Prior-free Data Acquisition for Accurate Statistical Estimation

YILING CHEN, Harvard University, USA
SHURAN ZHENG, Harvard University, USA

We study a data analyst's problem of acquiring data from self-interested individuals to obtain an accurate estimation of some statistic of a population, subject to an expected budget constraint. Each data holder incurs a cost, which is unknown to the data analyst, to acquire and report his data. The cost can be arbitrarily correlated with the data. The data analyst has an expected budget that she can use to incentivize individuals to provide their data. The goal is to design a joint acquisition-estimation mechanism to optimize the performance of the produced estimator, without any prior information on the underlying distribution of cost and data. We investigate two types of estimations: unbiased point estimation and confidence interval estimation.

Unbiased estimators: We design a truthful, individually rational, online mechanism to acquire data from individuals and output an unbiased estimator of the population mean when the data analyst has no prior information on the cost-data distribution and individuals arrive in a random order. The performance of this mechanism matches that of the optimal mechanism, which knows the true cost distribution, within a constant factor. The performance of an estimator is evaluated by its variance under the worst-case cost-data correlation.

Confidence intervals: We characterize an approximately optimal (within a factor 2) mechanism for obtaining a confidence interval of the population mean when the data analyst knows the true cost distribution at the beginning. This mechanism is efficiently computable. We then design a truthful, individually rational, online algorithm that is only worse than the approximately optimal mechanism by a constant factor. The performance of an estimator is evaluated by its expected length under the worst-case cost-data correlation.

CCS Concepts: • **Theory of computation → Algorithmic game theory and mechanism design**; Online algorithms; • **Mathematics of computing →** *Probabilistic inference problems.*

Additional Key Words and Phrases: data acquisition; prior-free; statistical estimation

ACM Reference Format:
Yiling Chen and Shuran Zheng. 2019. Prior-free Data Acquisition for Accurate Statistical Estimation. In *ACM EC '19: ACM Conference on Economics and Computation (EC '19), June 24–28, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3328526.3329564

1 INTRODUCTION

We study a data analyst's problem of estimating a population statistic (e.g. mean workout time in November) when data need to be acquired from self-interested data holders and the analyst has an expected budget constraint. Each data holder has a heterogeneous private cost to acquire and report his data (e.g. record duration of each workout in a month and report the total) and needs to be compensated at least by this cost to reveal his data. Individuals cannot fabricate their data if they decide to reveal it. Moreover, the values of the data and the private costs can be arbitrarily

EC '19, June 24–28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6792-9/19/06...$15.00
https://doi.org/10.1145/3328526.3329564

correlated in the population (e.g. those who work out regularly may use some fitness tracker which automatically records workout durations) and the correlation is unknown to the analyst a priori. A naive way for the analyst to acquire data in this setting is to offer a fixed compensation for each individual's data. But unless the payment level is higher than everyone's cost, in which case the analyst may run out of budget quickly and only be able to obtain a small sample, the collected sample will bias toward a low-cost subpopulation. Thus, the problem is how to design a joint pricing-estimation mechanism to get accurate estimations when data holders are strategic.

The problem of purchasing data for unbiased estimation of population mean was first formulated by Roth and Schoenebeck [26] and then further studied by Chen et al. [4]. Both works however assume that the cost distribution is known to the analyst and aim at obtaining an optimal unbiased estimator with minimum worst-case variance for population mean, where the worst-case is over all data-cost distributions consistent with the known cost distribution, subject to an expected budget constraint. The mechanism proposed by Roth and Schoenebeck [26] achieves optimality approximately when the cost distribution has piece-wise differentiable PDF, while the mechanism proposed by Chen et al. [4] achieves the exact optimality when the cost distribution is regular. Chen et al. [4] also extend the result to linear regression. The high-level idea of both mechanisms is to acquire a data point with reported cost c_i with a positive probability $A(c_i)$ (and some payment that is greater than or equal to c_i), then remove the sampling bias by re-weighting each collected data by $1/A(c_i)$, and finally average the re-weighted data to obtain an unbiased estimation (the Horvitz-Thompson Estimator). The assumption that the cost distribution is known allows the analyst to turn the mechanism design problem into a constrained optimization problem for finding an optimal allocation rule $A(c_i)$.

Our paper makes two novel contributions, both of which do away from the main limiting assumptions of the prior works. First, we consider a data analyst with no prior information on the data holders' costs. We design an online mechanism that outputs an unbiased estimation of population mean, with the same goal as in the prior works: minimize the variance of the unbiased estimator subject to a budget constraint. Our only assumption is that data holders show up in a uniformly random order. Here the challenge is that, in order to price well, the analyst needs to learn the cost distribution, but the pricing decisions need to be made for every arriving data holder. Our second contribution is to consider the bias-variance trade-off of the estimator. The previous works only consider unbiased estimators and the goal is to minimize the variance of the estimator. In this work, we allow the estimator to be biased and try to minimize the length of the confidence interval around the population mean, given a budget constraint. This necessarily requires us to reason about bias-variance trade-off together with data pricing, an aspect that, to the best of our knowledge, has not been explored in the literature. We design mechanisms for both the scenario where the analyst knows the cost distribution and the scenario where there is no prior information on costs.

1.1 Summary of Our Results and Techniques

Our work mainly addresses two questions:

(1) If the data analyst does not have any prior information on data holders' private costs (as well as their private data), is it possible to design an online data acquisition mechanism for unbiased estimation of population mean that is competitive with the optimal mechanism that knows the cost distribution a priori?
(2) Can we design an optimal joint acquisition-estimation mechanism for estimating confidence intervals of population mean, when cost distribution is known? Optimality here means minimum length of the confidence interval. When cost distribution is unknown, can we

> design an online joint acquisition-estimation mechanism for confidence intervals that is competitive with the optimal mechanism that knows the cost distribution a priori?

For the first question, we design an online mechanism that is only worse than the optimal mechanism by a constant factor. The only substantial assumption we make in our setting is that the data holders come in random order, so if there are n data holders in total, the cost-data distribution at each round is the discrete uniform distribution over the set of cost-data pairs of these n data holders. Our mechanism satisfies the budget constraint in expectation, with the guarantee that the data holders will always be willing to participate and truthfully report their costs.

Theorem 1.1 (Informal). *For the problem of purchasing data to get an unbiased estimator of population mean, assuming that the data holders come in random order, our online mechanism satisfies the following properties: (1) it is truthful and individually rational, (2) it satisfies the expected budget constraint, and (3) for any cost distribution, the variance of the produced unbiased estimator approaches that of the benchmark within a constant factor, where the benchmark is the optimal mechanism that knows the true cost distribution a priori.*

For the second question, we extend our mechanism to output a confidence interval (using sample mean and sample variance). The mechanism may introduce some bias to mean estimation in exchange for a lower variance, so that the length of the confidence interval is approximately optimized. We provide the characterization of the approximately optimal confidence interval mechanism when the cost distribution is known. This characterization allows us to efficiently compute the mechanism. We then design an online mechanism that matches the performance of the optimal mechanism that knows the cost distribution within a constant factor.

Theorem 1.2 (Informal). *For the problem of purchasing data to obtain a confidence interval, the approximately optimal mechanism that knows the cost distribution can be computed in polynomial time.*

This approximately optimal mechanism with known costs is constructed by analyzing the bias and variance trade-off for estimators for the mean. At any given bias level, by producing an estimator that has the lowest variance (for that bias level), we can construct a confidence interval using this biased mean estimation. We hence can design a mechanism to optimize for the length of the confidence interval. Since the optimal mechanism is difficult to compute, we approximate it to gain computational efficiency.

Theorem 1.3 (Informal). *For the problem of purchasing data to obtain a confidence interval, assuming that the data holders come in random order, our online mechanism has the following properties: (1) it is truthful and individually rational, (2) it satisfies the expected budget constraint, and (3) for any cost distribution, the performance of the produced confidence interval approaches that of the benchmark within a constant factor, where the benchmark is the optimal mechanism that knows the true cost distribution a priori.*

Our online mechanisms for both unbiased mean estimation and confidence interval are designed using approximately optimal mechanisms with known costs as building blocks. At any round i, the reported costs in previous rounds gives us an empirical cost distribution. We then apply the optimal mechanism for this cost distribution for data holder i. Each round's mechanism has a fraction of the total expected budget. Our online mechanisms allocate more budget for early rounds in a way so that the performance of the final produced estimator is only worse than the benchmark by a constant factor.

Most of the proofs of our results are omitted but can be found in the full version of the paper[1].

[1] https://arxiv.org/abs/1811.12655

1.2 Other Related Work

There is a growing interest in understanding statistical estimation and learning in environments with strategic agents. The works can be put in a few categories depending on the sources and types of strategic considerations.

In this work, as well as [26] and [4], agents do not derive utility or disutility from the estimation outcome, cannot fabricate their data, and have a cost for revealing their data. The mechanism uses payment to incentivize data revelation. [1] is similar on these fronts, but the work considers general supervised learning. They do not seek to achieve a notion of optimality. Instead, they take a learning-theoretic approach and design mechanisms to obtain learning guarantees (risk bounds). Cai et al. [3] focused on incentivizing individuals to exert effort to obtain high-quality data for the purpose of linear regression.

Another line of research examines data acquisition using differential privacy [6, 9–11, 23]. Agents care about their privacy and hence may be reluctant to reveal their data. The mechanism designer uses payments to balance the trade-off between privacy and accuracy. In this work, we implicitly assume that data holders to not have a privacy cost and hence they don't worry about potential leaking of their data by reporting their cost. In Section 6, we discuss the complication when data holders care about their privacy and their data and costs are correlated.

A third line of research studies settings where data holders may strategically misreport their data, there is no ground truth to verify the acquired data, and the analyst would like to design payment mechanisms to incentivize truthful data reporting for the purpose of regression or other analyses [14–16]. Because of the lack of verification, this line of work is closely related to the literature on peer prediction [20, 27].

In a fourth line of research, individuals' utilities directly depend on the inference or learning outcome (e.g. they want a regression line to be as close to their own data point as possible) and they can manipulate their reported data to influence the outcome. In these works, there often is no cost for reporting one's data and the data analyst doesn't use monetary payments. These works attempt to design or identify mechanisms (inference or learning processes) that are robust to potential data manipulations [5, 7, 8, 13, 18, 19, 24, 25].

1.3 Organization of the Paper

In the rest of the paper, we first formulate and characterize the optimal (or approximately optimal) mechanisms for the unbiased mean estimation and the confidence interval estimation when *the cost distribution is known* (Section 3 and Section 4 respectively). These mechanisms serve both as building blocks for developing our online mechanisms when the cost distribution is unknown and as benchmarks to which our online mechanisms are compared. Section 5 turns to the setting when the cost distribution is unknown. We develop online mechanisms with proven performance guarantees for both the unbiased mean estimation and the confidence interval estimation. We conclude with discussions and future directions in Section 6.

2 MODEL

Consider a data analyst who conducts a survey to estimate some statistic of a population of n people. In this work we focus on estimating the mean of some parameter of interest (e.g. alcohol consumption or BMI of an individual), denoted by z, and the confidence interval of the mean. Each individual incurs a cost c_i, unknown to the data analyst, to acquire and report his data z_i. The cost and data pair can be correlated (e.g. those who consume more alcohol may have a higher cost recording their consumption), and follows an unknown distribution $\mathcal{D}$ supported on $(C, \mathcal{Z})$. We assume that the cost is bounded by $\overline{C}$, i.e., $C \subseteq [0, \overline{C}]$. The parameter z is also bounded, and, without loss of generality, we assume z is between 0 and 1, i.e., $\mathcal{Z} \subseteq [0, 1]$. The data analyst has a budget $B = n\overline{B}$ that she can use to purchase data from the data holders.

We study an online setting where data holders arrive one by one to the survey, and no prior information on the distribution $\mathcal{D}$ (including the marginal distribution of the cost) is available before the survey. The analyst can gradually learn the distribution as data holders report their data. We make the following assumptions about the data sequence: (1) each individual only appears once, and (2) the data holders arrive in *a random order*, i.e., each permutation of the n people is equally likely. We use $(\mathbf{c}, \mathbf{z}) = (c_1, z_1), \ldots, (c_n, z_n)$ to denote a random sequence of costs and data points, and $\{(c_1, z_1), \ldots, (c_n, z_n)\}$ to represent a set of people's cost and data without the consideration of order.

When data holder i arrives, the analyst asks the data holder to report his cost. We use $\widehat{c_i}$ to denote the reported cost of data holder i. Based on the reported cost, the analyst may offer a price to acquire the data z_i. Formally, the analyst uses a *survey mechanism*, $M = (A, P)$, which consists of an allocation rule $A : C \longrightarrow [0, 1]$ and a payment rule $P : C \longrightarrow \mathbb{R}$. With probability $A(\widehat{c_i})$, the analyst offers payment $P(\widehat{c_i})$ to purchase data z_i. If the data holder accepts this payment, he gives his data z_i to the analyst. We assume that data holders do not misreport their data z_i. This assumption holds in situations when data can be verified once collected (e.g. medical records). The data holder walks away without revealing his data if $P(\widehat{c_i}) < c_i$. With probability $1 - A(\widehat{c_i})$, the analyst does not attempt to acquire the data.[2]

The analyst can adaptively choose a survey mechanism for each arriving data holder. We use $\mathbf{M} = (\mathbf{A}, \mathbf{P}) = (A^1, P^1), \ldots, (A^n, P^n)$ to represent a sequence of survey mechanisms. At round i, the analyst chooses an allocation rule A^i and a payment rule P^i based on all observed information before round i, denoted by $\mathcal{H}_{i-1}$. $\mathcal{H}_{i-1}$ includes the reported costs of the previous $i - 1$ data holders and data points that have been acquired. The survey mechanism (A^i, P^i) applies to the i-th arriving data holder. At the end of round n, the data analyst outputs an estimator $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ based on all observed information $\mathcal{H}_n$.

We want to design survey mechanisms that have the following incentive and budget properties:

Individual rationality: The utility of each data holder is always non-negative, i.e., $P^i(\widehat{c_i}) \geq \widehat{c_i}$ for all i and $\widehat{c_i}$.

Truthfulness in expectation: A data holder maximizes his expected utility by reporting his cost truthfully, i.e., $A^i(c_i)(P^i(c_i) - c_i) \geq A^i(\widehat{c_i})(P^i(\widehat{c_i}) - c_i)$ for all i and $\widehat{c_i} \neq c_i$.

Expected budget feasibility: $\mathbb{E}\left[\sum_{i=1}^{n} A^i(c_i) \cdot P^i(c_i)\right] \leq B = n \cdot \overline{B}$, where the expectation is taken over the random arriving order of the data holders and the internal randomness of the mechanism.

In this work we mainly investigate two types of estimation tasks: (1) get an unbiased estimator of the population mean, with the goal that the variance of the estimator is minimized; (2) find a confidence interval of the population mean, with the goal that the length of the confidence interval is minimized. As an estimator uses data obtained via survey mechanisms $\mathbf{M}$, it necessarily depends on $\mathbf{M}$. We now formally define unbiased estimator and confidence interval of population mean in our setting. The randomness of an estimator $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ comes in two parts: (1) the external randomness, which is the random order of $(\mathbf{c}, \mathbf{z})$, and (2) the internal randomness of the mechanisms $\mathbf{M}$. Our definitions require the estimators to be unbiased or a valid confidence interval for any realization of the external randomness.

[2] We describe survey mechanisms as direct-revelation mechanisms, where date holders report their costs. Any survey mechanism can be implemented as a posted-price mechanism, where a menu of (price, probability) pairs are presented and each data holder chooses one from the menu [4, 26].

Definition 2.1 (Unbiased estimator of population mean). *An estimator* $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ *is an unbiased estimator of the population mean* $\mathbb{E}[z] = \frac{1}{n}\sum_{i=1}^{n} z_i$ *if for any fixed sequence* $(\tilde{\mathbf{c}}, \tilde{\mathbf{z}})$,

$$\mathbb{E}[S(\mathbf{M}, (\tilde{\mathbf{c}}, \tilde{\mathbf{z}})] = \mathbb{E}[z],$$

where the expectation in $\mathbb{E}[S(\mathbf{M}, (\tilde{\mathbf{c}}, \tilde{\mathbf{z}}))]$ *is taken over the internal randomness of the mechanisms* $\mathbf{M}$.

Definition 2.2 (Confidence interval of population mean). *An estimator* $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ *is a confidence interval for the population mean* $\mathbb{E}[z] = \frac{1}{n}\sum_{i=1}^{n} z_i$ *with confidence level* γ *if it is an interval and for any fixed sequence* $(\tilde{\mathbf{c}}, \tilde{\mathbf{z}})$,

$$\Pr\left(\mathbb{E}[z] \in S(\mathbf{M}, (\tilde{\mathbf{c}}, \tilde{\mathbf{z}}))\right) \geq \gamma,$$

where the randomness is due to the internal randomness of the mechanisms $\mathbf{M}$.

Our goal is to design joint survey and estimation mechanisms, $(\mathbf{M}, S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$, such that the estimator $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ has good statistical performance on the population. For unbiased estimators, we prefer estimators with smaller variance. For confidence intervals, we prefer ones with smaller length. However, the performance of a mechanism on a population depends on the correlation between the population's cost and data, i.e. the distribution $\mathcal{D}$.[3] We hence take a worst-case analysis approach: measure the performance of a mechanism under worst-case cost-data correlation.

Definition 2.3 (Worst-case variance). *Given that the set of data holders' costs is* $C = \{c_1, \ldots, c_n\}$, *the worst-case variance of a point estimator* $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ *is defined as*

$$Var^*(S) = \max_{\mathcal{D} \text{ consistent with } C} Var_{\mathcal{D}}(S(\mathbf{M}, (\mathbf{c}, \mathbf{z})))$$

where the maximum is taken over all distributions $\mathcal{D}$ *consistent with the set of costs* C. *The randomness is due to the random order of* $(c_1, z_1), \ldots, (c_n, z_n)$ *and the internal randomness of the mechanism* $\mathbf{M}$.

Definition 2.4 (Worst-case expected length). *Given that the set of data holders' costs is* $C = \{c_1, \ldots, c_n\}$, *the worst-case expected length of a confidence interval* $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ *is defined as*

$$L^*(S) = \max_{\mathcal{D} \text{ consistent with } C} \mathbb{E}(|S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))|)$$

where $|S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))|$ *represents the length of the confidence interval. The maximum is taken over all distributions* $\mathcal{D}$ *consistent with the set of costs* C. *The randomness is due to the random order of* $(c_1, z_1), \ldots, (c_n, z_n)$ *and the internal randomness of the mechanisms* $\mathbf{M}$.

Roth and Schoenebeck [26] and Chen et al. [4] have also considered the design of joint survey and estimation mechanism for statistical estimation. The main differences between their model and our model are: (1) they assume the marginal cost distribution is known to the data analyst, while our data analyst doesn't have such information, (2) they have the same survey mechanism for all individuals, while we consider an online setting where the analyst can adaptively change the survey mechanism, and (3) they only consider the estimation of mean, while we also investigate the estimation of confidence intervals.

[3] For example, consider a mechanism that purchases each agent's data z_i with a constant probability $p = B/\sum_{i=1}^{n} c_i$ and payment c_i, then outputs $1/(pn)$ times the sum of all purchased data as an unbiased estimation of population mean. When z is always equal to 0, the variance will be zero; when z is always equal to 1, the variance will be $(1/p - 1)/n$.

3 PRELIMINARIES

In this section, we first show that we can easily extend known results on one-shot truthful mechanisms to achieve truthfulness and individual rationality for a sequence of survey mechanisms **M**. Then, we introduce the formulation proposed by Chen et al. [4] for obtaining the optimal unbiased estimator of population mean *when the cost distribution is known to the analyst.* Later in Section 5 we will use this known cost case as our benchmark for evaluating the performance of our optimal unbiased estimator when the cost distribution is unknown. While this optimal unbiased estimator has been studied by Chen et al. [4], their optimal mechanism requires the cost distribution to be regular. This is often not satisfied when we consider the costs of a finite set of data holders. We develop the characterization of the optimal unbiased estimator for arbitrary discrete cost distribution without any regularity assumption. We show that the optimal purchasing rule of cost-c data is decided by a quantity which we define as *regularized virtual costs* of the data.

3.1 Truthful and Individually Rational Survey Mechanisms

Since each data holder appears only once in our setting, requiring a sequence of survey mechanisms to be truthful and individually rational is equivalent to requiring that each (A^i, P^i) is truthful and individually rational, which can be achieved by a straight-forward extension of known results for truthful mechanisms.

The well-known Myerson's lemma states that monotonicity is the necessary and sufficient condition for an allocation rule to be truthful with some payment rule.

Lemma 3.1 (Myerson and Satterthwaite [22]). *An allocation rule $A(c)$ is the allocation rule of some truthful survey mechanism $(A(c), P(c))$ if and only if $A(c)$ is monotone non-increasing in c.*

The following lemmas from [4] are analogies of the original Myerson's Lemma, which are tailored for discrete cost distributions. Firstly, they show that given a fixed monotone non-increasing allocation rule $A(c)$ defined on a discrete cost set $\{c_1, \ldots, c_m\}$, there exists an optimal payment rule $P(c)$ that guarantees truthfulness and individual rationality.

Lemma 3.2 (Chen et al. [4], Claim 2 in Section B.1.2). *Let $A(c)$ be a monotone non-increasing allocation rule defined for set $\{c_1, \ldots, c_m\}$ with $c_1 \leq \cdots \leq c_m$. Define payment rule $P(c_i) = c_i + \frac{1}{A(c_i)} \sum_{j=i+1}^{m} A(c_j)(c_j - c_{j-1})$. Then $(A(c), P(c))$ is truthful and individually rational for all $c \in \{c_1, \ldots, c_m\}$, and any payment rule $P'(c)$ that guarantees the truthfulness and individual rationality of $(A(c), P'(c))$ must have $P'(c) \geq P(c)$ for all $c \in \{c_1, \ldots, c_m\}$.*

Furthermore for any cost distribution supported on $\{c_1, \ldots, c_m\}$, the expected total payment of $(A(c), P(c))$, with the optimal payment rule $P(c)$ defined in Lemma 3.2, can be equivalently represented in a simpler form in terms of virtual costs.

Definition 3.1 (Virtual costs). *Let $f(c)$ and $F(c)$ be the PDF and the CDF of a cost distribution $\mathcal{F}$ supported on $\{c_1, \ldots, c_m\}$ with $c_1 \leq \cdots \leq c_m$. Let $c_0 = 0$. The virtual cost function $\psi(c)$ of this cost distribution is defined as*

$$\psi(c_i) = c_i + \frac{c_i - c_{i-1}}{f(c_i)} F(c_{i-1})$$

for all $1 \leq t \leq m$.

We remark that it is very likely that $\psi(c_i)$ is not regular, i.e., is not a monotone non-decreasing function of c_i. Consider when the cost distribution is the uniform distribution over a finite set, $\psi(c_i) = c_i + (c_i - c_{i-1}) * (i - 1)$. If c_i is very close to c_{i-1}, then $\psi(c_i)$ will be roughly equal to c_i. If c_i is much larger than c_{i-1}, $\psi(c_i)$ can be close to $i * c_i$. So if there exist three consecutive costs, the difference between the first two costs is large, and the difference between the last two is small, the virtual costs will very likely be irregular. For example, $c_1 = 1$, $c_2 = 10$, $c_3 = 11$, then $\psi(c_2) = 19$ and $\psi(c_3) = 13$, so ψ is not monotone non-decreasing.

Lemma 3.3 (Chen et al. [4], Lemma 10 in Section B.1.2). *Let $A(c)$ be a monotone non-increasing allocation rule defined on set $\{c_1, \ldots, c_m\}$ with $c_1 \leq \cdots \leq c_m$. Let $P(c)$ be the optimal truthful and individually rational payment rule defined in Lemma 3.2. When cost follows a distribution $\mathcal{F}$ supported on $\{c_1, \ldots, c_m\}$, the expected total payment $\mathbb{E}_{c\sim\mathcal{F}}[A(c)P(c)]$ is equal to the expected virtual cost $\mathbb{E}_{c\sim\mathcal{F}}[A(c)\psi(c)]$ where $\psi(c)$ is the virtual cost function of $\mathcal{F}$.*

The above lemmas assume that costs are from a finite discrete set. Our benchmark mechanism where the analyst already knows all the costs satisfies this assumption. We'll use the above result to establish the performance of our benchmark mechanism. However, our mechanisms developed in this paper for the unknown cost case do not have any restriction on the set of possible costs. The allocation rules and the payment rules of our mechanisms are first computed on a discrete set using the above result, and then extended to all other values of cost. We show below that such extension preserves truthfulness and individual rationality.

Definition 3.2 (Extended allocation rule and payment rule). *Given a survey mechanism (A^d, P^d) that is defined on a discrete cost set $\{c_1, \ldots, c_m\}$ with $c_1 \leq \cdots \leq c_m$. The extended allocation rule and payment rule A, P are defined as follows*

$$A(c) = A^d(\lceil c \rceil), \quad P(c) = P^d(\lceil c \rceil), \text{ for all } c \in [0, c_m],$$

where $\lceil c \rceil$ is the minimum cost in $\{c_1, \ldots, c_m\}$ that is greater than or equal to c.

Lemma 3.4. *Let $A^d(c)$ be a monotone non-increasing allocation rule defined on set $\{c_1, \ldots, c_m\}$ with $c_1 \leq \cdots \leq c_m$. Let $P^d(c)$ be the optimal payment rule defined in Lemma 3.2. Then the extended allocation rule and payment rule of (A^d, P^d) is still truthful and individually rational.*

The proof of the lemma can be found in the full version of the paper.

3.2 Formulating the Optimal Unbiased Estimator with Known Costs

In this section, we formulate an optimization problem that solves the optimal survey mechanism for an unbiased estimator *when the cost distribution is known*, based on the results of [4]. The optimization problem is only slightly different from [4] in the objective function because in our setting, the agents are not i.i.d. drawn from the same distribution but come as a random permutation. The value of statistic z is assumed to be bounded and without loss of generality $0 \leq z \leq 1$.

Horvitz-Thompson Estimator: When we use truthful survey mechanisms $\mathbf{M} = (A^1, P^1), \ldots, (A^n, P^n)$ to purchase the data points, the data of agent i will be collected with probability $A^i(c_i)$. Define

$$\widehat{x}_i = \begin{cases} z_i, & \text{with probability } A^i(c_i) \\ 0, & \text{otherwise} \end{cases}$$

to be the observed data point which is set to zero if no purchase is made. Define $y_i = \frac{\widehat{x}_i}{A^i(c_i)}$. To get unbiased estimation, we use Horvitz-Thompson Estimator, which is the unique unbiased linear estimator in our setting (see [26] for more details),

$$S(\mathbf{M}, (\mathbf{c}, \mathbf{z})) = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Notice that an unbiased estimator always buys the data points with probability greater than 0, i.e., $A^i(c_i) > 0$ for all i and c_i. If $A^i(c_i) = 0$, the mechanism will never know the value of z_i and thus cannot be unbiased.

When the cost set $C = \{c_1, \ldots, c_n\}$ is known to the analyst, the optimal mechanism that uses the same survey mechanism for all data holders has been derived by Chen et al. [4]. They reduce the mechanism design problem to a min-max optimization problem. The optimal allocation rule that minimizes the worst-case variance of the Horvitz-Thompson Estimator can be formulated as the solution of an optimization problem $OPT(n, C, B)$, which is defined as follows,

$$
\begin{aligned}
OPT(n, C, B) = \arg\min_{A} \max_{\mathbf{z} \in [0,1]^n} \quad & \frac{1}{n^2}\left(\sum_{i=1}^{n} \frac{z_i^2}{A(c_i)} - \sum_{i=1}^{n} z_i^2\right) \qquad (1) \\
\text{s.t.} \quad & \sum_{i=1}^{n} A(c_i)\psi(c_i) \leq B \\
& A(c) \text{ is monotone non-increasing in } c \\
& 0 \leq A(c) \leq 1, \quad \forall c
\end{aligned}
$$

Here the objective function is changed from the original formulation in [4] so that it is equal to the worst-case variance of the Horvitz-Thompson Estimator in our setting. According to the law of total variance, $\mathrm{Var}(S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))) = \mathbb{E}[\mathrm{Var}(S|(\mathbf{c}, \mathbf{z}))] + \mathrm{Var}(\mathbb{E}[S|(\mathbf{c}, \mathbf{z})])$. Since the estimator is always unbiased for any order $(\mathbf{c}, \mathbf{z})$, the second term is zero. Furthermore, when conditioning on a sequence, $y_i = \frac{\widehat{x}_i}{A(c_i)}$ become independent when a fixed allocation rule A is used. Therefore the variance of the Horvitz-Thompson Estimator is equal to

$$
\begin{aligned}
\mathrm{Var}(S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))) =& \mathbb{E}_{(\mathbf{c}, \mathbf{z})}[\mathrm{Var}(S|(\mathbf{c}, \mathbf{z}))] = \mathbb{E}_{(\mathbf{c}, \mathbf{z})}\left[\frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(y_i|(\mathbf{c}, \mathbf{z})\right)\right] \\
=& \frac{1}{n^2} \cdot \mathbb{E}_{(\mathbf{c}, \mathbf{z})}\left[\sum_{i=1}^{n} \mathbb{E}[y_i^2|(\mathbf{c}, \mathbf{z})] - \mathbb{E}[y_i|(\mathbf{c}, \mathbf{z})]^2\right].
\end{aligned}
$$

For any arriving sequence $(\mathbf{c}, \mathbf{z})$, $\sum_{i=1}^{n} \mathbb{E}[y_i^2|(\mathbf{c}, \mathbf{z})] - \mathbb{E}[y_i|(\mathbf{c}, \mathbf{z})]^2$ stays the same, which is equal to $\sum_{i=1}^{n} \frac{z_i^2}{A(c_i)} - \sum_{i=1}^{n} z_i^2$. Therefore by maximizing over $\mathbf{z}$, we get the worst-case variance of the Horvitz-Thompson estimator

$$
\max_{\mathbf{z} \in [0,1]^n} \frac{1}{n^2}\left(\sum_{i=1}^{n} \frac{z_i^2}{A(c_i)} - \sum_{i=1}^{n} z_i^2\right).
$$

The last constraint $0 \leq A(c) \leq 1$ makes sure that A is an allocation rule of a survey mechanism. The second constraint is the sufficient and necessary condition for A to be the allocation rule of a truthful mechanism. The first constraint guarantees expected budget feasibility according to Lemma 3.3, where $\psi(c)$ is the virtual cost function.

To simplify the analysis, in this work we will use an approximation of the above optimization problem, which is defined as $APPROX(n, C, B)$ by removing the second term of the objective function of $OPT(n, C, B)$,

$$
\begin{aligned}
APPROX(n, C, B) = \arg\min_{A} \max_{\mathbf{z} \in [0,1]^n} \quad & \sum_{i=1}^{n} \frac{z_i^2}{A(c_i)} \qquad (2) \\
\text{s.t.} \quad & \sum_{i=1}^{n} A(c_i)\psi(c_i) \leq B \\
& A(c) \text{ is monotone non-increasing in } c \\
& 0 \leq A(c) \leq 1, \quad \forall c
\end{aligned}
$$

Lemma 3.5. *The worst-case variance of APPROX(n, C, B) is no more than the worst-case variance of OPT(n, C, B) plus $\frac{1}{n}$.*

Proof. As $z_i \in [0, 1]$, the second term of the objective function of $OPT(n, C, B) \leq 1/n$. □

3.3 Characterization of the Optimal Unbiased Estimator

The characterization of the optimal unbiased estimator when the cost distribution is known has been studied in both [4] and [26]. But neither of the solutions can be directly applied to our problem. Roth and Schoenebeck [26] require the PDF of the cost distribution to be piecewise differentiable except on a measure zero set, and Chen et al. [4] assume that the cost distribution is regular, i.e., the virtual cost function $\psi(c)$ is monotone non-decreasing in c. Although the optimization problem (2) has a convex objective function and thus can be solved efficiently by the convex optimization algorithms (see [2]), the closed-form solution is still non-trivial to derive. Below we give the exact characterization of the optimal solution, which has a very simple form and will further be used to derive the optimal confidence interval mechanism: the optimal allocation rule $A^i(c)$ is inversely proportional to the square root of the regularized virtual cost of c, which is defined as follows,

Definition 3.3 (Regularized virtual costs). *For a discrete uniform distribution supported on $\{c_1, \ldots, c_m\}$ with $c_1 \leq \cdots \leq c_m$ and its virtual costs function $\psi(c_1), \ldots, \psi(c_m)$. For every $i \leq k$, let $Avg(i, k)$ be the average of $\psi(c_i), \ldots, \psi(c_k)$. We define regularized virtual cost $\phi(c_i)$ as follows*

$$\phi(c_i) = \max\{\psi'(c_1), \ldots, \psi'(c_i)\},$$
$$\psi'(c_i) = \min_{k: k \geq i} Avg(i, k).$$

The form of the regularized virtual costs is very similar to the ironed virtual value used in revenue-maximizing auction design [21]. The idea is to replace the exact virtual cost (value) with the average virtual cost (value) on an interval that has non-regular virtual cost (value). But our proof is different because the underlying optimization problem is not the same.

Theorem 3.1. *The optimal solution of APPROX(n, C, B) is*

$$A(c_{(k)}) = \min\left\{1, \frac{\lambda}{\sqrt{\phi(c_{(k)})}}\right\}, \quad \textit{for all } 1 \leq k \leq n,$$

where $\phi(c)$ is the regularized virtual cost function when the cost distribution is the uniform distribution over C, and λ is chosen such that the budget constraint is satisfied with equality

$$\sum_{k=1}^{n} A(c_{(k)})\psi(c_{(k)}) = B.$$

The value of λ can be computed using binary search within time $O(\log |C|)$.

The proof of the theorem can be found in the full version of the paper, in which we demonstrate some properties of the regularized virtual cost function, and show that the KKT conditions are satisfied because of these properties.

4 OPTIMAL CONFIDENCE INTERVAL WITH KNOWN COSTS

In this section, we design purchasing mechanisms to get the best confidence interval of the statistic when the cost distribution is known to the analyst at the beginning of the survey. In this case, the optimal mechanism needs to find the optimal trade-off of the bias and the variance in order to minimize the length of the interval. We consider the class of confidence intervals that are defined around the sample mean, and the length of which is decided by a bias term and the sample variance. We will first introduce an extended survey mechanism that allows biased estimation, e.g., by ignoring the high-cost data. Then again we formulate an optimization problem and present the characterization of the optimal solution.

Before introducing the new estimator, we first show how to convert our unbiased estimator into a confidence interval. In this work, we use the most classic approach to construct confidence interval based on sample mean and sample variance.

Construct confidence interval using unbiased estimator: Consider an unbiased estimator $S(\mathbf{M}, (\mathbf{c}, \mathbf{z}))$ that uses survey mechanism $\mathbf{M} = (\mathbf{A}, \mathbf{P})$, and we want to construct a confidence interval for $\mathbb{E}[z] = \frac{1}{n} \sum_{i=1}^{n} z_i$. Again we use

$$\widehat{x_i} = \begin{cases} z_i, \text{ with probability } A^i(c_i) \\ 0, \text{ otherwise} \end{cases}$$

to denote the observed data point and define $y_i = \frac{\widehat{x_i}}{A^i(c_i)}$. Notice that the random variables $y_1, \ldots, y_n$ are not independent since the allocation rule A^i can depend on $c_1, \ldots, c_{i-1}$. But if we consider a fixed realization $(\tilde{\mathbf{c}}, \tilde{\mathbf{z}})$, the mechanisms $(A^1, P^1), \ldots, (A^n, P^n)$ will also be fixed. Then $y_1, \ldots, y_n$ become independent, because when the probability of purchasing each data point $A^1(c_1), \ldots, A^n(c_n)$ is fixed, whether to purchase each data point or not is independently decided. Therefore given a confidence level γ, we can construct a confidence interval of the expected mean $\mathbb{E}\left[\sum y_i/n \middle| (\tilde{\mathbf{c}}, \tilde{\mathbf{z}})\right]$ using the sample mean $\sum_{i=1}^{n} y_i/n$ and sample variance $\widehat{\sigma}^2 = \sum_{i=1}^{n} \left(y_i - \sum_{i=1}^{n} y_i/n\right)^2 /(n-1)$ according to Bernstein's inequality (see more details in the full version):

$$\left[\sum_{i=1}^{n} y_i/n - \frac{\alpha_\gamma}{\sqrt{n}} \cdot \widehat{\sigma}, \quad \sum_{i=1}^{n} y_i/n + \frac{\alpha_\gamma}{\sqrt{n}} \cdot \widehat{\sigma}\right]$$

where α_γ is a constant that is chosen to achieve confidence level γ. When the estimator is unbiased, $\mathbb{E}\left[\sum y_i/n \middle| (\tilde{\mathbf{c}}, \tilde{\mathbf{z}})\right] = \mathbb{E}[z]$ for all $(\tilde{\mathbf{c}}, \tilde{\mathbf{z}})$, this interval is just a confidence interval of $\mathbb{E}[z]$ with confidence level γ.

4.1 Confidence Interval and Bias-variance Tradeoff

This "unbiased" confidence interval does not necessarily have the minimum length. Observe that a small portion of high-cost data can drastically increase the variance of the unbiased estimator as $A(c)$ must be small. We can allow the mechanism to simply ignore these data points, i.e., to have $A^i(c_i) = 0$ for some i and set $y_i = 0$. This can probably reduce the variance of the estimator. However, doing so causes the estimator to be biased. Therefore, we need to increase the length of the confidence interval to compensate for this added bias.

For this reason, we extend the standard survey mechanism to incorporate biased estimation. The mechanism ignores a data point with probability $U(c)$. The added bias can be represented as a function of $U(c)$. Although the mechanisms now have a new component $U(c)$, the results in Section 3 can nevertheless be applied by seeing $(1 - U(c))A(c)$ as the allocation rule.

Survey mechanisms that allow bias: We add a new component $\mathbf{U} = (U^1, \ldots, U^n)$ into our allocation rule to allow biased estimation, where each U^i is a function of reported cost c_i. A mechanism that allows bias consists of $(\mathbf{A}, \mathbf{U}, \mathbf{P})$. When a data point with cost c_i comes at time i, the mechanism first flips a coin $\widehat{U_i}$ to decide whether to ignore this data point or not, and the probability of $\widehat{U_i}$ being 1 (which means ignoring the data) is equal to $U^i(c_i)$. If the data is ignored, a bias term will be added into the final estimation to compensate the error. If $\widehat{U_i} = 0$, then the mechanism purchases the data with probability $A^i(c_i) > 0$ and pays $P^i(c_i)$ if the data is purchased. Then the observed data $\widehat{x}_i$ follows

$$\widehat{x}_i = \begin{cases} z_i, \text{ with probability } (1 - U^i(c_i))A^i(c_i) \\ 0, \text{ with probability } (1 - U^i(c_i))(1 - A^i(c_i)) \\ \text{ignored, with probability } U^i(c_i). \end{cases}$$

We re-define y_i as

$$y_i = \begin{cases} \frac{\widehat{x}_i}{A^i(c_i)}, \text{ if } \widehat{U}_i = 0 \\ 0, \text{ if } \widehat{U}_i = 1. \end{cases}$$

Then for a fixed arriving sequence $(\tilde{\mathbf{c}}, \tilde{\mathbf{z}})$, the bias of estimator $\sum y_i/n$ is equal to

$$err = \mathbb{E}[z] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[y_i|(\tilde{\mathbf{c}}, \tilde{\mathbf{z}})] = \frac{1}{n}\sum_{i=1}^{n} z_i - \frac{1}{n}\sum_{i=1}^{n}(1 - \widehat{U}_i)\tilde{z}_i = \frac{1}{n}\sum_{i=1}^{n} \tilde{z}_i \cdot \widehat{U}_i.$$

Notice that this bias is not observable because the mechanism does not know the $\tilde{z}_i$ that is not purchased. But since $\tilde{z}_i$ is between 0 and 1 and we use worst-case analysis in this work, we can just assume $\tilde{z}_i$ equals its worst-case value 1. (This can be seen more clearly in our formulation of the optimization problem in the next section.) Then the confidence interval of $\mathbb{E}[z]$ with confidence level γ can be constructed as follows

$$\left[\sum_{i=1}^{n} y_i/n - \frac{\alpha_\gamma}{\sqrt{n}} \cdot \widehat{\sigma}, \quad \sum_{i=1}^{n} y_i/n + \frac{1}{n}\sum_{i=1}^{n} \widehat{U}_i + \frac{\alpha_\gamma}{\sqrt{n}} \cdot \widehat{\sigma}\right]$$

where $\widehat{U}_i$ is the indicator of whether the i-th data point is ignored and $\widehat{\sigma}^2$ is the sample variance of $y_1, \ldots, y_n$.

For convenience, in the rest of the paper, we write $U_c^i = U^i(c)$ for short, and in cases when the costs are indexed as $c_1, \ldots, c_n$ or $c_{(1)}, \ldots, c_{(n)}$, we use U_j^i to represent $U^i(c_j)$ or $U^i(c_{(j)})$.

4.2 Formulation of the Optimal Confidence Interval

We formulate a min-max optimization problem that approximately solves the optimal allocation rule. The expected length of the interval we construct is $2 \cdot \frac{\alpha_\gamma}{\sqrt{n}} \cdot \mathbb{E}[\widehat{\sigma}] + \mathbb{E}[err]$. Since the expectation of sample standard deviation $\mathbb{E}[\widehat{\sigma}]$ is difficult to compute, we estimate $\mathbb{E}[\widehat{\sigma}]$ with $\sqrt{\mathbb{E}\sum_{i=1}^{n} y_i^2/n}$. When $0 \le \mathbb{E}[y_i] \le 1$, the difference between $\mathbb{E}[\widehat{\sigma}]$ and $\sqrt{\mathbb{E}\sum_{i=1}^{n} y_i^2/n}$ is no more than $1 + O(1/n)$ (see the full version for more details).

The approximate expected length of the confidence interval can thus be written into a function of A and U and $\mathbf{z}$

$$2 \cdot \frac{\alpha_\gamma}{\sqrt{n}} \cdot \sqrt{\mathbb{E}\sum_{i=1}^{n} y_i^2/n} + \mathbb{E}[err] = 2 \cdot \frac{\alpha_\gamma}{\sqrt{n}} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n} \frac{(1 - U_i)z_i^2}{A(c_i)}} + \frac{1}{n}\sum_{i=1}^{n} z_i \cdot U_i.$$

Then we only need to take maximum over all possible $\mathbf{z}$ to get the worst-case expected length. Therefore, suppose the underlying cost set is $C = \{c_{(1)}, \ldots, c_{(n)}\}$ with $c_{(1)} \le \cdots \le c_{(n)}$, then the approximately optimal allocation rule (A^*, U^*) can be again formulated as the solution of

$OPT_{CI}(n, C, B)$, which is defined as

$$OPT_{CI}(n, C, B) = \arg\min_{A,U} \max_{z \in [0,1]^n} \quad \beta \cdot \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(1-U_i)z_i^2}{A(c_{(i)})}} + \frac{\sum_{i=1}^{n} z_i \cdot U_i}{n} \tag{3}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} (1-U_i) \cdot A(c_{(i)}) \psi(c_{(i)}) \le B$$

$$(1-U_c)A(c) \text{ is monotone non-increasing in } c$$

$$0 \le A(c) \le 1, \quad 0 \le U_c \le 1, \quad \forall c$$

where $\beta = 2 \cdot \frac{\alpha_Y}{\sqrt{n}}$.

Lemma 4.1. *Let L^* be the value of the objective function* (3) *when A^* and U^* is used. (L^* is an approximation of the worst-case expected length of the confidence interval produced by (A^*, U^*).) Then the difference between L^* and the worst-case expected length of the actual optimal confidence interval MIN is no more than $2\beta(1 + O(1/n)) = 4 \cdot \frac{\alpha_Y}{\sqrt{n}}(1 + O(1/n))$.*

The optimal solution of (3) is still difficult to solve. But if we replace the objective function by the sum of the squares of its two terms

$$APPROX_{CI}(n, C, B) = \arg\min_{A,U} \max_{z \in [0,1]^n} \quad \beta^2 \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{(1-U_i)z_i^2}{A(c_{(i)})} + \left(\frac{\sum_{i=1}^{n} z_i \cdot U_i}{n} \right)^2 \tag{4}$$

the optimal solution can be computed efficiently. The optimization problem with the new objective function will give a 2-approximation of (3) because $a^2 + b^2 \le (a+b)^2 \le 2(a^2 + b^2)$.

Lemma 4.2. *$APPROX_{CI}(n, C, B)$ will give a 2-approximation of $OPT_{CI}(n, C, B)$.*

In this work, we consider the bias-variance tradeoff in the setting of getting the shortest confidence interval. By changing the value of β in the objective function, our mechanism can also be applied to other estimation tasks that involve the bias-variance tradeoff. For example, if we want to minimize the mean squared error of the output estimator, or equivalently the risk of the estimator when the loss function is the squared error function, we can set $\beta = 1$ and apply the same mechanism.

4.3 Characterization of the Optimal Confidence Interval

So how should the optimal mechanism look like when some data can simply be ignored? It is natural to believe that the optimal mechanism should ignore the data points with the highest costs. This is corroborated in the following theorem that the optimal mechanism ignores all the data points with *regularized virtual costs* above a threshold H, and purchases (with probability) all the data points below the threshold. The characterization of the optimal A remains the same as the unbiased case.

Theorem 4.1. *The optimal solution of $APPROX_{CI}(n, C, B)$ is as follows:*

$$U_j = \begin{cases} 0, & \text{if } \phi(c_{(j)}) < H \\ p \in (0,1], & \text{if } \phi(c_{(j)}) = H \\ 1, & \text{if } \phi(c_{(j)}) > H \end{cases}$$

$$A(c_{(j)}) = \min\left\{1, \frac{\lambda}{\sqrt{\phi(c_{(j)})}}\right\}$$

where p is a constant in $(0, 1]$, and $\phi(c)$ is the regularized virtual cost function (Definition 3.3) when the cost distribution is the uniform distribution over C, and λ is chosen such that the budget constraint is satisfied with equality. The value of λ and H can be computed using binary search over set C within time $O(\log |C|)$.

The proof can be found in the full version of the paper. The optimal H can be found by binary search because it can be proved that the objective function is a convex function of $M = \sum_{j=1}^{n} U_j$, when A is optimized after U is decided. Let $g(M)$ be the the optimal value of the first term $\beta^2 \cdot \frac{1}{n} \sum_{j=1}^{n} \frac{1-U_j}{A(c_{(j)})}$ when $\sum_{j=1}^{n} U_j$ is set to M. Then the objective function is just $g(M) + \left(\frac{M}{n}\right)^2$. The second term $\left(\frac{M}{n}\right)^2$ is a convex function of M. We prove that $g(M)$ is also a convex function of M.

Lemma 4.3. *The function $g(M)$ is a convex function of M. Furthermore, let A_M be an optimal allocation rule when $\sum_{j=1}^{n} U_j = M$ and let $c_{(r)}$ be the largest cost that is not ignored with probability* 1. *Then for non-integer M that has $A_M(c_{(r)}) < 1$,*

$$\frac{\partial g(M)}{\partial M} = -\beta^2 \cdot \frac{2}{n} \cdot \frac{1}{A_M(c_{(r)})}.$$

which is an non-decreasing function of M.

Therefore the optimal H can be found by binary search over the optimal M such that $\frac{\partial g(M)}{\partial M} + \frac{2M}{n^2} \geq 0$ on the right and $\frac{\partial g(M)}{\partial M} + \frac{2M}{n^2} \leq 0$ on the left. The complete proof of the lemma can be found in the full version of the paper.

5 ONLINE MECHANISMS

We now move to the case when the cost distribution is unknown at the beginning. The idea of our mechanism is very simple: at time i, use the approximately optimal allocation rule $APPROX(\cdot)$ as if (1) there are i data holders with costs $c_1, \ldots, c_{i-1}$ and $\{\overline{C}\}$, and (2) the analyst's total budget for these i data points is proportional to $\sqrt{i}$, so that the average budget for each data point is proportional to $\frac{1}{\sqrt{i}}$. So the average budget is a decreasing function of i, which means we use more budget in the earlier stages when the estimation of the cost distribution is not accurate. We prove that for both of the unbiased estimator problem and the confidence interval problem, such an online mechanism will only be worse for a constant factor compared to the optimal mechanisms $OPT(n, C, B)$ that knows the cost distribution at the beginning.

Mechanism 1: Online Mechanism Outline

```
for i = 1,...,n do
    1) Let $A^i = APPROX\left(i,\ T_i,\ \xi_n B\sqrt{i}\right)$ be the optimal allocation rule when there are $i$ data
    holders with costs
        $T_i = \{c_1, \ldots, c_{i-1}, \overline{C}\}$,
    and the total budget for these $i$ data points is $\xi_n B\sqrt{i}$.
    2) Ask agent $i$ to report cost $c_i$ and purchase the data using (an approximation of) $(A^i, P^i)$,
    where $P^i$ is computed as in Lemma 3.2.
end for
Aggregate all collected to output an estimator.
```

The outline of our online mechanisms (for both the unbiased estimator and the confidence interval) is given in Mechanism 1.

We will describe the specifics of the online mechanism for the unbiased estimation of mean and the confidence interval respectively in the next two parts. For each, we'll prove that the online mechanism is optimal within a constant factor. We sketch the high-level idea of our proofs as follows.

(1) We compare both of our online algorithm and the benchmark with an intermediate mechanism
$$(A'^{,i}, P'^{,i}) = APPROX\left(i,\ \{c_1, \ldots, c_{i-1}, c_i\},\ \xi_n B\sqrt{i}\right)$$
at each step i. This intermediate mechanism $(A'^{,i}, P'^{,i})$ is basically the same as (A^i, P^i), but is "one-step-ahead". (A', P') is the optimal mechanism when the same amount of budget is assigned, but knows an additional piece of information, the value of c_i, beforehand.

(2) We show that the difference between $(A'^{,i}, P'^{,i})$ and (A^i, P^i) is no more than a constant factor. This is mainly due to the similarity of $(A'^{,i}, P'^{,i})$and (A^i, P^i). Since the two mechanisms only differ in one element in the cost set, the regularized virtual costs are not going to change a lot, and so is the allocation rule.

(3) Then we prove that if $(A'^{,i}, P'^{,i})$ is used at each round, the performance of the output estimator is no worse than a constant times the benchmark. Our budget allocation method and the random arriving order play crucial roles here. We give the basic idea of our budget allocation rule below.

A simplified modeling of the budget allocation. We're able to show that if we allocate average budget $\overline{b}_i$ at round i, the "loss" (of performance) occurred by our mechanism at round i is bounded by $\frac{r \cdot B}{i \cdot \overline{b}_i}$ times the "loss" occurred by the benchmark A^*, where r is a constant. So the optimal budget allocation is essentially the following problem

$$\begin{aligned} \min_{\overline{b}_i} \quad & r \cdot B \sum_{i=1}^{n} \frac{1}{i \cdot \overline{b}_i} \\ \text{s.t.} \quad & \sum_{i=1}^{n} \overline{b}_i \leq B, \\ & \overline{b}_i \geq 0. \end{aligned}$$

which gives $\overline{b}_i \propto \frac{1}{\sqrt{i}}$. In addition, it can be shown that the total "loss" of our mechanism is no more than a constant times the total "loss" of the benchmark when this budget allocation rule is applied.

Optimality of the online mechanisms. Our mechanisms are proved to be optimal within constant factors. But it remains open whether our constant factors are optimal. One possible method to improve the constant is to only collect agents' costs without purchasing any data at the beginning and start collecting data after a more accurate cost estimation is acquired. But such a method would weaken the incentive guarantee for the agents to report their true costs: if the agents' report costs are not going to affect their rewards, why would they report the true costs? It is also challenging to analyze the performance of such a method. As we've shown in Theorem 3.1 and Theorem 4.1, the optimal allocation rule is defined by the regularized virtual costs. It may not be easy to estimate the regularized virtual costs because (1) the regularized virtual cost function is a global property of the cost distribution; (2) the value of the virtual cost function is very sensitive to the PDF of the cost distribution, which appears in the denominator of the second term: $f(c_i)/(c_i - c_{i-1})$ in $\psi(c_i) = c_i + \frac{1}{f(c_i)/(c_i - c_{i-1})} F(c_{i-1})$ (see Definition 3.1). The analysis of our online mechanisms is possible because the intermediate mechanism $(A'^{,i}, P'^{,i})$ only differs from the optimal mechanism (A^i, P^i) in one element in the cost set, which doesn't affect the regularized virtual costs much.

However, our analysis cannot be easily extended to analyze the performance of the above mentioned alternate method as its difficult to bound the estimation error of the regularized virtual costs. The performance of this alternate method remains an open question.

5.1 Unbiased Estimator

We first introduce the benchmark to which we compare our online algorithm.

Definition 5.1. *Let $c_{(1)} \leq \cdots \leq c_{(n)}$ be the n data holders' costs ordered from smallest to largest. Suppose there is one more data holder with cost $\overline{C}$. We define the benchmark (A^*, P^*) to be the mechanism that purchases data from these $n + 1$ data holders, and minimizes the worst case variance when it knows the set of costs $\{c_{(1)}, c_{(2)}, \ldots, c_{(n)}, \overline{C}\}$ at the beginning.*

$$A^* = OPT(n + 1,\ \{c_{(1)}, c_{(2)}, \ldots, c_{(n)}, \overline{C}\},\ B). \tag{5}$$

This additional cost $\overline{C}$ can be interpreted as the loss of unknown upper bound of the possible cost. When the cost distribution is known, the mechanism knows the exact maximum cost $c_{(n)}$, and thus the optimal mechanism will never have a positive probability to buy a data point with cost higher than $c_{(n)}$. But when $c_{(n)}$ is unknown, the mechanism always has to buy any data point (with cost under $\overline{C}$) with a positive probability.

We show that our online mechanism satisfies all three constraints and its worst-case variance is roughly within a constant times the benchmark.

Theorem 5.1. *If we use Mechanism 1 with*

- $\xi_n = \frac{1}{4\sqrt{n}}$.
- *At round i, use the extended allocation rule and payment rule (Definition 3.2) of (A^i, P^i), where*

$$A^i = APPROX\left(i,\ T_i,\ \xi_n B\sqrt{i}\right)$$

and P^i is computed as in Lemma 3.2. Let the collected data be $\widehat{x_i}$.

- *Output unbiased estimator $S = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{x_i}}{A^i(c_i)}$ at last.*

Then the mechanism satisfies (1) truthfulness and individually rationality, (2) the expected total spending is no more than $B = n\overline{B}$, and (3) for any cost distribution $\{c_1, \ldots, c_n\}$ the worst-case variance of the final estimator S is no more than

$$16 \cdot \left(\left(1 + \frac{1}{n}\right)^2 \cdot Var^*(A^*) + \frac{1}{n} + \frac{1}{n\sqrt{n}} \cdot \frac{1}{A^*(\overline{C})}\right),$$

where A^* is the benchmark defined in Definition 5.1.

Discussion: We have the factor $\left(1 + \frac{1}{n}\right)^2$ in the first term of our upper bound because the benchmark mechanism has one more data point. It is no more than 4 when $n \geq 1$ and goes to 1 when n gets large. The second additive term $\frac{1}{n}$ is due to our estimation of $\mathrm{Var}(S)$ by $\mathbb{E}[S^2]$. We know that $\mathrm{Var}^*(A^*)$ is roughly $\frac{1}{n+1} \cdot \mathbb{E}[\frac{1}{A^*(c)}]$. So when the problem is non-trivial, we should have the average $\frac{1}{A^*(c)}$ much larger than 1, and $\frac{1}{n}$ will be small compared to $\mathrm{Var}^*(A^*)$. The last additive term $\frac{1}{n\sqrt{n}} \cdot \frac{1}{A^*(\overline{C})}$ is dominated by $\frac{1}{n}$. It is only comparable to $\mathrm{Var}^*(A^*)$ when $\sqrt{n} \leq \frac{1}{A^*(\overline{C}) \cdot \mathbb{E}[\frac{1}{A^*(c)}]}$

The complete proof can be found in the full version of the paper.

5.2 Confidence Interval

Our benchmark for the online mechanism is again the optimal mechanism that knows the cost distribution at the beginning and uses a single optimal mechanism (A^*, U^*, P^*) throughout the survey. We still add an additional cost $\overline{C}$ into the underlying cost set of the benchmark mechanism, in order to make the comparison possible without knowing the exact maximum cost.

Definition 5.2. *Let $c_{(1)} \leq \cdots \leq c_{(n)}$ be the n data holders' costs ordered from smallest to largest. Suppose there is one more data holder with cost $\overline{C}$. We define the benchmark (A^*, U^*, P^*) to be the mechanism that purchases data from these $n+1$ data holders, and minimizes the worst case variance when it knows the set of costs $\{c_{(1)}, c_{(2)}, \ldots, c_{(n)}, \overline{C}\}$ at the beginning.*

$$A^*, U^* = OPT_{CI}(n+1,\ \{c_{(1)}, c_{(2)}, \ldots, c_{(n)}, \overline{C}\},\ B). \tag{6}$$

Theorem 5.2. *If we use Mechanism 1 with*

- $\xi_n = \frac{1}{16\sqrt{n}}$.
- *At round i, compute*
$$(A^i, U^i) = APPROX_{CI}(i,\ T_i,\ \xi_n B\sqrt{i}).$$
Let $\mathbb{1}\left(U^i(c) \geq \frac{1}{2}\right)$ be the rule that completely ignores data with cost c if $U^i(c) \geq \frac{1}{2}$, and never ignores the data if $U^i(c) < \frac{1}{2}$. Then the mechanism purchases agent i's data using the extended allocation and payment rule of
$$\left(A^i, \mathbb{1}\left(U^i(c) \geq \frac{1}{2}\right), P^i\right),$$
where P^i is computed as in Lemma 3.2. Let the collected data be $\widehat{x_i}$.
- *Output confidence interval*
$$\left[\sum_{i=1}^n y_i/n - \frac{\alpha_\gamma}{\sqrt{n}} \cdot \widehat{\sigma},\quad \sum_{i=1}^n y_i/n + \frac{\sum_{i=1}^n \widehat{U_i}}{n} + \frac{\alpha_\gamma}{\sqrt{n}} \cdot \widehat{\sigma}\right],$$
where $y_i = \frac{\widehat{x_i}}{A^i(c_i)}$ and $\widehat{\sigma}^2$ is the sample variance of $y_1, \ldots, y_n$, and $\widehat{U_i}$ represents whether the i-th data point is ignored or not.

Then the mechanism is (1) truthful in expectation and individually rational; (2) satisfies the budget constraint $B = n\overline{B}$ in expectation; (3) and the for any cost distribution $\{c_1, \ldots, c_n\}$, the worst-case expected length of the output confidence interval is no more than

$$8\sqrt{10} \cdot L^* + \frac{2\sqrt{10}}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

where L^* is the approximate worst-case expected length of the benchmark as defined in Lemma 4.1.

Corollary 5.1. *The worst-case expected length of our mechanism's output confidence interval is no more than*

$$8\sqrt{10} \cdot MIN + \frac{2\sqrt{10} + 32\sqrt{10} \cdot \alpha_\gamma}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

where MIN is the worst-case expected length of the optimal confidence interval estimator.

Discussion: As we show in the previous section, L^* is roughly $\frac{1}{\sqrt{n}} \cdot \left(2\alpha_\gamma\sqrt{\mathbb{E}[\frac{1-U_c^*}{A^*(c)}]} + \sqrt{n} \cdot \mathbb{E}[U_c^*]\right)$. If the problem is non-trivial, we should have $\frac{1}{\sqrt{n}}$ dominated by L^*. The complete proof can be found in the full version of the paper .

6 DISCUSSIONS

In this work, we restrict our estimators to use only the collected data. When the data are correlated with the costs, the data analyst may gradually learn the cost-data correlation based on the collected pairs. This means that if a data holder arrives and reports his cost, the data analyst may form a prediction for his data based on his reported cost and the learned cost-data correlation, even if the data is not collected. This leads to an interesting open question: can the final estimation be improved if the data analyst makes use of such predicted data?

Allowing the data analyst to leverage on the cost-data correlation brings up an additional level of challenge when the data holders care about the privacy of their data. Such data holders may hesitate to report their costs, because reporting the cost itself reveals some information about his data. This makes it more challenging to achieve truthfulness in design an online mechanism.

Another open problem is whether it is possible to do better than the worst-case analysis. The optimality of our mechanism is based on the worst-case cost-data correlation. When the designer can gradually learn the cost-data correlation, is it possible to adjust the purchasing mechanism accordingly so that it achieves optimality with respect to the actual cost-data correlation?

More generally, it would be interesting to develop mechanisms for other more complicated statistical estimation tasks, such as hypothesis testing.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CCF-1718549. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Jacob Abernethy, Yiling Chen, Chien-Ju Ho, and Bo Waggoner. 2015. Low-Cost Learning via Active Data Procurement. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation (EC '15).* ACM, New York, NY, USA, 619–636. https://doi.org/10.1145/2764468.2764519

[2] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization.* Cambridge University Press, New York, NY, USA.

[3] Y. Cai, C. Daskalakis, and C. H. Papadimitriou. 2015. Optimum Statistical Estimation with Strategic Data Sources. In *Proceedings of the 28th Conference on Learning Theory.* 280–296.

[4] Yiling Chen, Nicole Immorlica, Brendan Lucier, Vasilis Syrgkanis, and Juba Ziani. 2018. Optimal Data Acquisition for Statistical Estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC '18).* ACM, New York, NY, USA, 27–44.

[5] Yiling Chen, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah. 2018. Strategyproof Linear Regression in High Dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC '18).* ACM, New York, NY, USA, 9–26. https://doi.org/10.1145/3219166.3219175

[6] Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. 2015. Truthful Linear Regression. In *Proceedings of the 28th Conference on Learning Theory.* 448–483.

[7] O. Dekel, F. Fischer, and A. D. Procaccia. 2010. Incentive Compatible Regression Learning. *J. Comput. System Sci.* 76, 8 (2010), 759–777.

[8] Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. 2017. Strategic Classification from Revealed Preferences. (2017).

[9] Lisa Fleischer and Yu-Han Lyu. 2012. Approximately Optimal Auctions for Selling Privacy when Costs are Correlated with Data. *CoRR* abs/1204.4031 (2012). http://arxiv.org/abs/1204.4031

[10] Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. 2014. Buying Private Data without Verification. *CoRR* abs/1404.6003 (2014). http://arxiv.org/abs/1404.6003

[11] Arpita Ghosh and Aaron Roth. 2011. Selling Privacy at Auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC '11).* ACM, New York, NY, USA, 199–208. https://doi.org/10.1145/1993574.1993605

[12] John Gurland and Ram C. Tripathi. 1971. A Simple Approximation for Unbiased Estimation of the Standard Deviation. *The American Statistician* 25, 4 (1971), 30–32. http://www.jstor.org/stable/2682923

[13] M. Hardt, N. Megiddo, C. H. Papadimitriou, and M. Wootters. 2016. Strategic Classification. In *7th.* 111–122.

[14] Yang Liu and Yiling Chen. 2016. Learning to incentivize: Eliciting effort via output agreement. *IJCAI* (2016).

[15] Yang Liu and Yiling Chen. 2017. Sequential Peer Prediction: Learning to Elicit Effort using Posted Prices.. In *AAAI.*

[16] Yang Liu and Yiling Chen. 2018. Surrogate Scoring Rules and a Dominant Truth Serum for Information Elicitation. *CoRR* abs/1802.09158 (2018). arXiv:1802.09158 http://arxiv.org/abs/1802.09158

[17] Andreas Maurer and Massimiliano Pontil. 2009. Empirical Bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740* (2009).

[18] R. Meir, S. Almagor, A. Michaely, and J. S. Rosenschein. 2011. Tight bounds for strategyproof classification. In *10th.* 319–326.

[19] R. Meir, A. D. Procaccia, and J. S. Rosenschein. 2012. Algorithms for Strategyproof Classification. *Artificial Intelligence* 186 (2012), 123–156.

[20] N. Miller, P. Resnick, and R. Zeckhauser. 2005. Eliciting informative feedback: The peer-prediction method. *Management Science* 51, 9 (2005), 1359–1373.

[21] Roger B Myerson. 1981. Optimal auction design. *Mathematics of operations research* 6, 1 (1981), 58–73.

[22] Roger B Myerson and Mark A Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of economic theory* 29, 2 (1983), 265–281.

[23] Kobbi Nissim, Salil Vadhan, and David Xiao. 2014. Redrawing the Boundaries on Purchasing Data from Privacy-sensitive Individuals. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science (ITCS '14).* ACM, New York, NY, USA, 411–422. https://doi.org/10.1145/2554797.2554835

[24] J. Perote and J. Perote-Peña. 2003. The impossibility of strategy-proof clustering. *Economics Bulletin* 4, 23 (2003), 1–9.

[25] J. Perote and J. Perote-Peña. 2004. Strategy-proof estimators for simple regression. *Mathematical Social Sciences* 47 (2004), 153–176.

[26] Aaron Roth and Grant Schoenebeck. 2012. Conducting Truthful Surveys, Cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12).*

[27] Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C. Parkes. 2016. Informed Truthfulness in Multi-Task Peer Prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation (EC '16).*