A Framework for Sample Efficient Interval Estimation with Control Variates

Shengjia Zhao
Stanford University

Christopher Yeh
Stanford University

Stefano Ermon
Stanford University

Abstract

We consider the problem of estimating confidence intervals for the mean of a random variable, where the goal is to produce the smallest possible interval for a given number of samples. While minimax optimal algorithms are known for this problem in the general case, improved performance is possible under additional assumptions. In particular, we design an estimation algorithm to take advantage of side information in the form of a control variate, leveraging order statistics. Under certain conditions on the quality of the control variates, we show improved asymptotic efficiency compared to existing estimation algorithms. Empirically, we demonstrate superior performance on several real world surveying and estimation tasks where we use the output of regression models as the control variates.

1 Introduction

Many real world problems require estimation of the mean of a random variable from unbiased samples. In high risk applications, the estimation algorithm should output a confidence interval, with the guarantee that the true mean belongs to the interval with e.g. 99% probability. The classic tools for this task are concentration inequalities, such as the Chernoff, Bernstein, or Chebychev inequalities (Hoeffding, 1962; Bernstein, 1924; Vershynin, 2010).

However, for many tasks, obtaining unbiased samples is expensive. For example, when estimating demographic quantities, such as income or political preference, drawing unbiased samples can require field survey. To make the situation worse, often we seek more granular estimates, e.g. for each individual district or county, meaning that we need to draw a sufficient number of samples for every district or county. Another difficulty can arise when we need high accuracy (i.e. a small confidence interval), since confidence intervals produced by typical concentration inequalities have size $O(1/\sqrt{\text{number of samples}})$. In other words, to reduce the size of a confidence interval by a factor of 10, we need 100 times more samples.

This dilemma is generally unavoidable because concentration inequalities such as Chernoff or Chebychev are minimax optimal: there exist distributions for which these inequalities cannot be improved. No-free-lunch results (Van der Vaart, 2000) imply that any alternative estimation algorithm that performs better (i.e. outputs a confidence interval with smaller size) on some problems, must perform worse on other problems. Nevertheless, we can identify a subset of problems where better estimation algorithms are possible.

We consider the class of problems where we have side information. We formalize side information as a random variable with known expectation and whose value is close (with high probability) to the random variable we want to estimate. Following the terminology in the Monte Carlo simulation literature, we call this side information random variable a "control variate" (Lemieux, 2017). Instead of the original estimation task, we can estimate the expected difference between the original random variable and the control variate. The hope is that the distribution of this difference is concentrated around 0. Some estimation algorithms output very good (small sized) confidence intervals for distributions concentrated around 0 compared to classic methods such as Chernoff bounds.

Many practical problems have very good control variates. One important class of problems is when we have a predictor for the random variable, and we can use the prediction as the control variate. For example, if we would like to estimate a neighborhood's average voting pattern, then we might have a prediction function for political preference from Google Street View images of that neighborhood (Gebru et al., 2017); if we would like

Proceedings of the 23$^{\text{rd}}$International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

to estimate the average asset wealth of households in a certain geographic region, we might have a regressor that predicts income from satellite images (Jean et al., 2016). These classifiers could be trained on past data (e.g. previous year survey results) or similar datasets, or they could even be crafted by hand.

For these problems we propose concentration bounds based on order statistics. In particular, we draw a connection between the recently proposed concept of *resilience* (Steinhardt, 2018) and concentration bounds on order statistics. We show how to use these concentration inequalities to design estimation algorithms that output better confidence intervals when we have a good control variate (e.g. output confidence intervals of size $O(1/\text{number of samples})$).

Our proposed estimation algorithm always produces valid confidence intervals, i.e. the true mean belongs to the interval with a specified probability. The only risk is that when the control variate is poor, the confidence interval could be worse (larger) than classic baselines such as Chernoff inequalities. We empirically show superior performance of the proposed estimation algorithm on three real world tasks: bounding regression error, estimating average wealth with satellite images, and estimating the covariance between wealth and education level.

2 Problem Setup

Our objective is to estimate the mean of some random variable Y taking values in $\mathcal{Y} \subseteq \mathbb{R}^d$. Given i.i.d. samples $y_1, \dots, y_m \sim Y$ and some choice of confidence level $\zeta \in (0,1)$, an estimation algorithm outputs $\hat{\mu} \in \mathbb{R}^d$ and an confidence interval size $c \in \mathbb{R}^+$. The estimation algorithm must satisfy

$$\Pr\left[\|\hat{\mu} - \mathbb{E}[Y]\| > c\right] \leq \zeta \tag{1}$$

where the probability Pr is with respect to the random samples of $y_1, \dots, y_m$ and any additional randomness in the execution of the (randomized) algorithm, and $\|\hat{\mu} - \mathbb{E}[Y]\|$ is any choice of semi-norm.

We will first focus on one dimensional problems where $Y \in \mathbb{R}$, and choose the norm $\|\hat{\mu} - \mathbb{E}[Y]\| = |\hat{\mu} - \mathbb{E}[Y]|$, then extend several results to more general setups.

2.1 Baseline Estimators and Optimality

The classical approach is to estimate $\mathbb{E}[Y]$ with the empirical mean $\mu_{\text{mean}} = \frac{1}{m}\sum_{i=1}^m y_i$, and its estimation error $|\mu_{\text{mean}} - \mathbb{E}[Y]|$ can be controlled using concentration inequalities.

To obtain concentration inequalities, we need some assumptions on $\mathcal{Y}$. For example, if there is some very small probability that $Y = +\infty$, then $\mathbb{E}[Y]$ is unbounded, but μ_{mean} can be finite, which makes it impossible to bound $|\mu_{\text{mean}} - \mathbb{E}[Y]|$. Common assumptions include sub-Gaussian, bounded moments, sub-exponential, etc (Vershynin, 2010). We will consider the sub-Gaussian and bounded moments assumptions in this paper.

Sub-Gaussian: If we assume $\forall t \in \mathbb{R}$, $\mathbb{E}[e^{t(Y - \mathbb{E}[Y])}] \leq e^{\sigma^2 t^2/2}$, then Chernoff-Hoeffding is the classic concentration inequality for confidence interval estimation (Hoeffding, 1962)

$$\Pr\left[|\hat{\mu} - \mathbb{E}[Y]| > \sqrt{\frac{2\sigma^2}{m}\log\frac{2}{\zeta}}\right] \leq \zeta.$$

In particular, when Y is supported on $[a,b]$ we have

$$\Pr\left[|\hat{\mu} - \mathbb{E}[Y]| > \sqrt{\frac{(b-a)^2}{2m}\log\frac{2}{\zeta}}\right] \leq \zeta.$$

Bounded Moments: Another common assumption is that the k-th order moment of Y is bounded, i.e. $\mathbb{E}[|Y - \mathbb{E}[Y]|^k] \leq \sigma^k$ for some $\sigma > 0$.

Under bounded moment assumptions, several concentration inequalities are known, such as the Chebychev inequality, Kolmogorov inequality (Hájek and Rényi, 1955) and Bernstein inequality (Bernstein, 1924). For example, when Y has a bounded second order moment, the Chebychev inequality states the following:

$$\Pr\left[|\hat{\mu} - \mathbb{E}[Y]| \geq \frac{\sigma}{\sqrt{\zeta m}}\right] \leq \zeta \tag{2}$$

It is known that these bounds are (asymptotically in m) minimax optimal. There exist random variables Y that satisfy the respective assumptions of each inequality, and the inequality cannot be improved (Hoeffding, 1962). Therefore, to further improve these estimation algorithms, additional assumptions will be necessary.

2.2 Control Variates

Suppose $\tilde{Y}$ is another random variable jointly distributed with Y and that we know its mean $\mathbb{E}[\tilde{Y}]$ (or have a very accurate estimate of it). We also have samples drawn from their joint distribution $(y_1, \tilde{y}_1), \dots, (y_m, \tilde{y}_m) \sim Y, \tilde{Y}$.

For $\tilde{Y}$ to be useful for our application, its value needs to be close to Y. In other words, $Y - \tilde{Y}$ should be a random variable that is concentrated around 0. The purpose of this variable $\tilde{Y}$ is similar to a "control variate" in the Monte Carlo community, but we use it here for a different task of interval estimation.

For example, in our household wealth estimation example, let Y denote the household-level wealth in a randomly sampled village, and let $\tilde{Y}$ be the predicted household-level wealth based on the satellite image of that village. If our predictor is accurate (i.e. $Y \approx \tilde{Y}$ with high probability), then $\tilde{Y}$ could be an effective control variate for Y. In addition, we can also estimate $\mathbb{E}[\tilde{Y}]$ very accurately because a very large number of satellite images are available with little cost (Wulder et al., 2012), so obtaining samples from $\tilde{Y}$ (without the corresponding Y) is inexpensive.

We would like to design an estimation algorithm that takes as input the samples $(y_1, \tilde{y}_1), \ldots, (y_m, \tilde{y}_m)$ and the known value of $\mathbb{E}[\tilde{Y}]$, and outputs a $\hat{\mu}$ along with an estimation error $c \in \mathbb{R}^+$ such that Eq.(1) holds. The hope is that because of the additional information $\tilde{Y}$ we are no longer restricted by the mini-max bounds on the confidence interval size c, and can output confidence intervals of smaller size.

3 Control Variates Interval Estimation

3.1 Estimation by Order Statistics

For convenience we denote $Z = Y - \tilde{Y}$ and $z_i = y_i - \tilde{y}_i$. Given the samples $z_1, \ldots, z_m$ we can compute their order statistics, which are the m samples in sorted order such that $Z_{(1)} \leq \cdots \leq Z_{(m)}$.

We first study the one dimensional real valued case (i.e. $Z \in \mathcal{Z} = \mathbb{R}$).

Our control variate estimation algorithm consists of three steps: given input samples $z_1, \ldots, z_m$, $\mathbb{E}[\tilde{Y}] \in \mathbb{R}$ and desired confidence $\zeta \in (0, 1)$

1) Choose some value of $k \in \{0, \ldots, m-1\}$. Set

$$\hat{\mu} = \mathbb{E}[\tilde{Y}] + \frac{Z_{(m-k)} + Z_{(1+k)}}{2} \tag{3}$$

2) Let r be the smallest value such that the following concentration inequalities are true:

$$\begin{aligned} \Pr\left[Z_{(m-k)} < \mathbb{E}[Z] - r\right] &\leq \zeta/2 \\ \Pr\left[Z_{(1+k)} > \mathbb{E}[Z] + r\right] &\leq \zeta/2 \end{aligned} \tag{4}$$

Algorithm 1 computes such a value of r.

3) Output $\hat{\mu}$ as the estimate of $\mathbb{E}[Y]$ and $r + \frac{Z_{(m-k)} - Z_{(1+k)}}{2}$ as the confidence interval size.

The following proposition guarantees the correctness of the control variate estimation algorithm.

Proposition 1. *Let $\hat{\mu}$ be defined as in Eq.(3). If Eq.(4) is satisfied for some $\zeta \in (0, 1)$ and $c > 0$, then*

$$\Pr\left[|\hat{\mu} - \mathbb{E}[Y]| > r + \frac{Z_{(m-k)} - Z_{(1+k)}}{2}\right] \leq \zeta. \tag{5}$$

Proof of Proposition 1. See Appendix. □

Note that in Eq.(5), the confidence interval size consists of two parts: $\frac{Z_{(m-k)} - Z_{(1+k)}}{2}$ and c. We will show that c is much smaller (in fact, has asymptotically better rates) compared to baselines previously discussed in Section 2.1. If we have a good control variate $\tilde{Y}$ (close to Y), then $\frac{Z_{(m-k)} - Z_{(1+k)}}{2}$ will be small. On the other hand, if the control variate $\tilde{Y}$ is poor, this algorithm could produce worse (larger) confidence intervals. This trade-off is unavoidable because of “no-free-lunch results” (Van der Vaart, 2000).

3.2 Order Statistics Concentration Inequalities

We will now show several bounds on the order statistics in Eq.(4). There are several known bounds (De Haan and Ferreira, 2007; Castillo, 2012) that can be applied to satisfy Eq.(4). However, these bounds are distribution dependent: we must assume that Z either is distributed as some known distribution (e.g. Gaussian), or asymptotically converges to some fixed distribution (e.g. Gumbel, Weibull or Frechet using the Fisher–Tippett–Gnedenko theorem) when m is very large (Fisher and Tippett, 1928).

Our contribution is to associate bounds on the order statistics with the notion of resilience used in the robust statistics literature. Many distribution independent conditions for resilience are known, which imply *distribution independent bounds on the order statistics*.

We first reproduce the definition of resilience from Steinhardt (2018) with a small modification.

Definition 1. *Let $s : [0, 1] \to \mathbb{R}^+$ be any function. We say a random variable $Z \in \mathcal{Z}$ is s-resilient from above if for any $\epsilon \in [0, 1]$ and measurable $B \subseteq \mathcal{Z}$ such that $\Pr[Z \in B] \geq 1 - \epsilon$, we have*

$$\mathbb{E}[Z | Z \in B] - \mathbb{E}[Z] \leq s(\epsilon).$$

It is s-resilient from below if

$$\mathbb{E}[Z] - \mathbb{E}[Z | Z \in B] \leq s(\epsilon).$$

We say that Z is s-resilient if it is both s-resilient from above and s-resilient from below.

When a random variable is s-resilient, we are essentially bounding the probability that Z takes a value much larger (or smaller) than $\mathbb{E}[Z]$. To see this, suppose for some $\epsilon \in (0,1)$, we choose $B = \{Z : Z > \mathbb{E}[Z] + s(\epsilon)\}$, then $\mathbb{E}[Z|Z \in B] - \mathbb{E}[Z] > s(\epsilon)$, then by our requirement of resilience, it must be that $\Pr[Z \in B] < 1-\epsilon$. Similar to other types of assumptions such as sub-Gaussian or bounded moments, resilience is also an assumption on how much a random variable can differ from its expectation. What is special about resilience is that it is particularly convenient for showing bounds on the order statistics.

Without loss of generality, we also assume that $s(\epsilon)$ is monotonically non-decreasing. If a random variable is $\hat{s}$-resilient, but $\hat{s}$ is not monotonically non-decreasing, we can define $s(\epsilon) = \inf_{\epsilon \leq \epsilon' \leq 1} \hat{s}(\epsilon')$, which is monotonically non-decreasing. It is easy to show that Z is s-resilient if and only if it is $\hat{s}$-resilient.

Many assumptions on the random variable Z can be converted into assumptions on resilience. For example, in the following Lemma 1, (1) is trivial to show, (2) is proved in (Steinhardt, 2018), and (3) we prove in the appendix.

Lemma 1. *The following random variables are resilient:*

1. ***Bounded:*** *If $\mathcal{Z} \subseteq [a,b]$, then Z is $(b - \mathbb{E}[Z])$-resilient from above and $(\mathbb{E}[Z] - a)$-resilient from below. It is $(b-a)$-resilient.*
2. ***Bounded Moments:*** *If $\mathbb{E}\left[|Z - \mathbb{E}[Z]|^l\right] \leq \sigma^l$ for some $l > 1$, then Z is s-resilient for $s(\epsilon) = \frac{\sigma}{(1-\epsilon)^{1/l}}$.*
3. ***Sub-Gaussian:*** *If $Z - \mathbb{E}[Z]$ is σ^2 sub-Gaussian, then Z is s-resilient for $s(\epsilon) = \sqrt{2\sigma \log \frac{1}{1-\epsilon}} + \sqrt{2\pi}\sigma$.*

Given the definition and examples of resilience, we can now use the following theorem to provide bounds on the order statistics when resilience holds.

Theorem 1. *Let $Z_{(1)}, \ldots, Z_{(m)}$ denote the order statistics of m independent samples of a random variable Z. If Z is s-resilient, $\forall \zeta \in (0,1), T \in \mathbb{N}$ and $0 \leq k < m/2$, then letting r be the output of Algorithm 1, we have*

$$\Pr\left[Z_{(m-k)} \leq \mathbb{E}[Z] - r\right] \leq \zeta \qquad (6)$$

$$\Pr\left[Z_{(1+k)} \geq \mathbb{E}[Z] + r\right] \leq \zeta \qquad (7)$$

If Z is s-resilient from above/below, then only Eq.(6)/Eq.(7) holds.

Proof of Theorem 1. See Appendix. □

Note that the algorithm has an additional hyperparameter T which is the number of iterations. We can choose any value of $T \in \mathbb{N}$, but choosing a large T always leads to a better confidence interval compared to choosing a smaller T. We observe in the experiments that choosing any $T \geq 10$ is optimal (up to floating point errors). Note that the choice of T does not affect the asymptotic performance in Corollary 1.

Algorithm 1 Order Statistics Confidence Interval

```
Input: sample size $m \in \mathbb{N}^+$; order $k < m/2$; confidence $\zeta \in (0,1)$; resilience $s : (0,1) \to \mathbb{R}^+$; and number of iterations $T \in \mathbb{N}$
  Set $v_0 = \left(\frac{\zeta}{(m+1)^k}\right)^{\frac{1}{m-k}}$
  for $i = 1, \ldots, T$ do
    Set $v_i = \left(\frac{\zeta}{(m(1-v_{i-1})+1)^k}\right)^{\frac{1}{m-k}}$
  end for
  Set $r = s(v_T)(v_T^{-1} - 1)$
  Return $v_T$ and $r$
```

Theorem 1 involves expressions with good constants. This obscures the asymptotic performance of the bound, which we reveal in the following corollary.

Corollary 1. *If Z is $b_1(1-\epsilon)^a + b_2$ resilient for any constants $a \in [-1,0]$ and $b_1, b_2 \in \mathbb{R}^+$, then there exists $\lambda > 0$ such that for sufficiently large m*

$$\Pr\left[\mathbb{E}[Z] \leq Z_{(1+k)} - \lambda \frac{\log\frac{1}{\zeta} + k \log m}{m^{1+a}}\right] \leq \zeta$$

$$\Pr\left[\mathbb{E}[Z] \geq Z_{(m-k)} + \lambda \frac{\log\frac{1}{\zeta} + k \log m}{m^{1+a}}\right] \leq \zeta.$$

Proof of Corollary 1. See Appendix. □

Based on different assumptions on resilience in Lemma 1, we can further simplify Corollary 1 and obtain more concrete bounds in Section 3.4.

Finally, in Lemma 1, a random variable Z bounded in $[a,b]$ is s-resilient, but s depends on $\mathbb{E}[Z]$ which is unknown. Therefore, we cannot evaluate s in Algorithm 1. One option is to use the weaker conclusion that Z is $(b-a)$-resilient and obtain a bound with worse constants (we choose this option in our asymptotic rate analysis in Section 3.4).

In practice it is possible to compute a bound with better constants: we only compute v_T in Algorithm 1 which does not depend on s, and use the following improved bound.

Corollary 2. *Let $Z_{(1)}, \ldots, Z_{(m)}$ denote the order statistics of m independent samples of a random variable Z. If Z is bounded in $[a,b]$, $\forall \zeta \in (0,1), T \in \mathbb{N}$ and $0 \leq k < m/2$, then letting v_T be computed as in*

Algorithm 1, we have

$$\Pr\left[\mathbb{E}[Z] \le a + v_T(Z_{(1+k)} - a)\right] \le \zeta$$
$$\Pr\left[\mathbb{E}[Z] \ge b - v_T(b - Z_{(m-k)})\right] \le \zeta.$$

Proof of Corollary 2. See Appendix. □

Instead of the standard estimation procedure in Section 2.2, we directly output

$$\left[a + v_T(Z_{(1+k)} - a),\ b - v_T(b - Z_{(m-k)})\right]$$

as our confidence interval for $\mathbb{E}[Z]$ (that holds with $1 - 2\zeta$ probability) in the experiments.

3.3 Multi-Dimensional Extension

We will extend the above results to multi-dimensional estimation problems. We first provide a multi-dimensional definition of resilience that extends Definition 1.

Definition 2. *Let Z be a random variable on $\mathcal{Z} \subseteq \mathbb{R}^d$, and $\|\cdot\|, \|\cdot\|_*$ be a pair of dual norms on $\mathbb{R}^d$. We say Z is s-resilient if $\forall v \in \mathbb{R}^d$ such that $\|v\|_* = 1$, and for all measurable $B \subseteq \mathcal{Z}$ such that $\Pr[B] \ge 1 - \epsilon$, we have*

$$\mathbb{E}[\langle Z, v\rangle | B] - \mathbb{E}[\langle Z, v\rangle] \le s(\epsilon).$$

As before, resilience in multiple dimensions also implies concentration bounds on the order statistics. The following theorem is the analog of Theorem 1.

Theorem 2. *Let $Z_{(1)}, \ldots, Z_{(m)}$ be independent samples of Z ordered such that $\|Z_{(1)}\| \le \cdots \le \|Z_{(m)}\|$. If Z is s-resilient, then for any r output by Algorithm 1 and for any $\zeta \in (0, 1)$, we have*

$$\Pr\left[\|Z_{(m-k)}\| \le \|\mathbb{E}[Z]\| - r\right] \le \zeta.$$

Proof of Theorem 2. See Appendix. □

3.4 Rate Comparison

Table 1 summarizes the asymptotic performance of our method compared to the baselines. Even though we consider the low sample setup (i.e. m is small), these asymptotic rates still provide insight into the trade-off between different methods.

In particular, as shown in Proposition 1 our method has two terms that determine the confidence interval: r and $\frac{Z_{(m-k)} - Z_{(1+k)}}{2}$.

$$\Pr\left[|\hat{\mu} - \mathbb{E}[Y]| > r + \frac{Z_{(m-k)} - Z_{(1+k)}}{2}\right] \le \zeta.$$

The latter term we will denote as B and it is determined by the quality of the control variate. For baseline methods there is only a single term c that determines the size of the confidence interval:

$$\Pr\left[|\hat{\mu} - \mathbb{E}[Y]| > c\right] \le \zeta.$$

In Table 1, we show that under each class of assumptions, r in our proposed algorithm always has a better rate compared to c (i.e. it is smaller when m, ζ are sufficiently large). Whether the improvement can justify the additional term B determines whether our algorithm performs well in practice.

4 Related Work

Extreme value theory (De Haan and Ferreira, 2007; Castillo, 2012) studies the probability of rare events or large deviation. Most results are asymptotic (De Haan and Ferreira, 2007) and are not applicable to our setup assuming small sample size. Several non-asymptotic results are also used in our proofs such as Eq.(8).

The notion of resilience (Steinhardt et al., 2017; Steinhardt, 2018) is most commonly used in analyzing robust estimation. Our paper draws the connection between resilience and order statistics concentration bounds. We hope this connection can be further exploited in future work to transfer results between the two fields of research.

A line of research related to ours is semi-supervised transfer learning and domain adaptation (Daumé III et al., 2010; Donahue et al., 2013; Kumar et al., 2010; Ding et al., 2018; Lopez-paz et al., 2012; Saito et al., 2019). In both setups, we have a pretrained classifier or regressor; in the target domain, there is a small amount of labeled data. The difference is in the objective: domain adaptation use the labeled data to fine-tune our classifier or regressor, while our objective is confidence interval estimation on the target domain. The different objectives lead to different sets of tools and desiderata.

5 Experiments

5.1 Certifying Regression Performance

Our first task is to upper and lower bound the difference between the output of a regression function $\tilde{Y}$ and the true target attribute Y. For this task, our goal is to bound the expected error $Z = Y - \tilde{Y}$ of the regression function. We are not directly interested in the expected value of the target attribute $\mathbb{E}[Y]$; instead we only want to show bounds $\text{LB} \le \mathbb{E}[Z] \le \text{UB}$. In addition, instead of a single global accuracy, we might care about accuracy for sub-groups in the data (e.g., based on some feature or sensitive attribute). In other words, let U be some random variable taking a finite set of values, we want to know the regression error

Conditions on Z	Bounded in $[a, b]$	Finite $\mathbb{E}[Z^2]$
Mean $\frac{1}{m}\sum_i Z_i$	$\Theta\left(\sqrt{\frac{1}{m}\log\frac{1}{\zeta}}\right)$ (Chernoff)	$\Theta\left(\frac{1}{\sqrt{m\zeta}}\right)$ (Chebyshev)
Maximum (Minimum) $Z_{(m)}, Z_{(1)}$	$B + O\left(\frac{1}{m}\log\frac{1}{\zeta}\right)$	$B + O\left(\frac{1}{\sqrt{m}}\log\frac{1}{\zeta}\right)$
k-th largest (smallest) $Z_{(m-k)}, Z_{(1+k)}$	$B + O\left(\frac{k\log m}{m}\log\frac{1}{\zeta}\right)$	$B + O\left(\frac{k\log m}{\sqrt{m}}\log\frac{1}{\zeta}\right)$

Table 1: Summary of asymptotic size of the confidence interval for estimation algorithms using different concentration inequalities. B is some value that corresponds to the bias of our method because of the $\frac{Z_{(m-k)}-Z_{(1+k)}}{2}$ term in Proposition 1.

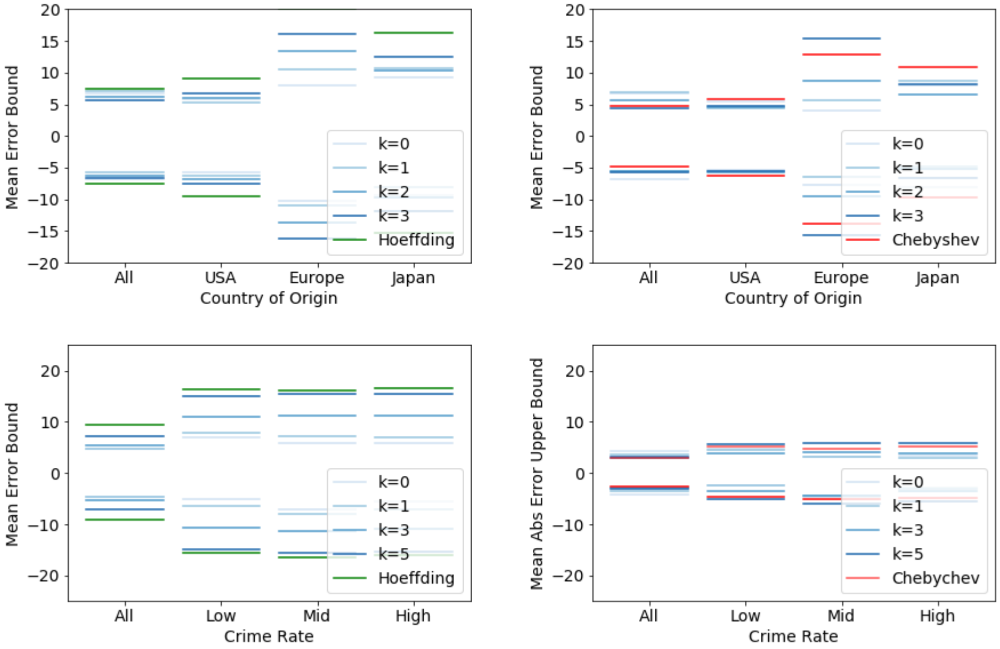

Figure 1: Confidence intervals (both upper and lower bounds) on $\mathbb{E}[Z]$ from different estimation algorithms. For our algorithm, we try different values of k (the k-th largest / smallest). For the MPG dataset, we also evaluate the regression error conditioned on different country of origin (top). For housing dataset, we evaluate the regression error conditioned on different crime rate level (bottom). In most of the experiments, our estimation algorithm performs better (outputs smaller confidence intervals). Chebychev sometimes performs better, especially with large sample size. Hoeffding always performs worse in this setup.

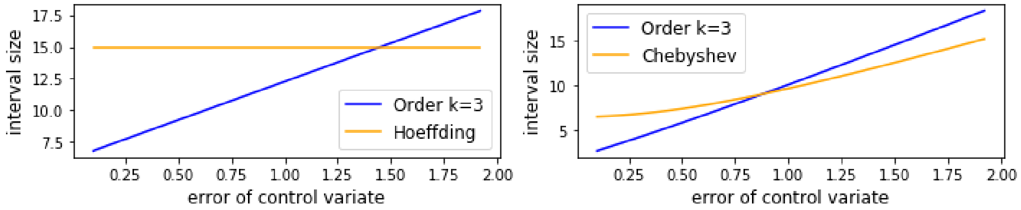

Figure 2: Size of the confidence interval as a function of the error of the control variate. We scale up (>1.0) or scaled down (<1.0) the error (i.e. $Z_{\text{new}} = \alpha Z_{\text{old}}$ for $\alpha \in [0, 2]$). When the control variate has smaller error the confidence interval is significantly better compared to Hoeffding or Chebychev.

$\mathbb{E}[Z|U = u]$ for each value of $U = u$. This can be important for fairness or identifying particular failure cases.

If the regression function is accurate, Z should be concentrated around 0, making it feasible to obtain better bounds with order statistics. We compare the bounds of Section 3.2 with the baseline bounds of Section 2.1. Code is available at https://github.com/ermongroup/ControlVariateBound

Datasets: We use two classic regression datasets in the UCI repository (Asuncion and Newman, 2007): Auto MPG, where the task is to predict the miles per gallon (MPG) of a vehicle based on 10 features; and Boston housing price prediction, where the task is to predict the housing price from 13 features.

Assumptions: As explained in Section 2.1, all estimation algorithms require some assumptions. Here, we either assume bounded support or bounded variance. Optimal choice of these assumptions usually relies on domain knowledge about the problem and is beyond the scope of this paper. Here, we simply assume that the error is bounded by $\pm b/2$ where b is the maximum MPG in the entire dataset. The reason is that any regression algorithm can trivially output $b/2$ and achieve this error. In the bounded variance case, we first compute an upper bound on $\mathbb{E}[Z^2]$ that holds with $1 - \zeta/2$ probability by Hoeffding inequality as an upper bound on the variance.

Results: The results are shown in Figure 1. Our order statistics bound works better in general if the number of test samples m is small. Here both datasets contain approximately 100 test samples, and our bound performs on-par with Chebychev and better than Hoeffding. Our bound also performs better when the control variate is more accurate (i.e. Z is concentrated around zero). This is empirically verified in Figure 2.

Our bound also depends on the choice of k. In general, with more data choosing a larger value of k is preferable, and vice versa.

5.2 Poverty Estimation Task

We apply our estimation algorithm on a real world task, where we estimate the average household-level asset wealth across provinces of countries in 23 sub-Saharan African countries. We used DHS Survey collected between 2009-16 and constructed an average household asset wealth index for 19,669 villages following the procedure described in Jean et al. (2016).

Setup: We emulate the setup where we have survey results from several countries, and train a regressor (a convolutional neural network) to predict asset wealth from satellite images (multispectral Landsat and nighttime VIIRS). To estimate the average asset wealth for a new country, we apply this regressor to satellite images from that country; we use the output of the regressor as a control variate. More specifically, we randomly pick 80% of the countries to train the regressor, and test the performance of our estimation algorithm on the remaining countries. We also use cross validation to more accurately evaluate our performance.

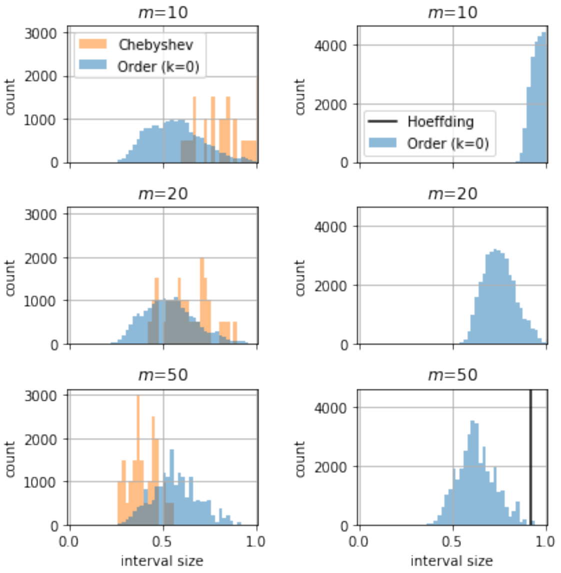

Figure 3: Histograms of 99%-confidence interval sizes for the average household-level asset wealth within a province, for 36 provinces across 13 countries in DHS surveys of sub-Saharan Africa. 1000 random subsets of size m are sampled from each province. We assume that household-level asset wealth is a random variable either with finite variance (left), or bounded in $[0, 1]$ (right). The histograms show interval sizes pooled over all 36 test provinces. For small sample size per province (i.e. <20 samples per province) our method achieves smaller confidence intervals. With more samples, the bias of our estimation algorithm dominates and our method performs worse.

Assumptions: As in Section 5.1, for estimation with bounded random variables, we upper- and lower-bound household wealth by the maximum and minimum wealth across the entire dataset. For estimation with bounded moment random variables, we use the empirical standard deviation estimated across the entire country, multiplied by an additional margin of 1.5×. Because of the small sample size, Chernoff bounding the standard deviation as in Section 5.1 is infeasible.

Results: The results are shown in Figure 3. Although our regression model is trained on all 23 countries in our dataset, we only test our method on the 36 provinces across 13 countries from which we have at least 90 labeled survey examples. Compared to Chernoff bound or Chebychev our estimation algorithm perform better when sample size is small, and worse when sample size is large. Because of the difficulty of predicting wealth from satellite images, the control variate is not very accurate, and further improvements are possible with improved prediction accuracy.

5.3 Covariance Estimation

The DHS surveys also include other demographic variables besides household asset wealth. Policy makers may be interested in the covariance of these demographic quantities, such as between maternal education level and household asset wealth.

More formally let W be the random variable that represents the average level of maternal education in a village, and U represent the average household asset wealth in the same village. The random variable we actually want to estimate is $Y = UW - U\mathbb{E}[W]$ because

$$\mathbb{E}[Y] = \mathbb{E}[UW] - \mathbb{E}[U]\mathbb{E}[W] = \mathrm{Cov}(U, W)$$

We assume that W is a quantity that is easy to survey, or has available data, so $\mathbb{E}[W]$ is known. This is commonly the case, when certain demographic variables are more widely surveyed than others. As before, we can train a regressor to predict U from satellite images, and we denote it's output as $\tilde{U}$. Our control variate is then $\tilde{Y} = \tilde{U}W - \tilde{U}\mathbb{E}[W]$. By using these new definitions for Y and $\tilde{Y}$ we can apply our estimation algorithm.

Setup: The setup is identical to Section 5.2, where we train the asset wealth regression model on 80% of the countries and test our estimation algorithm on the remaining countries. However, we only have maternal education survey data on 9 countries, so we only estimate confidence intervals of covariance in these countries. Maternal education level is measured at each household on an integer scale from 0 to 3, then averaged within each village. All the other assumptions are also identical to before.

Result: The results are shown in Figure 4. As expected, our estimation algorithm achieves superior performance compared to baseline estimators when the sample size m is small.

6 Conclusion

In this paper we propose a framework for estimating the confidence interval given a control variate random variable as side information. We show that under certain conditions on the control variate, the estimation algorithms out-performs classic minimax optimal estimation algorithms both asymptotically and empirically.

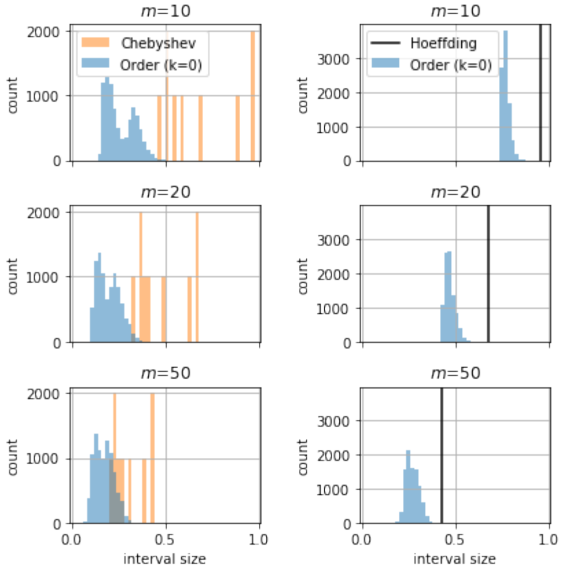

Figure 4: Histograms of confidence interval sizes at $\zeta = 0.01$ for the covariance between maternal education and household asset wealth, for 9 countries in DHS surveys of sub-Saharan Africa. 1000 random subsets of size m are sampled from each province. The confidence interval derived from order statistics outperforms both the Hoeffding interval and the Chebychev interval.

A major weakness of the estimator is diminished performance when we have a large number of samples. Because of no-free-lunch results, trade-offs are unavoidable, but it is an interesting direction of future research to find either better trade-offs or prove its impossibility.

7 Acknowledgments

This research was supported by TRI, Intel, NSF (#1651565, #1522054, #1733686), and ONR. The survey datasets are generously provided by Marshall Burke at Stanford University.

References

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. 1962.

Sergei Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Christiane Lemieux. *Control Variates*, pages 1–8. American Cancer Society, 2017. ISBN 9781118445112.

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113, 2017.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Jacob Steinhardt. *Robust Learning: Information Theory and Algorithms.* PhD thesis, Stanford University, 2018.

J Hájek and A Rényi. Generalization of an inequality of kolmogorov. *Acta Mathematica Hungarica*, 6(3-4): 281–283, 1955.

Michael A. Wulder, Jeffrey G. Masek, Warren B. Cohen, Thomas R. Loveland, and Curtis E. Woodcock. Opening the archive: How free data has enabled the science and monitoring promise of landsat. *Remote Sensing of Environment*, 122:2 – 10, 2012. ISSN 0034-4257. Landsat Legacy Special Issue.

Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction.* Springer Science & Business Media, 2007.

Enrique Castillo. *Extreme value theory in engineering.* Elsevier, 2012.

Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press, 1928.

Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv preprint arXiv:1703.04940*, 2017.

Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59. Association for Computational Linguistics, 2010.

Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised domain adaptation with instance constraints. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 668–675, 2013.

Abhishek Kumar, Avishek Saha, and Hal Daume. Coregularization based semi-supervised domain adaptation. In *Advances in neural information processing systems*, pages 478–486, 2010.

Zhengming Ding, Nasser M Nasrabadi, and Yun Fu. Semi-supervised deep domain adaptation via coupled neural networks. *IEEE Transactions on Image Processing*, 27(11):5214–5224, 2018.

David Lopez-paz, Jose M. Hernández-lobato, and Bernhard Schölkopf. Semi-Supervised Domain Adaptation with Non-Parametric Copulas. In *Advances in Neural Information Processing Systems 25*, pages 665–673, December 2012.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised Domain Adaptation via Minimax Entropy. In *2019 IEEE International Conference on Computer Vision (ICCV)*, April 2019.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.