ individually, they maximize the product over the larger space $\mathcal{P}_{n_1} \times \ldots \times \mathcal{P}_{n_k}$. The compatibility of the $\boldsymbol{p}_i$ implies that in fact $\boldsymbol{p} \in \mathcal{P}_n$; therefore it maximizes $\mathcal{L}[\boldsymbol{X}|G = y, \boldsymbol{p}]$ over $\mathcal{P}_n$ also. □

Next, it will be useful to view each $\boldsymbol{p}$ as inducing a decomposition of $\boldsymbol{X}$ into chunks $\boldsymbol{X} = (C_1|\ldots|C_k)$ within which all p_i are equal. Let $q_1 \ldots q_k$ be the values of the p_i within each of the chunks, and let $r_1 \ldots r_k$ be the chunk densities $r_j := r(C_j)$.

We now tackle a special case of Theorem 2.

Lemma 4. *If $r(\boldsymbol{X}^i)$ is maximized by $i = n$ then the optimal $\boldsymbol{p}^*$ is $p_1^*, \ldots p_n^* = \max\{r(\boldsymbol{X}^n), 1/2\}$.*

Proof. To see this, let $\boldsymbol{p}^* = \operatorname{argmax}_{\boldsymbol{p} \in \mathcal{P}_n}$. If $r := r(\boldsymbol{X}^i)$ then there are two cases to consider. First if $r > 1/2$ then, since $r_1 \leqslant r$ and $\sum_{j=1}^{k} \frac{|C_j|}{n} r_j = r$, it follows that either $r_1 = \ldots = r_k = r$ or there is some j for which $r_j \leqslant r_{j+1}$. But since $\boldsymbol{p}^*$ is decreasing, $q_j \geqslant q_{j+1}$. Therefore by Lemma 2 either $\mathcal{L}[C_j|G = y, (q_j, \ldots, q_j)]$ or $\mathcal{L}[C_{j+1}|G = y, (q_{j+1}, \ldots, q_{j+1})]$ can be increased by either lowering q_j towards r_j or raising q_{j+1} towards r_{j+1}. In either case $q_1, \ldots, q_k$ remain nonincreasing and by Equation (1) the likelihood $\mathcal{L}[\boldsymbol{X}|G = y, \boldsymbol{p}^*]$ increases also, contradicting the optimality of $\boldsymbol{p}^*$. Therefore $r_1 = \ldots = r_k = r$, in which case by Lemma 2 and Lemma 3 $\mathcal{L}[\boldsymbol{X}|G = y, \boldsymbol{p}^*]$ is maximized by $p_1^* = \ldots = p_n^* = r$.

Finally consider $r \leqslant 1/2$. Since r is the maximum density, $r_1 \leqslant 1/2$ also. And since $\boldsymbol{p} \in \mathcal{P}_n$ we have that $q_1, \ldots q_k \geqslant 1/2$. If $q_1 > 1/2$ then since $r_1 \leqslant 1/2$ we have by Lemma 2 that $\mathcal{L}[C_1|G = y, (q_1, \ldots, q_1)]$ is increased by lowering q_1 to the value of q_2. By Equation (1) this increases $\mathcal{L}[\boldsymbol{X}|G = y, \boldsymbol{p}^*]$, contradicting the optimality of $\boldsymbol{p}^*$. Therefore $q_1 = 1/2$, and so $q_1, \ldots q_k = 1/2$, as desired. □

We finally prove Theorem 2 by induction.

Proof of Theorem 2. For the base case take $n = 1$. By Lemma 4 it is of the desired form.

Next suppose that for all $\boldsymbol{X}'$ such that $|\boldsymbol{X}'| < n$, the optimal probability vector has the desired form. Given some $\boldsymbol{X}$ of length n, let i be the index maximizing the prefix density $r(\boldsymbol{X}^i)$ of y in $\boldsymbol{X}$, and call this maximum density r. Splitting it into $\boldsymbol{X} = (\boldsymbol{X}^i|\overline{\boldsymbol{X}}^i)$, by Lemma 4 we have that $\boldsymbol{X}^i$ has likelihood maximized by $\boldsymbol{p}^*$ for which $p_1^* = \ldots = p_i^* = r$. By the inductive hypothesis, we also know that the $\overline{\boldsymbol{p}}^*$ maximizing likelihood for $\overline{\boldsymbol{X}}^i$) is of the desired form. Therefore $(\boldsymbol{p}^*|\overline{\boldsymbol{p}}^*)$ is of the desired form also.

It remains to show that $(\boldsymbol{p}^*|\overline{\boldsymbol{p}}^*)$ is nonincreasing and maximizes $\mathcal{L}[X|G = y, p]$. Clearly for all $\boldsymbol{p} \in \boldsymbol{p}^*$ we have $p = r$ the maximum prefix density for $\boldsymbol{X}$. If r' is the maximum prefix density for $\overline{\boldsymbol{X}}^i$, then by hypothesis $p' \leqslant r'$ for all $p' \in \overline{\boldsymbol{p}}^*$. But $r \geqslant r'$, since otherwise $\boldsymbol{X}^i$ together with the prefix of $\overline{\boldsymbol{X}}^i$ witnessing r' would be a prefix of $\boldsymbol{X}$ with a density larger than r, a contradiction. Therefore $(\boldsymbol{p}^*|\overline{\boldsymbol{p}}^*)$ is nonincreasing, and so by Lemma 3 it also maximizes $\mathcal{L}[X|G = y, \boldsymbol{p}]$. □

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 the *pessimistic MLE* rule, denoted $f_{\mathrm{MLE}\downarrow}$, as $f_{\mathrm{MLE}\downarrow}(\boldsymbol{X}) = \lim_{\epsilon \to 0} \operatorname{argmax}_{y \in \{0,1\}} \mathcal{L}_{\downarrow}^{\epsilon}[\boldsymbol{X}; G = y]$. 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.

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. $1 - X_n$).*

Proof. Fix $\epsilon \in (0, 1/2)$. First, we demonstrate that if $\boldsymbol{p}^\star \in \operatorname{argmin}_{\boldsymbol{p} \in \mathcal{P}_n^\epsilon} \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}]$, then $p_i^\star \in \{1-\epsilon, 1/2\}$ for all $i \in [n]$. Suppose for contradiction that this is not the case. Let $S = \{k, \dots, j\}$ be a maximal contiguous block of indices of $\boldsymbol{p}^\star$ with some value $a \notin \{1-\epsilon, 1/2\}$. Note that the contribution of this block to $\mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}^\star]$ is equal to $a^{\sum_{i \in S} \mathbb{1}[X_i = y]} \cdot (1-a)^{\sum_{i \in S} \mathbb{1}[X_i \neq y]}$. It is easy to check that this is a convex function of a in $[0, 1]$.

Define $p_0^\star = 1 - \epsilon$ and $p_{n+1}^\star = 1/2$. Then, we have that $p_{k-1}^\star > a > p_{j+1}^\star$ and $a \notin \{1-\epsilon, 1/2\}$. Hence, it is feasible to increase or decrease a slightly. Since one of these two operations must reduce the likelihood, we have a contradiction.

Thus, in computing $\mathcal{L}_\downarrow^\epsilon[\boldsymbol{X}; G = y]$, it is sufficient to minimize over $\boldsymbol{p}$ which consist of only $1-\epsilon$ and $1/2$. Hence, we have that $\mathcal{L}_\downarrow^\epsilon[\boldsymbol{X}; G = y]$ is equal to

$$\min_{k \in [n] \cup \{0\}} (1-\epsilon)^{\sum_{i=1}^k \mathbb{1}[X_i = y]} \cdot \epsilon^{\sum_{i=1}^k \mathbb{1}[X_i \neq y]} \cdot (1/2)^{n-k}.$$

Note that the estimate y^* maximizing this objective can be found in linear time. We now study the value of y^* as $\epsilon \to 0$.

When $\epsilon \leqslant 2^{-n}$, the ϵ term in the equation dominates; that is, the pessimistic likelihood is achieved by maximizing the exponent of ϵ, and subject to that, maximizing the exponent of $1/2$. Thus, if y appears more often then $1-y$, then $\mathcal{L}_\downarrow^\epsilon[\boldsymbol{X}; G = 1-y]$ will have a higher exponent of ϵ (and therefore, will be lower) than $\mathcal{L}_\downarrow^\epsilon[\boldsymbol{X}; G = y]$. In this case, the rule will return y. Finally, suppose that both 0 and 1 appear exactly $n/2$ times. If $\boldsymbol{X}$ ends with k occurrences of y, then we will have $\mathcal{L}_\downarrow^\epsilon[\boldsymbol{X}; G = y] = (1-\epsilon)^{n/2-k} \cdot (\epsilon)^{n/2} \cdot (1/2)^k$ whereas $\mathcal{L}_\downarrow^\epsilon[\boldsymbol{X}; G = 1-y] = (1-\epsilon)^{n/2} \cdot (\epsilon)^{n/2}$, leading the rule to return $1-y$, as desired. □

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, it is often unlikely that a scoring rule can be *instance optimal* for a given objective and rules can often be incomparable, one rule achieving a better objective value on a judgment string while worse on another. To handle this, we change 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 $\max_{\boldsymbol{X} \in \{0,1\}^n} \operatorname{dist}^\epsilon(f_{\boldsymbol{w}}(\boldsymbol{X}); \boldsymbol{X})$ 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, $\operatorname{dist}^\epsilon(y; \boldsymbol{X})$ 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 $k \leqslant 2\lfloor n/3 \rfloor + 1$, $\sum_{i=1}^{k} \{\mathbb{1}[X_i = 0] - \mathbb{1}[X_i = 1]\} \leqslant \lfloor n/3 \rfloor$. Next, since 1 has a majority among the first $2\lfloor n/3 \rfloor + 1$ bits, $\sum_{i=1}^{2\lfloor n/3 \rfloor + 1} \{\mathbb{1}[X_i = 0] - \mathbb{1}[X_i = 1]\} \leqslant -1$. or $k > 2\lfloor n/3 \rfloor + 1$,

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

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

4.1 Other Objectives

We also investigate optimal scoring rules with respect to the optimistic and pessimistic MLE rules $f_{\mathrm{MLE}\uparrow}$ and $f_{\mathrm{MLE}\downarrow}$. Section 3.2 posits that $f_{\mathrm{MLE}\uparrow}$ and $f_{\mathrm{MLE}\downarrow}$ can equivalently be viewed as either maximizing their respective likelihoods for the chosen estimate or as minimizing their respective likelihoods for the unchosen estimate. However when it comes to identifying the worst-case optimal scoring rule across judgment profiles, this equivalence ceases to hold. Thus, we need to derive an optimal scoring rule for each case.

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

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

For example, $\boldsymbol{w}_\circ^\uparrow$ 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* $\boldsymbol{w}_\circ^\uparrow = (1, 0, \ldots, 0)$ *is optimal.*

Proof. Consider the even and odd length alternating judgment vectors $\boldsymbol{X}^e := (0,1,0,1,\ldots,0,1)$ and $\boldsymbol{X}^o := (0,1,0,1,\ldots,0,1,0)$. For any fixed $\epsilon > 0$, observe that by Theorem 2 for both $\boldsymbol{X}^{alt} = \boldsymbol{X}^e, \boldsymbol{X}^o$ we have that $\mathcal{L}_\uparrow[\boldsymbol{X}^{alt}; G = 1] = 2^{-n}$, and $\mathcal{L}_\uparrow[\boldsymbol{X}^{alt}; G = 0] = (1-\epsilon)2^{-n+1}$. These judgment vectors are noteworthy because no matter how $f_{\boldsymbol{w}}$ breaks ties, it must choose either 0 or 1 and incur either 2^{-n} or $(1-\epsilon)2^{-n+1}$ as its maximum likelihood for $\boldsymbol{X}^{alt}$. If $Q(\boldsymbol{w}) := \min_{\boldsymbol{X}} \mathcal{L}_\uparrow[\boldsymbol{X}; G = f_w(\boldsymbol{X})]$ is the objective we seek to maximize, then for all $\boldsymbol{w}$ we therefore have that $Q(\boldsymbol{w}) \leqslant (1-\epsilon)2^{-n+1}$.

Now consider the score $\boldsymbol{w}^* = (1, 0, \ldots, 0)$. For any judgment vector $\boldsymbol{X}$, the scoring rule chooses $f_{\boldsymbol{w}^*}(\boldsymbol{X}) = X_1$. Then taking $p = (1-\epsilon, 1/2, \ldots, 1/2)$ we see that $\mathcal{L}_\uparrow[\boldsymbol{X}; f_{\boldsymbol{w}^*}(\boldsymbol{X})] \geqslant \mathcal{L}[\boldsymbol{X}; G = X_1, p] = (1-\epsilon)2^{-n+1}$. Minimizing over all $\boldsymbol{X}$ yields $Q(\boldsymbol{w}^*) \geqslant (1-\epsilon)2^{-n+1}$, and so $\boldsymbol{w}^*$ 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 $\boldsymbol{w}_\circ^\downarrow = \boldsymbol{w}_\times^\downarrow = (1, \dots, 1, 1/2)$ are optimal for $\epsilon \leqslant 2^{-n}$, and coincide with the rule of Theorem 3.*

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

Theorem 7. *For each n, there is a choice of $\boldsymbol{w}_\times^\uparrow$ that is nonincreasing.*

Proof. We proceed by arguing that for any score with an increasing pair of entries, the score which flips these entries to be decreasing performs at least as well with respect to the objective. By applying this repeatedly we find that there is a decreasing optimal scoring rule.

To see this, consider a score $\boldsymbol{w}$ with increasing pair of indices $i < j$ for which $w_i < w_j$, and take $\boldsymbol{w}'$ to be this score which flips these two entries; that is, $\boldsymbol{w}'_i = \boldsymbol{w}_j$, $\boldsymbol{w}'_i = \boldsymbol{w}_j$, and $\boldsymbol{w}'_k = \boldsymbol{w}_k$ for all $k \neq i, j$. Again let $Q(\boldsymbol{w}) := \max_{\boldsymbol{X}} \mathcal{L}_\uparrow[\boldsymbol{X}; G = 1 - f_w(\boldsymbol{X})]$ be the objective quantity which our optimal scoring rule $\boldsymbol{w}_\times^\uparrow$ minimizes; we argue that $Q(\boldsymbol{w}') \leqslant Q(\boldsymbol{w})$.

Let $\boldsymbol{X}'$ be any vector of expert judgments, and let $y := f_w(\boldsymbol{X}')$ denote the estimate chosen by $\boldsymbol{w}$ for judgment vector $\boldsymbol{X}'$. We will construct an $\boldsymbol{X}$ for which $\mathcal{L}_\uparrow[\boldsymbol{X}'; G = 1 - f_{w'}(\boldsymbol{X}')] \leqslant \mathcal{L}_\uparrow[\boldsymbol{X}; G = 1 - f_w(\boldsymbol{X})]$. First, if $f_{w'}(\boldsymbol{X}') = y$ then take $\boldsymbol{X} = \boldsymbol{X}'$. This is the case if $X'_i = X'_j$, since $\boldsymbol{w} \cdot \mathbb{1}[X'_k = y] = \boldsymbol{w}' \cdot \mathbb{1}[X'_k = y]$. If $X'_i = y$ and $X'_j = 1 - y$ then $\boldsymbol{w}' \cdot \mathbb{1}[X'_k = y] \geqslant \boldsymbol{w} \cdot \mathbb{1}[X'_k = y]$ and so $f_{w'}(\boldsymbol{X}') = y$ also.

Therefore when $f_{w'}(\boldsymbol{X}') = 1 - y$ we have that $X'_i = 1 - y$ and $X'_j = y$. In this case let $\boldsymbol{X}$ be $\boldsymbol{X}'$ with X_i and X_j flipped; that is, $X_i := X'_j = y$ and $X_j := X'_i = 1 - y$ and $X_k := X'_k$ otherwise. Note that $f_w(\boldsymbol{X}) = f_{w'}(\boldsymbol{X}') = 1 - y$, since the scores are identical; $\boldsymbol{w} \cdot \mathbb{1}[X_k = 1 - y] = \boldsymbol{w}' \cdot \mathbb{1}[X'_k = 1 - y]$. Let $\boldsymbol{p}'$ be the probability vector which witnesses $\mathcal{L}_\uparrow[\boldsymbol{X}'; G = 1 - f_{w'}(\boldsymbol{X}')] = \mathcal{L}_\uparrow[\boldsymbol{X}'; G = y]$, so that it equals

$$(1 - p'_i) \cdot p'_j \cdot \prod_{X'_k = 1 - y; k \neq i,j} p'_k \prod_{X'_k = y; k \neq i,j} (1 - p'_k).$$

Then since it is a maximum we have that $\mathcal{L}_\uparrow[\boldsymbol{X}; G = 1 - f_w(\boldsymbol{X})] \geqslant \mathcal{L}[\boldsymbol{X}; G = y, \boldsymbol{p}']$, which equals

$$p'_i \cdot (1 - p'_j) \cdot \prod_{X'_k = 1 - y; k \neq i,j} p'_k \prod_{X'_k = y; k \neq i,j} (1 - p'_k).$$

But since $p'_i \geqslant p'_j$ and $(1 - p'_j) \geqslant (1 - p'_i)$, this is itself greater than or equal to $\mathcal{L}_\uparrow[\boldsymbol{X}'; G = y]$. Therefore for every $\boldsymbol{X}'$ there is some $\boldsymbol{X}$ such that $\mathcal{L}_\uparrow[\boldsymbol{X}'; G = 1 - f_{w'}(\boldsymbol{X}')] \leqslant \mathcal{L}_\uparrow[\boldsymbol{X}; G = 1 - f_w(\boldsymbol{X})]$. Taking the max over all $\boldsymbol{X}$ then yields $Q(\boldsymbol{w}') \leqslant Q(\boldsymbol{w})$, as desired.

Since every score can be made decreasing by applying finitely many such flips, we have that for every score $\boldsymbol{w}$ there is a decreasing score $\boldsymbol{w}^*$ for which $Q(\boldsymbol{w}^*) \leqslant Q(\boldsymbol{w})$. Therefore there is an optimal score which is decreasing. □

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.

5.1 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 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 Appendix B (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[4]: MatrixFactorization (MF), Iterative SVD (ISVD), and Soft Impute (SI). We evaluate the rules on two datasets from the PrefLib library [13], 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 Biding (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 ($h \in \{0.5, 0.8, 0.9\}$).

	f_{dist}	$f_{\mathrm{MLE}\uparrow}$	$f_{\mathrm{MLE}\downarrow}$	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 ($d = \mathrm{M}, h = 0.5$)	94.8	94.8	94.8	94.8	90.3	94.8	94.8	96.5	94.5	**96.8**
CONF ($d = \mathrm{M}, h = 0.8$)	**95.3**	95.2	**95.3**	**95.3**	92.0	**95.3**	95.2	93.0	91.0	93.5
CONF ($d = \mathrm{M}, h = 0.9$)	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 green and blue, respectively. Simulations for parameter values omitted in the table led to qualitatively similar conclusions.

[4] `https://github.com/iskandr/fancyimpute`

5.2 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.

5.3 Results

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

1. The scoring rules using vectors $\boldsymbol{w}_{\times}^{\uparrow}$ and $\boldsymbol{w}_{\circ}^{\uparrow}$ 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 $\boldsymbol{w}_{\times}^{\uparrow}$). Similarly, $f_{\text{MLE}\downarrow}$ performed (slightly) better than f_{dist} in only one dataset (CLINTON-OCT), and for several datasets it produced significantly worse results.

2. f_{dist}, $f_{\text{MLE}\uparrow}$, and the scoring rule using $\boldsymbol{w}^*$ perform comparably well, though each excelled in different datasets.

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 ($\boldsymbol{w}_{\circ}^{\uparrow}$).

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 appendix: axiomatic (Appendix A), Bayesian (Appendix B), and randomized (Appendix C).

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.

References

[1] Abramowitz, B.; and Anshelevich, E. 2018. Utilitarians without utilities: Maximizing social welfare for graph problems using only ordinal preferences. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 894–901.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

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

Appendix

A Axiomatic Approach

The axiomatic approach to preference aggregation develops desirable properties and classifies different aggregation rules by the axioms that they satisfy. Of particular interest are characterization theorems which, given a list of axioms, describe exactly the class of aggregation rules that satisfy them. In our case, we are interested in aggregation rules that transform an ordered list of bits into a bit.

In this appendix, we introduce a few simple axioms, and show that they characterize a set of three aggregation rules. Our main axiom demands that if the rule gives the same output on two strings, then it also gives this output when interleaving the two strings into one. Notably, the distortion-optimal rule we identified in Theorem 1 fails this axiom. We work with aggregation rules f that are defined for any number n of experts; previously, we assumed n to be fixed. Let us write $\{0,1\}^+ = \cup_{n\in\mathbb{N}} \{0,1\}^n$ for the set of all finite-length bitstrings. We study rules $f : \{0,1\}^+ \to \{0,1,\perp\}$, where $\perp$ represents a tie.

The following are four natural axioms that we may want such a rule to satisfy.

1. *Resoluteness:* $f(x) \neq \perp$ for all $\boldsymbol{X} \in \{0,1\}^+$. That is, f should never return a tie.

2. *Foundation*: $f(1) = 1$, $f(0) = 0$, $f(10) \in \{1,\perp\}$, and $f(01) \in \{0,\perp\}$. That is, if there is just one expert, f should agree with the expert. In the case of two experts, f should either return either the judgment of the more accurate expert, or a tie. Note that when resoluteness is also imposed, we have $f(1) = f(10) = 1$ and $f(0) = f(01) = 0$.

3. *Interleaving Consistency:* For all $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \{0,1\}^+$, if $f(\boldsymbol{X}_1) = f(\boldsymbol{X}_2) \in \{0,1\}$, then $f(\boldsymbol{X}) = f(\boldsymbol{X}_1) = f(\boldsymbol{X}_2)$ for all strings $\boldsymbol{X}$ that can be obtained by interleaving $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Here, interleaving means combining bits from the two strings in any way that maintains the order of the bits in each string. For example, if $\boldsymbol{X}_1 = 1101$ and $\boldsymbol{X}_2 = 01$, then $\boldsymbol{X} = 110011$ is an interleaving of them, since it can be obtained by taking the first two bits of $\boldsymbol{X}_1$, then the first bit of $\boldsymbol{X}_2$, then the third bit of $\boldsymbol{X}_1$, then the second bit of $\boldsymbol{X}_2$, and then the fourth bit of $\boldsymbol{X}_1$. Intuitively, if we see two separate ranked sets of judgments, and in each case we are convinced of the same ground truth estimate, then putting all judgments together and while preserving the accuracy rankings should not change our estimate.

These three axioms together imply a neutrality property, which requires that f cannot be biased to either 0 or to 1.

4. *Neutrality:* For each $\boldsymbol{X} \in \{0,1\}^+$, if $f(\boldsymbol{X}) \in \{0,1\}$, then $f(\overline{\boldsymbol{X}}) = 1 - f(\boldsymbol{X})$; and if $f(\boldsymbol{X}) = \perp$, then $f(\overline{\boldsymbol{X}}) = \perp$. Here, $\overline{\boldsymbol{X}}$ denotes the string where all the bits of $\boldsymbol{X}$ are flipped. Intuitively, the rule says that if labels '0' and '1' are swapped, then the answer should be relabeled accordingly.

Proposition 1. *Resoluteness, foundation, and interleaving consistency imply neutrality.*

Proof. Suppose f satisfies resoluteness, foundation, and interleaving consistency. Suppose for a contradiction that f fails neutrality, and hence there is a string $\boldsymbol{X} \in \{0,1\}^+$ such that $f(\boldsymbol{X}) = f(\overline{\boldsymbol{X}})$ (using resoluteness). Without loss of generality, assume that $f(\boldsymbol{X}) = f(\overline{\boldsymbol{X}}) = 1$. By interleaving the strings $\boldsymbol{X}$ and $\overline{\boldsymbol{X}}$, we can obtain the string $(01)^{|\boldsymbol{X}|}$. By interleaving consistency, $f((01)^{|\boldsymbol{X}|}) = 1$. On the other hand, foundation implies that $f(01) = 0$. By interleaving $|\boldsymbol{X}|$ copies of 01, we obtain $f((01)^{|\boldsymbol{X}|}) = 0$ from interleaving consistency, a contradiction. □

While resoluteness, foundation, and neutrality are relatively mild, interleaving consistency is a strong axiom. For example, it implies a number of other intuitive consistency axioms listed below.

5. *Addition Consistency:* For all $x \in \{0,1\}^+$ and $y \in \{0,1\}$, if $f(x) = y$, and x' is obtained by inserting a y-bit in x and possibly inserting a $(1-y)$-bit at a later position in x, then $f(x') = y$.

6. *Subtraction Consistency:* For all $x \in \{0,1\}^+$ (with $|x| \geqslant 2$) and $y \in \{0,1\}$, if $f(x) = y$, and x' is obtained by removing a $(1-y)$-bit from x and possibly removing a y-bit from a later position in x, then $f(x') = y$.

7. *Swap Consistency:* For all $x \in \{0,1\}^+$ and $y \in \{0,1\}$, if $f(x) = y$, and x' is obtained by replacing a $(1-y)$-bit with a y-bit in x, then $f(x') = y$.

8. *Shift Consistency:* For all $x \in \{0,1\}^+$ and $y \in \{0,1\}$, if $f(x) = y$, and x' is obtained by shifting a y-bit to the left in x or shifting a $(1-y)$-bit to the right in x, then $f(x') = y$.

Proposition 2. *Foundation and interleaving consistency imply addition consistency, subtraction consistency, swap consistency, and shift consistency.*

Proof. Addition consistency simply requires interleaving strings 0, 1, 01, or 10 with x. Subtraction consistency is equivalent to addition consistency. Swap consistency is implied by addition consistency and subtraction consistency (e.g. swapping a 0 with a 1 can be seen as adding a 1 and subtracting a 0). The same goes for shift consistency (e.g. changing $x0zk$ to $xz0k$ can be seen as first adding 10 to obtain $x01z0k$, and then subtracting 01 to obtain $xz0k$; both operations must preserve the answer as 1 if it was originally 1). □

In fact, interleaving consistency is so strong that (in the presence of the other axioms), it excludes all but three aggregation rules.

Theorem 8. *Rule $f : \{0,1\}^+ \to \{0,1,\perp\}$ satisfies resoluteness, foundation, and interleaving consistency if and only if f is one of the three following rules.*

1. *Rule f_1: Output the first bit.*

2. *Rule f_2: Output the more frequent bit. If there is a tie, output the first bit.*

3. *Rule f_3: Output the more frequent bit. If there is a tie, output the negation of the last bit.*

Proof. It is clear that f_1, f_2, and f_3 satisfy our axioms. For the other direction, suppose that f is a rule satisfying the axioms. By Proposition 2, we can assume that f also satisfies addition, subtraction, swap, and shift consistency. Note that the smallest string on which our axioms do not directly provide an answer is 100 (or equivalently, 001). It turns out that if f outputs the most accurate bit on this string (even though the two less accurate bits disagree with it), then f must always output the first bit, i.e., be the rule f_1.

Lemma 5. *If $f(011) = 0$, then $f = f_1$.*

Proof. First, we show that $f(01^k) = 0$ for all $k \geqslant 2$, by induction on k. The base case of $k = 2$ holds by assumption. Suppose $f(01^k) = 0$ for some $k \geqslant 2$. We want to show that $f(01^{k+1}) = 0$. Suppose for a contradiction that this is not the case. By resoluteness, this means $f(01^{k+1}) = 1$. Then, by neutrality, $f(10^{k+1}) = 0$. Let $\boldsymbol{X}_1 = 01^k$ and $\boldsymbol{X}_2 = 10^{k+1}$. Note that $\boldsymbol{X} = 100(10)^k$ can be obtained by interleaving $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Since $f(\boldsymbol{X}_1) = f(\boldsymbol{X}_2) = 0$, we have $f(\boldsymbol{X}) = 0$ by interleaving consistency. However, note that $\boldsymbol{X}$ can also be obtained by interleaving k copies of 10, and one copy of 100. By foundation, we have $f(10) = 1$, and by neutrality and our assumption, we have $f(100) = 1$. Hence, by interleaving consistency, we get $f(\boldsymbol{X}) = 1$, which is a contradiction. Thus, $f(01^k) = 0$ for all k.

We now show that $f = f_1$. By neutrality, it is enough to show that $f(0\boldsymbol{X}) = 0$ for all $\boldsymbol{X} \in \{0,1\}^+$. Let $\boldsymbol{X} \in \{0,1\}^+$, and consider the string $01^{|\boldsymbol{X}|}$. From above, we know that $f(01^{|\boldsymbol{X}|}) = 0$. By swap consistency, if we replace 1's with 0's in the string, f continues to output 0. It follows that $f(0\boldsymbol{X}) = 0$. This establishes that $f = f_1$, as desired. □

It remains to consider the case where $f(011) = 1$. In this case, note that f is “going for quantity over quality”. We show that f must then always output the more frequent bit. How ties are broken is decided by what f outputs on the string 0110 (or 1001), which is the smallest string where both bits appear an equal number of times and the answer is not trivially decided by our axioms.

Lemma 6. *If $f(011) = 1$, then $f = f_2$ if $f(0110) = 0$, and $f = f_3$ if $f(0110) = 1$.*

Proof. First, we show that $f(1^{k-1}0^k) = 0$ for all $k \geqslant 2$, by induction on k. Our assumption that $f(011) = 1$, along with neutrality, already establishes this for $k = 2$. Suppose $f(1^{k-1}0^k) = 0$ for some $k \geqslant 2$. We want to prove that $f(1^k 0^{k+1}) = 0$. Suppose for contradiction that this is not true. Then, by resoluteness, $f(1^k 0^{k+1}) = 1$, and by neutrality, $f(0^k 1^{k+1}) = 0$. Let $\boldsymbol{X}_1 = 1^{k-1}0^k$ and $\boldsymbol{X}_2 = 0^k 1^{k+1}$. The reader can check that the string $\boldsymbol{X} = (10)^{5k-1}011$ can be obtained by interleaving two copies of $\boldsymbol{X}_1$ and three copies of $\boldsymbol{X}_2$. Hence, by interleaving consistency, $f(\boldsymbol{X}) = 0$. However, $\boldsymbol{X}$ can also be obtained by interleaving $5k-1$ copies of 10 and one copy of 011, and $f(10) = f(011) = 1$ by foundation and by assumption. Hence, by interleaving consistency, we also have $f(\boldsymbol{X}) = 1$, which is a contradiction.

Thus, we have established that $f(1^{k-1}0^k) = 0$ (and therefore $f(0^{k-1}1^k) = 1$) for all $k \geqslant 2$. This implies that $f(\boldsymbol{X}) = 0$ for any string $\boldsymbol{X}$ that contains strictly more 0's than 1's. This is because if $\boldsymbol{X}$ contains k 0's, then we can start with the string $1^{k-1}0^k$, and then subtract 1's and shift 1's to the right to obtain $\boldsymbol{X}$; by subtraction and shift consistency both operations preserve the answer as 0. Similarly, $f(\boldsymbol{X}) = 1$ for any string $\boldsymbol{X}$ with strictly more 1's than 0's.

It remains to deduce the output of f on *balanced* strings, i.e. those with equal number of 0's and 1's. Suppose first that $f(0110) = 0$. In this case, we show that $f = f_2$, so that on balanced strings, f returns the first bit. To do this, we show, by strong induction on k, that $f(01^k 0^{k-1}) = 0$ for all $k \geqslant 2$. We have already assumed the base case of $k = 2$. For strong induction, suppose the hypothesis holds for all integers up to some k, but fails at $k + 1$. Thus, we have $f(01^k 0^{k-1}) = 0$ but $f(01^{k+1}0^k) = 1$, and so by neutrality $f(10^{k+1}1^k) = 0$. We now take $\boldsymbol{X}_1 = 01^k 0^{k-1}$ and $\boldsymbol{X}_2 = 10^{k+1}1^k$, and leave it to the reader to verify that $\boldsymbol{X} = (1001)(10)^{k-2}(1001)(10)^{k-1}$ can be obtained by interleaving $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Hence, by interleaving consistency, $f(\boldsymbol{X}) = 0$. However, $f(10) = 1$ by foundation, and since $f(0110) = 0$, by neutrality $f(1001) = 1$. Since $\boldsymbol{X}$ can also be obtained by interleaving copies of these strings, by interleaving consistency, $f(\boldsymbol{X}) = 1$, which is a contradiction. Hence, we have established that $f(01^k 0^{k-1}) = 0$ for all $k \geqslant 2$. Then, it follows immediately that $f(\boldsymbol{X}) = 0$ for all strings $\boldsymbol{X}$ with an equal number of 0's and 1's with the first bit being 0 as all such strings (with say k 0's and k 1's) can be obtained by starting from $01^k 0^{k-1}$ and shifting some 1's to the right (invoking shifting consistency). By neutrality, we get $f(\boldsymbol{X}) = 1$ for all strings $\boldsymbol{X}$ with an equal number of 0's and 1's with the first bit being 1. Thus, $f = f_2$.

Finally, when $f(0110) = 1$, the proof that $f = f_3$ proceeds very similarly by first arguing that $f(0^{k-1}1^k 0) = 1$ for all $k \geqslant 2$ using strong induction, and then shifting some 1's to the left to obtain any string with an equal number of 0's and 1's with the last bit being 0. □

These lemmas prove the theorem. □

Interestingly, all three rules characterized here are scoring rules (see Section 2). We have previously seen rule f_3 under the name $f_{\mathrm{MLE}\downarrow}$ in Theorem 3, where we showed that it optimizes the pessimistic likelihood. We have also already met f_1 as an optimal scoring rule according to the optimistic likelihood objective (Theorem 5). On the other hand, the distortion-optimal rule f_{dist} from Theorem 1 is evidently not among the three rules characterized by Theorem 8. One obvious reason is that it is not resolute (e.g. $f_{\mathrm{dist}}(011) = \perp$). However, even if we drop resoluteness, this rule violates interleaving consistency, even when interleaving two copies of the same string. To see this, we can check that $f_{\mathrm{dist}}(10^3 1^3) = 0$. We can obtain the string $11(0^3 1^3)^2$ by interleaving two copies of $10^3 1^3$, but $f_{\mathrm{dist}}(11(0^3 1^3)^2) = 1$.

All axioms in Theorem 8 are necessary. If resoluteness is dropped, the rule that outputs the more frequent bit and a tie if the numbers of 0's and 1's are equal satisfies all axioms except resoluteness. The rule that outputs the opposite of the first bit satisfies all axioms except foundation. Finally, f_{dist} satisfies all axioms except interleaving consistency. Additionally, one can check that f_{dist} satisfies all the weakenings of interleaving consistency discussed in Proposition 2, showing that we cannot weaken interleaving consistency to addition consistency in Theorem 8.

It would be interesting to obtain an analog of Theorem 8 without resoluteness, to allow us to capture the natural rule that outputs the majority bit if it exists, and a tie otherwise. It would also be interesting the characterize the class of rules that satisfy the axioms of Theorem 8 except with interleaving consistency weakened to addition consistency. Finally, the distortion-optimal rule f_{dist} is well-motivated by our likelihood analysis, and it seems like a natural binary aggregation rule in its own right. Thus, it would be interesting to find an axiomatization of that rule, including the axioms we have used here. Perhaps we could add an axiom capturing the particularly interesting property that f_{dist} follows the opinion of the most accurate third of experts should they all agree, and follows the opinion of the least accurate two thirds of experts should those agree. Unfortunately, a preliminary exploration suggests that finding an axiomatization may be difficult: there are rules other than f_{dist} satisfying all the axioms we have mentioned, and using a SAT-solver-based analysis, we have found that even after adding many other properties of f_{dist} to the mix, we still have not ruled out all other rules.

B Bayesian Approach

In Sections 3 and 4 we took a conservative approach—our analysis was worst-case over the possible values of the probabilities p_i (the probabilities that X_i are correct). In this section we assume that for each variable X_i, probability p_i comes from a known prior distribution. We assume that for each i the distribution $\mathcal{X}_i$ is discrete with $|\mathcal{X}_i|$ possible values[5], $\mathcal{X}_i = \{(x_j, y_j) \mid j \in [|\mathcal{X}_i|]\}$. If $(x_j, y_j) \in \mathcal{X}_i$ then the probability that p_i equals y_j is x_j. By $\mathcal{X}_i^{\rightarrow}$ we denote the set of possible values in $\mathcal{X}_i$, i.e., $\mathcal{X}_i^{\rightarrow} = \{y_j \mid (x_j, y_j) \in \mathcal{X}_i\}$. Similarly as in the other parts of paper, we assume that we additionally know that $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_n$. Then, the expected maximum likelihood is defined as:

$$\mathcal{EL}[\boldsymbol{X}; G = g] = \frac{1}{\gamma} \cdot \sum_{(x_1,p_1)\in\mathcal{X}_1} x_1 \sum_{\substack{(x_2,p_2)\in\mathcal{X}_2\\ p_2\leqslant p_1}} x_2 \quad \ldots \sum_{\substack{(x_n,p_n)\in\mathcal{X}_n\\ p_n\leqslant p_{n-1}}} x_n \cdot \mathcal{L}[\boldsymbol{X}; G = g, \boldsymbol{p} = (p_1, \ldots, p_n)],$$

where γ is a normalization factor. Note that in contrast to the definition of the likelihood function (cf. Section 2), the above formula does not depend on the vector of probabilities $\boldsymbol{p}$.

By substituting the formula for $\mathcal{L}[\boldsymbol{X}; G = g, \boldsymbol{p} = (p_1, \ldots, p_n)]$, we get that:

$$\begin{aligned}\mathcal{EL}[\boldsymbol{X}; G = g] =& \sum_{(x_1,p_1)\in\mathcal{X}_1} x_1 \sum_{\substack{(x_2,p_2)\in\mathcal{X}_2\\ p_2\leqslant p_1}} x_2 \quad \ldots \sum_{\substack{(x_n,p_n)\in\mathcal{X}_n\\ p_n\leqslant p_{n-1}}} x_n \cdot \prod_{i=1}^{n} p_i^{\mathbb{1}[X_i=g]} \cdot (1-p_i)^{\mathbb{1}[X_i\neq g]} \\ =& \sum_{(x_1,p_1)\in\mathcal{X}_1} x_1 p_1^{\mathbb{1}[X_1=g]}(1-p_1)^{\mathbb{1}[X_1\neq g]} \quad \ldots \sum_{\substack{(x_n,p_n)\in\mathcal{X}_n\\ p_n\leqslant p_{n-1}}} x_n p_n^{\mathbb{1}[X_n=g]}(1-p_n)^{\mathbb{1}[X_n\neq g]}.\end{aligned}$$

Observe that $\mathcal{EL}[\boldsymbol{X}; G = g]$ can be computed by a dynamic programing based on a backwards induction. We build an array A, where for each $i \in [n]$ and $y_i \in \mathcal{X}_i^{\rightarrow}$ the value $\mathrm{A}[i, y_i]$ has the following meaning:

$$\begin{aligned}\mathrm{A}[i, y_i] = \sum_{\substack{(x_i,p_i)\in\mathcal{X}_i\\ p_i\leqslant y_i}} x_i p_i^{\mathbb{1}[X_i=g]}(1-p_i)^{\mathbb{1}[X_i\neq g]} \sum_{\substack{(x_{i+1},p_{i+1})\in\mathcal{X}_{i+1}\\ p_{i+2}\leqslant p_{i+1}}} x_{i+1} p_{i+1}^{\mathbb{1}[X_{i+1}=g]}(1-p_{i+1})^{\mathbb{1}[X_{i+1}\neq g]} \cdot \\ \ldots \quad \cdot \sum_{\substack{(x_n,p_n)\in\mathcal{X}_n\\ p_n\leqslant p_{n-1}}} x_n p_n^{\mathbb{1}[X_n=g]}(1-p_n)^{\mathbb{1}[X_n\neq g]}.\end{aligned}$$

The base step is:

$$\mathrm{A}[n, y_n] = \sum_{\substack{(x_n,p_n)\in\mathcal{X}_n\\ p_n\leqslant y_n}} x_n p_n^{\mathbb{1}[X_n=g]}(1-p_n)^{\mathbb{1}[X_n\neq g]},$$

the inductive step is:

$$\mathrm{A}[i, y_i] = \sum_{\substack{(x_i,p_i)\in\mathcal{X}_i\\ p_i\leqslant y_i}} x_i p_i^{\mathbb{1}[X_i=g]}(1-p_i)^{\mathbb{1}[X_i\neq g]} \max_{\substack{y_{i+1}\in\mathcal{X}_{i+1}^{\rightarrow}\\ y_{i+1}\leqslant p_i}} \mathrm{A}[i+1, y_{i+1}],$$

[5] This allows us to represent prior distributions concisely.

and the final value is:

$$\mathcal{EL}[\boldsymbol{X}; G = g] = \max_{y_1 \in \mathcal{X}_1^{\rightarrow}} \mathrm{A}[1, y_1].$$

Note that in many contexts the prior distributions $\mathcal{X}_i$ can be estimated from available data. For example, in our experiments for collaborative filtering, described in Section 5, we assumed that the prior for all the agents is the same, $|\mathcal{X}_i| = \mathcal{X}$ for all $i \in N$, and that it is estimated through the following simple procedure. Assume we want to predict the decision of an agent $i \in N$ for an issue $j \in I$. For each agent $i' \neq i$ we count how often (measured as a fraction) the decisions of i and i' coincide. We construct $\mathcal{X}$ by aggregating these fractions over all $i' \in N$, $i' \neq i$.

C Randomized Aggregation Rules

In this section, we consider randomized aggregation rules, which, given a judgment profile $\boldsymbol{X}$, return 1 with some probability $\theta \in [0, 1]$ and 0 with the remaining probability $1 - \theta$. Alternatively, we can think of the rule as simply returning $\theta \in [0, 1]$ instead of an estimate in $\{0, 1\}$ like deterministic aggregation rules. We first extend the definition of distortion to real-valued estimates in $[0, 1]$. For $\epsilon \in (0, 1/2)$ and $\theta \in [0, 1]$, define

$$\operatorname{dist}^{\epsilon}(\theta; \boldsymbol{X}) = \sup_{\boldsymbol{p} \in \mathcal{P}_n^{\epsilon}} \frac{\max\left(\mathcal{L}[\boldsymbol{X}|G = 0, \boldsymbol{p}], \mathcal{L}[\boldsymbol{X}|G = 1, \boldsymbol{p}]\right)}{(1 - \theta) \cdot \mathcal{L}[\boldsymbol{X}|G = 0, \boldsymbol{p}] + \theta \cdot \mathcal{L}[\boldsymbol{X}|G = 1, \boldsymbol{p}]}.$$

Note that $\operatorname{dist}^{\epsilon}(0; \boldsymbol{X})$ (resp. $\operatorname{dist}^{\epsilon}(1; \boldsymbol{X})$) still coincides with the definition of distortion from Section 3.1. The distortion-optimal randomized estimate is then defined as $\operatorname{argmin}_{\theta \in [0,1]} \operatorname{dist}^{\epsilon}(\theta; \boldsymbol{X})$; for comparison, recall that the distortion-optimal deterministic estimate was defined as $\operatorname{argmin}_{\theta \in \{0,1\}} \operatorname{dist}^{\epsilon}(\theta; \boldsymbol{X})$ in Section 3.1.

We are now ready to identify the distortion-optimal randomized estimate. Unfortunately, this is no longer independent of the choice of ϵ.

Theorem 9. *For $\epsilon \in (0, 1/2)$, $n \in \mathbb{N}$, and $\boldsymbol{X} \in \{0, 1\}^n$, we have*

$$\operatorname*{argmin}_{\theta \in [0,1]} \operatorname{dist}^{\epsilon}(\theta; \boldsymbol{X}) = \frac{1 - \left(\frac{\epsilon}{1-\epsilon}\right)^{s_{\boldsymbol{X}}(1)}}{2 - \left(\frac{\epsilon}{1-\epsilon}\right)^{s_{\boldsymbol{X}}(0)} - \left(\frac{\epsilon}{1-\epsilon}\right)^{s_{\boldsymbol{X}}(1)}}. \tag{2}$$

Proof. Note that

$$\begin{aligned}
\operatorname{dist}^{\epsilon}(\theta; \boldsymbol{X}) &= \sup_{\boldsymbol{p} \in \mathcal{P}_n^{\epsilon}} \frac{\max\left(\mathcal{L}[\boldsymbol{X}|G = 0, \boldsymbol{p}], \mathcal{L}[\boldsymbol{X}|G = 1, \boldsymbol{p}]\right)}{(1 - \theta) \cdot \mathcal{L}[\boldsymbol{X}|G = 0, \boldsymbol{p}] + \theta \cdot \mathcal{L}[\boldsymbol{X}|G = 1, \boldsymbol{p}]} \\
&= \max\left(\sup_{\boldsymbol{p} \in \mathcal{P}_n^{\epsilon}} \frac{\mathcal{L}[\boldsymbol{X}|G = 0, \boldsymbol{p}]}{(1 - \theta) \cdot \mathcal{L}[\boldsymbol{X}|G = 0, \boldsymbol{p}] + \theta \cdot \mathcal{L}[\boldsymbol{X}|G = 1, \boldsymbol{p}]}, \sup_{\boldsymbol{p} \in \mathcal{P}_n^{\epsilon}} \frac{\mathcal{L}[\boldsymbol{X}|G = 1, \boldsymbol{p}]}{(1 - \theta) \cdot \mathcal{L}[\boldsymbol{X}|G = 0, \boldsymbol{p}] + \theta \cdot \mathcal{L}[\boldsymbol{X}|G = 1, \boldsymbol{p}]} \right) \\
&= \max\left(\frac{1}{(1 - \theta) \cdot 1 + \theta \cdot \inf_{\boldsymbol{p} \in \mathcal{P}_n^{\epsilon}} \frac{\mathcal{L}[\boldsymbol{X}|G=1,\boldsymbol{p}]}{\mathcal{L}[\boldsymbol{X}|G=0,\boldsymbol{p}]}}, \frac{1}{(1 - \theta) \cdot \inf_{\boldsymbol{p} \in \mathcal{P}_n^{\epsilon}} \frac{\mathcal{L}[\boldsymbol{X}|G=0,\boldsymbol{p}]}{\mathcal{L}[\boldsymbol{X}|G=1,\boldsymbol{p}]} + \theta \cdot 1} \right) \\
&= \max\left(\frac{1}{(1 - \theta) \cdot 1 + \theta \cdot \frac{1}{\operatorname{dist}^{\epsilon}(1; \boldsymbol{X})}}, \frac{1}{(1 - \theta) \cdot \frac{1}{\operatorname{dist}^{\epsilon}(0; \boldsymbol{X})} + \theta \cdot 1} \right) \\
&= \max\left(\frac{1}{(1 - \theta) \cdot 1 + \theta \cdot \left(\frac{\epsilon}{1-\epsilon}\right)^{s_{\boldsymbol{X}}(0)}}, \frac{1}{(1 - \theta) \cdot \left(\frac{\epsilon}{1-\epsilon}\right)^{s_{\boldsymbol{X}}(1)} + \theta \cdot 1} \right),
\end{aligned}$$

where the last step follows from Lemma 1.

Our goal is to find the $\theta \in [0, 1]$, which minimizes this quantity. It is easy to show that this happens when both terms inside the max become equal, which happens when θ achieves the value stated in the theorem statement. □

Some observations about the distortion-optimal randomized estimate are in order.

1. Note that $s_{\boldsymbol{X}}(0)$ and $s_{\boldsymbol{X}}(1)$ both cannot be 0 (the first bit has strength at least 1). Hence, the denominator in Equation (2) is positive, and the ratio is thus well-defined.

2. If, for some $y \in \{0, 1\}$, $s_{\boldsymbol{X}}(y) = 0$ (i.e. $1 - y$ "weakly dominates" y in every prefix of $\boldsymbol{X}$), then the rule deterministically chooses $1 - y$.

3. When $s_{\boldsymbol{X}}(0) > 0$ and $s_{\boldsymbol{X}}(1) > 0$, we have that $\theta \in \left[\frac{1}{2}\frac{1-2\epsilon}{1-\epsilon}, \frac{1}{2}\frac{1-\epsilon}{1-2\epsilon}\right]$. In other words, when ϵ is small, θ will be very close to $1/2$ regardless of the judgment profile $\boldsymbol{X}$, unless an estimate is chosen with certainty as in the edge cases identified above.

D Optimal Scoring Rules

Section 4.1 characterizes optimal scores $\boldsymbol{w}^{\uparrow}_{\circ}$, $\boldsymbol{w}^{\downarrow}_{\circ}$, and $\boldsymbol{w}^{\downarrow}_{\times}$, but the form of $\boldsymbol{w}^{\uparrow}_{\times}$ in general remains unknown. Table 2 and Table 3 catalog optimal scores $\boldsymbol{w}^{\uparrow}_{\times}$ for small numbers of experts n.

The scores in Table 2 are found by considering all $\boldsymbol{X} \in \{0, 1\}^n$ ordered from large to small according to $\mathcal{L}_{\uparrow}[X|0]$, and considering the longest possible prefix W of these $\boldsymbol{X}$ which does not contain both $\boldsymbol{X}$ and its complement. The $\boldsymbol{w}$ are then computed by a linear program which enforces $\boldsymbol{w} \cdot \boldsymbol{X} \leqslant 1/2 - 1/2^j$ for all $\boldsymbol{X} \in W$ and for the smallest j which makes the program feasible. The $\boldsymbol{w}$ are scaled up to eliminate fractions.

The scores in Table 3 are optimal according to an approach to ties which is distinct from breaking them adversarially: namely, vectors for which the scoring rule ties are disregarded in the evaluation of the maximum:

$$\boldsymbol{w}^{\uparrow}_{\times} \in \arg \min_{\boldsymbol{w} \in \mathbb{R}^n_{\geqslant 0}} \max_{\boldsymbol{X} : \boldsymbol{w} \cdot \boldsymbol{X} \neq \frac{1}{2}} \mathcal{L}_{\uparrow}[\boldsymbol{X}; G = 1 - f_{\boldsymbol{w}}(\boldsymbol{X})].$$

It follows that these scoring rules are also optimal under best-case, "optimistic" tiebreaking. These $\boldsymbol{w}$ are computed by again ranking all vectors $\boldsymbol{X} \in \{0, 1\}^n$ according to $\mathcal{L}_{\uparrow}[X|0]$ and using a linear program to enforce that $\boldsymbol{w} \cdot \boldsymbol{X} \leqslant 1/2$ for as many $\boldsymbol{X}$ with the largest $\mathcal{L}_{\uparrow}[X|0]$ as possible. Finally the $\boldsymbol{w}$ are again scaled up to eliminate fractions.

n	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	w_{16}	w_{17}	w_{18}	w_{19}	w_{20}
1	1																			
2	2	1																		
3	1	0	0																	
4	1	0	0	0																
5	28	9	9	9	9															
6	7	3	2	2	2	0														
7	20	7	7	7	7	0	0													
8	27	27	14	14	14	0	0	0												
9	9	6	3	3	3	2	2	2	2											
10	9	6	3	3	3	2	2	2	2	0										
11	36	24	11	11	11	11	8	8	8	0	0									
12	6	6	6	2	2	2	2	2	2	1	1	0								
13	17	14	14	4	4	4	4	4	4	4	4	3	0							
14	111	53	53	25	25	25	25	25	25	25	14	14	14	14						
15	403	403	256	124	124	124	124	124	93	93	93	93	93	93	0					
16	81	81	72	72	36	36	31	31	31	15	15	15	15	15	15	15				
17	30	30	30	29	15	15	15	15	15	15	15	0	0	0	0	0	0			
18	217	217	181	181	62	62	62	62	62	62	62	62	62	62	62	62	62	62		
19	217	217	217	178	93	62	62	62	62	62	62	62	62	62	62	62	62	62	0	
20	217	217	186	176	62	62	62	62	62	62	62	62	62	62	62	62	62	62	0	0

Table 2: Optimal scores $\boldsymbol{w}^{\uparrow}_{\times}$ for the first 20 values of n, with ties broken pessimistically.

n	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	w_{16}	w_{17}	w_{18}	w_{19}	w_{20}
1	1																			
2	1	1																		
3	1	1	0																	
4	1	1	0	0																
5	1	1	0	0	0															
6	1	1	0	0	0	0														
7	1	1	0	0	0	0	0													
8	3	1	1	1	1	1	0	0												
9	3	1	1	1	1	1	0	0	0											
10	2	2	1	1	1	1	0	0	0	0										
11	2	2	1	1	1	1	0	0	0	0	0									
12	3	3	1	1	1	1	1	1	1	1	0	0								
13	3	3	1	1	1	1	1	1	1	1	0	0	0							
14	3	3	1	1	1	1	1	1	1	1	0	0	0	0						
15	2	2	2	1	1	1	1	1	1	0	0	0	0	0	0					
16	4	3	3	1	1	1	1	1	1	1	1	1	1	1	1	0				
17	4	3	3	1	1	1	1	1	1	1	1	1	1	1	1	0	0			
18	4	3	2	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0		
19	2	2	2	2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	
20	2	2	2	2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0

Table 3: Optimal scores $\boldsymbol{w}^{\uparrow}_{\times}$ for the first 20 values of n, with ties broken optimistically.