Strategyproof Mean Estimation from Multiple-Choice Questions

Anson Kahng [1] **Gregory Kehne** [2] **Ariel D. Procaccia** [3]

Abstract

Given n values possessed by n agents, we study the problem of estimating the mean by truthfully eliciting agents' answers to multiple-choice questions about their values. We consider two natural candidates for estimation error: mean squared error (MSE) and mean absolute error (MAE). We design a randomized estimator which is asymptotically optimal for both measures in the worst case. In the case where prior distributions over the agents' values are known, we give an optimal, polynomial-time algorithm for MSE, and show that the task of computing an optimal estimate for MAE is $\#\mathcal{P}$-hard. Finally, we demonstrate empirically that knowledge of prior distributions gives a significant edge.

1. Introduction

Organizations often desire accurate estimates of population statistics (e.g., the mean of a set of values) in settings where eliciting exact values from agents is costly or impractical. For instance, suppose that you sit on an admissions committee, and your committee's task is to accurately estimate the number of candidates who will accept their admission offer, perhaps in order to decide how many more admissions offers to extend (concretely, consider the problem of admitting students in two waves corresponding to early action and regular admission). This is a consequential problem; there are significant direct and indirect costs associated with having many more or fewer matriculants than intended.[1]

Without more information either about or from the admits, this is hopeless. One potential approach is to ask each admit to provide an estimate of the probability p_i that the admit matriculates, but this is problematic because each admit may not know their own exact probability of accepting the offer, and coming up with exact probabilities places nontrivial cognitive loads on participants. Therefore, it is more reasonable to ask a multiple-choice question of the form, "How likely are you to accept this offer?" with choices "High," "Medium," and "Low." The task for the university, then, is to reconstruct an accurate estimate of the number of students who will accept their offers based on the coarse-grained information yielded by these multiple-choice queries. Specifically, the question is

> "What collection of multiple-choice questions should you ask, and how should you interpret the answers so as to estimate the expected number of matriculants as accurately as possible?"

1.1. Our Approach and Results

Consider n agents, where each agent i has a value $p_i \in [0, 1]$ and $p := (p_1, \dots, p_n)$ is the vector of all agents' values. We are interested in estimating $\|p\|_1 = \sum_{i=1}^n p_i$.

We interpret multiple-choice questions as forming a partition of $[0, 1]$ into subsets $X_1, \dots, X_k$, and asking agent i for the index j such that $p_i \in X_j$. It is known (Lambert & Shoham, 2009) that in order for multiple-choice questions to elicit truthful responses, each X_j must itself be an interval. Moreover, intervals are easier to interpret than arbitrary subsets; for example, the choice "Low" can be defined as $p_i \in [0, 1/3]$. Therefore, we restrict our multiple-choice questions to this framework.

Our goal is to design an estimator that consists of a set of (possibly different) multiple-choice questions which are posed to the agents, together with a function that outputs an estimate of $\|p\|_1$ based on the agents' answers to the multiple-choice questions; we denote the output of the estimator by $q(p)$. We measure the accuracy of the estimator using the *mean squared error (MSE)* $\mathbb{E}[(\|p\|_1 - q(p))^2]$ or the *mean absolute error (MAE)* $\mathbb{E}[|\|p\|_1 - q(p)|]$.[2]

We consider two settings corresponding to different levels of

[1] Computer Science Department, Carnegie Mellon University [2] Department of Mathematical Sciences, Carnegie Mellon University [3] School of Engineering and Applied Sciences, Harvard University. Correspondence to: Anson Kahng <akahng@cs.cmu.edu>, Gregory Kehne <gkehne@andrew.cmu.edu>, Ariel D. Procaccia <arielpro@cs.cmu.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1] In fact, our work on this paper originated from thinking about ways to address this problem in our institution.

[2] Throughout, we evaluate additive error with respect to estimating the sum of the p_i, from which the additive error with respect to mean estimation can easily be derived.

information about the agent's values. When no information about p is known (worst case), we consider the problem of designing a randomized estimator with good worst-case performance (when averaged over the estimator's randomness). We give a single randomized estimator $\bar{q}$ which guarantees

$$\mathsf{mse}(\bar{q}) = O\left(\frac{n}{k^2}\right), \quad \mathsf{mae}(\bar{q}) = O\left(\frac{\sqrt{n}}{k}\right)$$

and demonstrate that this is asymptotically optimal for both measures of error.

In the second setting, each p_i is drawn from a known distribution P_i; we consider the problem of designing a deterministic estimator which performs well on average (over the randomness of the p_i). We present an MSE-optimal estimator, and show that the problem of devising an MAE-optimal estimator is $\#\mathcal{P}$-hard.

Finally, we conduct experiments in the latter setting of known distributions, in which we aim to quantify the benefit of tailoring the estimator to the distributions. We focus on MSE due to our computational results and show that the optimal estimator significantly outperforms a naïve estimator.

1.2. Related Work

Caragiannis et al. (2016) study strategyproof mean estimation in a related setting, where strategic agents supply samples in order to move the estimation of the mean close to their own value. In this setting, they ask if the sample median is the best truthful estimator of the population mean, which is not the case, and characterize worst-case optimal truthful estimators that provably outperform the median for distributions with bounded support.

More broadly, mechanism design for information elicitation has been widely studied in computer science and economics (Zohar & Rosenschein, 2008; Chen & Kash, 2011; Waggoner & Chen, 2014). Many prior works in mechanism design focus on eliciting truthful signals from agents, often through direct verification mechanisms like strictly proper scoring rules (Gneiting & Raftery, 2007; Brier, 1950; Good, 1952; Winkler, 1969) and prediction markets (Wolfers & Zitzewitz, 2004; Berg et al., 2008). In a related vein, Radanovic & Faltings (2014) developed a mechanism for truthful elicitation of continuous signals, but we consider the problem of reconstructing a continuous value from discrete reports.

Additionally, Soloviev & Halpern (2018) consider the problem of acquiring information with resource limitations, where budget constraints on the number of tests in their setting roughly map to constraints in our setting on the granularity of queries. However, their setting involves noisy tests of Boolean formula truth values, as opposed to estimating a population statistic.

Furthermore, by considering estimation error as it varies over a range of k, we can investigate the relationship between the elicitation of values and the accuracy of our estimate, a tradeoff which has been studied for intelligent decision-making systems in a sequential setting by Boutilier (2002).

Alternatively, the task of optimally estimating $\sum_i p_i$ in terms of MSE or MAE can be viewed as variants of k-means and k-medians clustering, respectively. On the one hand, it resembles a special case of clustering in that P is product distribution over $[0, 1]^n$ (as opposed to a general distribution over a metric space), and our 'clusters' C correspond to vectors p which yield the same vector of multiple-choice answers, and the C are constrained to have a product structure as well. On the other hand, it is distinct from clustering in that r reports scalar representative ℓ_1 norms for the clusters, and so we take the ℓ_1 norm of C before calculating error.

This line of inquiry is also related to the notion of approximate query processing (AQP) in the data mining and databases literature, which is the practice of answering expensive aggregation queries with limited resources. Multiple research groups have focused on the bounded-error estimation of aggregates (e.g., sums of values in a database), but mostly through sampling techniques as opposed to summarization approaches (Jagadish et al., 1998; Chaudhuri et al., 2007; 2001; Babcock et al., 2003). More recently, there has been some work on summarizing data distributions with histograms in order to minimize the ℓ_2 distance between the distribution and the histogram approximation (Acharya et al., 2015; Ding et al., 2016), but this only coincides with our MSE setting when we query exactly one agent.

2. The Model

We consider a set of agents $[n] = \{1, \ldots, n\}$, each with an associated number $p_i \in [0, 1]$.

Our goal is to devise a scheme for estimating the sum $\|p\|_1 = \sum_{i=1}^n p_i$ (or, equivalently, the mean) to minimize additive error. We may ask each agent i which of k intervals contains their p_i, and so our estimator chooses n partitions $\mathcal{B}_i := \{B_{i,1}, \ldots, B_{i,k}\}$ of $[0, 1]$, and for each i the function $b_i : [0, 1] \to [k]$ poses the question to agent i and returns their response; $b_i(p_i) = j$ if $p_i \in B_{i,j}$. We refer to $b(p) := (b_1(p_1), \ldots, b_n(p_n))$ as the *classifier*. Next the *aggregator* $r : [k]^n \to \mathbb{R}$ takes the agents' responses and estimates $\sum_i p_i$; we refer to the output of r as the *report*.

We assume that each agent has an associated random variable X_i, and that the p_i are bounded, real-valued properties of the probability distributions of these X_i. Agents are not predisposed towards truth-telling, but alongside the multiple-choice questions they are offered payouts which will be awarded as a function of both their answers and the outcome of a future draw from their X_i. Agents are assumed to know the correct answers to their questions and to seek to maximize their expected payout (over the randomness of the future draw). For a single agent this is the setting of Lambert & Shoham (2009), who study multiple-choice questions about a general class of distributions and properties, and classify the properties which are *elicitable*; those about which rational agents can be incentivized to tell the truth. This classification implies that when the property p_i is real-valued, the answers of multiple-choice questions which successfully elicit it must correspond to ranges of possible values of p_i; for us this mandates that the partitions $\mathcal{B}_i$ are partitions of $[0, 1]$ into intervals.

In the example of matriculation, the p_i are the means of Bernoulli random variables modeling the future event that student i ultimately attends or does not attend. (We model this attendance outcome as unaffected by which answer which the agent chooses.) When p_i is the mean of X_i, our objective is to approximate the mean draw across all of the X_i. For more general elicitable p_i it is to approximate the mean of these properties.

Let $c_i := b_i(p_i)$ denote the answer of each agent to their question. For $c \in [k]^n$, the box $C_c := \prod_i B_{i,c_i} = b^{-1}(c)$ is the set of $(p_1, \ldots, p_n)$ for which each agent i answers c_i. In terms of the classifier and aggregator, our goal may be restated as finding the *estimator*

$$q := r \circ b : [0, 1]^n \to \mathbb{R}$$

that minimizes expected error.

We consider two natural measures of error, mean squared error (MSE) and mean absolute error (MAE), which are defined to be

$$\mathsf{mse}(q) = \mathbb{E}\left[(\|p\|_1 - q(p))^2\right],$$
$$\mathsf{mae}(q) = \mathbb{E}\left[|\|p\|_1 - q(p)|\right].$$

When the P_i are adversarially chosen, these expectations are taken over the randomness of the estimator; when each p_i is drawn from a known distribution P_i, these expectations are taken over the product distribution P.

Finally, throughout the paper we denote the centroid of $C \subset \mathbb{R}^n$ by $\mu(C)$. More formally,

$$\mu(C) := \frac{1}{P(C)} \int_C p \, dP,$$

where P is a measure on $\mathbb{R}^n$.

3. Worst-Case Guarantees

In this section we consider the case where no knowledge of the p_i is assumed, and establish upper and lower bounds on the performance of deterministic and randomized estimators.

First, suppose that the estimator $q = r \circ b$ is deterministic. For fixed b, it is clear that r should report the sum corresponding to the center point of each box $C_c = \prod_i B_{i,c_i}$, since this minimizes the worst-case error across all $p \in C_c$. Accordingly, an adversary will seek the box with the largest ℓ_1 diameter, which can be identified by finding the $B_{i,j}$ of maximum diameter for each i. Therefore the worst-case optimal deterministic estimator chooses equipartitions $\mathcal{B}_i = \left\{[0, \frac{1}{k}], \ldots, [\frac{k-1}{k}, 1]\right\}$, reports the ℓ_1 norm of the center of each box, and satisfies

$$\max_p (\|p\|_1 - q(p))^2 = \frac{n^2}{4k^2}$$
$$\max_p |\|p\|_1 - q(p)| = \frac{n}{2k},$$

and this is clearly tight. This is the uniform estimator, and we will denote it $q_U := r_U \circ b_U$, where b_U partitions each $[0, 1]$ into equal-size subintervals, and $r_U(c)$ is the ℓ_1 norm of the center of each C_c.

A randomized estimator, however, can perform significantly better over worst-case inputs. Indeed, consider the following randomized estimator, which we denote by $\bar{q} = r \circ \bar{b}$. We construct a randomized classifier by choosing "shifts" $s_i \in [0, \frac{1}{k-1})$ for each i uniformly and independently. Take

$$\bar{b}_i(p_i) := j \qquad \text{s.t.} \qquad \frac{j-1}{k-1} \leq p_i + s_i < \frac{j}{k-1}.$$

Intuitively, this partitions $[0, 1]$ into k subintervals by taking the $k - 1$ thresholds $1/(k - 1), \ldots, 1$ and shifting them left by s_i. Then make the (deterministic) reports

$$r_i(j) := \frac{j-1}{k-1}$$

for $j \in [k]$, and define $\bar{b}(q) := (\bar{b}_1(p_1), \ldots, \bar{b}_n(p_n))$ and take the aggregator $r(c) := r_1(c_1) + \ldots + r_n(c_n)$. Putting these together yields the randomized estimator

$$\bar{q} := r \circ \bar{b},$$

which is illustrated in Figure 1.

Our main result for this section is the following theorem.

Theorem 1. *In the worst-case setting, the randomized estimator $\bar{q}$ satisfies*

$$\mathsf{mse}(\bar{q}) = O\left(n/k^2\right)$$
$$\mathsf{mae}(\bar{q}) = O\left(\sqrt{n}/k\right).$$

Moreover, these bounds are asymptotically optimal for both measures of error.

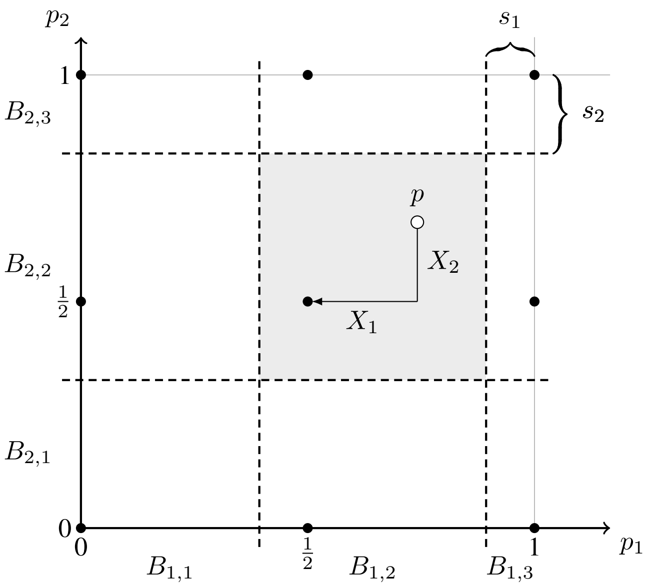


Figure 1: An illustration of $\bar{q}$ for $n = 2$ and $k = 3$. Here $(r_1 \circ \bar{b}_1)(p_1) = 1/2$ and $(r_2 \circ \bar{b}_2)(p_2) = 1/2$, and so $\bar{q}(p) = 1/2 + 1/2 = 1$. The box $C_{(2,2)}$ is highlighted.

The rest of the section is devoted to proving the theorem. We do so via several lemmas, starting with the upper bound.

Lemma 1. *The randomized estimator $\bar{q}$ satisfies*

$$\mathsf{mse}(\bar{q}) = O\left(n/k^2\right)$$
$$\mathsf{mae}(\bar{q}) = O\left(\sqrt{n}/k\right).$$

Proof. For fixed p, we begin by analyzing the estimator coordinate-by-coordinate. Take

$$X_i := p_i - (r_i \circ \bar{b}_i)(p_i)$$

to be the (signed) error of $\bar{q}$ in coordinate i, and let $j_i := \max\{j \in [k] : \frac{j-1}{k-1} \le p_i\}$. If

$$z_i := p_i(k-1) - (j_i - 1)$$

is the proportion of the way between multiples of $1/(k-1)$ that p_i falls, then $X_i = (z_i - 1)/(k-1)$ with probability z_i and $X_i = z_i/(k-1)$ with probability $1 - z_i$. Note that $\mathbb{E}[X_i] = 0$. By the definition of r and the independence of the X_i,

$$\begin{aligned}\mathsf{mse}(\bar{q}(p)) &= \mathbb{E}\left[\left(\sum_i X_i\right)^2\right] = \sum_i \mathrm{Var}\left[X_i\right] \\ &= \sum_i \frac{z_i^2(1-z_i) + z_i(1-z_i)^2}{(k-1)^2} \le \frac{n/4}{(k-1)^2} \\ &= O\left(\frac{n}{k^2}\right).\end{aligned}$$

We now turn to the MAE case. Note that X_i is bounded in some range of width $1/(k-1)$, and that

$$\mathsf{mae}(\bar{q}(p)) = \mathbb{E}[|\|p\|_1 - \bar{q}(p)|] = \mathbb{E}\left[\left|\sum_i X_i\right|\right].$$

Next we apply Hoeffding's inequality in order to upper bound this expectation. By Hoeffding (1963),

$$\Pr\left[\left|\sum_i X_i\right| \ge t\right] \le 2\exp\left\{\frac{-2t^2}{\frac{n}{(k-1)^2}}\right\},$$

and so choosing $t = m\sqrt{n}/(k-1)$ for $m \in \mathbb{N}$ yields

$$\Pr\left[\left|\sum_i X_i\right| \ge \frac{m\sqrt{n}}{(k-1)}\right] \le \frac{2}{e^{2m^2}}. \tag{1}$$

Finally, let $X := \sum_i X_i$ and observe that for any $\sigma > 0$,

$$\begin{aligned}\mathsf{mae}(\bar{q}(p)) &= \mathbb{E}\left[|X|\right] \\ &\le \sum_{m=1}^{\infty} \sigma m \Pr\left[|X| \in [\sigma(m-1), \sigma m]\right] \\ &\le \sum_{m=1}^{\infty} \sigma m \Pr\left[|X| \ge \sigma(m-1)\right].\end{aligned}$$

Taking $\sigma = \sqrt{n}/(k-1)$ and applying Equation (1),

$$\le \frac{\sqrt{n}}{k-1} \sum_{m=1}^{\infty} \frac{2m}{e^{2(m-1)^2}}.$$

This infinite series converges, which implies that $\mathsf{mae}(\bar{q}(p)) = O(\sqrt{n}/k)$, as desired. □

By Yao's minimax principle (Yao, 1977), in order to derive a lower bound for all randomized algorithms, it suffices to fix a distribution over inputs and lower bound the average performance of any deterministic algorithm over this randomized input. To this end, we will consider the uniform distribution over $[0,1]^n$, which we denote D, and lower bound the performance of any deterministic estimator over it. In doing so, we will prove the intuitive fact that the uniform estimator $q_U = r_U \circ b_U$ is optimal for D.

First we present a structural lemma about the optimal aggregator r for any fixed classifier b. Let $S(C, P)$ denote the probability distribution over $\mathbb{R}$ derived by taking the ℓ_1 norm of C:

$$\Pr[S \le x] = \Pr_P[\|p\|_1 \le x | p \in C].$$

We will repeatedly make use of the following insight, which follows from calculus and reflects one of Lloyd's optimality conditions for k-means clustering (Lloyd, 1982):

Lemma 2. *For $p \sim P$ and for fixed b, the MSE- and MAE-optimal aggregators $r_b^1(c)$ and $r_b^2(c)$ (respectively) report the means and medians (respectively) of $S(C_c, P)$ for all $c \in [k]^n$.*

With this in hand, we are ready to analyze the performance of q_U over D, which (by Yao's minimax principle) establishes the lower bound of Theorem 1.

Lemma 3. *If $p \sim D$ is drawn uniformly at random then the uniform estimator $q_U = r_U \circ b_U$ is optimal in terms of both MSE and MAE, and*

$$\mathsf{mae}(q_U) = \Omega\left(n/k^2\right),$$
$$\mathsf{mae}(q_U) = \Omega\left(\sqrt{n}/k\right).$$

Proof. This follows in two steps. We will first prove that the uniform strategy $q_U = r_U \circ b_U$ is optimal for MSE and compute $\mathsf{mse}(q_U)$ directly. Then we will prove that q_U is optimal for MAE, and finally lower bound $\mathsf{mae}(q_U)$.

To begin, note that under the uniform distribution $P = D$ and for fixed b, Lemma 2 implies that for both MSE and MAE, the optimal aggregator is the r_b which reports the ℓ_1 norms of the centers of the C_c. Therefore we may assume that all estimators use reports r_b which are optimal for their b, and argue that b_U is the best partitioning.

The MSE case boils down to a question of variance, and it turns out we can compute the error directly. As before, let

$$X_i := \sum_{j=1}^{k} \mathbb{1}_{\{p_i \in B_{i,j}\}} \left(p_i - \mu(B_{i,j})\right)$$

be the (signed) error in coordinate i, the difference between their actual p_i and the center $\mu(B_{i,j})$ of the interval $B_{i,j}$ containing their p_i. Then because the p_i are independent and

$$r_b(c) = \mu\left(S\left(C_c, D\right)\right) = \sum_i \mu_D(B_{i,c_i}),$$

we have that for $q = r_b \circ b$:

$$\begin{aligned}
\mathsf{mse}(q) = \mathbb{E}_D\left[\left(\|p\|_1 - q(p)\right)^2\right] &= \mathbb{E}_D\left[\left(\sum_i X_i\right)^2\right] \\
= \mathrm{Var}\left[\sum_i X_i\right] &= \sum_i \mathrm{Var}[X_i] = \sum_i \int_0^1 X_i^2 \, dx \\
&= \sum_i \sum_j \int_{B_{i,j}} (x - \mu_D(B_{i,j}))^2 \, dx \\
&= \sum_i \sum_j \frac{1}{12} \mathsf{diam}(B_{i,j})^3.
\end{aligned}$$

At this point, the method of Lagrange multipliers confirms that MSE is minimized when all $B_{i,j}$ are of equal diameter $1/k$, which yields precisely b_U. Therefore $q_U = r_U \circ b_U$ is optimal for D, and it has cost $\mathsf{mse}(q_U) = n/12k^2$.

We now turn to MAE, and use a differential argument to prove that q_U is MAE-optimal when $P = D$. We begin by showing that the MAE contributed by a box C_c is convex in each of the dimensions of C_c. This will let us argue that for any classifier b with partitions $\{\mathcal{B}_i\}$ in which one partition is unbalanced, meaning that for some i (say $i = 1$) and some j it is the case that $\mathsf{diam}(B_{1,j}) < \mathsf{diam}(B_{1,j+1})$, the $b^\star$ which equalizes their widths decreases the MAE: $\mathsf{mae}(r_{b^\star} \circ b^\star) < \mathsf{mae}(r_b \circ b)$. This then implies that $q_U = r_U \circ b_U$ is MAE-optimal, since it is the only b which cannot be equalized in this way.

To see that MAE contribution is convex in each dimension of C, let $C' := \prod_{i=2}^n [-w_i, w_i]$ and consider the box $C(t) := [-t, t] \times C'$. Then by Lemma 2 the optimal report for both C and C' is 0. Let the contribution to MAE by C with report r be denoted $e(C, r)$, and call the contribution with optimal report $e(C)$. Then by symmetry,

$$\begin{aligned}
e(C(t), 0) &= \int_{C(t)} \left|\sum_i p_i\right| \, dp \\
&= 2\int_0^t \int_{C'} \left| x + \sum_{i=2}^n p_i \right| \, dp \, dx \\
&= 2\int_0^t e(C', x) \, dx,
\end{aligned}$$

and therefore

$$\frac{de(C(t), 0)}{dt} = 2e(C', t).$$

The (omitted) proof of Lemma 2 shows that $\frac{de(C,r)}{dr} > 0$ for r greater than the optimal report, and so

$$\frac{d^2 e(C(t), 0)}{dt^2} > 0,$$

as desired.

In order to show that $b^\star$ improves upon b, note that their boxes C_c differ only for those of the form

$$\hat{C}_c := B_{1,j} \times \prod_{i=2}^n B_{i,c_i}, \qquad \tilde{C}_c := B_{1,j+1} \times \prod_{i=2}^n B_{i,c_i}$$

Pairing these up by c, it suffices to show that for each c,

$$e(\hat{C}_c^\star) + e(\tilde{C}_c^\star) < e(\hat{C}_c) + e(\tilde{C}_c),$$

where $C_c^\star$ are the boxes given by $b^\star$. This follows from the convexity of $e(C)$ in each dimension of C, established above. We conclude that q_U is MAE-optimal.

We next lower bound $\mathsf{mae}(q_U)$. Let $L \subseteq [n]$ and $H \subseteq [n]$ be the set of indices i with errors X_i that are negative and positive, respectively. It holds that

$$\begin{aligned}\mathsf{mae}(q_U) &= \mathbb{E}\left[\left|\sum_{i\in[n]}(p_i - q_U(p_i))\right|\right] \\ &= \mathbb{E}\left[\left|\sum_{i\in L} X_i + \sum_{i\in H} X_i\right|\right].\end{aligned}$$

We establish that L and H are, with constant probability, of sufficiently different sizes to lead to $\sqrt{n}/k$ error. First, note that because we are playing against a uniform adversary, the probability that each p_i is in L is $1/2$; the probability that each p_i is in H is symmetrically $1/2$. Because these are just Bernoulli random variables, applying the De Moivre-Laplace theorem (a version of the Central Limit Theorem) tells us that, as n becomes large, the sum of these Bernoulli random variables converges to a normal distribution with mean $n/2$ and standard deviation $\sqrt{n}/2$. Therefore, we know that with constant probability, the total number of agents in H is at least $\sqrt{n}$ from its mean of $n/2$, that is,

$$\Pr\left[||H| - \mathbb{E}[|H|]| \geq \sqrt{n}\right] = \beta$$

for a constant β. It follows that, with constant probability β, $||L| - |H|| \geq 2\sqrt{n}$; denote this event by $\mathcal{E}$. Therefore,

$$\begin{aligned}\mathsf{mae}(q_U) &= \mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|\right] \\ &\geq \mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right| \,\middle|\, \mathcal{E}\right]\cdot \Pr[\mathcal{E}] \qquad (2) \\ &= \mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right| \,\middle|\, \mathcal{E}\right]\cdot \beta.\end{aligned}$$

Now assume that $\mathcal{E}$ occurred Without loss of generality, assume that $|L| \leq |H|$ and, in particular, randomly break H up into two sets, H_1 and H_2, such that $|H_1| = |L|$ and $|H_2| \geq 2\sqrt{n}$. By construction of H_1, the sum of the errors in indices $i \in L \sqcup H_1$ is symmetric with mean 0. It holds that

$$\begin{aligned}&\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right| \,\middle|\, \mathcal{E}\right] \\ &\quad= \mathbb{E}\left[\left|\sum_{i\in L\cup H_1} X_i + \sum_{i\in H_2} X_i\right| \,\middle|\, \mathcal{E}\right] \\ &\quad\geq \Pr\left[\left|\sum_{i\in L\cup H_1} X_i\right| \geq 0 \,\middle|\, \mathcal{E}\right] \mathbb{E}\left[\left|\sum_{i\in H_2} X_i\right| \,\middle|\, \mathcal{E}\right] \\ &\quad\geq \frac{1}{2}\cdot 2\sqrt{n}\cdot\frac{1}{4k} = \Omega\left(\frac{\sqrt{n}}{k}\right).\end{aligned}$$

The desired bound follows by combining this with Equation (2). □

4. Estimation with Priors

In practice, it is useful to go beyond worst-case guarantees and ask to what degree knowing some additional information about the p_i can improve our ability to estimate their sum. Specifically, suppose that we have access to distributions $P_1, \ldots, P_n$ from which the p_i are drawn; we make the standard assumption (for computational complexity results) that these distributions are discrete. Can we design a more accurate estimator that takes prior knowledge into account?

For instance, in our running example of admissions (say for a Ph.D. program), one could easily come up with priors by applying machine learning to historical admissions data. The prior for a candidate would, of course, depend on relevant features such as their alma mater and research interests.

It is important to note that, in the setting with priors, the optimal strategyproof estimator is a *deterministic* partition into intervals. Indeed, note that the process for randomized mechanisms in the setting with priors consists of the following steps. First, the mechanism announces its distribution over multiple choice questions. Second, the mechanism chooses multiple choice questions and proposed payments from its distribution, and true p_i's are independently drawn from the known priors. Third, agents answer the multiple choice questions posed by the mechanism. Finally, payments to agents are made based on their answers and the true outcomes. Crucially, when agents answer the multiple-choice questions, they have already seen which questions were drawn from the mechanism's distribution. This means that a randomized truthful estimator must be a distribution over deterministic truthful estimator, which, per Lambert & Shoham (2009), are defined by partitions into intervals. But the error of such a distribution is just a convex combination of the errors of deterministic estimators in its support.

4.1. An Efficient Estimator for MSE

When estimation is evaluated according to MSE, it turns out that we can answer the foregoing question in the positive:

Theorem 2. *Given discrete priors P_i for each p_i, $i \in [n]$, there is a polynomial-time estimator that is optimal with respect to MSE.*

A key component of our analysis is the following structural insight:

Lemma 4. *Given an estimator $q = r_b \circ b$, where r_b reports optimally for a given classifier b, $\mathsf{mse}_P(q) = \sum_i \mathsf{mse}_{P_i}(q_i)$, where $q_i = r_{b_i} \circ b_i$.*

We will prove this by employing a result from the vector quantization literature. In quantization, roughly speaking, the task is to compress some signal using only a representative subset of its values in such a way that compression error is minimized. Vector quantization performs this task for vector-valued signals, and does so using an n-dimensional partition (Gray & Neuhoff, 1998). Since it typically seeks an MSE-error-minimizing vector representative $\vec{r}(c)$ for each box C_c in its n-dimensional partition, vector quantization may be seen as an instance of k^n-means clustering in $\mathbb{R}^n$ subject to the constraint that all clusters obey this product-of-partitions structure. In this setting, it is known (Jégou et al., 2011) that if the partitions are made along independent axes, then the MSE of the optimal vector quantizer is additive:

Lemma 5. *If $\vec{q}$ is a vector quantizer as described above which reports the centroids $\mu(C_c)$ for every C_c, and is given by $\vec{r}(c) = (r_1(c_1), \ldots, r_n(c_n))$ and $b(p) = (b_1(p_1), \ldots, b_n(p_n))$, then*

$$\mathbb{E}_P \left[\|p - \vec{q}(p)\|_2^2 \right] = \sum_i \mathbb{E}_{P_i} \left[(p_i - r_i(b_i(p_i)))^2 \right], \quad (3)$$

where each r_i reports the centroid $\mu(B_{i,j})$.

This is a direct consequence of independence of the p_i, together with the fact that the X_i have mean 0.

The key difference between our problem and the vector quantization setting is that we measure error with respect to the aggregate $\sum_i p_i$, which means errors with respect to individual agents can "cancel out." Nevertheless, the same structural insight holds, as we show now.

Proof of Lemma 4. We proceed in two steps. First we argue that for fixed classifier b, the optimal report r_b reports the ℓ_1 norm $r_b(c) = \|\mu_P(C_c)\|_1$ of the centroid of each box C_c. Next, we argue that this coincides with the error given on the right-hand side of Equation (3); the theorem then follows.

We begin by showing that the optimal aggregator reports $r_b(c) = \|\mu_P(C_c)\|_1$. As in Lemma 3, let $e(C, r)$ denote the contribution of $C = \prod_i B_i$ to MSE under report r. We proceed via a differential argument:

$$\begin{aligned} e(C, r) &= \int_C (\|p\|_1 - r)^2 \, dP \\ &= r^2 \int_C dP - 2r \int_C \|p\|_1 \, dP + \int_C \|p\|_1^2 \, dP \\ \frac{de(C, r)}{dr} &= 2r \int_C dP - 2 \int_C \|p\|_1 \, dP. \end{aligned}$$

Setting $\frac{de(C,r)}{dr} = 0$ yields the optimal report,

$$\begin{aligned} r^\star &= \frac{1}{P(C)} \int_C \|p\|_1 \, dP = \frac{1}{P(C)} \left(\sum_i \int_C p_i \, dP \right) \\ &= \frac{1}{\prod_i P_i(B_i)} \sum_i \left(\int_{B_i} p_i \, dP_i \prod_{j \neq i} \int_{B_j} dP_j \right) \\ &= \sum_i \frac{1}{P_i(B_i)} \int_{B_i} p_i \, dP_i = \sum_i \mu_{P_i}(B_i) \\ &= \|\mu_P(C)\|_1, \end{aligned}$$

where this last step is a standard property of the centroid.

Therefore for $q = r_b \circ b$, the error $\mathsf{mse}(q)$ takes the form

$$\begin{aligned} \mathsf{mse}_P(q) &= \sum_{c \in [k]^n} \int_{C_c} \left(\|p\|_1 - \|\mu(C_c)\|_1 \right)^2 \, dP \\ &= \sum_{c \in [k]^n} \int_{C_c} \left(\sum_i (p_i - \mu_{P_i}(B_{i,c_i})) \right)^2 \, dP. \end{aligned}$$

Since the centroid is an unbiased estimator,

$$\begin{aligned} &= \sum_{c \in [k]^n} \int_{C_c} \sum_i (p_i - \mu_{P_i}(B_{i,c_i}))^2 \, dP \\ &= \mathbb{E}_P \left[\|p - \vec{q}(p)\|_2^2 \right], \end{aligned}$$

as in the right-hand side of Equation (3). Therefore by Lemma 5,

$$\begin{aligned} &= \sum_i \mathbb{E}_{P_i} \left[(p_i - \mu(B_{i,c_i}))^2 \right] \\ &= \sum_i \mathsf{mse}_{P_i}(q_i). \end{aligned}$$

□

Proof of Theorem 2. Our goal is to minimize $\mathsf{mse}_P(q)$, and Lemma 4 implies that this can be accomplished by individually minimizing $\mathsf{mse}_{P_i}(r_{b_i} \circ b_i)$ for each $i \in [n]$. Since the P_i are discrete distributions, finding the b_i which minimizes $\mathsf{mse}_{P_i}(r_{b_i} \circ b_i)$ is precisely an instance of one-dimensional Euclidean k-means. It is well-known that this can be solved efficiently via dynamic programming using a recurrence described by Jensen (1969). Therefore, given priors P_i we can derive an estimator q which minimizes $\mathsf{mse}(q)$: we first find $b_1^\star, \ldots, b_n^\star$ which minimize $\mathsf{mse}_{P_i}(r_{b_i}^2 \circ b_i)$ for each $i \in [n]$, where $r_{b_i}^2$ is the optimal aggregator for b_i, then take the optimal partitions

$$b(p) = (b_1^\star(p_1), \ldots, b_n^\star(p_n)).$$

Lemma 4 then guarantees that this $q = r_b^2 \circ b$ is optimal. □

4.2. Hardness for MAE

In contrast to the case of MSE with priors, devising an MAE-optimal mean estimation strategy is $\#\mathcal{P}$-hard.

To be concrete, the computational problem is defined as follows: Given a collection of discrete prior distributions $P_1, \ldots, P_n$ and a positive integer k, we are asked for a collection of partitions $b_i : [0,1] \to [k]$ and an aggregator $r : [k]^n \to \mathbb{R}$ which together minimize $\mathsf{mae}(r \circ b) = \mathbb{E}_P \left[|\|p\|_1 - r \circ b(p)| \right]$, where $P = \prod_i P_i$.

Theorem 3. *Given discrete priors P_i for each p_i, $i \in [n]$, the problem of computing an optimal estimator with respect to MAE is $\#\mathcal{P}$-hard.*

To prove the theorem, we reduce from the following problem. Given a rational $x \in [0,1]$ and nonnegative integer weights $\alpha_1, \ldots, \alpha_n$, WEIGHTED-BINOMIAL-MEDIAN (WBM) asks for a median of the random variable $Z := \sum_{i=1}^n \alpha_i \,\mathsf{Bernoulli}(x)$, where the $\mathsf{Bernoulli}(x)$ are independent (and identically distributed). In the full version of the paper,[3] we prove that WBM is $\#\mathcal{P}$-Hard.

Proof of Theorem 3. We reduce from WBM. If $k = 1$ then the reduction is immediate: if each of the P_i is a scaled down copy of $\alpha_i \mathsf{Bernoulli}(x)$, then finding the optimal report for the random variable $\sum_i P_i$ amounts to finding the (scaled down) median of $\sum_i \alpha_i \mathsf{Bernoulli}(x)$.

More generally, given an instance of WBM described by $(x, \alpha_1, \ldots, \alpha_n)$, we will construct an instance of our problem, MAE-ESTIMATOR, for any $k \geq 2$ for which determining optimal partitions and reporting scheme will solve our instance of WBM.

Our P_i will be discrete distributions given by

$$\Pr\left[p_i = \frac{1}{2k}\right] = \frac{1-x}{k} \tag{4}$$

$$\Pr\left[p_i = \frac{1 + \delta \frac{\alpha_i}{\sum_t \alpha_t}}{2k}\right] = \frac{x}{k} \tag{5}$$

$$\Pr\left[p_i = \frac{2j-1}{2k}\right] = \frac{1}{k} \qquad \text{for } j = 2, \ldots, k\,. \tag{6}$$

We will choose δ small enough such that the optimal partition of each of the P_i necessarily groups the atoms described in Equation (4) and Equation (5) together, and gives each of the atoms of Equation (6) its own interval in the partition. To find such a δ, first consider the "good" case when the partitions are of this form. In this case, there are k^n total boxes, each with weight $1/k^n$. Within each box C, the distribution of ℓ_1 norms has range upper bounded by $\delta/(2k)$. Within each C, the range of this distribution is an upper bound on the ℓ_1 distance between any atom in C and the optimal report for C. Therefore, a loose upper bound on total MAE is

$$\sum_{c \in [k]^n} P(C_c) \frac{\delta}{2k} = \frac{\delta}{2k}. \tag{7}$$

On the other hand, consider the "bad" case when at least one of the partitions groups either two of the Equation (6) atoms together or the Equation (5) atom together with at least one of the Equation (6) atoms. Assume without loss of generality that the $i = 1$ partitioning is "bad". We will focus on the case when an Equation (5) and at least one Equation (6) atom are grouped together (because it is an interval, necessarily $j = 2$ is included), since in the best case it is the least costly scenario. Because of the product structure of the boxes induced by the partitions, for every pair of vectors u and u' in the support of P of the form

$$u = \left(\frac{1 + \delta \frac{\alpha_i}{\sum_j \alpha_j}}{2k}, u^- \right) \qquad u' = \left(\frac{3}{2k}, u^- \right),$$

where $u^- \sim \prod_{j=2}^n P_j$, necessarily u and u' are contained in the same box. Therefore among each pair of u and u', at least

$$M_{u^-} = \frac{\min\{x, 1-x\}}{k} \prod_{j=2}^n P_j(u_j^-)$$

mass must travel $\|u'\|_1 - \|u\|_1$ to the estimate for their shared box, which yields a lower bound on the error of

$$\sum_{u^-} \left(\frac{1-\delta}{k} M_{u^-} \right) = \frac{(1-\delta)\min\{x, 1-x\}}{k^2}. \tag{8}$$

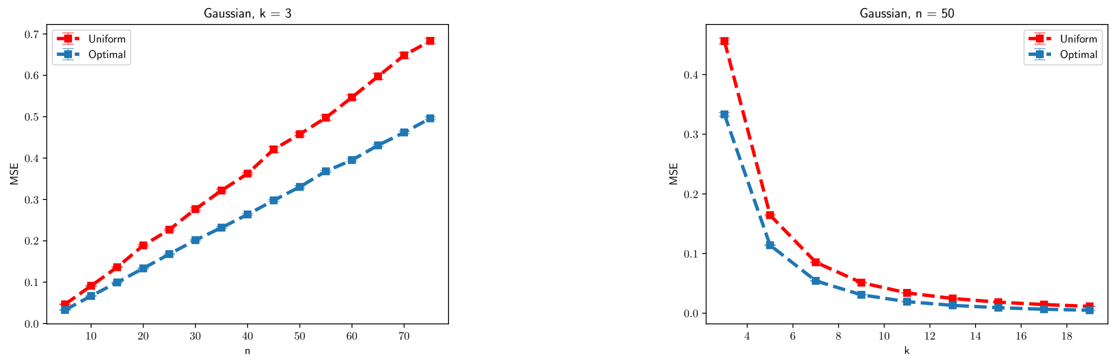
By Equations (7) and (8), choosing a $\delta < \frac{\min\{x, 1-x\}}{k}$ guarantees that the optimal partitioning for our instance is the "good" partitioning, and so all of the Equation (4) and Equation (5) atoms appear in the same box $C^\star := \prod_i B_{i,1}$.

Recall that by Lemma 2, the MAE-minimizing estimate for a fixed box C is a median of the distribution of ℓ_1 norms of the vectors $u \in C$ according to P. Therefore MAE-ESTIMATOR finds some MAE-optimal report $r^\star$ for the box $C^\star$, which by Equation (4) and Equation (5) implies that $\frac{r^\star - n/2k}{\delta}$ is a median of $\sum_i \alpha_i \mathsf{Bernoulli}(x)$, solving the given instance of WBM. $\square$

5. Experiments

In Section 4 we showed that when prior distributions are known, the optimal estimator with respect to the MSE can be computed in polynomial time. However, it is reasonable to ask to what degree incorporating these prior distributions leads to more accurate estimation schemes for plausible

[3] The full version is available at http://procaccia.info/publications.



(a) Fixed $k = 3$, ranging over n

(b) Fixed $n = 50$, ranging over k

Figure 2: MSE of the uniform and optimal estimators, averaged over 100 distributions sampled from the Gaussian family. Bars are standard error of the mean.

families of prior distributions. In this section we aim to answer this question. We focus on the MSE as our measure of error, because Theorem 3 shows that an optimal estimator with respect to MAE is hard to compute.

In more detail, we compare the MSE-optimal prior-sensitive estimator of Theorem 2 to the deterministic worst-case optimal strategy described in Section 3, which does not incorporate knowledge of prior distributions. This is the uniform estimator, which for all $i \in [n]$ partitions the ith interval into equal intervals of length $1/k$. It is pointless to use the more elaborate randomized estimator of Theorem 1, because when instances are drawn from a distribution, randomization does not help: for any randomized estimator there is a deterministic estimator that performs at least as well.

We compare the two estimation schemes (optimal and uniform) on discrete distributions drawn from three families: uniform, Gaussian, and bimodal. Recall that in our model prior distributions are comprised of m atoms. For each choice of number of agents n, number of intervals k, and family of distributions, we generate instances as follows: we sample n discrete prior distributions of m uniformly weighted atoms, one distribution per agent. For the family of uniform distributions, each sampled distribution is formed by drawing m samples from $U(0, 1)$; for the Gaussian family, each sampled distribution is formed by drawing m samples from a truncated Gaussian over the domain $[0, 1]$ with mean μ and standard deviation σ drawn i.i.d. from $U(0, 1)$; and for the bimodal family, each sampled distribution is formed by drawing m samples from an equal mixture of two truncated Gaussians over the domain $[0, 1]$, each again with μ and σ drawn i.i.d. from $U(0, 1)$. We then independently sample a number of points from each of the agents' discrete distributions to form a collection of draws $p \sim P$, and evaluate the performance of both the uniform and MSE-optimal estimators on these draws.

Figure 2 shows sample averages of MSE for both the uniform and optimal estimators applied to distributions from the Gaussian family, for a range of n and for fixed $k = 3$, or for a range of k and fixed $n = 50$. The MSE for each pair of generated distribution P and estimator is measured as an average over 1000 draws $p \sim P$. For a fixed value of k, as n increases, the optimal estimator significantly outperforms the uniform estimator, suggesting that knowledge of the distribution gives a benefit in practice. For additional figures and simulation details, see the full version of the paper.

6. Discussion

Although we have assumed throughout that $p_i \in [0, 1]$, the results in Section 3 can be generalized to general bounded p_i. Similarly, the optimal strategy described in Section 4.1 holds for prior distributions over unbounded $p_i \in \mathbb{R}$.

There are also promising avenues for future work to extend our results to richer settings. For instance, in Section 4 we assume that the prior distributions P_i are given. However, it would be interesting to consider the setting in which the P_i are initially unknown but gradually discovered over rounds of questions; i.e., a learning setting where the elicitation scheme learns the prior distributions P_i in the course of accurately estimating the mean of the p_i.

Moreover, while we have focused on estimating $\|p\|_1$, one may ask if it is possible to estimate other functions of p. For instance, can the median of the p_i be efficiently and accurately estimated in this multiple-choice question setting?

Future work may also explore settings in which the error metric is asymmetric: For instance, it may be more costly to underestimate than overestimate the size of a matriculating class due to space and resource constraints, and an optimal estimator would take this cost asymmetry into account.

Acknowledgments

This work was partially supported by the National Science Foundation under grants CCF-2007080, IIS-2024287 and CCF-1733556; and by the Office of Naval Research under grants N00014-16-1-3075 and N00014-17-1-2428.

References

Acharya, J., Diakonikolas, I., Hegde, C., Li, J. Z., and Schmidt, L. Fast and near-optimal algorithms for approximating distributions by histograms. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pp. 249–263, 2015.

Babcock, B., Chaudhuri, S., and Das, G. Dynamic sample selection for approximate query processing. In *Proceedings of the 13th International Conference on Management of Data (SIGMOD)*, pp. 539–550, 2003.

Berg, J., Forsythe, R., Nelson, F., and Rietz, T. Results from a dozen years of election futures markets research. In Plott, C. R. and Smith, V. L. (eds.), *Handbook of Experimental Economics Results*, pp. 742–751. Elsevier, 2008.

Boutilier, C. A POMDP formulation of preference elicitation problems. In *Proceedings of the 18th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 239–246, 2002.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Caragiannis, I., Procaccia, A., and Shah, N. Truthful univariate estimators. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 127–135, 2016.

Chaudhuri, S., Das, G., Datar, M., Motwani, R., and Narasayya, V. Overcoming limitations of sampling for aggregation queries. In *Proceedings of the 17th IEEE International Conference on Data Engineering (ICDE)*, pp. 534–542, 2001.

Chaudhuri, S., Das, G., and Narasayya, V. Optimized stratified sampling for approximate query processing. *ACM Transactions on Database Systems (TODS)*, 32(2): article 9, 2007.

Chen, Y. and Kash, I. A. Information elicitation for decision making. In *Proceedings of the 10th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 175–182, 2011.

Ding, B., Huang, S., Chaudhuri, S., Chakrabarti, K., and Wang, C. Sample + seek: Approximating aggregates with distribution precision guarantee. In *Proceedings of the 26th International Conference on Management of Data (SIGMOD)*, pp. 679–694, 2016.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Good, I. J. Rational decisions. *Journal of the Royal Statistical Society*, 14(1):107–114, 1952.

Gray, R. and Neuhoff, D. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, 1998.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jagadish, H. V., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K. C., and Suel, T. Optimal histograms with quality guarantees. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, pp. 24–27, 1998.

Jégou, H., Douze, M., and Schmid, C. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

Jensen, R. E. A dynamic programming algorithm for cluster analysis. *Operations Research*, 17(6):1034–1057, 1969.

Lambert, N. and Shoham, Y. Eliciting truthful answers to multiple-choice questions. In *Proceedings of the 10th ACM Conference on Economics and Computation (EC)*, pp. 109–118, 2009.

Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

Radanovic, G. and Faltings, B. Incentives for truthful information elicitation of continuous signals. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 770–776, 2014.

Soloviev, M. and Halpern, J. Y. Information acquisition under resource limitations in a noisy environment. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 6443–6450, 2018.

Waggoner, B. and Chen, Y. Output agreement mechanisms and common knowledge. In *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pp. 220–226, 2014.

Winkler, R. L. Scoring rules and the evaluation of probability assessors. *Journal of the American Statistical Association*, 64(327):1073–1078, 1969.

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

Yao, A. C. Probabilistic computations: Towards a unified measure of complexity. In *Proceedings of the 17th Symposium on Foundations of Computer Science (FOCS)*, pp. 222–227, 1977.

Zohar, A. and Rosenschein, J. S. Mechanisms for information elicitation. *Artificial Intelligence*, 172(16-17): 1917–1939, 2008.