Fisher Information for Distributed Estimation under a Blackboard Communication Protocol

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür
Stanford University, Stanford, CA 94305
Email: {lpb, yjhan, aozgur}@stanford.edu

***Abstract*—We consider the problem of learning high-dimensional discrete distributions and structured (e.g. Gaussian) distributions in distributed networks, where each node in the network observes an independent sample from the underlying distribution and can use k bits to communicate its sample to a central processor. We consider a blackboard communication model, where nodes can share information interactively through a public blackboard but each node is restricted to write at most k bits on the final transcript. We characterize the impact of the communication constraint k on the minimax risk of estimating the underlying distribution under ℓ^2 loss, and develop minimax lower bounds that apply in a unified way to many common statistical models. This is achieved by explicitly characterizing the Fisher information from the blackboard transcript.**

I. Introduction

Estimating a distribution from samples is a fundamental unsupervised learning problem that has been studied in statistics since the late nineteenth century [12]. Consider the following distribution estimation model

$$X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} P$$

where we would like to estimate the unknown distribution P under ℓ^2 loss. Unlike the traditional statistical setting where samples $X_1, \cdots, X_n$ are available to the estimator as they are, in this paper we consider a distributed setting where each observation X_i is available at a different node in a network and has to be communicated to a central processor by using k bits.

We consider a very general blackboard communication protocol [11] where all nodes communicate via a publicly shown blackboard while the total number of bits each node can write in the final transcript Y is limited by k. When one node writes a message bit on the blackboard, all other nodes can see the content of the message bit, and depending on the written bit, another node can take the turn to write a message on the blackboard. Upon receiving the final transcript Y, the central processor produces an estimate $\hat{P}$ of the distribution P based on the transcript Y and known procotol $\Pi \in \Pi_{\mathsf{BB}}$. The goal is to jointly design the protocol Π and the estimator $\hat{P}(\cdot)$ so as to minimize the worst case squared ℓ^2 risk, i.e., to characterize

$$\inf_{\Pi} \inf_{\hat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \|\hat{P} - P\|_2^2,$$

where $\mathcal{P}$ denotes the class of distributions that P belongs to. We study two different instances of this estimation problem:

1) High-dimensional discrete distributions: in this case we assume that $P = (p_1, \cdots, p_d)$ is a discrete distribution with known support size d and $\mathcal{P}$ denotes the probability simplex over d elements. By "high-dimensional" we mean that the support size d of the underlying distribution may be comparable to the sample size n.
2) Structured distributions: in this case, we assume that we have some additional information regarding the structure of the underlying distribution or density. In particular, we assume that the underlying distribution or density can be parametrized such that

$$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} P_\theta$$

where $\theta \in \Theta \subset \mathbb{R}^d$. In this case, estimating the underlying distribution amounts to estimating the parameters of this distribution and we are interested in the following parameter estimation problem under squared ℓ^2 risk

$$\inf_{\Pi} \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2,$$

where $\hat{\theta}(\cdot)$ is an estimator of θ.

Statistical estimation in distributed settings has gained increasing popularity over the recent years motivated by the fact that modern data sets are often distributed across multiple machines and processors, and bandwidth and energy limitations in networks and within multiprocessor systems often impose significant bottlenecks on the performance of algorithms. There are also an increasing number of applications in which data is generated in a distributed manner and the data (or features of it) are communicated over bandwidth-limited links to central processors. In particular, recent works [5], [6], [7] focus on a special case of the distributed parameter estimation problem described above, when the underlying distribution is known to have Gaussian structure, i.e. $P_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ with σ^2 known and $\theta \in \Theta = \mathbb{R}^d$, often called the Gaussian location model. On the other hand, [3] focuses on the first problem described above, distributed estimation of high-dimensional discrete distributions, under ℓ^1 loss.

In the authors' recent work [1], we showed that the results of these papers (results of [5], [6], [7] and the corresponding results of [3] under ℓ^2 loss) can be recovered in a unified framework which focuses on Fisher information from a single quantized sample under the following two simpler communication models:

- Independent protocols: each node independently quantizes its observation into k bits and communicates it to the central processor.
- Sequential protocols: node j for $1 \leq j \leq n$ quantizes its observation to k bits after it observes the quantized samples broadcasted by nodes $1, 2 \ldots, j-1$.

Characterizing Fisher information from a single sample quantized to k-bits was sufficient to address distributed estimation under these simpler protocols via an immediate application of the chain rule for Fisher information. However, applying the chain rule for Fisher information does not yield the desired answer in the case of blackboard protocols, as the interaction between nodes introduces

978-1-5386-9291-2/19/$31.00 ©2019 IEEE

dependency between the bits written by different nodes on the blackboard. In this follow-up paper, we show how to characterize and bound the Fisher information from the blackboard transcript Y, and use these bounds to recover the results of [5], [6], [7], [3] under a blackboard protocol, which is also the model considered in these works except [5]. This further demonstrates the power and flexibility of the Fisher information approach to developing minimax lower bounds in distributed estimation problems. With the exception of our prior work [1], all prior work on related problems (including [5], [6], [7], [3], [2], [13]) build on an alternative well-established technique for developing minimax lower bounds on statistical estimation problems that relies on converting the estimation problem to a carefully constructed hypothesis testing problem via the use of Fano's inequality. Even though the Cramér-Rao bound which yields a lower bound on the estimation error of an unbiased estimator in terms of the Fisher information is one of the most classical results in statistics, we are not aware of any other Fisher information based approaches to distributed estimation problems.

II. The Blackboard Communication Protocol

Suppose there are n nodes and each node has access to one sample X_i such that

$$X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} P_\theta$$

where $\theta \in \Theta \subset \mathbb{R}^d$. Nodes communicate their samples to a central node (and each other) via a publicly shown blackboard, and the total number of bits each node can write in the final transcript Y is limited by k bits. When one node writes a message (bit) on the blackboard, all other nodes can see the content of the message. Formally, a blackboard communication protocol $\Pi \in \Pi_{\mathsf{BB}}$ can be viewed as a binary tree [11], where each internal node v of the tree is assigned a deterministic label $l_v \in [n]$ indicating the identity of the node to write the next bit on the blackboard if the protocol reaches tree node v; the left and right edges departing from v correspond to the two possible values of this bit and are labeled by 0 and 1 respectively. Because all bits written on the blackboard up to the current time are observed by all nodes, the nodes can keep track of the progress of the protocol in the binary tree. The value of the bit written by node l_v (when the protocol is at node v of the binary tree) can depend on the sample X_{l_v} observed by this node (and implicitly on all bits previously written on the blackboard encoded in the position of the node v in the binary tree). Therefore, this bit can be represented by a function $b_v(x) = p_v(1|x) \in [0,1]$, which we associate with the tree node v; node l_v transmits 1 with probability $b_v(X_{l_v})$ and 0 with probability $1-b_v(X_{l_v})$. Note that a proper labeling of the binary tree together with the collection of functions $\{b_v(\cdot)\}$ (where v ranges over all internal tree nodes) completely characterizes all possible (possibly probabilistic) blackboard communication strategies for the nodes. See Figure 1 for an example.

The k-bit communication constraint for each node can be viewed as a labeling constraint for the binary tree; for each $i \in [n]$, each possible path from the root node to a leaf node can visit exactly k internal nodes with label i. In particular, the depth of the binary tree is nk and there is one-to-one correspondence between all possible transcripts $y \in \{0,1\}^{nk}$ and paths in the tree. Let $\tau(y)$ denote the set of nodes v that are traversed by the path associated with transcript y.

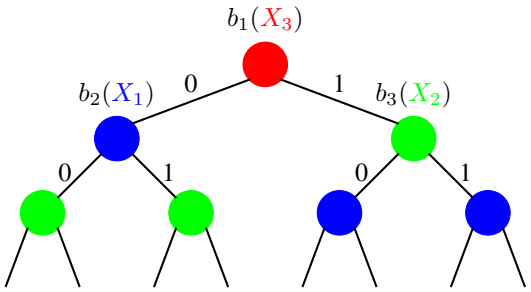

Fig. 1. A valid blackboard protocol for 3 nodes each writing a single bit on the blackboard.

Let $b_{v,y}(x_{l_v}) = b_v(x_{l_v})$ if the path associated with y takes the "1" branch after node v, and $b_{v,y}(x_{l_v}) = 1 - b_v(x_{l_v})$ otherwise. The probability distribution of Y can be written as

$$\mathbb{P}(Y=y) = \mathbb{E}\left[\prod_{v\in\tau(y)} b_{v,y}(X_{l_v})\right]$$

so that by the independence of the X_i,

$$\begin{aligned}\mathbb{P}(Y=y) &= \prod_{i=1}^{n} \mathbb{E}\left[\prod_{v\in\tau(y):\, l_v=i} b_{v,y}(X_i)\right] \\ &= \prod_{i=1}^{n} \mathbb{E}\left[p_{i,y}(X_i)\right]\end{aligned}$$

where $p_{i,y}(x_i) = \prod_{v\in\tau(y):\, l_v=i} b_{v,y}(x_i)$. Along with this characterization of the probability distribution of Y, we will need the following Lemma that appears in [2].

Lemma 1. *For each $j = 1, \ldots, d$,*

$$\sum_y \prod_{i\neq j} \mathbb{E}[p_{i,y}(X_i)] = 2^k \ .$$

III. Fisher Information

Recall that in the scalar case (when $\theta \in \mathbb{R}$), the Fisher information from a sample $X \sim P_\theta$, where P_θ has density $f(x|\theta)$ with respect to some base measure ν, can be written in terms of the score function

$$S_\theta(x) = \frac{\partial}{\partial\theta} \log f(x|\theta) \ .$$

The Fisher information from X for estimating θ is then

$$I_X(\theta) = \mathbb{E}[S_\theta(X)^2]$$

where $\mathbb{E}$ denotes taking the expectation with respect to $f(x|\theta)$.

In the vector case, when $\theta \in \mathbb{R}^d$, we have a score function for each component θ_i,

$$S_{\theta_i}(x) = \frac{\partial}{\partial\theta_i} \log f(x|\theta) \ ,$$

and a score function vector

$$S_\theta(X) = (S_{\theta_1}(X), \ldots, S_{\theta_d}(X)) \ .$$

In this paper we will be interested in the sum of the Fisher informations from the transcript Y for each component θ_i, i.e.

$$I_Y(\theta) = \sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i} \log \mathbb{P}(Y=y)\right)^2\right] \ .$$

In order to lower bound the minimax risk under the blackboard model, we will proceed by characterizing and upper bounding $I_Y(\theta)$.

Proposition 1. *The score for component θ_i can be written as*

$$\frac{\partial}{\partial \theta_i} \log \mathbb{P}(Y=y) = \sum_{j=1}^{n} \frac{\mathbb{E}[S_{\theta_i}(X_j) p_{j,y}(X_j)]}{\mathbb{E}[p_{j,y}(X_j)]} \ .$$

Proof. The main difficulty here is that we need to interchange integration over the sample space $\mathcal{X}$ and differentiation with respect to θ_i:

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \log \mathbb{P}(Y=y) &= \sum_{j=1}^{n} \frac{\partial}{\partial \theta_i} \log \mathbb{E}\left[p_{j,y}(X_j)\right] \\
&= \sum_{j=1}^{n} \frac{\frac{\partial}{\partial \theta_i} \mathbb{E}\left[p_{j,y}(X_j)\right]}{\mathbb{E}\left[p_{j,y}(X_j)\right]} \\
&= \sum_{j=1}^{n} \frac{\frac{\partial}{\partial \theta_i} \int_{\mathcal{X}} p_{j,y}(x_j) f(x_j|\theta) d\nu(x_j)}{\mathbb{E}\left[p_{j,y}(X_j)\right]} \\
&= \sum_{j=1}^{n} \frac{\int_{\mathcal{X}} p_{j,y}(x_j) \frac{\partial}{\partial \theta_i} f(x_j|\theta) d\nu(x_j)}{\mathbb{E}\left[p_{j,y}(X_j)\right]} \\
&= \sum_{j=1}^{n} \frac{\mathbb{E}[S_{\theta_i}(X_j) p_{j,y}(X_j)]}{\mathbb{E}[p_{j,y}(X_j)]} \ .
\end{aligned}$$

This interchange can be justified when $f(x|\theta)$ satisfies the regularity conditions from [1]; namely that $\sqrt{f(x|\theta)}$ is continuously differentiable with respect to each θ_j for ν-almost all $x \in \mathcal{X}$, and that $\mathbb{E}[S_{\theta_j}(X)^2]$ exists and is continuous in θ_j. See also [8]. □

The Fisher information from Y for estimating the component θ_i is then

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log \mathbb{P}(Y=y)\right)^2\right] \\
&\quad = \sum_{j,k,y} \mathbb{P}(Y=y) \frac{\mathbb{E}[S_{\theta_i}(X_j) p_{j,y}(X_j)] \mathbb{E}[S_{\theta_i}(X_k) p_{k,y}(X_k)]}{\mathbb{E}[p_{j,y}(X_j)] \mathbb{E}[p_{k,y}(X_k)]} \ .
\end{aligned}$$

Note that when $j \neq k$ the terms within this summation are zero:

$$\begin{aligned}
&\sum_{y} \mathbb{P}(Y=y) \frac{\mathbb{E}[S_{\theta_i}(X_j) p_{j,y}(X_j)] \mathbb{E}[S_{\theta_i}(X_k) p_{k,y}(X_k)]}{\mathbb{E}[p_{j,y}(X_j)] \mathbb{E}[p_{k,y}(X_k)]} \\
&\quad = \mathbb{E}\left[S_{\theta_i}(X_j) S_{\theta_i}(X_k) \sum_{y} \prod_{l=1}^{n} p_{l,y}(X_l)\right] \\
&\quad = \mathbb{E}\left[S_{\theta_i}(X_j) S_{\theta_i}(X_k)\right] = 0 \ . \qquad (1)
\end{aligned}$$

The step in (1) follows since $\prod_{l=1}^{n} p_{l,y}(x_l)$ describes the probability that $Y = y$ for fixed samples $x_1, \ldots, x_n$, and thus $\sum_y \prod_{l=1}^{n} p_{l,y}(x_l) = 1$.

Returning to the Fisher information from Y we have that

$$\begin{aligned}
&\sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log \mathbb{P}(Y=y)\right)^2\right] \\
&\quad = \sum_{y} \mathbb{P}(Y=y) \sum_{j,d} \left(\frac{\mathbb{E}[S_{\theta_i}(X_j) p_{j,y}(X_j)]}{\mathbb{E}[p_{j,y}(X_j)]}\right)^2 \ . \qquad (2)
\end{aligned}$$

Let $\mathbb{E}_{j,y}$ denote taking expectation with respect to the new density

$$\frac{p_{j,y}(x_j) f(x_j|\theta)}{\mathbb{E}[p_{j,y}(X_j)]} \ .$$

Now we can simplify (2) as

$$\begin{aligned}
I_Y(\theta) &= \sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log \mathbb{P}(Y=y)\right)^2\right] \\
&= \sum_{y} \mathbb{P}(Y=y) \sum_{j} \|\mathbb{E}_{j,y}[S_\theta(X_j)]\|^2 \ . \qquad (3)
\end{aligned}$$

IV. Upper Bounds on Fisher Information

Using the expression for the Fisher information from (3), we will develop upper bounds for this Fisher information under assumptions on the tail of the distribution of $S_\theta(X)$. In this first bound we only assume that the projection of the score function vector $S_\theta(X)$ onto any unit vector has finite variance.

Theorem 1. *Suppose that*

$$\mathbb{E}[\langle u, S_\theta(X) \rangle^2] \leq I_0$$

for some constant I_0 and any unit vector $u \in \mathbb{R}^d$. Then

$$\sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log \mathbb{P}(Y=y)\right)^2\right] \leq n 2^k I_0 \ .$$

Proof. Picking

$$u = \frac{\mathbb{E}_{j,y}[S_\theta(X)]}{\|\mathbb{E}_{j,y}[S_\theta(X)]\|} \ ,$$

the Cauchy-Schwarz inequality implies

$$\begin{aligned}
&\mathbb{E}[p_{j,y}(X)] \|\mathbb{E}_{j,y}[S_\theta(X)]\|^2 \\
&\quad = \frac{1}{\mathbb{E}[p_{j,y}(X)]} \left(\mathbb{E}[\langle u, S_\theta(X) \rangle p_{j,y}(X)]\right)^2 \\
&\quad \leq \frac{1}{\mathbb{E}[p_{j,y}(X)]} \mathbb{E}[\langle u, S_\theta(X) \rangle^2] \mathbb{E}[p_{j,y}(X)^2] \\
&\quad \leq \frac{1}{\mathbb{E}[p_{j,y}(X)]} \mathbb{E}[\langle u, S_\theta(X) \rangle^2] \mathbb{E}[p_{j,y}(X)] \\
&\quad = \mathbb{E}[\langle u, S_\theta(X) \rangle^2] \ .
\end{aligned}$$

Together with (3) this gives

$$\begin{aligned}
&\sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log \mathbb{P}(Y=y)\right)^2\right] \\
&\quad = \sum_{y} \mathbb{P}(Y=y) \sum_{j} \|\mathbb{E}_{j,y}[S_\theta(X_j)]\|^2 \\
&\quad \leq \sum_{j,y} I_0 \prod_{i \neq j} \mathbb{E}[p_{i,y}(X)] \\
&\quad = I_0 n 2^k \ .
\end{aligned}$$

The last equality follows from Lemma 1. □

Recall that for $p \geq 1$, the Ψ_p Orlicz norm of a random variable X is defined as

$$\|X\|_{\Psi_p} = \inf\{K \in (0, \infty) \mid \mathbb{E}[\Psi_p(|X|/K)] \leq 1\}$$

where $\Psi_p(x) = \exp(x^p) - 1$. A random variable with finite $p = 1$ Orlicz norm is sub-exponential, while a random variable with finite $p = 2$ Orlicz norm is sub-Gaussian [9]. Our second theorem shows that when the Ψ_p Orlicz norm of the projection of $S_\theta(X)$ onto any unit vector is bounded by some finite constant, the Fisher information can increase at most polynomially with k with order $k^{\frac{2}{p}}$.

Theorem 2. *Suppose that*

$$\|\langle u, S_\theta(X)\rangle\|_{\Psi_p} \le N$$

for some constant N, $p \ge 1$, and any unit vector $u \in \mathbb{R}^d$. Then

$$\sum_{i=1}^{d} \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log \mathbb{P}(Y=y)\right)^2\right] \le 4N^2 n k^{\frac{2}{p}} .$$

Proof. Again letting

$$u = \frac{\mathbb{E}_{j,y}[S_\theta(X)]}{\|\mathbb{E}_{j,y}[S_\theta(X)]\|}$$

we have

$$\begin{aligned}
2 &\ge \mathbb{E}[\exp((|\langle u, S_\theta(X)\rangle|/N)^p)] \\
&\ge \mathbb{E}[p_{j,y}(X)\exp((|\langle u, S_\theta(X)\rangle|/N)^p)] \\
&\ge \mathbb{E}[p_{j,y}(X)]\mathbb{E}_{j,y}[\exp((|\langle u, S_\theta(X)\rangle|/N)^p)] \\
&\ge \mathbb{E}[p_{j,y}(X)]\exp\left(\left|\frac{1}{N}\mathbb{E}_{j,y}[\langle u, S_\theta(X)\rangle]\right|^p\right) \\
&\ge \mathbb{E}[p_{j,y}(X)]\exp\left(\left(\frac{1}{N}\mathbb{E}_{j,y}[\langle u, S_\theta(X)\rangle]\right)^p\right)
\end{aligned}$$

and

$$\|\mathbb{E}_{j,y}[S_\theta(X_j)]\| \le N\left(\log\frac{2}{\mathbb{E}[p_{j,y}(X)]}\right)^{\frac{1}{p}} .$$

Continuing from (3),

$$\begin{aligned}
&\sum_{i=1}^{d}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\mathbb{P}(Y=y)\right)^2\right] \\
&\quad\le N^2\sum_y \mathbb{P}(Y=y)\sum_j\left(\log\frac{2}{\mathbb{E}[p_{j,y}(X)]}\right)^{\frac{2}{p}} \\
&\quad= N^2\sum_{y,j}\left(\prod_{i\ne j}\mathbb{E}[p_{i,y}(X)]\right)\mathbb{E}[p_{j,y}(X)]\left(\log\frac{2}{\mathbb{E}[p_{j,y}(X)]}\right)^{\frac{2}{p}}
\end{aligned} \tag{4}$$

Finally, by upper bounding (4) with the upper concave envelope ϕ of $x \mapsto x\left(\log\frac{2}{x}\right)^{\frac{2}{p}}$ on $[0,1]$, and then using both Lemma 1 and Jensen's inequality:

$$\begin{aligned}
&\sum_{i=1}^{d}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\mathbb{P}(Y=y)\right)^2\right] \\
&\quad\le N^2\sum_{y,j}\left(\prod_{i\ne j}\mathbb{E}[p_{i,y}(X)]\right)\phi\left(\mathbb{E}[p_{j,y}(X)]\right) \\
&\quad\le N^2\sum_j 2^k\phi\left(\sum_y\frac{1}{2^k}\mathbb{P}(Y=y)\right) \\
&\quad= N^2 n(k+1)^{\frac{2}{p}} \qquad (5) \\
&\quad\le 4N^2 n k^{\frac{2}{p}} .
\end{aligned}$$

In line (5) we have used the fact that ϕ matches the function $x \mapsto x\left(\log\frac{2}{x}\right)^{\frac{2}{p}}$ for $0 < x \le \frac{1}{2}$. □

V. Minimax Lower Bounds

Using the upper bounds on Fisher information developed in Theorems 1 and 2, we will apply the van Trees inequality [10] to achieve a minimax lower bound on the risk for the underlying distributed estimation problem. We will see that when the score function has finite variance (as in Theorem 1), the lower bound decreases exponentially with k; and when the score function has some sub-Gaussian structure (as in Theorem 2 with $p = 2$), the lower bound decays like $1/k$.

Theorem 3. *Suppose $\Theta = [-B, B]^d$. For any estimator $\hat\theta(Y)$ and communication protocol $\Pi \in \Pi_{\mathsf{BB}}$, if $S_\theta(X)$ satisfies the hypotheses in Theorem 1 then*

$$\sup_{\theta\in\Theta}\mathbb{E}[\|\hat\theta-\theta\|^2] \ge \frac{d^2}{I_0 2^k n + \frac{d\pi^2}{B^2}},$$

and if $S_\theta(X)$ satisfies the hypotheses in Theorem 2 then

$$\sup_{\theta\in\Theta}\mathbb{E}[\|\hat\theta-\theta\|^2] \ge \frac{d^2}{4N^2 k^{\frac{2}{p}} n + \frac{d\pi^2}{B^2}} .$$

Proof. Consider the squared error risk in estimating θ:

$$\mathbb{E}[\|\theta-\hat\theta\|^2] = \sum_{i=1}^{d}\mathbb{E}[(\theta_i-\hat\theta_i)^2] .$$

In order to lower bound this risk, we will use the van Trees inequality from [10]. Suppose we have a prior μ_i for the parameter θ_i. The van Trees inequality for the component θ_i gives

$$\begin{aligned}
&\int_{-B}^{B}\mathbb{E}[(\hat\theta_i(Y)-\theta_i)^2]\mu_i(\theta_i)d\theta_i \\
&\qquad\ge \frac{1}{\int_{-B}^{B}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\mathbb{P}(Y=y)\right)^2\right]\mu_i(\theta_i)d\theta_i + I(\mu_i)}
\end{aligned} \tag{6}$$

where $I(\mu_i) = \int_{-B}^{B}\frac{\mu_i'(\theta)^2}{\mu_i(\theta)}d\theta$ is the Fisher information from the prior. Note that the required regularity condition that

$$\mathbb{E}\left[\frac{\partial}{\partial\theta_i}\log\mathbb{P}(Y=y)\right] = 0$$

follows trivially since the expectation over Y is just a finite sum:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\theta_i}\log\mathbb{P}(Y=y)\right] &= \sum_m\frac{\partial}{\partial\theta_i}\mathbb{P}(Y=y) \\
&= \frac{\partial}{\partial\theta_i}\sum_m\mathbb{P}(Y=y) = 0 .
\end{aligned}$$

The prior μ_i can be chosen to minimize this Fisher information and achieve $I(\mu_i) = \pi^2/B^2$ [8]. Let $\mu(\theta) = \prod_i\mu_i(\theta_i)$. By

summing over each component,

$$\begin{aligned}
&\int_\Theta \sum_{i=1}^d \mathbb{E}[(\theta_i - \hat{\theta}_i)^2]\mu(\theta)d\theta \\
&\geq \sum_{i=1}^d \frac{1}{\int_\Theta \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log \mathbb{P}(Y=y)\right)^2\right]\mu(\theta)d\theta + \frac{\pi^2}{B^2}} \qquad (7)\\
&= d\sum_{i=1}^d \frac{1}{d}\frac{1}{\int_\Theta \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log \mathbb{P}(Y=y)\right)^2\right]\mu(\theta)d\theta + \frac{\pi^2}{B^2}} \\
&\geq d\,\frac{1}{\sum_{i=1}^d \frac{1}{d}\int_\Theta \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log \mathbb{P}(Y=y)\right)^2\right]\mu(\theta)d\theta + \frac{\pi^2}{B^2}} \qquad (8)\\
&= \frac{d^2}{\int_\Theta I_Y(\theta)\mu(\theta)d\theta + \frac{d\pi^2}{B^2}}\ .
\end{aligned}$$

Therefore,

$$\sup_{\theta\in\Theta}\mathbb{E}[\|\hat{\theta}(Y)-\theta\|^2] \geq \frac{d^2}{\sup_{\theta\in\Theta} I_Y(\theta) + \frac{d\pi^2}{B^2}}\ . \qquad (9)$$

The inequaltiy (8) follows from Jensen's inequality via the convexity of $x \mapsto 1/x$ for $x > 0$, and the inequality (7) follows both from this convexity and (6). □

In the proof above we could have used the multivariate version of the van Trees inequality [10] to arrive at the same result, but we have used the single-variable version in each coordinate instead in order to simplify the required regularity conditions.

A. Applications to Common Statistical Models

Theorem 3 gives a lower bound on the minimax risk for the distributed estimation of θ under many common statistical models. We summarize some of these results in the following corollaries. The lower bound for the Gaussian location model in Corollary 1 reproduces the lower bound from [2] and is valid with fewer samples n. It also matches the achievability result from [7]. For the distribution estimation problem, Corollary 2 matches both the lower bound and achievability result from [2]. We see that tight bounds can be shown by directly characterizing the Fisher information from the blackboard transcript.

Corollary 1 (Gaussian location model)**.** *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$ with $[-B, B]^d \subset \Theta$. For any $\Pi \in \Pi_{\mathsf{BB}}$ and $nB^2\min\{k,d\} \geq d\sigma^2$, we have*

$$\sup_{\theta\in\Theta}\mathbb{E}[\|\hat{\theta}(Y)-\theta\|^2] \geq C\sigma^2\max\left\{\frac{d^2}{nk}, \frac{d}{n}\right\}$$

for any estimator $\hat{\theta}$ where $C > 0$ is a universal constant independent of n, k, d, σ^2, B.

The condition that $nB^2\min\{k,d\} \geq d\sigma^2$ is a weak condition that ensures that we can ignore the second term in the denominator of (9). For fixed B, σ, this condition is weaker than just assuming that n is at least order d, which is required for consistent estimation anyways. We will make a similar assumption in the subsequent corollaries.

Corollary 2 (distribution estimation)**.** *Suppose that $\mathcal{X} = \{1, \ldots, d+1\}$ and that*

$$f(x|\theta) = \theta_x\ .$$

Let Θ be the probability simplex with $d+1$ variables. For any $\Pi \in \Pi_{\mathsf{BB}}$ and $n\min\{2^k, d\} \geq d^2$, we have

$$\sup_{\theta\in\Theta}\mathbb{E}[\|\hat{\theta}(Y)-\theta\|^2] \geq C\max\left\{\frac{d}{n2^k}, \frac{1}{n}\right\}$$

for any estimator $\hat{\theta}$ where $C > 0$ is a universal constant independent of n, k, d.

In this final corrolary, we give an example where the score function is sub-exponential, and therefore Theorem 2 with $p = 1$ yields a quadratic dependence on k. It is unknown whether or not this is order-wise optimal.

Corollary 3 (Gaussian covariance estimation)**.** *Suppose that $X \sim \mathcal{N}(0, \mathrm{diag}(\theta_1, \ldots, \theta_d))$ with $[\sigma^2_{\min}, \sigma^2_{\max}]^d \subset \Theta$. Then for $n\left(\sigma^2_{\max} - \sigma^2_{\min}\right)^2\min\{k^2, d\} \geq d\sigma^4_{\min}$, we have*

$$\sup_{\theta\in\Theta}\mathbb{E}[\|\hat{\theta}-\theta\|^2] \geq C\sigma^4_{\min}\max\left\{\frac{d^2}{nk^2}, \frac{d}{n}\right\}$$

for any communication protocol $\Pi \in \Pi_{\mathsf{BB}}$ and any estimator $\hat{\theta}$, where $C > 0$ is a universal constant independent of $n, k, d, \sigma_{min}, \sigma_{max}$.

References

[1] Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür, "A Geometric Characterization of Fisher Information from Quantized Samples with Applications to Distributed Statistical Estimation," *Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing*, 2018.

[2] Yanjun Han, Ayfer Özgür and Tsachy Weissman, "Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints," *Proceedings of Machine Learning Research*, 75:1-26, 2018.

[3] Yanjun Han, Pritam Mukherjee, Ayfer Özgür and Tsachy Weissman, "Distributed Statistical Estimation of High-Dimensional and Nonparametric Distributions," *Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT)*.

[4] Shun-ichi Amari, "On Optimal Data Compression in Multiterminal Statistical Inference," *IEEE Transactions on Information Theory*, 57(9):5577-5587, 2011.

[5] Yuchen Zhang, John Duchi, Micheal I Jordan, and Martin J Wainwright, "Information-Theoretic Lower Bounds for Distributed Statistical Estimation with Communication Constraints," *Advances in Neural Information Processing Systems*, pp. 2328-2336, 2013.

[6] Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and 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*, pp 1011–1020, ACM, 2016.

[7] Ankit Garg, Tengyu Ma, and Huy Nguyen, "On communication cost of distributed statistical estimation and dimensionality," *Advances in Neural Information Processing Systems*, pp. 2726–2734, 2014.

[8] A. A. Borovkov. "Mathematical Statistics," Gordon and Breach Science Publishers, Amsterdam, 1998.

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

[10] Richard D. Gill and Boris Y. Levit. "Applications of the van Trees inequality: a Bayesian Cramér-Rao Bound," *Bernoulli* 1(1/2), pp 059-079, 1995.

[11] E. Kushilevitz and N. Nisan. "Communication Complexity," Cambridge University Press, 1997.

[12] K. Pearson. "Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material." *Philosophical Trans. of the Royal Society of London*, 186:343–414, 1895.

[13] Jayadev Acharya, Clément L. Canonne, Himanshu Tyagi. "Inference under Information Constraints I: Lower Bounds from Chi-Square Contraction," *arXiv preprint*, *arXiv*:1812.11476, 2018.

Acknowledgements

This work was supported in part by NSF award CCF-1704624 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.