A Geometric Characterization of Fisher Information from Quantized Samples with Applications to Distributed Statistical Estimation

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

***Abstract*— Consider the Fisher information for estimating a vector $\theta \in \mathbb{R}^d$ from the quantized version of a statistical sample $X \sim f(x|\theta)$. Let M be a k-bit quantization of X. We provide a geometric characterization of the trace of the Fisher information matrix $I_M(\theta)$ in terms of the score function $S_\theta(X)$. When $k = 1$, we exactly solve the extremal problem of maximizing this geometric quantity for the Gaussian location model, which allows us to conclude that in this model, a half-space quantization is the one-bit quantization that maximizes $Tr(I_M(\theta))$. Under assumptions on the tail of the distribution of $S_\theta(X)$ projected onto any unit vector in $\mathbb{R}^d$, we give upper bounds demonstrating how $Tr(I_M(\theta))$ can scale with k. We apply these results to find lower bounds on the minimax risk of estimating θ from multiple quantized samples of X, for example in a distributed setting where the samples are distributed across multiple nodes and each node has a total budget of k-bits to communicate its sample to a centralized estimator. Our bounds apply in a unified way to many common statistical models including the Gaussian location model and discrete distribution estimation, and they recover and generalize existing results in the literature with simpler and more transparent proofs.**

I. Introduction

Fisher information plays a central role in the standard statistical problem of estimating some parameter θ, that can take its value from a set $\Theta \subseteq \mathbb{R}^d$, given a statistical sample $X \in \mathcal{X}$. In this work, we study the effects of quantization of the sample X on the Fisher information for estimating θ, and the related question of how to efficiently represent X with a given number of bits so as to maximize the Fisher information it provides about θ. Quantization of data is of interest in many different settings. For example, in many machine learning systems data is distributed across different machines or generated in a distributed fashion, and it needs to be communicated efficiently while preserving maximal information about an underlying parameter of interest. In such modern applications, θ can have a very large dimension, i.e. $\theta \in \mathbb{R}^d$, where d can potentially be very large. In this vector case, one quantity of interest is the sum of the Fisher informations for estimating each individual component θ_i for $i = 1, \dots, d$, or equivalently the trace of the Fisher information matrix for estimating θ.

In this paper, we provide a geometric characterization of the trace of the Fisher information matrix for estimating θ from a k-bit quantized sample of X. This characterization has a natural geometric interpretation in terms of the score-function $S_\theta(X)$, and enables us to prove upper bounds on the trace of the Fisher information matrix that hold for any k-bit quantization strategy. When these upper bounds are used together with the van Trees inequality, they easily lend themselves to lower bounds on the minimax squared error risk of estimating θ in a distributed setting, where there are multiple nodes each observing an independent and identically distributed sample from the distribution of X, and each node has k-bits to communicate its sample to a centralized estimator. The central estimator then estimates the underlying parameter θ from the k-bit messages it receives from the nodes. The messages can be communicated independently, or in an interactive fashion according to a blackboard protocol using private/public randomness. We recover and generalize existing results in the literature [13], [14], [4], [1] for this setting under different statistical models (including the Gaussian location model and discrete distribution estimation) in a unified way with simpler and more transparent proofs.

In a flavor similar to [1], our upper bounds on the trace of the Fisher information matrix reveal that the the tail of the distribution of the score function $S_\theta(X)$ dictates the dependence on the quantization rate k. If the projection of the score function vector onto any unit vector has finite variance at most I_0, then the trace of the Fisher information for estimating θ from the quantized sample is upper bounded by $I_0 2^k$. Furthermore, if the projection of the score function vector onto any unit vector has finite Ψ_p Orlicz norm N, then we show that the trace of the Fisher information for estimating θ from the quantized sample is $O\left(N^2 k^{\frac{2}{p}}\right)$ with a small absolute constant that is independent of d. This implies that when the score function is sub-Gaussian, which is the case for the Gaussian location model, the Fisher information increases linearly in k. On the other hand, when the score-function is sub-exponential, the trace of the Fisher information matrix increases at most as k^2. These qualitatively different scalings for the trace of Fisher information matrix translate to qualitatively different minimax bounds for the associated distributed estimation problems.

Finally, we give a "most Fisher-informative bit" result that demonstrates that for the Gaussian location model, where $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$, the one-bit quantization that maximizes the trace of the Fisher information for estimating the mean parameter θ is given by two half-spaces whose defining hyperplane intersects the true mean θ. This is reminiscent of Borell's isoperimetric result in Gauss space that implies that a half-space quantization maximizes the mutual information

between a Gaussian random vector and a one-bit quantized version of the same random vector that has been corrupted with additive Gaussian noise [9],[11].

There has been some previous work in analyzing Fisher information from a quantized scalar random variable such as [5],[6],[7],[8]. Here we instead consider the arbitrary quantization of a random vector, and are able to study the trade-off between the quantization rate k and the number of parameter components d that we are trying to estimate. Our work follows [1], which introduces a similar geometric approach to obtain minimax bounds for distributed parameter estimation under communication constraints.

II. Fisher information from a quantized sample

Let P_θ be a family of probability measures on $\mathcal{X}$ parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$. Suppose each P_θ is dominated by some base measure ν and that each P_θ has density $f(x|\theta)$ with respect to ν. Let $X \in \mathcal{X}$ be a single sample drawn from $f(x|\theta)$. Any (potentially randomized) k-bit quantization strategy for X can be expressed in terms of the conditional probabilities

$$b_m(x) = p(m|x) \quad \text{for} \quad m \in [1:2^k], \quad x \in \mathcal{X}.$$

We assume that there is a well-defined joint probability distribution with density

$$f(x, m|\theta) = f(x|\theta)p(m|x)$$

and that $p(m|x)$ is a regular conditional distribution. For a given $\theta \in \mathbb{R}^d$ and quantization strategy, denote the likelihood that the quantization M takes a specific value m by $p(m|\theta)$. Let

$$\begin{aligned} S_\theta(m) &= (S_{\theta_1}(m), \ldots, S_{\theta_d}(m)) \\ &= \left(\frac{\partial}{\partial \theta_1} \log p(m|\theta), \ldots, \frac{\partial}{\partial \theta_d} \log p(m|\theta) \right) \end{aligned}$$

be the score of this likelihood. In an abuse of notation, we will also denote the score of the likelihood $f(x|\theta)$ by

$$\begin{aligned} S_\theta(x) &= (S_{\theta_1}(x), \ldots, S_{\theta_d}(x)) \\ &= \left(\frac{\partial}{\partial \theta_1} \log f(x|\theta), \ldots, \frac{\partial}{\partial \theta_d} \log f(x|\theta) \right) . \end{aligned}$$

The Fisher information matrix for estimating θ from M is

$$I_M(\theta) = \mathbb{E}[S_\theta(M)^T S_\theta(M)]$$

and likewise the Fisher information matrix for estimating θ from an X is

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

We will assume throughout that $f(x|\theta)$ satisfies the following regularity conditions:

(1) For each j and fixed $\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_{j_d}$, the function $\sqrt{f(x|\theta)}$, thought of as a function of θ_j, is continuously differentiable with respect to θ_j at μ-almost every $x \in \mathcal{X}$.

(2) For each j and all θ, the expected value $[I_X(\theta)]_{jj} = \mathbb{E}[S_{\theta_j}(X)^2]$ exists and is continuous in θ_j.

These two conditions justify interchanging differentiation and integration as in

$$\begin{aligned} \frac{\partial}{\partial \theta_j} p(m|\theta) &= \frac{\partial}{\partial \theta_j} \int f(x|\theta)p(m|x)d\nu(x) \\ &= \int \frac{\partial}{\partial \theta_j} f(x|\theta)p(m|x)d\nu(x) \end{aligned}$$

for each j, and they also ensure that $p(m|\theta)$ is continuously differentiable with respect to each θ_j (see Lemma 1, Section 26 in [15]). We will also assume, without loss of generality, that $f(x|\theta) > 0$. For each fixed θ, this can be done by restricting the domain $\mathcal{X}$ to only include those x values such that $f(x|\theta) > 0$ when taking an expectation, or equivalently, by defining $S_\theta(x) = 0$ whenenver $f(x|\theta) = 0$. In the same way we will assume that $p(m|\theta) > 0$.

Our first two lemmas establish a geometric interpretation of the trace $Tr(I_M(\theta))$. The first lemma is a slight variant of Theorems 1 and 2 from [3], and our debt to that work is clear.

Lemma 1: The (i,i)-th entry of the Fisher information matrix $I_M(\theta)$ is

$$[I_M(\theta)]_{i,i} = \mathbb{E}\left[\mathbb{E}\left[S_{\theta_i}(X)|M \right]^2 \right] .$$

The inner conditional expectation is with respect to the distribution $f(x|\theta)$, while the outer expectation is over the conditioning random variable M.

Proof:

$$\begin{aligned} \mathbb{E}\left[S_{\theta_i}(X)|m \right] &= \int S_{\theta_i}(x) \frac{f(x|\theta)p(m|x)}{p(m|\theta)} d\nu(x) \\ &= \int \frac{\frac{\partial}{\partial \theta_i} f(x|\theta)}{f(x|\theta)} \frac{f(x|\theta)p(m|x)}{p(m|\theta)} d\nu(x) \\ &= \frac{1}{p(m|\theta)} \int \frac{\partial}{\partial \theta_i} f(x|\theta)p(m|x)d\nu(x) \\ &= \frac{1}{p(m|\theta)} \frac{\partial}{\partial \theta_i} \int f(x|\theta)p(m|x)d\nu(x) \\ &= S_{\theta_i}(m) \end{aligned}$$

Squaring both sides and taking the expectation over M gives

$$\mathbb{E}[S_{\theta_i}(M)^2] = \mathbb{E}\left[\mathbb{E}\left[S_{\theta_i}(X)|M \right]^2 \right]$$

where the left-hand side is by definition $[I_M(\theta)]_{i,i}$. ∎

Lemma 2: The trace of the Fisher information matrix $I_M(\theta)$ can be written as

$$\begin{aligned} Tr(I_M(\theta)) &= \sum_{i=1}^{d} [I_M(\theta)]_{i,i} \\ &= \sum_m p(m|\theta) \| \mathbb{E}[S_\theta(X)|m] \|^2 . \end{aligned} \tag{1}$$

Proof: By Lemma 1,

$$\begin{aligned}\sum_{i=1}^{d}[I_M(\theta)]_{i,i} &= \sum_{i=1}^{d}\mathbb{E}\left[\mathbb{E}\left[S_{\theta_i}(X)|M\right]^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{d}\mathbb{E}\left[S_{\theta_i}(X)|M\right]^2\right] \\ &= \mathbb{E}\left[\|\mathbb{E}[S_\theta(X)|M]\|^2\right] \\ &= \sum_m p(m|\theta)\|\mathbb{E}[S_\theta(X)|m]\|^2 \ .\end{aligned}$$

■

In order to get some geometric intuition for the quantity (1), consider a special case where the quantization is deterministic and the score function $S_\theta(x)$ is a bijection. In this case, the quantization map partitions the space $\mathcal{X}$ into disjoint quantization bins, and this induces a corresponding partitioning of the score functions values $S_\theta(x)$. Each vector $\mathbb{E}[S_\theta(X)|m]$ is then the centroid of the set of $S_\theta(x)$ values corresponding to quantization bin m (with respect to the induced probability distribution on $S_\theta(X)$). Lemma 2 shows that $Tr(I_M(\theta))$ is equal to the average magnitude squared of these centroid vectors.

A. Upper Bounds on $Tr(I_M(\theta))$

In this section, we give two upper bounds on $Tr(I_M(\theta))$. The proofs appear in Appendix A. The first theorem upper bounds $Tr(I_M(\theta))$ in terms of the variance of $S_\theta(X)$ when projected onto any unit vector.

Theorem 1: If for any $\theta \in \Theta$ and any unit vector $u \in \mathbb{R}^d$,

$$\text{var}(\langle u, S_\theta(X)\rangle) \leq I_0 \ ,$$

then

$$Tr(I_M(\theta)) \leq \min\{Tr(I_X(\theta)),\ 2^k I_0\} \ .$$

The upper bound $Tr(I_M(\theta)) \leq Tr(I_X(\theta))$ follows easily from the data processing inequality for Fisher information [10]. The theorem shows that when I_0 is finite, $Tr(I_M(\theta))$ can increase at most exponentially in k.

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 [16]. Our second theorem shows that when the Ψ_p Orlicz norm of the projection of $S_\theta(X)$ onto any unit vector is finite, $Tr(I_M(\theta))$ can increase at most like $k^{\frac{2}{p}}$.

Theorem 2: If for any $\theta \in \Theta$, some $p \geq 1$, and any unit vector $u \in \mathbb{R}^d$,

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

then

$$Tr(I_M(\theta)) \leq \min\{Tr(I_X(\theta)),\ CN^2 k^{\frac{2}{p}}\}$$

where $C = \frac{8}{e^2} + 4$.

We next apply the above two results to common statistical models. We will see that neither of these bounds is strictly stronger than the other and depending on the statistical model, one may yield a tighter bound than the other.

B. Applications to Common Statistical Models

Example 1 (Gaussian location model): Consider the Gaussian location model $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$ where we are trying to estimate the mean θ of a d-dimensional Gaussian random vector with fixed covariance $\sigma^2 I_d$. In this case,

$$S_\theta(x) = \frac{1}{\sigma^2}(\theta - x)$$

so that $S_\theta(X) \sim \mathcal{N}(0, 1/\sigma^2)$. Therefore

$$\|\langle u, S_\theta(X)\rangle\|_{\Psi_2} = \Theta\left(\frac{1}{\sigma}\right)$$

for any unit vector $u \in \mathbb{R}^d$, so by Theorem 2,

$$Tr(I_M(\theta)) = O\left(\frac{k}{\sigma^2}\right) \ . \tag{2}$$

Example 2 (variance of a Gaussian): Now suppose $X = (X_1, \ldots, X_d) \sim \mathcal{N}(0, \text{diag}(\theta_1, \ldots, \theta_d))$ and $\Theta \subseteq [\sigma^2_{\min}, \sigma^2_{\max}]^d$ with $\sigma_{\min} > 0$. The components of the score function are

$$S_{\theta_i}(x) = \frac{x_i^2}{2\theta_i^2} - \frac{1}{2\theta_i} \ .$$

Therefore each independent component $S_{\theta_i}(X)$ is chi-squared distributed with one degree of freedom and

$$\|\langle u, S_\theta(X)\rangle\|_{\Psi_1} = O\left(\frac{1}{\sigma^2_{\min}}\right) \ .$$

Using Theorem 2,

$$Tr(I_M(\theta)) = O\left(\left(\frac{k}{\sigma^2_{\min}}\right)^2\right) \ .$$

Example 3 (distribution estimation): Suppose that $\mathcal{X} = \{1, \ldots, d+1\}$ and that

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

Let $\theta_1, \ldots, \theta_d$ be the free parameters of interest and suppose they can vary from $\frac{1}{4d} \leq \theta_i \leq \frac{1}{2d}$. Note that

$$\theta_{d+1} = 1 - \sum_{i=1}^{d}\theta_i \ .$$

We have

$$S_{\theta_i}(x) = \begin{cases} \frac{1}{\theta_i} & , \ x = i \\ -\frac{1}{\theta_{d+1}} & , \ x = d+1 \\ 0 & , \ \text{otherwise} \end{cases}$$

for $i = 1, \ldots, d$. For any unit vector $u = (v_1, \ldots, v_d)$,

$$\begin{aligned}
&\text{var}(\langle u, S_\theta(X)\rangle) \\
&\quad = \sum_{x=1}^{d+1} \theta_x \left(\sum_{i=1}^{d} v_i S_{\theta_i}(x) \right)^2 \\
&\quad = \theta_{d+1} \frac{1}{\theta_{d+1}^2} \left(\sum_{i=1}^{d} v_i \right)^2 + \sum_{x=1}^{d} \theta_x \left(\sum_{i=1}^{d} v_i S_{\theta_i}(x) \right)^2 \\
&\quad \leq 2d + \sum_{x=1}^{d} \theta_x v_x^2 \frac{1}{\theta_x^2} \leq 6d \, .
\end{aligned}$$

By Theorem 1,

$$Tr(I_M(\theta)) = O(d2^k) \, . \tag{3}$$

Example 4 (product Bernoulli model): Consider $X = (X_1, \ldots, X_d) \sim \prod_{i=1}^{d} \text{Bern}(\theta_i)$. With this model,

$$S_{\theta_i}(x) = \begin{cases} \frac{1}{\theta_i} & , \; x_i = 1 \\ \frac{-1}{1-\theta_i} & , \; x_i = 0 \, . \end{cases}$$

If $\Theta = [1/2 - \epsilon, 1/2 + \epsilon]^d$ for some $0 < \epsilon < 1/2$, i.e. the model is relatively dense, then $\text{var}(\langle u, S_\theta(X)\rangle)$ and $\|\langle u, S_\theta(X)\rangle\|_{\Psi_2}^2$ are both $\Theta(1)$. In this case Theorem 1 gives

$$Tr(I_M(\theta)) = O(2^k)$$

while Theorem 2 gives

$$Tr(I_M(\theta)) = O(k) \, .$$

In this situation Theorem 2 gives the better bound. On the other hand, if $\Theta = [(1-\epsilon)\frac{1}{d}, (1+\epsilon)\frac{1}{d}]^d$, i.e. the model is sparse, then $\text{var}(\langle u, S_\theta(X)\rangle) = \Theta(d)$ and $\|\langle u, S_\theta(X)\rangle\|_{\Psi_2}^2 = \Theta(d^2)$. In this case Theorem 1 gives

$$Tr(I_M(\theta)) = O(d2^k)$$

while Theorem 2 gives

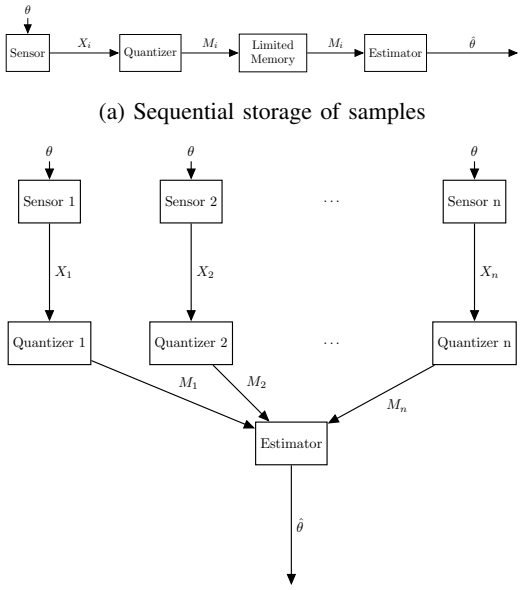
$$Tr(I_M(\theta)) = O(d^2 k) \, .$$

In the sparse case $Tr(I_X(\theta)) = \Theta(d^2)$, so only the bound from Theorem 1 is nontrivial. It is interesting that Theorem 2 is able to use the sub-Gaussian structure in the first case to yield a better bound – but in the second case, when the tail of the score function is essentially not sub-Gaussian, Theorem 1 yields the better bound.

III. Distributed Estimation of High-Dimensional Distributions and Parameters

In this section we will apply Theorems 1 and 2 to statistical estimation with multiple quantized samples. Let $X_1, \ldots, X_n$ be i.i.d. generated by the distribution $f(x|\theta)$. We consider three different models for quantizing these samples:

- Independent Quantization: under this model, each sample is independently quantized to k-bits. Formally, each sample X_i, for $i = 1, \ldots, n$, is encoded to a k-bit string M_i by a possibly randomized quantization strategy, denoted by $q_i(x_i) : \mathcal{X} \to [1 : 2^k]$, which can be expressed in terms of the conditional probabilities
$$p(m_i|x_i) \quad \text{for } m_i \in [1 : 2^k], \; x_i \mathcal{X}.$$
- Sequential Quantization: under this model, we assume that samples arrive sequentially and the quantization of the sample X_i can depend on the previously stored quantized samples $M_1, \ldots, M_{i-1}$ corresponding to the previously observed samples $X_1, \ldots, X_{i-1}$. Formally, each sample X_i, for $i = 1, \ldots, n$, is encoded to a k-bit string M_i by a set of possibly randomized quantization strategies $\{q_{m_1,\ldots,m_{i-1}}(x_i) : \mathcal{X} \to [1 : 2^k] : m_1, \ldots, m_{i-1} \in [1 : 2^k]\}$, where each strategy $q_{m_1,\ldots,m_{i-1}}(x_i)$ can be expressed in terms of the conditional probabilities
$$\begin{aligned} &p(m_i|x_i; m_1, \ldots, m_{i-1}) \\ &\quad \text{for } m_i \in [1 : 2^k] \text{ and } x_i \in \mathcal{X}. \end{aligned}$$



(a) Sequential storage of samples

(b) Distributed communication of samples

Fig. 1: Quantization for storage and communication

These two models are motivated by a scenario where a continuous stream of samples is captured sequentially and each sample is stored in digital memory by using k bits/sample. See Figure 1a. In the first case, each samples is quantized independently of the other samples (even though the quantization strategies for different bits can be different and jointly optimized ahead of time), while under the second model the quantization of each sample X_i can depend on the information $M_1, \ldots, M_i$ stored in the memory of the system at time i.

These two models can be equivalently used to model a distributed estimation setting where there are n sensors, each observing an independent sample from the underlying statistical model. See Figure 1b. Each sensor has k bits to communicate its sample to a centralized estimator. Under

the independent model, each sensor communicates its sample independently from the other nodes by transmitting a k-bit message. Under the sequential model, sensors are ordered and communication occurs in a sequential fashion; sensor i observes the k-bit messages $M_1, \ldots, M_i$ communicated by previous sensors and therefore its k-bit message M_i can depend on the previously transmitted messages $M_1, \ldots, M_i$. Note that this second model allows for some limited interaction between the sensors while the first one does not. For the distributed estimation scenario in Figure 1b, one can also consider a fully interactive communication model for the sensors, which we describe next.

- Blackboard Model: all sensors communicate via a publicly shown blackboard while the total number of bits each sensor can write in the final transcript Y is limited by k bits. When one sensor writes a message (bit) on the blackboard, all other sensors can see the content of the message. Formally, a blackboard communication protocol Π_{BB} can be viewed as a binary tree [18], where each internal node v of the tree is assigned a deterministic label $l_v \in [n]$ indicating the identity of the sensor to write the next bit on the blackboard if the protocol reaches 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 sensors 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) 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 node v; sensor 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 nodes) completely characterizes all possible (possibly probabilistic) communication strategies for the sensors. 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. Note that there is also one-to-one correspondence between $y \in \{0,1\}^{nk}$ and the k-bit messages $m_1, \ldots, m_n$ transmitted by the n sensors. In particular, the transcript $y \in \{0,1\}^{nk}$ contains the same amount of information as $m_1, \ldots, m_n$, since given the transcript y (and the protocol) one can infer $m_1, \ldots, m_n$ and vice versa (for the opposite direction note that the protocol specifies which sensor transmits first so given $m_1, \ldots, m_n$ one can follow the path in the protocol tree).

Under all the three quantization/communication models above, the end goal is to produce an estimate θ of the underlying parameter θ from the k-bit quantizations/messages $M_1, \ldots, M_n$ stored in the system/received by the central node. As usual, the encoding strategies or the protocols used in each case can be jointly optimized and agreed upon by all parties ahead of time.

We are interested in the quantity

$$I_{(M_1,\ldots,M_n)}(\theta)$$

under each model. (Note that under the blackboard model, the central estimator observes the transcript Y, however as we already argued $Tr(I_Y(\theta)) = Tr(I_{(M_1,\ldots,M_n)}(\theta))$.)

We have

$$\begin{aligned} Tr(I_{(M_1,\ldots,M_n)}(\theta)) &= \sum_{j=1}^{d} [I_{(M_1,\ldots,M_n)}(\theta)]_{j,j} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{d} [I_{M_i|(M_1,\ldots,M_{i-1})}(\theta)]_{j,j} \\ &= \sum_{i=1}^{n} \sum_{m_1,\ldots,m_{i-1}} p(m_1,\ldots,m_{i-1}|\theta) Tr(I_{M_i|(m_1,\ldots,m_{i-1})}(\theta)) \end{aligned} \tag{4}$$

due to the chain-rule for Fisher information. Under the independent model,

$$[I_{M_i|(m_1,\ldots,m_{i-1})}(\theta)]_{j,j} = [I_{M_i}(\theta)]_{j,j}.$$

Under the sequential and the blackboard models, conditioning on specific $m_1, \ldots, m_{i-1}$ only effects the distribution $p(m_i|\theta)$ by fixing the quantization strategy for X_i. Formally, for the sequential model

$$\begin{aligned} &\mathbb{P}(M_i = m_i|\theta; m_1, \ldots, m_{i-1}) \\ &\quad = \mathbb{P}(q_{m_1,\ldots,m_{i-1}}(X_i) = m_i|\theta; m_1, \ldots, m_{i-1}) \\ &\quad = \mathbb{P}(q_{m_1,\ldots,m_{i-1}}(X_i) = m_i|\theta), \end{aligned}$$

where the last step follows since $X_1, \ldots, X_{i-1}$ is independent of X_i and therefore conditioning of $m_1, \ldots, m_{i-1}$ does not change the distribution of X_i. A similar argument can be made for the blackboard model by observing that conditioning on messages $M_1, \ldots, M_{i-1}$ allows to effectively prune the encoding tree to a smaller tree which induces a new quantization strategy for M_i.

Since the bounds from Theorems 1 and 2 apply for any quantization strategy, they apply to each of the terms in (4), and the following statements hold under all three quantization models:

(i) Under the hypotheses in Theorem 1,

$$Tr(I_{M_1,\ldots,M_n}(\theta)) \leq nI_0 2^k \ .$$

(ii) Under the hypotheses in Theorem 2,

$$Tr(I_{M_1,\ldots,M_n}(\theta)) \leq nCN^2 k^{\frac{2}{p}} \ .$$

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 [17]. For concreteness, suppose $\Theta = [-B, B]^d$. Denote $M = (M_1, \ldots, M_n)$, and suppose we have a prior μ_i for the parameter θ_i. The van Trees inequality for the component θ_i gives

$$\int_{-B}^{B} \mathbb{E}[(\hat{\theta}_i(M) - \theta_i)^2]\mu_i(\theta_i)d\theta_i \geq \frac{1}{\int_{-B}^{B}[I_M(\theta)]_{i,i}\mu_i(\theta_i)d\theta_i + I(\mu_i)} \tag{5}$$

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 $\mathbb{E}[S_{\theta_i}(M)] = 0$ follows trivially since the expectation over M is just a finite sum:

$$\mathbb{E}[S_{\theta_i}(M)] = \sum_m \frac{\partial}{\partial \theta_i} p(m|\theta) = \frac{\partial}{\partial \theta_i} \sum_m p(m|\theta) = 0 \ .$$

The prior μ_i can be chosen to minimize this Fisher information and achieve $I(\mu_i) = \pi^2/B^2$ [15]. Let $\mu(\theta) = \prod_i \mu_i(\theta_i)$. By summing over each component,

$$\begin{aligned}
&\int_\Theta \sum_{i=1}^{d} \mathbb{E}[(\theta_i - \hat{\theta}_i)^2]\mu(\theta)d\theta \\
&\quad \geq \sum_{i=1}^{d} \frac{1}{\int_\Theta [I_M(\theta)]_{i,i}\mu(\theta)d\theta + \frac{\pi^2}{B^2}} && (6) \\
&\quad = d \sum_{i=1}^{d} \frac{1}{d} \frac{1}{\int_\Theta [I_M(\theta)]_{i,i}\mu(\theta)d\theta + \frac{\pi^2}{B^2}} \\
&\quad \geq d \ \frac{1}{\sum_{i=1}^{d} \frac{1}{d} \int_\Theta [I_M(\theta)]_{i,i}\mu(\theta)d\theta + \frac{\pi^2}{B^2}} && (7) \\
&\quad = \frac{d^2}{\int_\Theta Tr(I_M(\theta))\mu(\theta)d\theta + \frac{d\pi^2}{B^2}} \ .
\end{aligned}$$

Therefore,

$$\sup_{\theta \in \Theta} \mathbb{E}[\|\hat{\theta}(M) - \theta\|^2] \geq \frac{d^2}{\sup_{\theta \in \Theta} Tr(I_M(\theta)) + \frac{d\pi^2}{B^2}} \ . \tag{8}$$

The inequaltiy (7) follows from Jensen's inequality via the convexity of $x \mapsto 1/x$ for $x > 0$, and the inequality (6) follows both from this convexity and (5). Using the bounds we developed in Section II-B, the relation (8) gives a lower bound on the minimax risk for the distributed estimation of θ under common statistical models. We summarize these results in the following corollaries:

Corollary 1 (Gaussian location model): Let $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$ with $\Theta = [-B, B]^d$. Under all three quantization/communication models described above, for $nB^2 \min\{k, d\} \geq d\sigma^2$, we have

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

for any quantization strategy and 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 (8). 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 similar assumptions in the subsequent corollaries.

Corollary 2 (variance of a Gaussian): Let $X \sim \mathcal{N}(0, \mathrm{diag}(\theta_1, \ldots, \theta_d))$ with $\Theta = [\sigma^2_{\min}, \sigma^2_{\max}]^d$. Under all three quantization/communication models described above, for $n\left(\frac{\sigma^2_{\max} - \sigma^2_{\min}}{2}\right)^2 \min\{k^2, d\} \geq d\sigma^4_{\min}$, we have

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

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

Corollary 3 (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. Under all three quantization/communication models described above, for $n \min\{2^k, d\} \geq d^2$, we have

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

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

The lower bounds from Corollaries 1 and 3 match those from [1]. Corollary 3 matches the upper bound from the achievable scheme in [2], while Corollary 1 matches the upper bound from [14] if we are allowed to use the interactive or blackboard models. The bound in Corollary 2 is new, and it is an unknown whether or not it is order optimal.

IV. The Most Fisher-Informative Bit

Consider the Gaussian location model with $k = 1$ bit of information about a sample X. In this model, the distribution $f(x|\theta)$ is a Gaussian with known covariance $\sigma^2 I_d$ and mean θ. The score function is

$$S_\theta(X) = \frac{1}{\sigma^2}(\theta - X)$$

which is a Gaussian with covariance $\frac{1}{\sigma^2} I_d$ and mean zero. For general k, the problem of finding exactly which quantization function b maximizes $Tr(I_M(\theta))$ for some θ is a difficult optimization problem. However, when $k = 1$ we can show that $Tr(I_M(\theta))$ is maximized when the two quantization bins $b^{-1}(1)$ and $b^{-1}(2)$ are complementary half-spaces. We show this by explicitly determining the maximal value of

$\|\mathbb{E}[S_\theta(X)|m]\|$ from the right-hand side of (1) for fixed $t = p(m|\theta)$. For convenience, consider only deterministic quantization maps $b_m(x) \in \{0,1\}$. By the convexity of $\|\mathbb{E}[S_\theta(X)b_m(X)]\|$ in b_m, this suffices for the stochastic quantization case as well.

Let ϕ be the standard normal distribution. We have the following theorem (whose proof is omitted due to space constraints):

Theorem 3: Let $Y \sim \mathcal{N}(0, I_d)$. Over all subsets A with $\mathbb{E}[1_A(Y)] = t$, the magnitude of the vector $\mathbb{E}[Y|A]$ is maximized when A is a half-space. In this case,

$$\|\mathbb{E}[Y|A]\| = \frac{\phi(Q^{-1}(t))}{t}$$

where Q is the Q-function defined by

$$Q(\tau) = \int_\tau^\infty \phi(x)dx \; .$$

Note that Theorem 3 only solves the problem of finding one bin $b^{-1}(m)$ that is optimal, but it does not solve the problem of jointly optimizing all of the bins that must partition the space of possible X-values. However, when $k = 1$ there are only two quantization bins $b^{-1}(1)$ and $b^{-1}(2)$. When one bin is a half-space with probability t, the other bin must be its complementary half-space with probability $1 - t$. Furthermore, a quantization scheme that partitions the X-values into two half-spaces also partitions the $S_\theta(X)$-values into two half-spaces. Such a scheme will jointly maximize both terms in the sum

$$\begin{aligned} Tr(I_M(\theta)) &= \sum_{i=1}^d [I_M(\theta)]_{i,i} \\ &= t\|\mathbb{E}[S_\theta(X)|M=1]\|^2 \\ &\quad + (1-t)\|\mathbb{E}[S_\theta(X)|M=2]\|^2 \end{aligned}$$

so by Theorem 3,

$$\begin{aligned} Tr(I_M(\theta)) &= \frac{1}{t\sigma^2} g_1(Q^{-1}(t))^2 \\ &\quad + \frac{1}{(1-t)\sigma^2} g_1(Q^{-1}(1-t))^2 \; . \end{aligned}$$

This one-parameter expression has its maximum at $t = 1/2$ where

$$Tr(I_M(\theta)) = \frac{2}{\pi\sigma^2} \; .$$

Therefore the one-bit quantization scheme that maximizes the trace of the Fisher information is given by two half-spaces whose defining hyperplane in $S_\theta(X)$-space intersects the origin, and this corresponds to two half-spaces in X-space whose defining hyperplane intersects the true value of θ.

APPENDIX

A. Proofs of Theorems 1 and 2

Consider some m and fix its likelihood $t = p(m|\theta)$. We will proceed by upper-bounding $\|\mathbb{E}[S_\theta(X)|m]\|$ from the right-hand side of (1). Note that

$$\mathbb{E}[S_\theta(X)|m] = \frac{\mathbb{E}[S_\theta(X)b_m(X)]}{t}$$

where $\mathbb{E}[b_m(X)] = t$ and $0 \le b_m(x) \le 1$ for all $x \in \mathcal{X}$. We use $\langle \cdot, \cdot \rangle$ to denote the usual inner product.

For some fixed m and $t = p(m|\theta)$, let U be a d-by-d orthogonal matrix with columns $u_1, u_2, \ldots, u_d$ and whose first column is given by

$$u_1 = \frac{1}{\|\mathbb{E}[S_\theta(X)|m]\|}\mathbb{E}[S_\theta(X)|m] \; .$$

We have

$$\begin{aligned} t\mathbb{E}[S_\theta(X)|m] &= \int S_\theta(x) b_m(x) f(x|\theta) d\nu(x) \\ &= \int \left(\sum_{i=1}^d u_i \langle u_i, S_\theta(x) \rangle \right) b_m(x) f(x|\theta) d\nu(x) \\ &= \sum_{i=1}^d \left(\int \langle u_i, S_\theta(x) \rangle b_m(x) f(x|\theta) d\nu(x) \right) u_i \end{aligned}$$

and since $u_2, \ldots, u_d$ are all orthogonal to $\mathbb{E}[S_\theta(X)|m]$,

$$\mathbb{E}[S_\theta(X)|m] = \frac{1}{t} \left(\int \langle u_1, S_\theta(x) \rangle b_m(x) f(x|\theta) d\nu(x) \right) u_1 \; .$$

Therefore,

$$\|\mathbb{E}[S_\theta(X)|m]\| = \frac{1}{t}\mathbb{E}[\langle u_1, S_\theta(X) \rangle b_m(X)] \; . \tag{9}$$

1) Proof of Theorem 1: To finish the proof of Theorem 1, note that the upper bound $Tr(I_M(\theta)) \le Tr(I_X(\theta))$ follows easily from the data processing inequality for Fisher information [10]. Using (9) and the Cauchy-Schwarz inequality,

$$\begin{aligned} t\|\mathbb{E}[S_\theta(X)|m]\|^2 &= \frac{1}{t} \left(\mathbb{E}[\langle u_1, S_\theta(X) \rangle b_m(X)] \right)^2 \\ &\le \frac{1}{t} \mathbb{E}[\langle u_1, S_\theta(X) \rangle^2] \mathbb{E}[b_m(X)^2] \\ &\le \frac{1}{t} \mathbb{E}[\langle u_1, S_\theta(X) \rangle^2] \mathbb{E}[b_m(X)] \\ &= \mathbb{E}[\langle u_1, S_\theta(X) \rangle^2] \; . \end{aligned}$$

So if $\text{var}\langle u_1, S_\theta(X) \rangle \le I_0$, then because score functions have zero mean,

$$t\|\mathbb{E}[S_\theta(X)|m]\|^2 \le I_0 \; .$$

Therefore by Lemma 2,

$$Tr(I_M(\theta)) \le 2^k I_0 \; .$$

2) Proof of Theorem 2: Turning to Theorem 2, we now assume that for some $p \ge 1$ and any unit vector $u \in \mathbb{R}^d$, the random vector $\langle u, S_\theta(X) \rangle$ has finite Ψ_p norm less than or equal to N. For $p = 1$ or $p = 2$, this is the common assumption that $S_\theta(X)$ is sub-exponential or sub-Gaussian, respectively, as a vector.

In particular $\langle u_1, S_\theta(X) \rangle$ has $\|\langle u_1, S_\theta(X) \rangle\|_{\Psi_p} \le N$, and

$$
\begin{aligned}
2 &\geq \mathbb{E}[\exp((|\langle u_1, S_\theta(X)\rangle|/N)^p)] \\
&\geq \mathbb{E}[\exp((\langle u_1, S_\theta(X)\rangle/N)^p)] \\
&\geq \mathbb{E}[b_m(X)\exp((\langle u_1, S_\theta(X)\rangle/N)^p)] \\
&\geq t\mathbb{E}[\exp((\langle u_1, S_\theta(X)\rangle/N)^p)|m] \\
&\geq t\exp\left(\left(\frac{1}{N}\mathbb{E}[\langle u_1, S_\theta(X)\rangle|m]\right)^p\right)
\end{aligned}
$$

so that

$$
\mathbb{E}[\langle u_1, S_\theta(X)\rangle|m] \leq N\left(\log\left(\frac{2}{t}\right)\right)^{\frac{1}{p}} .
$$

Therefore by (9),

$$
\|\mathbb{E}[S_\theta(X)|m]\| \leq N\left(\log\left(\frac{2}{t}\right)\right)^{\frac{1}{p}} . \tag{10}
$$

By Lemma 2

$$
Tr(I_M(\theta)) = \sum_m p(m|\theta)\|\mathbb{E}[S_\theta(X)|m]\|^2 ,
$$

and therefore by (10)

$$
Tr(I_M(\theta)) \leq \sum_m N^2 p(m|\theta)\left(\log\left(\frac{2}{p(m|\theta)}\right)\right)^{\frac{2}{p}} .
$$

To bound this expression we will use the following properties:

(i) $x \mapsto x\left(\log\frac{2}{x}\right)^{\frac{2}{p}}$ is concave for $0 \leq x \leq 2e^{\frac{p-2}{p}}$

(ii) $x\left(\log\frac{2}{x}\right)^{\frac{2}{p}} \leq 2e^{-\frac{2}{p}}\left(\frac{2}{p}\right)^{\frac{2}{p}}$ for $0 \leq x \leq 1$

(iii) $\sum_m p(m|\theta) = 1$

There can be at most one m-value such that $p(m|\theta) \geq 2e^{\frac{p-2}{p}}$, so we'll call such a term m_0, separate it out, and treat it separately.

$$
\begin{aligned}
&\sum_m p(m|\theta)\left(\log\left(\frac{2}{p(m|\theta)}\right)\right)^{\frac{2}{p}} \\
&\quad\leq 2e^{-\frac{2}{p}}\left(\frac{2}{p}\right)^{\frac{2}{p}} + \sum_{m\neq m_0} p(m|\theta)\left(\log\left(\frac{2}{p(m|\theta)}\right)\right)^{\frac{2}{p}} \qquad (11) \\
&\quad\leq 2e^{-\frac{2}{p}}\left(\frac{2}{p}\right)^{\frac{2}{p}} + \\
&\qquad (2^k-1)\sum_{m\neq m_0}\frac{1}{2^k-1}p(m|\theta)\left(\log\left(\frac{2}{p(m|\theta)}\right)\right)^{\frac{2}{p}} \\
&\quad\leq 2e^{-\frac{2}{p}}\left(\frac{2}{p}\right)^{\frac{2}{p}} + \left(\log\left(2(2^k-1)\right)\right)^{\frac{2}{p}} \qquad (12) \\
&\quad\leq \frac{8}{e^2} + (k+1)^{\frac{2}{p}} \qquad (13)
\end{aligned}
$$

In (11) we separate out the possible m_0 such that $p(m_0|\theta) \geq 2e^{\frac{p-2}{p}}$ and then use property (ii) to bound that specific term. Then (12) follows from Jensen's inequality and the concavity described in property (i). Setting C large enough so that (13) is less than or equal to $Ck^{\frac{2}{p}}$ completes the proof.

II. Acknowledgement

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

References

[1] 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.

[2] 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)*.

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

[4] 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.

[5] Parvathinathan Venkitasubramaniam, Lang Tong, and Ananthram Swami, "Minimax Quantization for Distributed Maximum Likelihood Estimation," *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[6] Parvathinathan Venkitasubramaniam, Gökhan Mergen, Lang Tong, and Ananthram Swami, "Quantization for Distributed Estimation in Large Scale Sensor Networks," *International Conference on Intelligent Sensing and Information Processing,* pp 121-127, 2005.

[7] Alejandro Ribeiro and Georgios B. Giannakis, "Non-Parametric Distributed Quantization-Estimation Using Wireless Sensor Networks," ' *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

[8] Wai-Man Lam and Amy R Reibman, "Design of Quantizers for Decentralized Estimation Systems," *IEEE Transactions on Communications*, 41(11):1602-1605, 1993.

[9] Guy Kindler, Ryan O'Donnel, and David Witmer, "Remarks on the Most Informative Function Conjecture at fixed mean," *arXiv preprint*, *arXiv*:1506.03167, 2015.

[10] Ram Zamir, "A Proof of the Fisher Information Inequality via a Data Processing Argument,' *IEEE Transactions on Information Theory*, 44(3):1246-1250, 1998.

[11] Christer Borell, "Geometric bounds on the Ornstein–Uhlenbeck velocity process," Probability Theory and Related Fields, 70(1):1–13, 1985.

[12] H.V.Poor, "An Introduction to Signal Detection and Estimation," Springer-Verlag, New York, 1994.

[13] 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.

[14] 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.

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

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

[17] 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.

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