7-31-2020

Probabilistic Error Upper Bounds For Distributed Statistical Estimation

Matthew Jin
Dartmouth College

Probabilistic Error Upper Bounds For Distributed Statistical Estimation

Dartmouth Computer Science Technical Report TR2020-903

Matthew Jin*

Advisor: Amit Chakrabarti †

July 31, 2020

Abstract

The size of modern datasets has spurred interest in distributed statistical estimation. We consider a scenario in which randomly drawn data is spread across a set of machines, and the task is to provide an estimate for the location parameter from which the data was drawn. We provide a one-shot protocol for computing this estimate which generalizes results from Braverman et al. [2], which provides a protocol under the assumption that the distribution is Gaussian, as well as from Duchi et al. [4], which assumes that the distribution is supported on the compact set $[-1, 1]$. Like that of Braverman et al., our protocol is optimal in the case that the distribution is Gaussian.

1 Problem Setup

The rapid growth in the size of modern datasets has fueled interest in distributed statistical inference; the dataset is too large to fit on a single machine, and is thus separated across multiple machines. The amount of communication between machines is often the computational bottleneck for distributed learning algorithms, so often the goal is to design algorithms which minimize the communication cost.

We consider the coordinator communication model, in which all machines communicate with a coordinator machine which does not receive any data, and do not communicate with each other. As is typical in modern machine learning contexts, we consider a large dataset drawn from some probability distribution on the real numbers which is spread across a large number of machines. We consider algorithms which determine the location parameter of the distribution from which it is drawn.

More precisely, let μ be a d-dimensional probability distribution on $\mathbb{R}^d$, and let θ be the location parameter of μ, where $\theta \in \mathbb{R}^d$ and $\forall i \in \{1, 2, ..., d\}$, $|\theta_i| \leq \text{poly}(md)$. Let the scale parameter be σ.

*matthew.g.jin.17@dartmouth.edu

†ac@dartmouth.edu

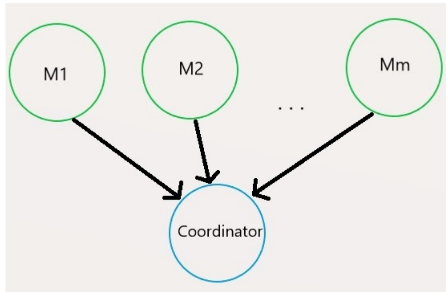


Figure 1: In the coordinator communication model, communication goes only from machines $M_1, ..., M_m$ to the machine designated as the coordinator.

For example, for the normal distribution, the location parameter is the mean and the scale parameter is the variance. For the Cauchy distribution, the location parameter is the median, and the scale parameter is denoted as such. Our goal is to compute an estimate $\hat{\theta}$ for θ, where each of m machines is given n i.i.d samples drawn from μ.

2 Existing Literature

An important special case of this problem, where μ is Gaussian, is considered at length by Braverman et al. [2], which demonstrated a lower bound and an upper bound optimal up to constant factor when the mean vector is dense. Under the same communication model, the protocol of Braverman et al. [2] has communication cost $O(md + d\log^2(mdn/\sigma))$ bits and expected mean-squared-error:

$$\mathbb{E}[|\hat{\theta} - \overline{\theta}|^2] = O\Big(\frac{1}{m}\Big).$$

In this work, we focus on generalizing the upper bound of Braveman et al. [2], providing a protocol which has similar communication cost for any distribution satisfying a certain set of conditions. Duchi et al. [4] considers both upper and lower bounds for this problem for the family $\mathcal{P}_d$ of distributions supported on the compact set $[-1, 1]^d$. Suresh et al. [8] also addresses the problem of distributed mean estimation under a limited communication budget, yet unlike this and the other mentioned works, it makes no assumption on the distribution of the dataset, and it is concerned with estimating the empirical mean of the data rather than that of the underlying statistical model. Garg et al. [5] studies the problem of d-dimensional distributed Gaussian mean estimation at the optimal minimax rate. They prove lower bounds of $\Omega(md/\log(m))$ and $\Omega(md)$ bits for achieving the minimax squared loss in the interactive and simultaneous settings respectively, and also provide an interactive protocol achieving the minimax squared loss with $O(md)$ bits of communication.

Lee et al. [6] seeks to mitigate the communication bottleneck for the problem of distributed high-dimensional sparse regression by devising a protocol that requires only a single round of communication. It does this by computing "debiased" local lasso estimators which are then averaged.

The coordinator communication model that we consider here has been studied with various applications in mind in a number of earlier works. For example, Arackaparambil et al. [1] and Cormode et al. [3] study the *functional monitoring problem*, in which k machines each receive a stream of tokens and communicate with a coordinator that wishes to continuously monitor a function on the union of those streams. These papers were largely concerned with devising protocols that minimize the number of bits communicated and monitor the function with small error.

3 BGMNW Estimation Protocol For Normal Distribution

The protocol from Braverman et al. [2] for estimating the mean parameter for the specific case in which μ is Gaussian is stated below. In the case of a multivariate Gaussian, the protocol below is repeated for each dimension.

Let $r = \lfloor \log(\frac{mdn}{\sigma}) \rfloor$, and $\{x\} = x - \lfloor x \rfloor$. Let $g(x) = \sum_{k=1}^{\infty} \frac{1}{k} e^{-2k^2\pi^2} \sin(2k\pi x)$. We present the algorithm for estimating $\overline{\theta} = \theta \frac{\sqrt{n}}{\sigma}$ in two phases. As with all real numbers, $\overline{\theta}$ has an integer part as well as a fractional one. The integer part of the parameter estimate can be directly computed from the output of Phase 1. Phase 2 computes the fractional part of the parameter estimate.

Data: A set of samples $\{X_j^{(k)}\}_{k=1}^{n}$, each drawn i.i.d. from $\mu = \mathcal{N}(\mu, \sigma^2)$, is given to the jth machine, $j \in \{1, ..., m\}$.

Machine i, $i \in \{1, 2, ..., m\}$: $X_i \leftarrow \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^{n} X_i^{(k)}$

Machine i, $i \in \{1, 2, ..., r\}$, sends the first r bits of X_i to the coordinator.

Machine i, $i \in \{r+1 ... m\}$: $R_i \leftarrow X_i - \lfloor X_i \rfloor$; $R_i' \leftarrow X_i + \frac{1}{5} - \lfloor X_i + \frac{1}{5} \rfloor$

Machine i, $i \in \{r+1 ... m\}$ sends bits B_i and B_i' to Coordinator, where B_i is 1 with probability R_i, 0 with probability $1 - R_i$, and B_i' is 1 with probability R_i', 0 with probability $1 - R_i'$.

$\gamma \leftarrow \frac{\sqrt{n}}{\sigma}$ times the median of the X_i's sent by machines $1, ..., r$

Algorithm: Phase 1: Normal Distribution

If the following conditions on $\{\gamma\}$:

- $\frac{1}{50} < \{\gamma\} < \frac{49}{50}$
- $|\{\gamma\} - \frac{1}{4}| \geq \frac{3}{100}$
- $|\{\gamma\} - \frac{3}{4}| \geq \frac{3}{100}$

are satisfied, then the following code is executed by Coordinator:

1 $Z \leftarrow \lfloor \gamma \rfloor$

2 $T \leftarrow \frac{1}{m-r} \sum_{i=r+1}^{m} B_i$

3 $S \leftarrow \{i \in \mathbb{Z} \big| |Z + \frac{i}{\sqrt{m}} - \gamma| < \frac{1}{100}\}$

4 **for** $i \in S$ **do**

5   **if** $|g(\frac{i}{\sqrt{m}}) - \pi(\frac{1}{2} - T)| \leq \frac{1}{\sqrt{m}}$ **then**

6   **return** $Z + \frac{i}{\sqrt{m}}$

7   **end**

8 **end**

9 **return** γ

Algorithm: Phase 2: Normal Distribution

If any condition is violated, Phase 2 is executed as written with four modifications:

1. Line 1 is replaced by $Z \leftarrow \lfloor \gamma + \frac{1}{5} \rfloor$
2. Line 2 is replaced by $T \leftarrow \frac{1}{m-r} \sum_{i=r+1}^{m} B'_i$.
3. Line 3 is replaced by $S \leftarrow \{i \in \mathbb{Z} \big| |Z + \frac{i}{\sqrt{m}} - (\gamma + \frac{1}{5})| < \frac{1}{100}\}$
4. Line 6 is replaced by **return** $Z + \frac{i}{\sqrt{m}} - \frac{1}{5}$.

These modifications shift the quantity being estimated in order to avoid problematic subintervals of $[0, 1]$; the shift is then undone to return the proper parameter estimate.

4 General Estimation Protocol

We now present a protocol for estimating the unknown location parameter θ of a distribution with probability distribution function f and known scale parameter σ. Because the mean parameter vector of a general d-dimensional may not necessarily be inferred via repetition of the protocol for each dimension, this analysis will assume $d = 1$. Moreover, we assume $n = 1$ for the protocol statement as well as the analysis. Note that these assumptions can be relaxed given any particular distribution by having each machine send a sufficient statistic computed from its n input samples and, if applicable, repeat the protocol for each of d dimensions. These modifications will yield an analogous analysis as that which will be demonstrated in this work. We make the following assumptions on μ:

- If $X \sim \mu$, $\forall \alpha > 0, \exists \beta > 0, \Pr[X > \theta + \alpha] \leq \frac{1}{2} - \beta$. Similarly, $\forall \alpha > 0, \exists \beta' > 0, \Pr[X < \theta - \alpha] \leq \frac{1}{2} - \beta'$.
- $|\theta| \leq poly(m)$
- $\sum_{i=-\infty}^{\infty} f(x)$ is convergent for $x \in \mathbb{R}$.
- $r = O(\log(m/\sigma)) = o(m)$

- $\exists c > 0, m/\sigma \geq m^c$
- If $X \sim \mu$ and $h(\theta) = \mathbb{E}[X - \lfloor X \rfloor]$: $h'(\theta)$ has finitely many zeros.
- $m \geq \frac{1}{\alpha^2}$, where α is the parameter associated with set J, defined below.

If $h'(\theta)$ has finitely many zeros, then, as we will argue later, there exists a distribution-dependent constant $\alpha > 0$ and a set of subintervals J of $[0, 1]$ which *α-respects* h'. This notion is defined as follows:

Definition 4.1. Let J be a set of disjoint open subintervals of $[0, 1]$. If $i = (a, b) \in J$, let $|i| = b - a$, and let $w = (1 - \alpha, 1] \cup [0, \alpha)$ be considered a single contiguous interval; indeed, it is as such if integral shifts of $[0, 1]$ are considered. We say that J *α-respects* h' if:

1. $\forall i = (a, b) \in J, i \neq w, [b, b + \alpha + (|i| + 2\alpha)]$ is disjoint of all elements of J.
2. $w \in J$ and $[\alpha, \alpha + \alpha + (2\alpha + 2\alpha)]$ is disjoint of all elements of J, or $\exists i \in J, w \subset i$. Define $|w| = 2\alpha$.
3. $\exists m_1 > 0, m_2 > 0 \forall x \in [0, 1] \backslash J$, $m_1 \leq |h'(x)| \leq m_2$.

Let J' denote the extension of J, which is defined:

Definition 4.2. If J α-respects h', let J' be the *extension* of J, defined as follows:

$$i = (a, b) \in J \wedge i \neq w \implies \big(a - \alpha, b + \alpha\big) \in J.'$$

Let the extension of w, denoted w', in J' be $(1 - \alpha - \alpha, 1] \cup [0, \alpha + \alpha)$; $|w'| = 4\alpha$. By construction,

1. $\forall i = (a, b) \in J', [b, b + |i|]$ is disjoint of all elements of J'.
2. $\exists m > 0 \forall x \in [0, 1] \backslash J'$, $0 < |h'(x)| \leq m$.

The reason for defining J and J' in this way will become more apparent when we analyze our protocol.

Suppose that the subintervals are indexed $1, ..., |J'|$. Let r be defined as in Section 3. Whenever J' is referred to in the protocol, it refers to the extension of the respectful set J, which is guaranteed to exist by the conditions imposed on the distribution.

As before, the integer part of the estimate can be directly computed after Phase 1, while Phase 2 computes the fractional part.

Data: For each machine $i \in \{1, ..., m\}$, an i.i.d. sample from μ $\{X_i\}$.

Machine i, $i \in \{1, 2, ..., m\}$: $X_i \leftarrow \frac{X_i}{\sigma}$

Machine i, $i \in \{1, 2, ..., r\}$, sends the first r bits of X_i to the coordinator.

Machine i, $i \in \{r + 1, ..., m\}$ allocates an array $R^{(i)}$ of size $|J'| + 1$

for *Machine* i, $i \in \{r + 1, ..., m\}$, $k \in J'$ **do**

  $R^{(i)}[k] \leftarrow X_i + |k| - \lfloor X_i + |k| \rfloor$

end

Machine i, $i \in \{r + 1, ..., m\}$: $R^{(i)}[|J'| + 1] \leftarrow X_i - \lfloor X_i \rfloor$

for *Machine* i, $i \in \{r + 1, ..., m\}$, $k \in R^{(i)}$: **do**

  Sends to the coordinator a bit $B_k^{(i)}$ which is 1 with probability $R^{(i)}[k]$ and 0 with probability $1 - R^{(i)}[k]$.

end

$\gamma \leftarrow$ the median of the X_i's sent by machines $1, ..., r$.

Algorithm: Phase 1: Generalized Distribution

As before, the coordinator executes Phase 2 to compute the fractional component of the parameter estimate. As described in Definition 5.1, $w = (1 - 2\alpha, 1] \cup [0, 2\alpha)$ is considered to be a contiguous subinterval such that $|w| = 4\alpha$. Recall the notation $\{x\} = x - \lfloor x \rfloor$. To simplify the notation, if k denotes a subinterval, let $B_k^{(i)}$ denote the bit sent by the ith machine which is associated with subinterval k. Note that if $\{\gamma\}$ is not found to belong to any interval of J', then it must belong to $\overline{J'} = [0, 1] \backslash J'$; the bits associated with $\overline{J'}$ are denoted $B_{|J'|+1}^{(i)}$. Finally, let $h(x) = \int_0^1 \sum_{i=-\infty}^{\infty} y f((y - x) + i) dy$; notice that $h(\theta) = \mathbb{E}[R_i]$. Note that we are working within an idealized model in requiring the computation of h within the algorithm definition. For real-world purposes, assume that f is such that h can be efficiently evaluated to arbitrary precision.

$Z \leftarrow$ null; $B \leftarrow \frac{1}{m-r} \sum_{i=r+1}^{m} B_{|J|+2}^{(i)}$
shift $\leftarrow 0$
for $k \in J'$ **do**
    if $\{\gamma\} \in k$ **then**
        shift $\leftarrow |k|$
        $B \leftarrow \frac{1}{m-r} \sum_{i=r+1}^{m} B_{k}^{(i)}$
    end
end
$Z \leftarrow \lfloor \gamma + \text{shift} \rfloor$
$S \leftarrow \{i \in \mathbb{Z} \big| |Z + \frac{i}{\sqrt{m}} - (\gamma + \text{shift})| < \alpha\}$
for $i \in S$ **do**
    if $|h(\frac{i}{\sqrt{m}}) - B| \leq \frac{m_2}{\sqrt{m}}$ **then**
        return $Z + \frac{i}{\sqrt{m}} - \mathit{shift}$
    end
end
return γ

Algorithm: Phase 2: Generalized Distribution

We observe that the total communication of this algorithm is

$$O(r^2 + |J|(m-r)) = O(\log^2(m/\sigma) + |J|m).$$

5 Generalized Algorithm Analysis

This section will first provide an expression for the derivative of the expected value of the fractional part of the input data. This derivative is closely tied to the notion of a set of subintervals of $[0,1]$ which $\alpha-$*respects* h', which in turn is key for the analysis of the protocol. Finally, we will demonstrate that the mean squared error of the parameter estimate computed by the generalized algorithm is $O(\frac{1}{m})$ after having rescaled the distribution to have scale parameter 1. For notational simplicty, throughout this section we will let θ refer to the location parameter of the rescaled distribution.

5.1 Derivation of h' and α-Respectful Intervals

In this subsection, we let the domain of h' be $[0,1]$, recognizing that the results generalize for h' for any integral shifts of that domain, as h' has period 1.

Theorem 5.1. *Let F be the antiderivative of f, let $h(\theta) = \mathbb{E}[R_i]$. If $\forall y \in \mathbb{R}$, $\sum_{i=-\infty}^{\infty} f(y+i)$ is convergent and f is continuous, then*

$$h'(\theta) = 1 - \sum_{i=-\infty}^{\infty} f(i-\theta).$$

Proof. Let $h(\theta) = \mathbb{E}[R_i]$. We have:

$$h(\theta) = \int_0^1 \sum_{i=-\infty}^{\infty} xf((x-\theta)+i)dx.$$

Because $\sum_{i=-\infty}^{\infty} f((x-\theta)+i)$ is convergent, and therefore the series is absolutely convergent due to nonnegativity of f, we have:

$$h(\theta) = \sum_{i=-\infty}^{\infty} \int_0^1 xf((x-\theta)+i)dx.$$

Let $u = x + i - \theta$. Then $x = u - i + \theta$. By a change of variables, we have:

$$h(\theta) = \sum_{i=-\infty}^{\infty} \int_{i-\theta}^{1+i-\theta} (u-i+\theta)f(u)du.$$

Let $g(u) = u \cdot f(u)$. Then

$$h(\theta) = \sum_{i=-\infty}^{\infty} \Big[\int_{i-\theta}^{1+i-\theta} g(u)du + (\theta - i) \int_{i-\theta}^{1+i-\theta} f(u)du \Big].$$

Let G, F respectively be the antiderivatives of g, f, which exist because f is continuous. Applying the Fundamental Theorem of Calculus, we have:

$$h'(\theta) = \sum_{i=-\infty}^{\infty} \Big(\frac{d}{d\theta}\Big[G(1+i-\theta) - G(i-\theta)\Big] + \frac{d}{d\theta}\Big((\theta - i)\Big[F(1+i-\theta) - F(i-\theta)\Big]\Big)\Big).$$

Applying chain rule, as well as substituting for g,

$$= \sum_{i=-\infty}^{\infty} [(i-\theta)f(i-\theta) - (1+i-\theta)f(1+i-\theta) + (\theta-i)f(i-\theta) - (\theta-i)f(1+i-\theta) + F(1+i-\theta) - F(i-\theta)]$$

$$= \sum_{i=-\infty}^{\infty} [-(1+i-\theta)f(1+i-\theta) - (\theta-i)f(1+i-\theta) + F(1+i-\theta) - F(i-\theta)]$$

$$= \sum_{i=-\infty}^{\infty} [F(1+i-\theta) - F(i-\theta) - f(1+i-\theta)].$$

Recognizing that F is a cumulative distribution function, we conclude

$$h'(\theta) = 1 - \sum_{i=-\infty}^{\infty} f(1+i-\theta) = 1 - \sum_{i=-\infty}^{\infty} f(i-\theta).$$

□

Theorem 5.1 allows an algorithm user to determine whether there exists a set J which α-respects h':

Definition 5.1. Let J be a set of disjoint open subintervals of $[0,1]$. If $i = (a,b) \in J$, let $|i| = b - a$, and let $w = (1-\alpha, 1] \cup [0, \alpha)$ be considered a single contiguous interval; indeed, it is as such if integral shifts of $[0,1]$ are considered. We say that J *α-respects* h' if:

1. $\forall i = (a,b) \in J, i \neq w, [b, b + \alpha + (|i| + 2\alpha)]$ is disjoint of all elements of J.
2. $w \in J$ and $[\alpha, \alpha+\alpha+(2\alpha+2\alpha)]$ is disjoint of all elements of J, or $\exists i \in J, w \subset i$. Define $|w| = 2\alpha$.
3. $\exists m_1 > 0, m_2 > 0 \forall x \in [0,1] \backslash J$, $m_1 \leq |h'(x)| \leq m_2$.

Let J' denote the extension of J, which is defined:

Definition 5.2. If J α-respects h', let J' be the *extension* of J, defined as follows:

$$i = (a,b) \in J \wedge i \neq w \implies \big(a - \alpha, b + \alpha\big) \in J.'$$

Let the extension of w, denoted w', in J' be $(1 - \alpha - \alpha, 1] \cup [0, \alpha + \alpha)$; $|w'| = 4\alpha$. By construction,

1. $\forall i = (a,b) \in J', [b, b + |i|]$ is disjoint of all elements of J'.
2. $\exists m > 0 \forall x \in [0,1] \backslash J'$, $0 < |h'(x)| \leq m$.

An important property of an α-respectful set J and its extension J' is that if γ is the median of the samples sent by machines $1, ..., r$, and the following properties hold:

- $\{\gamma\} \in [0,1] \backslash J'$
- $|\{\theta\} - \{\gamma\}| \leq \alpha$

then we have that $\forall x \in [\{\theta\} - \alpha, \{\theta\} + \alpha], \exists m > 0$,

$$0 < |h'(x)| \leq m.$$

If the algorithm user is able to verify that $h'(\theta)$ has finitely many zeros, then an α-respectful set J is guaranteed to exist:

Lemma 5.1. Suppose that h' has finitely many zeros in the interval $[0,1]$. Then there exists a set of subintervals which α-respects h'.

Proof. If there are finitely many zeros, then all of the zeros are isolated. Therefore, we can pick α and neighborhoods around the zeros to be sufficiently small that conditions (1) and (2) of Definition 5.1 are satisfied. Let J denote the set of these neighborhoods. For condition (3), by construction, all values of $|h'|$ on the set $[0,1] \backslash J$ are at least $m_1 > 0$, where m_1 is the minimum value of $|h'|$ over the set of endpoints of the subintervals of J. The lower bound therefore holds.

For the upper bound of condition (3): because J is the union of open sets, it itself is open; therefore, $[0,1] \backslash J$ is closed. Because h' is continuous, its image is also closed and bounded; the claim thus follows. ☐

5.2 Expected Mean Squared Error Derivation

We now demonstrate that $\mathbb{E}[|\hat{\theta}-\theta|^2] \le O\big(\frac{1}{m}\big)$. We first present a few definitions and facts which will be useful in this analysis. Define $B = \frac{1}{m-r}\sum_{i=r+1}^{m} B^{(i)}$. Because each $B^{(i)}$ is Bernoulli with some probability p:

$$\mathrm{Var}(B) = \mathbb{E}[|B-\mathbb{E}[B]|^2] = \frac{1}{(m-r)^2}p(1-p)(m-r) \le \frac{1}{m-r}.$$

Applying our assumption that $r = o(m)$,

$$\mathbb{E}[|B-\mathbb{E}[B]|^2] \le \frac{2}{m}.$$

Let $0 < s < 1$ be a constant. Define $\mathcal{E}_s$ to be the event that the median γ of $X_1, ..., X_r$ is within an additive s of θ. Then the following lemma holds:

Lemma 5.2. If $X \sim \mu$, and μ is such that $\forall s > 0 \exists \beta > 0, \Pr[X > \theta + s] \le \frac{1}{2} - \beta$. Then for some arbitrarily large constant $a > 0$,

$$\Pr[\mathcal{E}_s] \ge 1 - 2\Big(\frac{m}{\sigma}\Big)^{-2a\beta^2}$$

Proof. Let $X^{(i)} \sim \mu$ be the ith machine's sample. Let

$$Y^{(i)} = \begin{cases} 1 & X^{(i)} > \theta + \alpha \\ 0 & X^{(i)} \le \theta + \alpha \end{cases}$$

Define $\overline{Y} = \frac{1}{r}\sum_{i=1}^{r} Y^{(i)}$. Let E denote the event that $\gamma > \theta + \alpha$. Observe that E is equivalent to the event $\overline{Y} \ge \frac{1}{2}$ and that $\mathbb{E}[\overline{Y}] \le \frac{1}{2} - \beta$. Using this along with Hoeffding's inequality, we have:

$$\begin{aligned} \Pr[E] = \Pr\Big[\overline{Y} \ge \frac{1}{2}\Big] &= \Pr\Big[\overline{Y} - \mathbb{E}[\overline{Y}] \ge \frac{1}{2} - \mathbb{E}[\overline{Y}]\Big] \\ &\le \Pr[\overline{Y} - \mathbb{E}[\overline{Y}] \ge \beta] \\ &\le e^{-2r\beta^2} \end{aligned}$$

Recalling that $r = O(\log(\frac{m}{\sigma}))$, we thus have that for some arbitrarily large constant $a > 0$, $\Pr[E] \le \big(\frac{m}{\sigma}\big)^{-2a\beta^2}$. Analogous reasoning for the event in which $\gamma < \theta - \alpha$ yields the same probability. Therefore, $\Pr[\mathcal{E}_\alpha] \ge 1 - 2\big(\frac{m}{\sigma}\big)^{-2a\beta^2}$. ☐

Recall the definition:

$$h(x) = \int_0^1 \sum_{i=-\infty}^{\infty} y f((y-x)+i) dy$$

Define $\mathcal{F}$ to be the event that the coordinator identifies an i such that:

- $|Z + \frac{i}{\sqrt{m}} - \gamma| < \alpha$
- $|h(\frac{i}{\sqrt{m}}) - B| \le \frac{1}{\sqrt{m}}$

We will now prove a set of lemmas which will together demonstrate that $\mathbb{E}[|\hat{\theta}-\theta|^2] \leq O(\frac{1}{m})$. We will do this by first bounding $\mathbb{E}[|\hat{\theta}-\theta|^2|\mathcal{E}_\alpha]$. Note that by definition of J' and the logic of the algorithm, we know that $\{\gamma + \text{shift}\} \in [0,1]\backslash J'$. For notational simplicty, we will refer to $\gamma + \text{shift}$ as γ.

The following fact will be used in the lemmas to follow: we know that $\{\gamma\} \geq 3\alpha$ because as a result of shifting, $\{\gamma\} \in [0,1]\backslash J'$ and $w \in J'$. If $\mathcal{E}_\alpha$ occurs, the integral component of the estimate must therefore be correct and $|\{\theta\} - \{\gamma\}| \leq \alpha$.

Lemma 5.3. $\mathbb{E}[|\hat{\theta}-\theta|^2|\mathcal{E}_\alpha \wedge \mathcal{F}] \leq O\big(\frac{1}{m} + \mathbb{E}[|B-\mathbb{E}[B]|^2|\mathcal{E}_\alpha \wedge \mathcal{F}]\big)$

Proof. Because $\mathcal{F}$ occurs, the fractional component is $\frac{i}{\sqrt{m}}$, where i satisfies the described conditions.

$$\mathbb{E}[|\hat{\theta}-\theta|^2|\mathcal{E}_\alpha \wedge \mathcal{F}] = \mathbb{E}\left[\Big|\frac{i}{\sqrt{m}} - \{\theta\}\Big|^2 |\mathcal{E}_\alpha \wedge \mathcal{F}\right].$$

We have J α-respects h'; let J' be its extension. We are given that $|\frac{i}{\sqrt{m}} - \{\gamma\}| \leq \alpha$ and $\{\gamma\} \in [0,1]\backslash J'$. Therefore, $\exists m_1 > 0, m_2 > 0, \forall x$ between $\frac{i}{\sqrt{m}}$ and $\{\gamma\}$,

$$m_1 \leq |h'(x)| \leq m_2$$

where m_1, m_2 are defined as in Definition 5.1.

We can therefore apply the Mean Value Theorem:

$$\mathbb{E}[|\hat{\theta}-\theta|^2|\mathcal{E}_\alpha \wedge \mathcal{F}] \leq \frac{1}{m_1^2}\mathbb{E}[|h\Big(\frac{i}{\sqrt{m}}\Big) - h(\{\theta\})|^2|\mathcal{E}_\alpha \wedge \mathcal{F}].$$

By triangle inequality,

$$\mathbb{E}[|\hat{\theta}-\theta|^2|\mathcal{E}_\alpha \wedge \mathcal{F}] \leq \frac{1}{m_1^2}\mathbb{E}\left[\Big(\Big|h\Big(\frac{i}{\sqrt{m}}\Big) - B\Big| + |B - h(\{\theta\})|\Big)^2 |\mathcal{E}_\alpha \wedge \mathcal{F}\right].$$

Because $\mathcal{F}$ occurs and $h(\{\theta\}) = \mathbb{E}[R_i] = \mathbb{E}[B_i] = \mathbb{E}[B]$,

$$\mathbb{E}[|\hat{\theta}-\theta|^2|\mathcal{E}_\alpha \wedge \mathcal{F}] \leq \frac{1}{m_1^2}\mathbb{E}\left[\Big(\frac{m_2}{\sqrt{m}} + |B - \mathbb{E}[B]|\Big)^2 \Big|\mathcal{E}_\alpha \wedge \mathcal{F}\right].$$

Recalling that $\forall a, b \in \mathbb{R}$, $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\mathbb{E}[|\hat{\theta}-\theta|^2|\mathcal{E}_\alpha \wedge \mathcal{F}] \leq \frac{1}{m_1^2}\left(\frac{2m_2^2}{m} + 2\mathbb{E}\big[B - \mathbb{E}[B]|\mathcal{E}_\alpha \wedge \mathcal{F}\big]^2\right).$$

□

Now, consider the following lemma bounding the expected error in the event that $\mathcal{F}$ does not occur:

Lemma 5.4. $\mathbb{E}[|\hat{\theta}-\theta|^2\big|\mathcal{E}_\alpha \wedge \neg\mathcal{F}] \leq O(1) \cdot \mathbb{E}[|B-\mathbb{E}[B]|^2\big|\mathcal{E}_\alpha \wedge \neg\mathcal{F}]$

Proof. Given $\mathcal{E}_\alpha \wedge \neg\mathcal{F}$, we have that $\forall i \in \mathbb{Z}$ such that $|\frac{i}{\sqrt{m}} - \{\gamma\}| \leq \alpha$, $\left|h\left(\frac{i}{\sqrt{m}}\right) - B\right| > \frac{1}{\sqrt{m}}$.

We will first demonstrate the following subclaim:

Claim: Let m_1 is the constant of Definition 5.1. Let $m_1' = \min(1, m_1)$. Then, given $\mathcal{E}_\alpha \wedge \neg\mathcal{F}$, $|\mathbb{E}[B] - B| \geq m_1'\alpha$.

Proof. Assume to the contrary $\mathcal{E}_\alpha \wedge \neg\mathcal{F} \wedge |\mathbb{E}[B] - B| < m_1'\alpha$. Recalling that $\mathbb{E}[B] = h(\theta) = h(\{\theta\})$, we have that $|h(\{\theta\}) - B| < m_1'\alpha$. Equivalently, $B \in \big(h(\{\theta\}) - m_1'\alpha, h(\{\theta\}) + m_1'\alpha\big)$. Because $\{\gamma\} \in [0,1]\backslash J'$ by design and $\mathcal{E}_\alpha$ has occurred, we have that $\forall x \in [\{\theta\} - \alpha, \{\theta\} + \alpha]$, $\exists m_1 > 0, m_2 > 0, m_1 \leq |h'(x)| \leq m_2$.

Because h' is nowhere zero in $[\{\theta\} - \alpha, \{\theta\} + \alpha]$, either h' is either uniformly positive or uniformly negative in that interval. Suppose that it is uniformly positive.

By Mean Value Theorem, $\frac{h(\{\theta\}-\alpha)-h(\{\theta\})}{-\alpha} \leq m_2$, and therefore $h(\{\theta\} - \alpha) \geq h(\{\theta\}) - m_2\alpha$. Similarly, $\frac{h(\{\theta\}+\alpha)-h(\{\theta\})}{\alpha} \geq m_1$, and therefore $h(\{\theta\} + \alpha) \geq h(\{\theta\}) + m_1\alpha$. By Intermediate Value Theorem, $h(x)$ takes on all values in the range $[h(\{\theta)\} - m_2\alpha, h(\{\theta\}) + m_1\alpha]$ when $x \in [\{\theta\} - \alpha, \{\theta\} + \alpha]$. Recalling that $m_1' \leq m_1 \leq m_2$, we conclude that $\exists y \in [\{\theta\} - \alpha, \{\theta\} + \alpha]$, $h(y) = B$. Analogous reasoning for the case in which h' is uniformly negative yields the same result.

Therefore, let $y \in [\{\theta\} - \alpha, \{\theta\} + \alpha]$ such that $h(y) = B$. Given that $m \geq \frac{1}{\alpha^2}$, $\exists i \in \mathbb{Z}$ such that $|\frac{i}{\sqrt{m}} - \{\theta\}| \leq \alpha \wedge |\frac{i}{\sqrt{m}} - y| \leq \frac{1}{\sqrt{m}}$. We can thus say that $|h(\frac{i}{\sqrt{m}}) - h(y)| \leq m_2|\frac{i}{\sqrt{m}} - y| \leq \frac{m_2}{\sqrt{m}}$, contradicting that $\mathcal{F}$ did not occur. □

Recalling that $\mathcal{E}_\alpha$ occurs and $\gamma = \hat{\theta}$,

$$|\hat{\theta} - \theta|^2 \leq \alpha^2 \leq \frac{1}{m_1'^2}|\mathbb{E}[B] - B|^2.$$

We thus conclude that $\mathbb{E}[|\hat{\theta} - \theta|^2|\mathcal{E}_\alpha \wedge \neg\mathcal{F}] \leq O(1) \cdot \mathbb{E}[|B - \mathbb{E}[B]|^2|\mathcal{E}_\alpha \wedge \neg\mathcal{F}]$, as desired. □

Lemma 5.5. $\mathbb{E}[|\hat{\theta} - \theta|^2|\mathcal{E}_\alpha] \leq O\big(\frac{1}{m}\big)$.

Proof. By definition,

$$\mathcal{E}[|\hat{\theta} - \theta|^2|\mathcal{E}_\alpha] = \mathbb{E}[|\hat{\theta} - \theta|^2|\mathcal{E}_\alpha \wedge \mathcal{F}]\Pr[\mathcal{F}] + \mathbb{E}[|\hat{\theta} - \theta|^2|\mathcal{E}_\alpha \wedge \neg\mathcal{F}]\Pr[\neg\mathcal{F}]$$

Applying Lemmas 5.3 and 5.4:

$$\begin{aligned}\mathbb{E}[|\hat{\theta} - \theta|^2|\mathcal{E}_\alpha] &\leq O\Big(\frac{1}{m} + \mathbb{E}[|B - \mathbb{E}[B]|^2|\mathcal{E}_\alpha \wedge \mathcal{F}]\Big)\Pr[\mathcal{F}] + O(1) \cdot \mathbb{E}[|B - \mathbb{E}[B]|^2\big|\mathcal{E}_\alpha \wedge \neg\mathcal{F}]\Pr[\neg\mathcal{F}] \\ &\leq O\Big(\frac{1}{m}\Big) + O(1) \cdot \mathbb{E}[|B - \mathbb{E}[B]|^2|\mathcal{E}_\alpha]\end{aligned}$$

We have that $\Pr[\mathcal{E}_\alpha] \geq 1 - 2\big(\frac{m}{\sigma}\big)^{-2a\beta^2}$ for a arbitrarily large, as well as the problem assumption that $\exists c > 0, m/\sigma \geq m^c$. The latter assumption in turn implies that $m > m^{1-c} \geq \sigma \implies m/\sigma > 1$.

Therefore, a can be chosen such that $\Pr[\mathcal{E}_\alpha] \geq \frac{1}{2}$.

$$\mathbb{E}[|B - \mathbb{E}[B]|^2 | \mathcal{E}_\alpha] \leq \frac{\mathbb{E}[|B - \mathbb{E}[B]|^2]}{\Pr[\mathcal{E}_\alpha]} \leq 2\mathbb{E}[|B - \mathbb{E}[B]|^2].$$

Therefore, recalling that $\mathbb{E}[|B - \mathbb{E}[B]|^2] \leq \frac{2}{m}$,

$$\mathbb{E}[|B - \mathbb{E}[B]|^2 | \mathcal{E}_\alpha] \leq O\Big(\frac{1}{m}\Big).$$

Finally, we conclude $\mathbb{E}[|\hat{\theta} - \theta|^2 | \mathcal{E}_\alpha] \leq O\big(\frac{1}{m}\big)$. □

Theorem 5.2. $\mathbb{E}[|\hat{\theta} - \theta|^2] \leq O\big(\frac{1}{m}\big)$

Proof. Let $|\theta| \leq U = poly(m)$. By Lemma 5.2, $\Pr[\neg\mathcal{E}_\alpha] \leq 2 \cdot \big(\frac{m}{\sigma}\big)^{-2a\beta^2}$, for some arbitrarily large constant $a > 0$. We have that:

$$\begin{aligned}
\mathbb{E}[|\hat{\theta} - \theta|^2] &= \mathbb{E}[|\hat{\theta} - \theta|^2 | \mathcal{E}_\alpha] \Pr[\mathcal{E}_\alpha] + \mathbb{E}[|\hat{\theta} - \theta|^2 | \neg\mathcal{E}_\alpha] \Pr[\neg\mathcal{E}_\alpha] \\
&\leq \mathbb{E}[|\hat{\theta} - \theta|^2 | \mathcal{E}_\alpha] + \frac{2 \cdot 4U^2}{(m/\sigma)^{2a\beta^2}} \\
&\leq \mathbb{E}[|\hat{\theta} - \theta|^2 | \mathcal{E}_\alpha] + \frac{poly(m)}{(m/\sigma)^{2a\beta^2}}
\end{aligned}$$

By assumption, $\exists c > 0, m/\sigma \geq m^c$. Therefore,

$$\mathbb{E}[|\hat{\theta} - \theta|^2] \leq \mathbb{E}[|\hat{\theta} - \theta|^2 | \mathcal{E}_\alpha] + \frac{poly(m)}{m^{2ca\beta^2}}.$$

Taking a to be adequately large, we have

$$\mathbb{E}[|\hat{\theta} - \theta|^2] \leq \mathbb{E}[|\hat{\theta} - \theta|^2 | \mathcal{E}_\alpha] + \frac{1}{m}.$$

Applying Lemma 5.5, we conclude

$$\mathbb{E}[|\hat{\theta} - \theta|^2] \leq O\Big(\frac{1}{m}\Big).$$

□

Corollary 5.1. Recall that in this section, we referred to θ as the location parameter of the distribution rescaled to have scale 1. If $\overline{\theta}$ denotes this location parameter and θ denotes the location parameter of the original, unscaled distribution, we have $\frac{\theta}{\sigma} = \overline{\theta}$. We conclude:

$$\mathbb{E}[|\hat{\theta} - \theta|^2] \leq O\Big(\frac{\sigma^2}{m}\Big).$$

6 Cauchy Distribution Analysis

We will now demonstrate that there exists an α-respectful set for a Cauchy distribution. The Cauchy distribution is notable in statistics as the canonical example of a distribution with undefined mean and variance and in physics for being the distribution of the X-intercept of a laser spinning with uniformly distributed angle. It is also one of the few stable distributions.

Let f denote the Cauchy probability density function, and let its location parameter be θ. We will make the simplifying assumptions that the number of samples per machine is 1 and that the scale parameter $\gamma = 1$.

Lemma 6.1. Let $h(\theta) = \mathbb{E}[R_i]$. Then $h'(\theta) = 1 - \frac{1}{\pi} \sum\limits_{i=-\infty}^{\infty} \frac{1}{(1+i-\theta)^2+1}$.

Proof.

$$\mathbb{E}[R_i] = \frac{1}{\pi} \int\limits_0^1 \sum_{i=-\infty}^{\infty} \frac{x}{1+(x+i-\theta)^2} dx$$

Noting that $\sum\limits_{i=-\infty}^{\infty} \frac{x}{1+(x+i-\theta)^2}$ is convergent and therefore the series is absolutely convergent due to nonnegativity of f,

$$\mathbb{E}[R_i] = \frac{1}{\pi} \sum_{i=-\infty}^{\infty} \int\limits_0^1 \frac{x}{1+(x+i-\theta)^2} dx.$$

Let $u = x + i - \theta$. Then $x = u - i + \theta$.

$$\begin{aligned}
\mathbb{E}[R_i] &= \frac{1}{\pi} \sum_{i=-\infty}^{\infty} \int\limits_{i-\theta}^{1+i-\theta} \frac{u-i+\theta}{1+u^2} du \\
&= \frac{1}{\pi} \sum_{i=-\infty}^{\infty} \int\limits_{i-\theta}^{1+i-\theta} \frac{u}{1+u^2} du - \frac{1}{\pi} \sum_{i=-\infty}^{\infty} \int\limits_{i-\theta}^{1+i-\theta} \frac{i-\theta}{1+u^2} du \\
&= \frac{1}{\pi} \sum_{i=-\infty}^{\infty} \Big|_{i-\theta}^{1+i-\theta} \frac{1}{2} \log(u^2+1) - \frac{i-\theta}{\pi} \sum_{i=-\infty}^{\infty} \Big|_{i-\theta}^{1+i-\theta} \tan^{-1}(u) \\
&= \frac{1}{\pi} \sum_{i=-\infty}^{\infty} \Big[\frac{1}{2} \log((1+i-\theta)^2+1) - \frac{1}{2} \log((i-\theta)^2+1)\Big] - \frac{i-\theta}{\pi} \sum_{i=-\infty}^{\infty} [\tan^{-1}(1+i-\theta) - \tan^{-1}(i-\theta)]
\end{aligned}$$

We have the final result:

$h(\theta) = \mathbb{E}[R_i] = \frac{1}{\pi} \sum\limits_{i=-\infty}^{\infty} \frac{1}{2} \log \frac{(1+i-\theta)^2+1}{(i-\theta)^2+1} - (i-\theta)[\tan^{-1}(1+i-\theta) - \tan^{-1}(i-\theta)]$.

We then have that:

$h'(\theta) = \frac{1}{\pi} \sum\limits_{i=-\infty}^{\infty} \Big[- \frac{1}{(1+i-\theta)^2+1} - \tan^{-1}(i-\theta) + \tan^{-1}(i-\theta+1)\Big]$.

Observing that the $\tan^{-1}$ terms telescope, canceling except at the index limits, we have:

$h'(\theta) = 1 - \frac{1}{\pi} \sum\limits_{i=-\infty}^{\infty} \frac{1}{(1+i-\theta)^2+1}$. □

Theorem 6.1. *Let $S = [0,1] - ([0.2, 0.3] \cup [0.7, 0.8])$. If X_i is drawn from a Cauchy distribution with median $\theta \in S$ and scale parameter $\gamma = 1$, then $\exists$ constant $C_1 > 0$ such that $\forall \theta$,*

$$C_1 \leq |h'(\theta)| \leq 1$$

Proof. By Lemma 6.1, we have

$$h'(\theta) = 1 - \frac{1}{\pi}\sum_{i=-\infty}^{\infty} \frac{1}{(1+i-\theta)^2+1}$$
$$= 1 - \frac{1}{\pi}\Big[\sum_{i=-10000}^{10000} \frac{1}{(1+i-\theta)^2+1} + \sum_{|i|\geq 10001} \frac{1}{(1+i-\theta)^2+1}\Big].$$

From Maple, we have that $\forall\theta \in S$,

$$1\times 10^{-3} \leq \Big|1 - \frac{1}{\pi}\sum_{i=-10000}^{10000} \frac{1}{(1+i-\theta)^2+1}\Big| \leq 4\times 10^{-3}. \quad (1)$$

Using the monotonicity of the tails of the distribution, we can bound the sums by the integral:

$$\sum_{|i|\geq 10001} \frac{1}{(1+i-\theta)^2+1} \leq \int_{-\infty}^{-10000} \frac{1}{(1+x-\theta)^2+1}dx + \int_{10000}^{\infty} \frac{1}{(1+x-\theta)^2+1}dx$$
$$= \Big|_{x=-\infty}^{x=-10000} \tan^{-1}(1+x-\theta) + \Big|_{x=10000}^{x=\infty} \tan^{-1}(1+x-\theta)$$
$$= \tan^{-1}(-9999-\theta) + \frac{\pi}{2} + \frac{\pi}{2} - \tan^{-1}(10001-\theta).$$

Observing that $\tan^{-1}$ is a monotonically increasing over $\mathbb{R}$, $\forall\theta \in S$, we have:

$$\pi + \tan^{-1}(-10000) - \tan^{-1}(10001) \leq \sum_{|i|\geq 10001} \frac{1}{(1+i-\theta)^2+1} \leq \pi + \tan^{-1}(-9999) - \tan^{-1}(10000).$$
$$1\times 10^{-4} \leq \sum_{|i|\geq 10001} \frac{1}{(1+i-\theta)^2+1} \leq 2\times 10^{-4}$$

We finally conclude that $\forall\theta \in S$,

$$1\times 10^{-4} \leq h'(\theta) \leq 1.$$

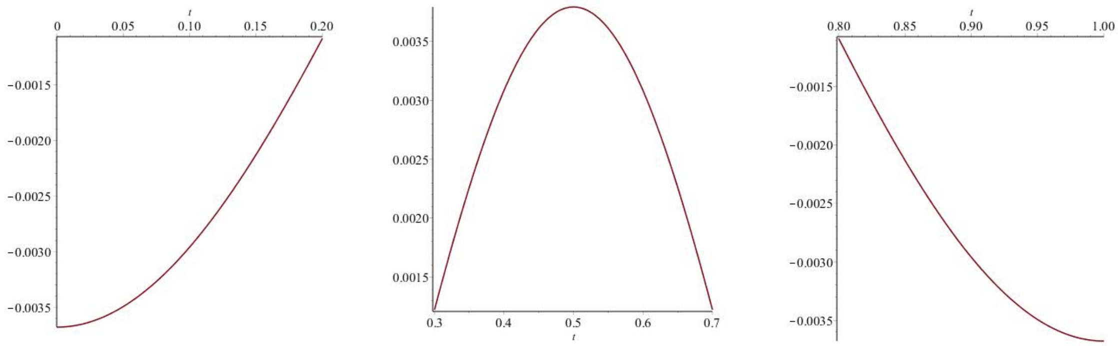
□

Letting $\alpha = \frac{1}{100}$, we see that $S \cup \{(\frac{98}{100}, 1] \cup [0, \frac{2}{100})\}$ indeed α-respects h'.

Finally, note that if $X \sim \mu$, $\Pr[X > \theta + \alpha] = 1 - F(\theta+\alpha) = 1 - \frac{1}{\pi}\tan^{-1}(\alpha) - \frac{1}{2} = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}(\alpha)$. Therefore, for arbitrary α, take $\beta = \frac{1}{\pi}\tan^{-1}(\alpha)$: we thus have that

$$\forall\alpha > 0, \exists\beta > 0, \Pr[X > \theta + \alpha] \leq \frac{1}{2} - \beta.$$

Analogous reasoning demonstrates that $\forall\alpha > 0, \exists\beta' > 0, \Pr[X < \theta - \alpha] \leq \frac{1}{2} - \beta'$.



(a) Interval (0,0.2) (b) Interval (0.3,0.7) (c) Interval (0.8,1)

Figure 2: Plots of $1 - \frac{1}{\pi}\sum_{i=-10000}^{10000} \frac{1}{(1+i-\theta)^2+1}$ demonstrating Inequality 1.

Appendices

A Illustration of Inequality 1

References

[1] C. J. Arackaparambil, J. Brody, and A. Chakrabarti. "Functional monitoring without monotonicity." *Proceedings of the 36th International Colloquium on Automata, Languages and Programming, 2009, pp. 95–106.*

[2] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, David P. Woodruff. "Communication lower bounds for statistical estimation problems via a distributed data processing inequality." *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*

[3] Cormode, Graham, S. Muthukrishnan, and Ke Yi. "Algorithms for distributed functional monitoring." *ACM Transactions on Algorithms (TALG) 7.2 (2011): 21.*

[4] Duchi, J. C., Jordan, M. I., Wainwright, M. J. and Zhang, Y. (2014). "Optimality guarantees for distributed statistical estimation." *arXiv preprint arXiv:1405.0782.*

[5] Ankit Garg, Tengyu Ma, and Huy L. Nguyen. "On Communication Cost of Distributed Statistical Estimation and Dimensionality." *Advances in Neural Information Processing Systems.*

[6] Jason D Lee, Yuekai Sun, Qiang Liu, and Jonathan E Taylor. "Communication- efficient sparse regression: a one-shot approach." *arXiv preprint arXiv:1503.04337.*

[7] Phillips, Jeff M., Elad Verbin, and Qin Zhang. "Lower bounds for number-in-hand multiparty communication complexity, made easy." Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms. *Society for Industrial and Applied Mathematics, 2012.*

[8] Suresh, Ananda Theertha, et al. "Distributed mean estimation with limited communication." *arXiv preprint arXiv:1611.00429* (2016).