Truthful Univariate Estimators

Ioannis Caragiannis CARAGIAN@CEID.UPATRAS.GR
University of Patras, Greece

Ariel D. Procaccia ARIELPRO@CS.CMU.EDU
Carnegie Mellon University, USA

Nisarg Shah NKSHAH@CS.CMU.EDU
Carnegie Mellon University, USA

Abstract

We revisit the classic problem of estimating the population mean of an unknown single-dimensional distribution from samples, taking a game-theoretic viewpoint. In our setting, samples are supplied by strategic agents, who wish to pull the estimate as close as possible to their own value. In this setting, the sample mean gives rise to manipulation opportunities, whereas the sample median does not. Our key question is whether the sample median is the best (in terms of mean squared error) *truthful* estimator of the population mean. We show that when the underlying distribution is symmetric, there are truthful estimators that *dominate* the median. Our main result is a characterization of worst-case optimal truthful estimators, which provably outperform the median, for possibly asymmetric distributions with bounded support.

1. Introduction

A central problem in statistics deals with estimating the *population mean* of an unknown distribution D given n samples $x_1, \ldots, x_n$ drawn i.i.d. from the distribution (Fisher, 1925; Lehmann & Casella, 1998). Classic results show that, not surprisingly, the sample mean $(1/n)\sum_i x_i$ is a good estimator for the population mean. It is unbiased and often minimizes a popular risk function—the *mean squared error* (MSE)—that measures the error of an estimator by the expected square of the distance between the estimate and the population mean (Lehmann & Casella, 1998).

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

However, from the game-theoretic viewpoint, the sample mean suffers from serious incentive issues when the input samples are provided by (and are the private information of) self-interested individuals. To illustrate this, imagine a setting where we want to set a common temperature throughout a building with many occupants. Each occupant has an ideal temperature in mind, and our goal is to set the temperature to be the mean of the distribution representing the occupants' ideal temperatures.[1] We ask the occupants to fill out a survey, which is completed by n (arguably random) respondents, and use the sample mean to set the common temperature. The difficulty is that, for example, an occupant who prefers a relatively high temperature can easily manipulate the outcome by lying about her ideal temperature and reporting a much higher value, thereby pulling the sample mean closer to her desired ideal temperature.

This simple example is representative of a much broader phenomenon. Indeed, machine learning based algorithms now govern many aspects of our daily lives. Their power stems from data, but when it comes from strategic sources and incentives are not correctly aligned, the data will inevitably be contaminated, and any statistical properties the algorithms possess in the non-strategic setting will be worthless. This has led a number of researchers in machine learning, algorithmic game theory, and economics to explore the role of incentives in machine learning and statistics; see Section 1.2 for a brief survey. The current paper contributes to this line of research by addressing the issue of strategic behavior in the ubiquitous statistical problem of estimating the mean of a single-dimensional distribution.

Going back to our illustrative example, one approach that is known to prevent any incentives for manipulation is to take the *median* (rather than the mean) of the elicited samples. Why? If an occupant's ideal temperature is higher than

[1] This is in the spirit of the Predicted Mean Vote (PMV) model of thermal comfort (Fanger, 1970).

the median, submitting a value higher than her ideal temperature will have no effect, and submitting a value lower than the ideal temperature can only reduce the median, thus pushing it farther away from the occupant's ideal temperature. A symmetric argument works if the occupant's ideal temperature is lower than the median. It is clear that no occupant can report false inputs to bring the sample median closer to her ideal temperature—regardless of what the other occupants report. We refer to estimators that exhibit such immunity to manipulation as *truthful estimators*.

Remarkably, the truthfulness of the sample median sometimes comes at a small cost in terms of loss of statistical efficiency. For instance, if the underlying distribution is a Gaussian with variance σ^2, the MSE of the sample mean is σ^2/n (which is known to be optimal among all estimators), and the MSE of the sample median is larger by only a small constant factor—approximately $(\pi/2)\cdot\sigma^2/n$ (Fisher, 1925). This observation motivates our main research question, which is simple yet previously unasked:

> *Is the sample median the most statistically efficient (in terms of mean squared error) truthful estimator of the population mean?*

1.1. Our Results

The starting point of our analysis is a powerful, classic result at the intersection of game theory and social choice (Moulin, 1980): the only way to elicit and aggregate n private inputs truthfully (subject to additional mild requirements) is to take the median of $2n-1$ values—the n elicited values and $n-1$ predetermined "phantom" values. Originally proved in a non-statistical domain completely different from ours, this result applies to our domain and yields a characterization of truthful estimators. However, at first glance the role of phantom values is unclear in our domain of statistical estimation.

Our first result (Proposition 3.4) surprisingly shows that there is a way to set the values of the phantoms, in the absence of any information about the underlying distribution, that is guaranteed to be at least as good as the sample median on any symmetric distribution in terms of mean squared error, and strictly better under mild conditions.

Our main result (Theorem 4.5) extends the analysis to possibly asymmetric distributions but with a known bounded support, which we assume to be $[0,1]$ without loss of generality. In this case, there do not exist truthful estimators that are always better than the sample median (and vice versa). We thus use the well-established principle of minimax mean squared error (Perron & Marchand, 2002). Theorem 4.5 fully characterizes truthful estimators achieving the minimax mean squared error of $1/16$ in the limit as the number of samples grows. Interestingly, the sample median achieves a far worse mean squared error of $1/4$.

1.2. Related Work

As mentioned above, there are several papers that study fundamental learning tasks in settings that involve strategic agents, either as providers of sampled data or as experts responding to the data (Dalvi et al., 2004; Perote & Perote-Peña, 2004; Dekel et al., 2010; Meir et al., 2012; Horel et al., 2014; Cai et al., 2015; Cummings et al., 2015; Hardt et al., 2016). Here we discuss the most closely related ones.

Dekel et al. (2010) study truthful algorithms for regression learning, where the labels for examples (which are sampled from an underlying distribution) come from strategic agents, who want the output to be as close as possible to their own beliefs. Dekel et al. are interested in minimizing a given loss function with respect to examples drawn from the underlying distribution (rather than with respect to a summary statistic, like in our case). They give special attention to the case where each agent controls a single point, and focus on deriving conditions that guarantee that empirical risk minimization is truthful. Mapping these results to our setting (by setting the function class to be constant functions, and the loss function to be absolute loss or squared loss) would simply imply that the median is truthful, while the mean is not.

Similarly to Dekel et al. (2010), Perote and Perote-Peña (2004) design truthful estimators for linear regression when samples are controlled by strategic agents, but in a non-statistical setting. In contrast, in our setting the family of all truthful estimators is well understood (Moulin, 1980).

Another recent example is the paper of Cai et al. (2015), who study a truthful estimation setting, focusing also on problems like linear regression. Their setting is fundamentally different from ours on both the conceptual and technical levels. Indeed, they imagine a statistician who wants to incentivize workers to exert an effort, and the workers want to maximize payment minus effort. In contrast, in our setting the individuals are interested in the outcome of the estimation process itself—this allows us to obtain positive results without the use of monetary incentives.

2. Model and Truthful Estimators

In our model there exists an unknown distribution D over the reals. There are n individuals, and each individual i holds a sample x_i drawn from D. Our goal is to design a univariate estimator $g : \mathbb{R}^n \to \mathbb{R}$ that takes these samples as input, and returns an estimate for a parameter θ (e.g., the population mean) of the underlying distribution. Hereinafter, the estimators are univariate.

In the absence of additional structure, this is a classic problem that has been studied extensively in the statistics and machine learning literature. We study a setting where each

individual i is strategic and her goal is to bring the final estimate close to her sample x_i (which is her private value). In order to facilitate this, she may manipulate and report a number $\hat{x}_i$ different from x_i. The behavior of each individual i can be explained through a utility function $u_i : \mathbb{R} \to \mathbb{R}$ (unknown to the estimator) that describes the utility derived as a function of the value of the final estimate. A typical assumption in the game theory literature is that these single-dimensional utilities are *single-peaked* around the private value x_i (Nisan et al., 2007). Informally, this means that the farther the estimate from x_i (in either direction) the lower the utility derived. Formally, for all $a \leq b \leq x_i$ and for all $a \geq b \geq x_i$, we must have $u_i(a) \leq u_i(b) \leq u_i(x_i)$.

In this case, our goal is to find a *truthful estimator* of θ that incentivizes the individuals to report their samples truthfully, assuming they have single-peaked utilities. The attention to truthful estimators can be justified via the well-known *revelation principle* (Gibbard, 1973): for every non-truthful estimator, there exists an estimator that takes the inputs, performs the optimal manipulation on behalf of the individuals, and then passes the manipulated input to the non-truthful estimator. This new estimator is clearly truthful, and gives the same output as the non-truthful estimator does with manipulated inputs. This general principle allows us to restrict our search space to that of truthful estimators. Formally, we say that an estimator $g : \mathbb{R}^n \to \mathbb{R}$ is *truthful* (a.k.a. *strategyproof*) if for all $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, $i \in \{1, \ldots, n\}$, and $x_i' \in \mathbb{R}$, we have $u_i(g(\mathbf{x})) \geq u_i(g(\mathbf{x}_{-i}, x_i'))$ for all utility functions u_i that are single-peaked around x_i. Here, $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

Moulin (1980) provides a full characterization of truthful aggregation functions in single-dimensional social choice settings. Specifically, he is interested in settings where the single-peaked preferences of the individuals are private but fixed (i.e., there is no underlying distribution). For example, one can imagine the peaks x_i to be positions of the employees of a startup on the question of how high their common wage should be; in this case, the function g would return a joint decision regarding the wage. Below we reformulate the results of Moulin in the language of estimators.

The main result of Moulin's paper is actually a characterization of truthful estimators that satisfy two additional intuitive conditions. We say that an estimator g is *range-respecting* if for all $\mathbf{x} \in \mathbb{R}^n$, we have $\min(\mathbf{x}) \leq g(\mathbf{x}) \leq \max(\mathbf{x})$.[2] We say that an estimator is *anonymous* if permuting the samples does not change the output. In other words, the estimator is a function of the *set* of samples, and does not depend on their indices. Moulin (1980) identified (actually, characterized) the class of truthful, anonymous, and range-respecting estimators as *generalized medians*.

Definition 2.1. *A generalized median of n samples is parametrized by $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R} \cup \{+\infty, -\infty\}$ (called phantoms), and is given by $\texttt{med}(x_1, \ldots, x_n, \alpha_1, \ldots, \alpha_{n-1})$, where $\texttt{med}$ is the standard median. That is, every generalized median is the median of $2n-1$ values: n samples and $n-1$ predetermined phantoms.*

Lemma 2.2 (Moulin (1980), reformulated)**.** *An estimator is truthful, anonymous, and range-respecting if and only if it is a generalized median.*

Moulin's characterization is extremely powerful due to its generality: it does not depend on the parameter to be estimated, the loss function used to measure the error, or the underlying distribution (in fact, as mentioned above, it is formulated in a setting where the private values are *not* drawn from a distribution). We discuss its other potential applications in Section 5.

In order to develop intuition for Moulin's characterization, we give an example of a family of truthful estimators.

Example 2.3 (Order Statistics as Generalized Medians)**.** Given $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, let $x_{(1)}, \ldots, x_{(n)}$ be a permutation such that $x_{(i)} \leq x_{(i+1)}$ for $i \in \{1, \ldots, n-1\}$. Then, $x_{(i)}$ is known as the i^{th} order statistic. Estimator $g(\mathbf{x}) = x_{(i)}$ is a generalized median in which $n-i$ phantoms are placed at $-\infty$ and $i-1$ phantoms are placed at ∞. For odd n, the sample median $\texttt{med}(\mathbf{x}) = x_{(n+1)/2}$ is recovered by placing an equal number of phantoms at $-\infty$ and ∞. For even n, the standard definition (see Cabrera et al., 1994) of the median $(x_{(n)} + x_{(n+1)})/2$ is not truthful. We need to use either the left-median $\texttt{med}_\ell(\mathbf{x}) = x_{(n/2)}$ or the right-median $\texttt{med}_r(\mathbf{x}) = x_{(n/2+1)}$, obtained by placing $n/2$ phantoms at $-\infty$ and $n/2-1$ at ∞, or vice versa.

While generalized medians can estimate certain parameters very well (e.g., sample median is often a good estimator of the population median), it is unclear how well they can estimate the population mean of the underlying distribution. In this paper we focus on studying and designing truthful estimators for estimating the population mean. We measure the loss by the *mean squared error*.

More formally, let us denote the mean of a distribution D by $\mu(D)$. Given n i.i.d. samples drawn from D, the mean squared error (MSE) of estimator g for estimating $\mu(D)$ is given by

$$\texttt{MSE}(g, D) = \mathbb{E}_{x_1, \ldots, x_n \sim D}[(g(x_1, \ldots, x_n) - \mu(D))^2].$$

For many distributions, the best estimator for the population mean in terms of MSE is the sample mean, but it is not truthful. Our goal is to pinpoint the optimal truthful estimators.

[2] This leads to Pareto efficiency in the social choice setting of Moulin (1980), but is an equally intuitive axiom for estimators.

3. Symmetric Distributions

In this section, we study the case of symmetric distributions. This is an interesting case because for symmetric distributions, the population mean coincides with the population median, making the sample median an attractive choice. Surprisingly (to us), our result in this section (Proposition 3.4) shows that the median is, in fact, *not* an optimal truthful estimator.

Definition 3.1. *A distribution D with PDF $f : \mathbb{R} \to \mathbb{R}$ is called symmetric if there exists a point $\mu \in \mathbb{R}$ such that $f(\mu - \delta) = f(\mu + \delta)$ for all $\delta \in \mathbb{R}$.*

Note that the point of symmetry μ is also the population mean and the population median of D. This enforces a simple structure on the performance of order statistics: order statistics that are closer to the median *dominate* those that are farther.

Definition 3.2. *Estimator g_1 is said to be dominated by estimator g_2 for a family of distributions $\mathcal{D}$ if $\texttt{MSE}(g_1, D) \geq \texttt{MSE}(g_2, D)$ for all $D \in \mathcal{D}$.*

Domination is an extremely strong comparison between two estimators. Note that not every pair of estimators are comparable in terms of domination (see the discussion after Proposition 3.4)—but order statistics are comparable, as the following lemma shows. Its proof is in the appendix.

Lemma 3.3. *For the family $\mathcal{D}^{sym}$ of symmetric distributions and for $t \leq (n-1)/2$, the t^{th} order statistic is dominated by the $(t+1)^{th}$ order statistic, and the $(n-t+1)^{th}$ order statistic is dominated by the $(n-t)^{th}$ order statistic. Consequently, for odd n the median and for even n the left and the right medians dominate all order statistics for $\mathcal{D}^{sym}$.*

In the proof of the lemma, note that the median achieves a *strictly lower* MSE than all other order statistics if the PDF of the underlying symmetric distribution is positive everywhere. Lemma 3.3 establishes that the optimal placement of phantoms in $\{-\infty, \infty\}$ is given by the median. Can we, however, outperform the median by placing phantoms on (finite) real values? At first glance, this seems impossible. Recall that we need to fix our estimator—and thus the locations of the phantoms—before we see the samples. In contrast to the original social choice setting (Moulin, 1980), where the phantoms have an intrinsic meaning (the n inputs represent individual preferences and the $n-1$ phantoms represent societal preferences), in our setting the phantoms are purely a tool for better estimating the population mean.[3] So the key question is: *where should we place the phantoms in the absence of any information about the distribution?* We show that for an even n, placing a single phantom anywhere in $\mathbb{R}$ and $n/2 - 1$ phantoms at $-\infty$ and ∞ each *dominates* the (left/right) median, and thus, by Lemma 3.3, every order statistic; the proof is given in the appendix.

Proposition 3.4. *Let $\mathcal{D}^{sym}$ denote the family of symmetric distributions and $n \in \mathbb{N}$ be even. For $\mathbf{x} = (x_1, \ldots, x_n)$ and $\alpha \in \mathbb{R}$, define $g_\alpha(\mathbf{x}) = \texttt{med}(x_1, \ldots, x_n, \alpha)$. Then, for every $\alpha \in \mathbb{R}$, g_α dominates all the order statistics for $\mathcal{D}^{sym}$.*

Once again, in the proof of the proposition, we have a strict inequality in the form $\texttt{MSE}(g_\alpha, D) < \texttt{MSE}(\texttt{med}_\ell, D) = \texttt{MSE}(\texttt{med}_r, D)$ if the PDF of D is positive everywhere.

Lemma 3.3 and Proposition 3.4 paint a rather clear picture for the case of symmetric distributions: The median dominates all the order statistics (generalized medians with no real-valued phantoms), and for the case of even n, every generalized median with a single real-valued phantom (and an equal number of phantoms on $-\infty$ and ∞) dominates the median. As an example, Figure 3(a) (in the appendix) shows the improvement of g_α over the median as a function of α, when the underlying distribution is a standard Gaussian. Figure 3(b) (in the appendix) shows the maximum improvement as a function of n.

Ideally we would like to extend the analysis to generalized medians with multiple real-valued phantoms. But it turns out that these rules are incomparable to the median in the strong sense of domination. For example, consider the generalized median of n samples in which all the $n-1$ phantoms are placed at 0. It has a strictly lower MSE than the median when the true mean happens to be at 0, but a strictly higher MSE if the true mean happens to be sufficiently far from 0.

Therefore, we adopt a weaker yet well-established notion of comparison between estimators, which relies on their *worst-case* error (Perron & Marchand, 2002).

Definition 3.5. *The maximum mean squared error of an estimator g for a family of distributions $\mathcal{D}$ is defined as $\texttt{MMSE}(g, \mathcal{D}) = \sup_{D \in \mathcal{D}} \texttt{MSE}(g, D)$. An estimator is called a minimax estimator for $\mathcal{D}$ if it minimizes the MMSE on $\mathcal{D}$ among all estimators. We say that an estimator is a minimax truthful estimator for $\mathcal{D}$ if it is truthful, and it minimizes the MMSE on $\mathcal{D}$ among all truthful estimators.*

One caveat is that taking the worst-case over the family of all symmetric distributions results in an infinite error for every estimator. Hence, we take the worst-case over subfamilies created by fixing a symmetric distribution and translating it along the real line. Formally, for a symmetric distribution $D \in \mathcal{D}^{\text{sym}}$, define the family of distributions

$$\mathcal{T}_D = \{H \in \mathcal{D}^{\text{sym}} \mid \exists \theta \in \mathbb{R} \text{ s.t. } \forall x \in \mathbb{R}, \\ F_H(x - \theta) = F_D(x - \mu(D))\},$$

[3] In situations such as the thermal comfort voting example from the introduction, the phantoms can represent societal preferences about energy consumption, etc. However, the goal of this paper is to study whether they can be used purely to achieve better statistical guarantees.

where F_D and F_H denote the CDFs of D and H, respectively. Note that θ is the mean of H, and is sufficient to identify H within $\mathcal{T}_D$.

Fix an arbitrary $D \in \mathcal{D}^{\text{sym}}$. While every order statistic has a constant MSE for all distributions $H \in \mathcal{T}_D$, the MSE of a generalized median that has finite phantoms varies with the mean $\mu(H)$. Let g be a generalized median of n samples. Let n_- and n_+ denote the number of its phantoms in $\mathbb{R} \cup \{-\infty\}$ and in $\mathbb{R} \cup \{+\infty\}$, respectively. Let $a = n - n_-$ and $b = n_+ + 1$. Note that the MSE of g converges to the MSE of the a^{th} and b^{th} order statistics when $\mu(H) \to \infty$ and $\mu(H) \to -\infty$, respectively. Thus, its worst-case MSE is at least as much as the MSE of the respective order statistic. Comparing the order statistics using Lemma 3.3 yields the next result.

Proposition 3.6. *For every symmetric distribution $D \in \mathcal{D}^{sym}$, the following estimators are minimax truthful estimators for the family of distributions $\mathcal{T}_D$:*

- *the median (if n is odd), and*
- *the left median, the right median, and all generalized medians with one phantom in $\mathbb{R}$ and $n/2-1$ phantoms on ∞ and $-\infty$ each (if n is even).*

As with Lemma 3.3 and Proposition 3.4, the domination of the estimators in Proposition 3.6 becomes strict if the PDF of the underlying distribution D is positive everywhere.

4. Distributions with Bounded Support

In the previous section we focused on the case of symmetric distributions, which help the sample median achieve a low MSE by making it unbiased. Yet, Theorem 3.4 showed that generalized medians of the form $g(\mathbf{x}) = \texttt{med}(\mathbf{x}, \alpha)$ dominate the median. At the same time, using a single real-valued phantom α among n samples limits the advantage when n is large. In this section, we study a setting with possibly asymmetric distributions, and show that there exist generalized medians that significantly outperform the median, even for large n.

Specifically, we focus on distributions with a known bounded support $[m, M]$ where $m, M \in \mathbb{R}$ (that is, distributions D such that $\Pr_{X \sim D}(X \in [m, M]) = 1$), and compare estimators based on their worst-case mean squared error (MMSE) *in the limit as the number of samples n goes to infinity.* We can immediately make a few simplifications to our setting. First, as the mean squared error is proportional to $(M - m)^2$, we can assume the support is $[0, 1]$ without loss of generality. Second, for a generalized median rule we can move all its phantoms in $[-\infty, 0)$ to 0 and all its phantoms in $(1, \infty]$ to 1 without changing its output. Thus, without loss of generality we assume that all the phantoms lie in $[0, 1]$.

A known support also introduces a subtle challenge. It limits the flexibility of the individuals to lie as they can no longer report a false value outside the support. This might allow additional mechanisms to become truthful, preventing us from directly using Moulin's characterization (Lemma 2.2). However, it can be easily shown that the characterization also holds in this case. The proof is essentially identical to the proof by Moulin, but requires a few minor modifications.

Lemma 4.1. *Suppose that the individuals have private values in a known interval $[m, M]$, and cannot report a false value outside this interval. Then, an estimator is truthful, anonymous, and range-respecting if and only if it is a generalized median.*

We are now ready to formalize our setting. Note that a generalized median is only defined for a fixed n. Hence, technically, for analyzing the limiting case of $n \to \infty$ we need to use a sequence of generalized medians $\{g_n\}_{n \geq 1}$. To prevent the rule in the sequence from changing dramatically with each n, we restrict the sequence to be *congruent*.

Definition 4.2. *Let g be a generalized median of n samples in $[0, 1]$, and let P denote the set of its phantoms. The empirical cumulative distribution function (ECDF) of g is the (right-continuous) function $G : [0, 1] \to [0, 1]$ such that $G(x) = |\{\alpha \in P | \alpha \leq x\}|/|P|$ for all $x \in [0, 1]$. We say that the sequence of generalized medians $\{g_n\}_{n \geq 1}$ with ECDFs $\{G_n\}_{n \geq 1}$ is congruent if $\lim_{n \to \infty} G_n = G$ (pointwise) for a function G, and in this case, we call G the limiting ECDF of $\{g_n\}_{n \geq 1}$.*

Hereinafter, we assume a sequence of generalized medians to be congruent unless specified otherwise. The next definition describes an ECDF that will be of special interest.

Definition 4.3. *We say that a sequence of generalized medians has uniform phantoms in the limit if its limiting ECDF G satisfies $G(x) = x$ for all $x \in [0, 1]$.*

Given a family of distributions $\mathcal{D}$ and a sequence of generalized medians $\{g_n\}_{n \geq 1}$, let us define the *maximum limiting mean squared error* as

$$\texttt{MLMSE}(\{g_n\}_{n \geq 1}, \mathcal{D}) = \sup_{D \in \mathcal{D}} \lim_{n \to \infty} \texttt{MSE}(g_n, D).$$

The definition is subject to the existence of the limit; our analysis ensures the existance under a mild assumption.

Our main result in this section identifies sequences of generalized medians that minimize MLMSE over distributions with bounded support. First, let us develop some intuition with an illustrative example.

Example 4.4. Suppose our goal is to minimize MLMSE over the family of distributions $\mathcal{D}_{\{0,1\}}$ with support $\{0, 1\}$ (instead of $[0, 1]$). Take a sequence of generalized medians

$\{g_n\}_{n\geq 1}$ with limiting ECDF G. Let us analyze its limiting mean squared error on the distribution D_p such that $\Pr_{X\sim D_p}[X=0] = 1-p$ and $\Pr_{X\sim D_p}[X=1] = p$.

Let us take n i.i.d. samples from D_p. As $n \to \infty$, the fractions of 0's and 1's will converge to $1-p$ and p, respectively, due to the law of large numbers. Thus, in the limit the output of g_n will belong to $G^{-1}(p)$ with probability 1.[4] Thus, achieving zero limiting MSE on D_p requires $p \in G^{-1}(p)$, i.e., $G(p) = p$. Applying this argument to every $D_p \in \mathcal{D}_{\{0,1\}}$, we get that $\texttt{MLMSE}(\{g_n\}_{n\geq 1}, \mathcal{D}_{\{0,1\}}) = 0$ if and only if $G(x) = x$ for all $x \in [0,1]$, that is, having uniform phantoms in the limit is the only way to achieve consistency[5] for every distribution in $\mathcal{D}_{\{0,1\}}$.

In contrast, let us consider the median. More technically, let g_n to be the median if n is odd, and the left (or the right) median if n is even. The limiting ECDF G satisfies $G(x) = 1/2$ for $x < 1$, and $G(1) = 1$. Clearly, the median does not have uniform phantoms in the limit, and it is easy to check that the MLMSE of the median for $\mathcal{D}_{\{0,1\}}$ is $1/4$. The worst-case distribution is $D_{1/2}$, for which the median always belongs to $\{0,1\}$, whereas the mean is $1/2$.

We now turn to our main result. Let $\mathcal{D}^{\text{inc}}_{[0,1]}$ denote the set of distributions with support $[0,1]$ and a strictly increasing CDF, or, equivalently, strictly positive density. We restrict the CDF to be strictly increasing for technical reasons, but note that any distribution can be approximated to an arbitrarily high precision by distributions with strictly increasing CDF. It turns out that the median still is not optimal in terms of MLMSE, while having uniform phantoms in the limit is. However, there are two key differences from the case of $\mathcal{D}_{\{0,1\}}$: (i) the optimal MLMSE is now positive, that is, it is impossible to be consistent for every distribution in $\mathcal{D}^{\text{inc}}_{[0,1]}$ subject to truthfulness, and (ii) it is possible to achieve the optimal MLMSE without having uniform phantoms in the limit. The theorem gives a full characterization of the optimal limiting ECDFs.

Theorem 4.5. *For a congruent sequence of generalized medians $\{g_n\}_{n\geq 1}$ with limiting ECDF G, we have $\texttt{MLMSE}\big(\{g_n\}_{n\geq 1}, \mathcal{D}^{\text{inc}}_{[0,1]}\big) = (R(G))^2$, where*

$$R(G) = \sup_{x\in[0,1]} \max\Big(x\cdot(1-G(x)), (1-x)\cdot G(x)\Big).$$

Further, we have that $R(G) \geq 1/4$, implying that the optimal MLMSE is $1/16$.

Let us first visualize this result. The gray region highlighted in Figure 1 shows the region in which a limiting ECDF G must lie for $R(G) = 1/4$. Theorem 4.5 shows that this is the necessary and sufficient condition to achieve the optimal MLMSE of $1/16$ on $\mathcal{D}^{\text{inc}}_{[0,1]}$. While the limiting ECDF of the uniform phantoms rule (the solid blue line) lies in this region, the limiting ECDF of the median (the dashed red line) lies in the region $R(G) = 1/2$, thus resulting in a worse MLMSE of $1/4$.

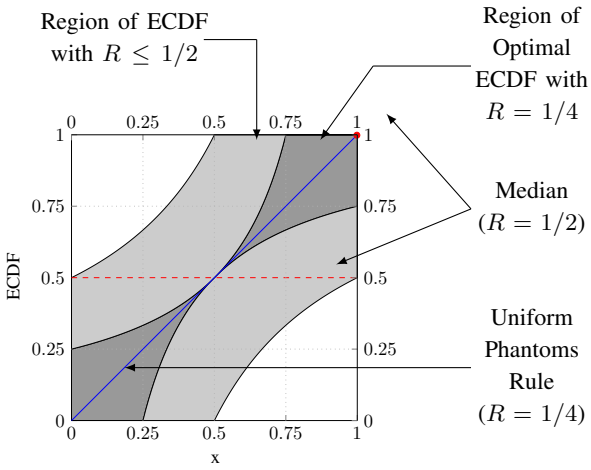

Figure 1. Region of optimal ECDF ($R = 1/4$), the ECDF of the uniform phantoms rule (which falls in $R = 1/4$, and the ECDF of the median (which falls under $R \leq 1/2$, but does not entirely fall under $R = 1/4$).

Proof of Theorem 4.5. The proof proceeds as follows. First, we show that the limit of $n \to \infty$ in the definition of MLMSE always exists in our case, and compute it analytically. We then use this to identify the worst-case distributions (in the MLMSE definition) for a given sequence of generalized medians with limiting ECDF G, and show that the worst-case limiting mean squared error (MLMSE) is exactly $R^2(G)$.

Fix a congruent sequence of generalized medians $\{g_n\}_{n\geq 1}$ with limiting ECDF G, and a distribution $D \in \mathcal{D}^{\text{inc}}_{[0,1]}$. Let G_n denote the ECDF of the $n-1$ phantoms in g_n, and let F_n denote the empirical CDF of n samples drawn i.i.d. from D. The empirical CDF of the $2n-1$ points—consisting of n samples from D and $n-1$ phantoms—is given by $(n\cdot F_n + (n-1)\cdot G_n)/(2n-1)$, which converges to $H = (1/2)\cdot(F+G)$ (pointwise) as $n \to \infty$ due to the law of large numbers. Hence, the median of the $2n-1$ points, which is the output of the rule g_n, converges to the point $x^*(D) = H^{-1}(1/2)$ as $n \to \infty$.[6] We drop D from the notation when it is clear from the context. We have shown that as $n \to \infty$, the output of the rule converges to x^*. Thus, the limit of the mean squared error exists, and is

[4] Note that, in general, $G^{-1}(\cdot)$ could be a (non-singleton) set.

[5] Consistency refers to the property of achieving zero error as $n \to \infty$ (Vapnik, 1998).

[6] The convergence to the unique point $x^*(D)$ is due to the fact that the strictly increasing F (and a weakly increasing G) result in a strictly increasing H. By H^{-1}, we mean the left-continuous inverse given by $H^{-1}(p) = \inf_{x\in[0,1]} H(x) \geq p$ for $p \in [0,1]$.

equal to $(x^* - \mu)^2$, where μ is the mean of distribution D.

We are now ready to identify the worst-case distributions in the MLMSE definition. First, let us introduce the left-continuous versions of the functions F and G. Define $F^L(x) = \lim_{y \to x^-} F(y)$ and $G^L(x) = \lim_{y \to x^-} G(y)$ for all $x \in [0, 1]$.

Next, we show the upper bound: $(x^* - \mu)^2 \le R^2(G)$, i.e., $|x^* - \mu| \le R(G)$. To achieve this, we find an upper and a lower bound on μ in terms of x^*. Let X be a random variable with distribution D.

1. As $\Pr_{X \sim D}[X \ge x^*] = 1 - F^L(x^*)$, we have $\mathbb{E}[X] = \mu \ge (1 - F^L(x^*)) \cdot x^*$.
2. As $\Pr_{X \sim D}[X \le x^*] = F(x^*)$, we have $\mathbb{E}[X] = \mu \le x^* \cdot F(x^*) + 1 \cdot (1 - F(x^*))$.

Together, these bounds imply

$$\begin{aligned}
&|x^* - \mu| \\
&\le \max\left(x^* - (1 - F^L(x^*))\, x^*, x^* \cdot F(x^*) + 1 - F(x^*) - x^*\right) \\
&= \max\left(x^* \cdot F^L(x^*), (1 - x^*) \cdot (1 - F(x^*))\right). \qquad (1)
\end{aligned}$$

Recall that $x^* = \inf\{x \in [0, 1] : ((F + G)/2)(x) \ge 1/2\}$. Due to the right-continuity of F and G, we can replace the $\inf$ by $\min$. Hence, we get

$$\frac{F(x^*) + G(x^*)}{2} \ge \frac{1}{2} \Rightarrow 1 - F(x^*) \le G(x^*).$$

On the other hand, we have that for all $x < x^*$, $(F(x) + G(x))/2 < 1/2$. Hence, taking the limit $x \to (x^*)^-$ from below, we get

$$\frac{F^L(x^*) + G^L(x^*)}{2} \le \frac{1}{2} \Rightarrow F^L(x^*) \le 1 - G^L(x^*).$$

Substituting these bounds into Equation (1), we get

$$|x^* - \mu| \le \max\left(x^* \cdot (1 - G^L(x^*)), (1 - x^*) \cdot G(x^*)\right).$$

Now, clearly $(1 - x^*) \cdot G(x^*) \le \sup_{x \in [0,1]} (1 - x) \cdot G(x) \le R(G)$. To show that $x^* \cdot (1 - G^L(x^*)) \le R(G)$, recall that $G^L(x^*) = \lim_{y \to (x^*)^-} G(y)$. Thus,

$$\begin{aligned}
x^* \cdot (1 - G^L(x^*)) &= \lim_{y \to (x^*)^-} y \cdot (1 - G(y)) \\
&\le \sup_{y \in [0,1]} y \cdot (1 - G(y)) \le R(G).
\end{aligned}$$

Thus, in both cases we have $|x^* - \mu| \le R(G)$, as required. This concludes the proof of the upper bound.

For the lower bound, we want to prove that $\texttt{MLMSE}(\{g_n\}_{n \ge 1}, \mathcal{D}^{\text{inc}}_{[0,1]}) \ge \sup_{x \in [0,1]} \max(x \cdot (1 - G(x)), (1 - x) \cdot G(x))$. We prove this inequality for each $x \in [0, 1]$. Fix an $x \in [0, 1]$.

Bound 1: $\texttt{MLMSE}(\{g_n\}_{n \ge 1}, \mathcal{D}^{\text{inc}}_{[0,1]}) \ge x \cdot (1 - G(x))$. We in fact prove the stronger lower bound of $x \cdot (1 - G^L(x))$. If $G^L(x) = 1$ or $x = 0$, the bound holds trivially. Thus, assume $G^L(x) < 1$ and $x > 0$. If $G^L(x) = 0$, then the lower bound we want to prove is x. In this case, without loss of generality we can assume $x = \sup\{y \in [0, 1] : G(y) = 0\}$. Note that it retains $G^L(x) = 0$, and the desired lower bound x does not decrease.

Now, in both cases we can assume that $G^L(x) < 1$ and $G(x + d) > 0$ for all $d > 0$. For $k \in \mathbb{N}$, choose

$$\delta_k = \min\left(G\left(x + 1/k\right), 1/k, (1 - G^L(x))/2\right).$$

Note that $\delta_k > 0$. Consider the distribution D_k with CDF F_k obtained as follows. (i) Place a probability mass of $t_0 = 1 - G^L(x) - \delta_k$ on the point 0. Note that our choice of δ_k ensures $t_0 > 0$. (ii) Distribute probability $\delta_k/2$ uniformly in the interval $(0, x)$. (iii) Place a probability mass of $t_x = G^L(x)$ on point x. (iv) Distribute probability $\delta_k/2$ uniformly in the interval $(x, 1)$.

Let us analyze $x^*(D_k)$. For any $y < x$, we have

$$\frac{F_k(y) + G(y)}{2} \le \frac{t_0 + \delta_k/2 + G^L(x)}{2} < \frac{1}{2}.$$

Hence, $x^*(D_k) \ge x$. Further,

$$\frac{F_k(x + 1/k) + G(x + 1/k)}{2} \ge \frac{1 - \delta_k/2 + G(x + 1/k)}{2} \ge \frac{1}{2},$$

where the last inequality follows because we chose $\delta_k \le G(x + 1/k)$. Hence, $x^*(D_k) \le x + 1/k$.

Since $x^*(D_k) \in [x, x + 1/k]$, we have $\lim_{k \to \infty} x^*(D_k) = x$. Further, $\lim_{k \to \infty} \mu(D_k) = x \cdot G^L(x)$. Hence,

$$\begin{aligned}
\texttt{MLMSE}(\{g_n\}_{n \ge 1}, \mathcal{D}^{\text{inc}}_{[0,1]}) &\ge \lim_{k \to \infty} (x^*(D_k) - \mu(D_k))^2 \\
&= x^2 \cdot (1 - G^L(x))^2.
\end{aligned}$$

Bound 2: $\texttt{MLMSE}(\{g_n\}_{n \ge 1}, \mathcal{D}^{\text{inc}}_{[0,1]}) \ge (1 - x) \cdot G(x)$. If $G(x) = 0$, the bound holds trivially. Assume $G(x) > 0$. If $G(x) = 1$, then the lower bound we want to prove is $1 - x$. In this case, we can replace x with $x = \min\{y \in [0, 1] : G(y) = 1\}$. As the $\min$ exists (G is right-continuous), this retains $G(x) = 1$ and makes the lower bound stronger.

Thus, in either case we can assume $G(x - d) < 1$ for all $d > 0$. Now, choose

$$\delta_k = \min\left(1 - G\left(x - 1/k\right), 1/k, G(x)/2\right).$$

Note that $\delta_k > 0$ because $G(x - 1/k) < 1$ and $G(x) > 0$. Consider the distribution D_k with CDF F_k obtained as follows. i) Distribute probability $\delta_k/2$ uniformly in the interval $(0, x)$. ii) Place a probability mass of $t_x = 1 - G(x)$

on the point x. iii) Distribute probability $\delta_k/2$ uniformly in the interval $(x, 1)$. iv) Place a probability mass of $t_1 = G(x) - \delta_k$ on point 1. Once again, $t_1 > 0$ due to our choice of δ_k.

Let us now analyze $x^*(D_k)$. Note that

$$\frac{F_k(x-1/k) + G(x-1/k)}{2} < \frac{\delta_k/2 + G(x-1/k)}{2} \leq \frac{1}{2},$$

where the last inequality holds because we chose $\delta_k \leq 1 - G(x-1/k)$. This implies $x^*(D_k) \geq x - 1/k$. Further,

$$\frac{F_k(x) + G(x)}{2} \geq \frac{\delta_k/2 + t_x + G(x)}{2} \geq \frac{1}{2},$$

implying that $x^*(D_k) \leq x$. Hence, once again $x^*(D_k) \in [x-1/k, x]$ implies $\lim_{k\to\infty} x^*(D_k) = x$. We also have $\lim_{k\to\infty} \mu(D_k) = x \cdot (1 - G(x)) + 1 \cdot G(x)$. Hence,

$$\begin{aligned}\texttt{MLMSE}(\{g_n\}_{n\geq 1}, \mathcal{D}^{\text{inc}}_{[0,1]}) &\geq \lim_{k\to\infty} (x^*(D_k) - \mu(D_k))^2 \\ &= (1-x)^2 \cdot (G(x))^2.\end{aligned}$$

This concludes the proof of $\texttt{MLMSE}(\{g_n\}_{n\geq 1}, \mathcal{D}^{\text{inc}}_{[0,1]}) = R^2(G)$. Finally, using $x = 1/2$ in the definition of $R(G)$, we get $R(G) \geq (1/2) \cdot \max(1 - G(1/2), G(1/2)) \geq 1/4$, which concludes the proof of the theorem. ■

We remark that, while the possible discontinuity of G (and thus of $(F+G)/2$) complicates the proof of Theorem 4.5, it is necessary in order for the analysis to be broad enough to incorporate the limiting ECDF of the median, which is indeed discontinuous.

The uniform phantoms rule suggests an intuitive way of placing the phantoms in the support, and although Theorem 4.5 establishes its superiority over the sample median in the worst-case over distributions, we expect it to outperform the sample median on many distributions of practical interest. For example, Figure 2 shows that for the truncated normal distribution with support $[0, 1]$, the uniform phantoms rule performs better for most combinations of values of the mean μ and the standard deviation σ. Interestingly, the relative performance is largely determined by σ, large values being preferable for the uniform phantoms rule. Thus, in practice knowledge about the standard deviation can help guide the placement of the phantoms.

5. Discussion

This paper studies a basic univariate estimation framework with mean squared error risk function, finds domination relationships among truthful estimators, and identifies the truthful estimators that are minimax optimal. While some of the simplifying assumptions made in the analysis are crucial, a few others can easily be dropped. For example,

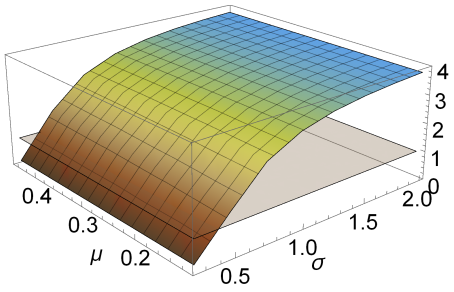

Figure 2. The ratio of the MSE of the sample median to the MSE of the uniform phantoms rule under the truncated normal distribution with mean μ, standard deviation σ, and support $[0, 1]$ as $n \to \infty$. The gray plane indicates a ratio of 1. We plot $\mu \in (0, 0.5)$ because the case of $\mu > 0.5$ is symmetric.

we assume estimators to be range-respecting, but it is easy to drop this assumption, as this only changes the characterization of truthful estimators (Theorem 2.2) from generalized medians with $n-1$ phantoms to those with $n+1$ phantoms. On the other hand, dropping anonymity can be tricky as it now allows exponentially many phantoms—one for each subset of indices (Moulin, 1980).

It would be natural to extend our work to more general estimation settings, such as estimation with multi-dimensional distributions, estimation of parameters other than the population mean, use of risk functions other than the mean squared error, etc. It would also be interesting to study whether a truthful estimator can make use of additional available information, e.g., the variance of the underlying distribution (which does not help place phantoms in our unbounded support case), or a prior distribution over the estimated parameter (which can guide the placement of phantoms even in our unbounded support case).

Finally, we point out a subtle relation between the truthful and non-truthful cases, in the context of minimax estimation and distributions with bounded range. Hodges and Lehmann (1950) showed that the minimax estimator (without the truthfulness requirement) for the population mean of a distribution with a known range $[0, 1]$, on samples $x_1, \ldots, x_n$, returns

$$\frac{\sum_i x_i}{n} \cdot \frac{\sqrt{n}}{\sqrt{n}+1} + \frac{1}{2} \cdot \frac{1}{\sqrt{n}+1}.$$

Interestingly, observe that this is equivalent to a "generalized mean" estimator that takes the mean of the given n samples along with $\sqrt{n}$ phantoms, all placed at $1/2$. This stands in contrast to the characterization of minimax truthful estimators given in Theorem 4.5, which requires the phantoms to be uniformly spaced throughout the support (in the limit). Nonetheless, the remarkable fact that phantoms, which arise in generalized medians purely from the truthfulness requirement, play a larger role in minimax estimation calls for a deeper investigation.

Acknowledgments

This work was supported in part by NSF grants CCF-1215883, CCF-1525932, and IIS-1350598; and by a Sloan Research Fellowship.

References

Cabrera, J., Maguluri, G., and Singh, K. An odd property of the sample median. *Statistics and Probability Letters*, 19(4):349–354, 1994.

Cai, Y., Daskalakis, C., and Papadimitriou, C. H. Optimum statistical estimation with strategic data sources. In *Proceedings of the 28th Conference on Computational Learning Theory (COLT)*, pp. 280–296, 2015.

Cummings, R., Ioannidis, S., and Ligett, K. Truthful linear regression. In *Proceedings of the 28th Conference on Computational Learning Theory (COLT)*, pp. 448–483, 2015.

Dalvi, N., Domingos, P., Mausam, Sanghai, S., and Verma, D. Adversarial classification. In *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 99–108, 2004.

Dekel, O., Fischer, F., and Procaccia, A. D. Incentive compatible regression learning. *Journal of Computer and System Sciences*, 76(8):759–777, 2010.

Fanger, P. O. *Thermal Comfort: Analysis and applications in environmental engineering*. McGraw-Hill, 1970.

Fisher, R. A. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(05):700–725, 1925.

Gibbard, A. Manipulation of voting schemes. *Econometrica*, 41:587–602, 1973.

Hardt, M., Megiddo, N., Papadimitriou, C. H., and Wootters, M. Strategic classification. In *Proceedings of the 7th Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 111–122, 2016.

Hodges Jr, J. L. and Lehmann, E. L. Some problems in minimax point estimation. *The Annals of Mathematical Statistics*, pp. 182–197, 1950.

Horel, T., Ioannidis, S., and Muthukrishnan, S. Budget feasible mechanisms for experimental design. In *Proceedings of the 11th Latin American Symposium Theoretical Informatics (LATIN)*, pp. 719–730, 2014.

Lehmann, E. L. and Casella, G. *Theory of Point Estimation*. Springer Verlag, 1998.

Meir, R., Procaccia, A. D., and Rosenschein, J. S. Algorithms for strategyproof classification. *Artificial Intelligence*, 186:123–156, 2012.

Moulin, H. On strategy-proofness and single-peakedness. *Public Choice*, 35:437–455, 1980.

Nisan, N., Roughgarden, T., Tardos, É., and Vazirani, V. (eds.). *Algorithmic Game Theory*. Cambridge University Press, 2007.

Perote, J. and Perote-Peña, J. Strategy-proof estimators for simple regression. *Mathematical Social Sciences*, 47: 153–176, 2004.

Perron, F. and Marchand, E. On the minimax estimator of a bounded normal mean. *Statistics and Probability Letters*, 58:327–333, 2002.

Vapnik, V. N. *Statistical learning theory*. Wiley New York, 1998.