Journal of Machine Learning Research 21 (2020) 1-30 Submitted 9/19; Revised 6/20; Published 10/20

Lower Bounds for Learning Distributions under Communication Constraints via Fisher Information

Leighton Pate Barnes LPB@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Yanjun Han YJHAN@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Ayfer Özgür AOZGUR@STANFORD.EDU
Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA

Editor: XuanLong Nguyen

Abstract

We consider the problem of learning high-dimensional, nonparametric 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 three different models for communication. Under the independent model, each node communicates its sample to a central processor by independently encoding it into k bits. Under the more general sequential or blackboard communication models, nodes can share information interactively 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. We develop minimax lower bounds that apply in a unified way to many common statistical models and reveal that the impact of the communication constraint can be qualitatively different depending on the tail behavior of the score function associated with each model. A key ingredient in our proofs is a geometric characterization of Fisher information from quantized samples.

Keywords: Fisher information, statistical estimation, communication constraints, learning distributions

1. Introduction

Estimating a distribution from samples is a fundamental unsupervised learning problem that has been studied in statistics since the late nineteenth century (Pearson, 1895). Consider the following distribution estimation model

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

where we would like to estimate the unknown distribution P under squared 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 and has to be communicated to a central processor by using k bits. We consider three different types of communication protocols:

1. Independent communication protocols Π_{Ind}: each node sends a k-bit string M_i simultaneously (independent of the other nodes) to the central processor and the final transcript is $Y = (M_1, \ldots, M_n)$;

2. Sequential communication protocols Π_{Seq}: the nodes sequentially send k-bit strings M_i, where quantization of the sample X_i can depend on the previous messages $M_1, \ldots, M_{i-1}$;

3. Blackboard communication protocols Π_{BB} (Kushilevitz and Nisan, 1997): 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 and depending on the written bit another node can take the turn to write a message on the blackboard.

Upon receiving the transcript Y, the central processor produces an estimate $\hat{P}$ of the distribution P based on the transcript Y and known procotol Π which can be of type Π_{Ind}, Π_{Seq}, or Π_{BB}. Our goal is to jointly design the protocol Π and the estimator $\hat{P}(Y)$ so as to minimize the worst case squared ℓ^2 risk, i.e., to characterize

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

where $\mathcal{P}$ denotes the class of distributions which P belongs to. We study three 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. Non-parametric densities: in this case $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} f$, where f is some density that possesses some standard Hölder continuity property (Nemirovski, 2000).

3. Parametric 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

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

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

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

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. For example, there is a recent line of works (Zhang et al., 2013; Braverman et al., 2016; Garg et al., 2014) which focus on the distributed parameter estimation problem where the underlying distribution has a Gaussian structure, i.e., $P_\theta = \mathcal{N}(\theta, I_d)$ with $\theta \in \Theta \subseteq \mathbb{R}^d$, often called the Gaussian location model. The high-dimensional discrete distribution estimation problem is studied in Diakonikolas et al. (2017); Han et al. (2018a), where extensions to distributed property testing are studied in Acharya et al. (2019b,a). These works show that the dependence of the estimation performance on k can be qualitatively different: the estimation error decays linearly in k for the Gaussian location model, while for distribution estimation/testing it typically decays exponentially in k. This difference was first studied in Han et al. (2018b) which develops geometric lower bounds for distributed estimation, where the Gaussian mean estimation problem and distribution estimation problem admit different geometric structures. However, the arguments heavily rely on hypothesis testing and the specific geometric objects remain implicit.

Another closely-related thread is the locally private estimation problem, which shares many similarities with communication-constrained problems (Duchi and Rogers, 2019). We refer to Duchi et al. (2013) for a general treatment of estimation problems under locally differentially private (LDP) constraints, while optimal schemes (and lower bounds) for estimating discrete distributions are proposed in Kairouz et al. (2016); Wang et al. (2016); Ye and Barg (2018); Acharya et al. (2019c). Similar to the previous discussions, strong or distributed data-processing inequalities (Duchi et al., 2013; Xu and Raginsky, 2017) are typically used in scenarios where the linear/quadratic dependence on the pricacy parameter ε is tight, and explicit modeling becomes necessary in scenarios where the tight dependence on ε is exponential.

In this paper, we approach all distributed estimation problems under communication constraints in a unified way. Specifically, we propose an explicit geometric object related to the Fisher information, and develop a framework that characterizes the Fisher information for estimating an underlying unknown parameter from a quantized sample. Equivalently, we ask the question: how can we best represent $X \sim P_\theta$ with k bits so as to maximize the Fisher information it provides about the underlying parameter θ? This framework was first introduced in Barnes et al. (2018, 2019), and there has been some previous work in analyzing Fisher information from a quantized scalar random variable such as Venkitasubramaniam et al. (2006, 2005); Ribeiro and Giannakis (2005); Lam and Reibman (1993). Different from these works, here we consider the arbitrary quantization of a random vector and are able to study the impact of the quantization rate k along with the dimension d of the underlying statistical model on the Fisher information. As an application of our framework, we use upper bounds on Fisher information to derive lower bounds on the minimax risk of the distributed estimation problems discussed above and recover many of the existing results in the literature (Zhang et al., 2013; Braverman et al., 2016; Garg et al., 2014; Han et al.,

2018a), which are known to be rate-optimal. Our technique is significantly simpler and more transparent than those in Zhang et al. (2013); Braverman et al. (2016); Garg et al. (2014); Han et al. (2018a). In particular, the strong/distributed data processing inequalities used in Zhang et al. (2013); Braverman et al. (2016); Garg et al. (2014) are typically technical and seem to be only applicable to models where the fundamental dependence of the minimax risk on the quantization rate k is linear, e.g., the Gaussian location model. Moreover, our approach points out that the Fisher information is the same as the explicit geometric object from Han et al. (2018b), and we recover most of the results from that work via this simpler approach. We also extend the results of Han et al. (2018b) to derive minimax lower bounds for statistical models with sub-exponential score functions, which is useful, for example, when we are interested in estimating the variance of a Gaussian distribution.

1.1 Organization of the Paper

In the next section, we introduce the problem of characterizing Fisher information from a quantized sample. We present a geometric characterization for this problem and derive two upper bounds on Fisher information as a function of the quantization rate. We also evaluate these upper bounds for common statistical models. In Section 3, we formulate the problem of distributed learning of distributions under communication constraints with independent, sequential and blackboard communication protocols. We use the upper bounds on Fisher information from Section 2 to derive lower bounds on the minimax risk of distributed estimation of discrete and parametric distributions. There we also provide a more detailed comparison of our results with those in the literature. Finally, in Section 4 we discuss extending these results to non-parametric density estimation.

2. Fisher information from a quantized sample

Let $\{P_\theta\}_{\theta\in\Theta}$ be a family of probability measures on the measurable space $(\mathcal{X},\mathcal{A})$ parameterized by $\theta\in\Theta\subseteq\mathbb{R}^d$. Suppose that $\{P_\theta\}_{\theta\in\Theta}$ is a dominated family and that each P_θ has density $f(x|\theta)$ with respect to some dominating measure 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

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

We assume that $p(m|x)$ is a regular conditional probability. Under any given P_θ and quantization strategy, denote by $p(m|\theta)$ the likelihood that the quantized sample M takes a specific value m. Let

$$\begin{aligned}S_\theta(m)&=(S_{\theta_1}(m),\dots,S_{\theta_d}(m))\\&=\left(\frac{\partial}{\partial\theta_1}\log p(m|\theta),\dots,\frac{\partial}{\partial\theta_d}\log p(m|\theta)\right)\in\mathbb{R}^d\end{aligned}$$

be the vector-valued score function of M under P_θ. With a slight abuse of notation, we also denote the score function of X under P_θ as

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

Consequently, the Fisher information matrices of estimating θ from M and from X are defined as

$$\begin{aligned} I_M(\theta) &= \mathbb{E}[S_\theta(M) S_\theta(M)^T], \\ I_X(\theta) &= \mathbb{E}[S_\theta(X) S_\theta(X)^T], \end{aligned}$$

respectively.

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

(1) The function $\theta \mapsto \sqrt{f(x|\theta)}$ is continuously differentiable coordinate-wise at ν-almost every $x \in \mathcal{X}$;

(2) For all θ, the Fisher information matrix $I_X(\theta)$ exists and is continuous coordinate-wise in θ.

These two conditions justify interchanging differentiation and integration as in

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

and they also ensure that $p(m|\theta)$ is continuously differentiable with respect to θ coordinate-wise (Borovkov, 1998, Section 26, Lemma 1).

The following two lemmas establish a geometric interpretation of the trace $\mathsf{Tr}(I_M(\theta))$, and are slight variants of Theorems 1 and 2 from ichi Amari (2011).

Lemma 1 *For $i \in [d]$, the (i,i)-th entry of the Fisher information matrix $I_M(\theta)$ is*

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

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

Proof For any $m \in [2^k]$, we have

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

Squaring both sides and taking expectation with respect to M completes the proof. ■

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

$$\mathsf{Tr}(I_M(\theta)) = \sum_{m \in [2^k]} p(m|\theta) \|\mathbb{E}[S_\theta(X)|m]\|_2^2 \, . \tag{1}$$

Proof By Lemma 1,

$$\begin{aligned}
\sum_{i=1}^{d} [I_M(\theta)]_{i,i} &= \sum_{i=1}^{d} \mathbb{E}\left[\mathbb{E}\left[S_{\theta_i}(X)|M\right]^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} \mathbb{E}\left[S_{\theta_i}(X)|M\right]^2\right] \\
&= \mathbb{E}\left[\|\mathbb{E}[S_\theta(X)|M]\|_2^2\right] \\
&= \sum_{m} p(m|\theta) \|\mathbb{E}[S_\theta(X)|m]\|_2^2 \, .
\end{aligned}$$

■

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 between $\mathcal{X}$ and $\mathbb{R}^d$. 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 $\mathsf{Tr}(I_M(\theta))$ is equal to the average squared magnitude of these centroid vectors.

2.1 Upper Bounds on $\mathsf{Tr}(I_M(\theta))$

In this section, we give two upper bounds on $\mathsf{Tr}(I_M(\theta))$ depending on the different tail behaviors of $S_\theta(X)$, with proofs deferred to Appendix A. The first theorem upper bounds $\mathsf{Tr}(I_M(\theta))$ 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 $u \in \mathbb{R}^d$,*

$$\mathsf{Var}(\langle u, S_\theta(X) \rangle) \leq I_0 \, ,$$

then

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

The first upper bound $\mathsf{Tr}(I_M(\theta)) \leq \mathsf{Tr}(I_X(\theta))$ follows easily from the data processing inequality for Fisher information (Zamir, 1998). The second upper bound in Theorem 1 shows that when I_0 is finite, the trace $\mathsf{Tr}(I_M(\theta))$ can increase at most exponentially in k.

Our second theorem upper bounds $\mathsf{Tr}(I_M(\theta))$ in terms of the Ψ_p Orlicz norm of $S_\theta(X)$ when projected onto any unit vector. 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

$$\Psi_p(x) = \exp(x^p) - 1 \ .$$

Note that a random variable with finite Ψ_1 Orlicz norm is sub-exponential, while a random variable with finite Ψ_2 Orlicz norm is sub-Gaussian (Vershynin, 2010).

Theorem 2 *If for any $\theta \in \Theta$ and any unit vector $u \in \mathbb{R}^d$,*

$$\|\langle u, S_\theta(X)\rangle\|_{\Psi_p}^2 \leq I_0$$

holds for some $p \geq 1$, then

$$\mathsf{Tr}(I_M(\theta)) \leq \min\{\mathsf{Tr}(I_X(\theta)),\ Ck^{\frac{2}{p}} I_0\},$$

where $C = 4$.

Theorem 2 shows that when the score function $S_\theta(X)$ has a lighter tail, then the trace $\mathsf{Tr}(I_M(\theta))$ can increase at most polynomially in k at the rate $O(k^{\frac{2}{p}})$.

2.2 Applications to Common Statistical Models

We next apply the above two results to common statistical models. We will see that depending on the statistical model, either bound may be tighter. The proofs of Corollaries 1 through 4 appear in Appendix B. In the next section, we show that Corollaries 1, 3, 4 yield tight results for the minimax risk of the corresponding distributed estimation problems.

For the Gaussian location model, Corollary 1 follows by showing that the score function associated with this model has finite Ψ_2 Orlicz norm and applying Theorem 2.

Corollary 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,*

$$\mathsf{Tr}(I_M(\theta)) \leq \min\left\{\frac{d}{\sigma^2}, C\frac{k}{\sigma^2}\right\} \tag{2}$$

where

$$C = \frac{32}{3} \ .$$

For covariance estimation in the independent Gaussian sequence model, Corollary 2 follows by showing that the score function associated with this model has finite Ψ_1 Orlicz norm and applying Theorem 2.

Corollary 2 (Gaussian covariance estimation) *Suppose* $X \sim \mathcal{N}(0, \mathrm{diag}(\theta_1, \ldots, \theta_d))$ *and* $\Theta \subseteq [\sigma_{\min}^2, \sigma_{\max}^2]^d$ *with* $\sigma_{\max} > \sigma_{\min} > 0$. *In this case,*

$$\mathsf{Tr}(I_M(\theta)) \leq \min\left\{\frac{d}{2\sigma_{\min}^4}, C\left(\frac{k}{\sigma_{\min}^2}\right)^2\right\}$$

where

$$C = \frac{16(\log 4 + 2(2+\sqrt{2}))^2}{(\log 2)^2} \ .$$

For distribution estimation, Corollary 3 is a consequence of Theorem 1 along with characterizing the variance to the score function associated with this model.

Corollary 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* $\frac{1}{4d} \leq \theta_i \leq \frac{1}{2d}$. *In this case,*

$$\mathsf{Tr}(I_M(\theta)) \leq 6\min\{d^2, d2^k\} \ . \tag{3}$$

For the product Bernoulli model, the tightness of Theorems 1 and 2 differ in different parameter regions, as shown in the following Corollary 4.

Corollary 4 (Product Bernoulli model) *Suppose that* $X \sim \prod_{i=1}^d \mathrm{Bern}(\theta_i)$. *If* $\Theta = [1/2 - \varepsilon, 1/2 + \varepsilon]^d$ *for some* $0 < \varepsilon < 1/2$, *i.e., the model is relatively dense, then*

$$\mathsf{Tr}(I_M(\theta)) \leq C\min\{d, k\}$$

for some constant C *that depends only on* ε. *If* $\Theta = [(\frac{1}{2} - \varepsilon)\frac{1}{d}, (\frac{1}{2} + \varepsilon)\frac{1}{d}]^d$, *i.e., the model is relatively sparse, then*

$$\mathsf{Tr}(I_M(\theta)) \leq \frac{2d}{\frac{1}{2} - \varepsilon}\min\{d, 2^k\} \ .$$

In the product Bernoulli model,

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

Hence, when $\Theta = [1/2 - \varepsilon, 1/2 + \varepsilon]^d$, $\mathsf{Var}(\langle u, S_\theta(X)\rangle)$ and $\|\langle u, S_\theta(X)\rangle\|_{\Psi_2}^2$ are both $\Theta(1)$. In this case, Theorem 1 gives

$$\mathsf{Tr}(I_M(\theta)) = O(2^k),$$

while Theorem 2 gives

$$\mathsf{Tr}(I_M(\theta)) = O(k) \ .$$

In this situation Theorem 2 gives the better bound. On the other hand, if $\Theta = [(\frac{1}{2} - \varepsilon)\frac{1}{d}, (\frac{1}{2} + \varepsilon)\frac{1}{d}]^d$, then $\mathsf{Var}(\langle u, S_\theta(X)\rangle) = \Theta(d)$ and $\|\langle u, S_\theta(X)\rangle\|_{\Psi_2}^2 = \Theta(d^2)$. In this case Theorem 1 gives

$$\mathsf{Tr}(I_M(\theta)) = O(d2^k),$$

while Theorem 2 gives

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

In the sparse case $\mathsf{Tr}(I_M(\theta)) \leq \mathsf{Tr}(I_X(\theta)) = \Theta(d^2)$, so only the bound from Theorem 1 is non-trivial. 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.

The upper bounds on the Fisher information matrix in the examples of this section are all sharp within multiplicative constants, with the exception of Corollary 2 whose sharpness is unknown. This in turn also implies the tightness of Theorems 1 and 2 (at least in the case $p = 2$). There are two ways to show the tightness. First, for the statistical models studied in Section 3.3, the tightness holds whenever the risk lower bounds for the associated estimation problem are matched by an communication/estimation scheme. In fact, a smaller upper bound on the Fisher information matrix would imply a risk lower bound higher than what can be achieved, which cannot happen. Second, we may also directly look at the Fisher information matrix with explicit quantization schemes. For example, in the Gaussian location model in Corollary 1, we may let $k' = \min\{k, d\}$ and use the quantization scheme $m = (m_1, \cdots, m_{k'}) \in \{0,1\}^{k'}$ with $m_i = \mathbb{1}(X_i \geq 0)$. Then for $\theta = 0$, we may explicit compute the Fisher information matrix as

$$I_M(\theta) = \mathrm{diag}\left(\frac{2}{\pi\sigma^2}, \cdots, \frac{2}{\pi\sigma^2}, 0, \cdots, 0\right),$$

and therefore $\mathsf{Tr}(I_M(\theta)) = 2\min\{k, d\}/(\pi\sigma^2)$, matching the claimed upper bound of Corollary 1.

3. Distributed Parameter Estimation

In this section, we apply the results in the previous section to the distributed estimation of parameters of an underlying statistical model under communication constraints. In this section we only focus on parametric models, while in the next section we generalize to non-parametric models by parametric reduction. The main technical exercise involves the application of Theorems 1 and 2 to statistical estimation with multiple quantized samples where the quantization of different samples can be independent or dependent as dictated by the communication protocol.

3.1 Problem Formulation

Let

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

where $\theta \in \Theta \subset \mathbb{R}^d$. We consider three different types of protocols for communicating each of these samples with k bits to a central processor that is interested in estimating the underlying parameter θ:

(1) Independent communication protocols Π_{Ind}: each sample is independently quantized to k-bits and then communicated. Formally, for $i \in [n]$, each sample X_i is encoded

to a k-bit string M_i by a possibly randomized quantization strategy, denoted by $q_i : \mathcal{X} \to [2^k]$, which can be expressed in terms of the conditional probabilities

$$p(m_i|x_i) \text{ for } m_i \in [2^k],\ x_i \in \mathcal{X}.$$

(2) Sequential communication protocols Π_{Seq}: samples are communicated sequentially by broadcasting the communication to all nodes in the system including the central processor. Therefore, the quantization of the sample X_i can depend on the previously transmitted quantized samples $M_1, \ldots, M_{i-1}$ corresponding to samples $X_1, \ldots, X_{i-1}$ respectively. Formally, each sample X_i, for $i \in [n]$, is encoded to a k-bit string M_i by a set of possibly randomized quantization strategies $\{q_{m_1,\ldots,m_{i-1}} : \mathcal{X} \to [2^k] : m_1, \ldots, m_{i-1} \in [2^k]\}$, where each strategy $q_{m_1,\ldots,m_{i-1}}(x_i)$ can be expressed in terms of the conditional probabilities

$$p(m_i|x_i; m_1, \ldots, m_{i-1}) \text{ for } m_i \in [1 : 2^k] \text{ and } x_i \in \mathcal{X}.$$

While these two models can be motivated by a distributed estimation scenario where the topology of the underlying network can dictate the type of the protocol (see Figure 1a) to be used, they can also model the quantization and storage of samples arriving sequentially at a single node. For example, consider a scenario where a continuous stream of samples is captured sequentially and each sample is stored in digital memory by using k bits/sample. In the independent model, each sample would be quantized independently of the other samples (even though the quantization strategies for different samples can be different and jointly optimized ahead of time), while under the sequential model the quantization of each sample X_i would depend on the information $M_1, \ldots, M_{i-1}$ stored in the memory of the system at time i. This is illustrated in Figure 1b.

We finally introduce a third type of communication protocol that allows nodes to communicate their samples to the central processor in a fully interactive manner while still limiting the number of bits used per sample to k bits. Under this model, each node can see the previously written bits on a public blackboard, and can use that information to determine its quantization strategy for subsequently transmitted bits. This is formally defined below.

(3) Blackboard communication protocols Π_{BB}: 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 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 (Kushilevitz and Nisan, 1997), 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) communication strategies for the nodes.

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 a one-to-one correspondence between all possible transcripts $y \in \{0,1\}^{nk}$ and paths in the tree. Note that there is also a one-to-one correspondence between $y \in \{0,1\}^{nk}$ and the k-bit messages $m_1, \ldots, m_n$ transmitted by the n nodes. 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 this direction note that the protocol specifies the node to transmit first, so given $m_1, \ldots, m_n$ one can deduce the path followed in the protocol tree).

Under all three communication protocols above, the ultimate goal is to produce an estimate $\hat{\theta}$ of the underlying parameter θ from the nk bit transcript Y or equivalently the collection of k-bit messages $M_1, \ldots, M_n$ observed by the estimator. Note that the encoding strategies/protocols used in each case can be jointly optimized and agreed upon by all parties ahead of time. Formally, we are interested in the following parameter estimation problem under squared ℓ_2^2 risk

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

where $\hat{\theta}(M_1, \ldots, M_n)$ is an estimator of θ based on the quantized observations. Note that with an independent communication protocol, the messages $M_1, \ldots, M_n$ are independent, while this is no longer true under the sequential and blackboard protocols.

3.2 Main Results for Distributed Parameter Estimation

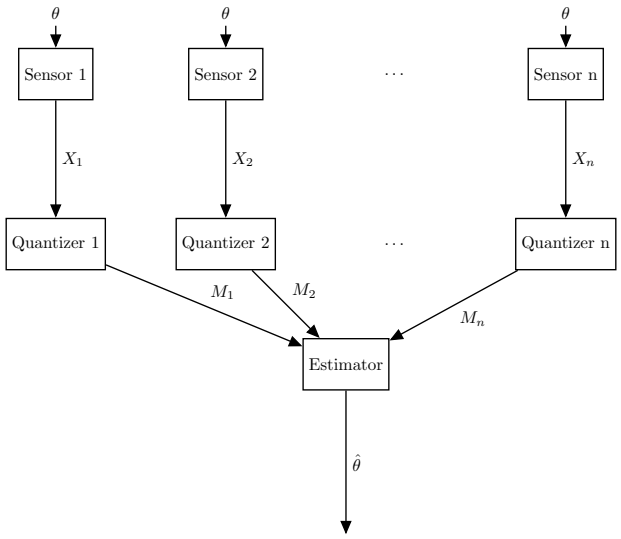
We next state our main theorem for distributed parameter estimation. We will show in the next subsection that this theorem can be applied to obtain tight lower bounds for distributed estimation under many common statistical models, including the discrete distribution estimation and the Gaussian mean estimation.

Theorem 3 *Suppose $[-B, B]^d \subset \Theta$. For any estimator $\hat{\theta}(M_1, \ldots, M_n)$ and communication protocol $\Pi \in \Pi_{\mathsf{Ind}}$, Π_{Seq}, or Π_{BB}, if $S_\theta(X)$ satisfies the hypotheses in Theorem 1 then*

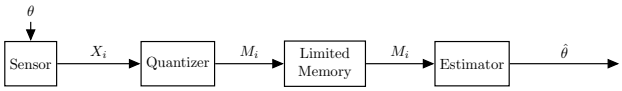
$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq \frac{d^2}{I_0 2^k n + \frac{d\pi^2}{B^2}},$$

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

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq \frac{d^2}{C I_0 k^{2/p} n + \frac{d\pi^2}{B^2}},$$



(a) Distributed communication of samples



(b) Storing a stream of samples

Figure 1: Two applications that require quantization of samples. The quantization strategy can be independent or sequential.

where $C = 4$.

We next prove Theorem 3 for $\Pi \in \Pi_{\mathsf{Ind}}$ or Π_{Seq} by a straightforward application of the Van Trees Inequality combined with the conclusions of Theorems 1 and 2. The proof for $\Pi \in \Pi_{\mathsf{BB}}$ requires more work and is deferred to the Appendix C.

Proof of Theorem 3 We are interested in the quantity

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

under each model. We have

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

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 the sequential model, 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,

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

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. 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 both quantization models:

(i) Under the hypotheses in Theorem 1,

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

(ii) Under the hypotheses in Theorem 2,

$$\mathsf{Tr}(I_{M_1,\ldots,M_n}(\theta)) \leq nCI_0 k^{\frac{2}{p}} \ .$$

Consider the squared error risk in estimating θ:

$$\mathbb{E}\|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^{d} \mathbb{E}[(\theta_i - \hat{\theta}_i)^2] \, .$$

In order to lower bound this risk, we will use the van Trees inequality (Gill and Levit, 1995). Suppose we have a prior μ_i for the parameter θ_i. For convenience denote $M = (M_1, \ldots, M_n)$. 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 of the prior μ_i. Note that the required regularity condition $\mathbb{E}[S_{\theta_i}(M)] = 0$ used in (5) holds 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 $I(\mu_i) = \pi^2/B^2$ (Borovkov, 1998). Let $\theta = (\theta_1, \cdots, \theta_d)$ and $\mu(\theta) = \prod_i \mu_i(\theta_i)$. By the van Trees inequality (5) on one dimension,

$$\begin{aligned}
\int_{[-B,B]^d} \mathbb{E}[(\hat{\theta}_i(M) - \theta_i)^2] \mu(\theta) d\theta &= \int_{[-B,B]^{d-1}} \left(\int_{-B}^{B} \mathbb{E}[(\hat{\theta}_i(M) - \theta_i)^2] \mu_i(\theta_i) d\theta_i \right) \prod_{j \neq i} [\mu_j(\theta_j) d\theta_j] \\
&\geq \int_{[-B,B]^{d-1}} \frac{1}{\int_{-B}^{B} [I_M(\theta)]_{i,i} \mu_i(\theta_i) d\theta_i + I(\mu_i)} \prod_{j \neq i} [\mu_j(\theta_j) d\theta_j] \\
&\geq \frac{1}{\int_{[-B,B]^{d-1}} \left[\int_{-B}^{B} [I_M(\theta)]_{i,i} \mu_i(\theta_i) d\theta_i + I(\mu_i) \right] \prod_{j \neq i} [\mu_j(\theta_j) d\theta_j]} \\
&= \frac{1}{\int_{[-B,B]^d} [I_M(\theta)]_{i,i} \mu(\theta) d\theta + \pi^2/B^2},
\end{aligned}$$

where the last inequality follows from the convexity of $x \mapsto 1/x$ for $x > 0$. Consequently,

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

where the second inequality is again due to the convexity of $x \mapsto 1/x$ for $x > 0$. Therefore,

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta}(M) - \theta\|^2 \geq \frac{d^2}{\sup_{\theta \in \Theta} \mathsf{Tr}(I_M(\theta)) + \frac{d\pi^2}{B^2}} \, . \tag{6}$$

We could have equivalently used the multivariate version of the van Trees inequality (Gill and Levit, 1995) 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.

Combining (6) with (i) and (ii) proves the theorem.

3.3 Applications to Common Statistical Models

Using the bounds in Section 2.2, Theorem 3 gives lower bounds on the minimax risk for the distributed estimation of θ under common statistical models. We summarize these results in the following corollaries.

Corollary 5 (Gaussian location model) *Let $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$ with $[-B, B]^d \subset \Theta$. For $nB^2 \min\{k, d\} \geq d\sigma^2$, we have*

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq C\sigma^2 \max\left\{\frac{d^2}{nk}, \frac{d}{n}\right\}$$

for any communication protocol Π of type Π_{Ind}, Π_{Seq}, or Π_{BB} and any estimator $\hat{\theta}$, where $C > 0$ is a universal constant independent of n, k, d, σ^2, B.

Note that the condition $nB^2 \min\{k, d\} \geq d\sigma^2$ in the above corollary is a weak condition that ensures that we can ignore the second term in the denominator of (6). For fixed B, σ, this condition is weaker than just assuming that n is at least order d, which is required for a consistent estimation anyways. We will make similar assumptions in the subsequent corollaries.

The corollary recovers the results in Zhang et al. (2013); Garg et al. (2014) (without logarithmic factors in the risk) and the corresponding result from Han et al. (2018b) without the condition $k \geq \log d$. An estimator using the blackboard communication protocol that achieves this result is given in Garg et al. (2014).

Corollary 6 (Gaussian covariance estimation) *Suppose that $X \sim \mathcal{N}(0, \operatorname{diag}(\theta_1, \ldots, \theta_d))$ with $[\sigma_{\min}^2, \sigma_{\max}^2]^d \subset \Theta$. Then for $n\left(\sigma_{\max}^2 - \sigma_{\min}^2\right)^2 \min\{k^2, d\} \geq d\sigma_{\min}^4$, we have*

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq C\sigma_{\min}^4 \max\left\{\frac{d^2}{nk^2}, \frac{d}{n}\right\}$$

for any communication protocol Π of type Π_{Ind}, Π_{Seq}, or Π_{BB} and any estimator $\hat{\theta}$, where $C > 0$ is a universal constant independent of $n, k, d, \sigma_{min}, \sigma_{max}$.

The bound in Corollary 6 is new, and it is unknown whether or not it is order optimal.

Corollary 7 (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 $n \min\{2^k, d\} \geq d^2$, we have

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq C \max\left\{\frac{d}{n2^k}, \frac{1}{n}\right\}$$

for any communication protocol Π *of type* Π_{Ind}, Π_{Seq}, *or* Π_{BB} *and any estimator* $\hat{\theta}$, *where* $C > 0$ *is a universal constant independent of* n, k, d.

This result recovers the corresponding result in Han et al. (2018b) and matches the upper bound from the achievable scheme developed in Han et al. (2018a) (when the performance of the scheme is evaluated under ℓ_2^2 loss rather than ℓ_1).

Corollary 8 (Product Bernoulli model) *Suppose that* $X = (X_1, \ldots, X_d) \sim \prod_{i=1}^d Bern(\theta_i)$. *If* $\Theta = [0,1]^d$, *then for* $n \min\{k, d\} \geq d$ *we have*

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq C \max\left\{\frac{d^2}{nk}, \frac{d}{n}\right\}$$

for any communication protocol Π *of type of type* Π_{Ind}, Π_{Seq}, *or* Π_{BB} *and any estimator* $\hat{\theta}$, *where* $C > 0$ *is a universal constant independent of* n, k, d.

If $\Theta = \{(\theta_1, \ldots, \theta_d) \in [0,1]^d : \sum_{i=1}^d \theta_i = 1\}$, *then for* $n \min\{2^k, d\} \geq d^2$, *we get instead*

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta} - \theta\|_2^2 \geq C \max\left\{\frac{d}{n2^k}, \frac{1}{n}\right\} .$$

The corollary recovers the corresponding result from Han et al. (2018b) and matches the upper bound from the achievable scheme presented in the same paper.

4. Distributed Estimation of Non-Parametric Densities

Finally, we turn to the case where the X_i are drawn independently from some probability distribution on $[0,1]$ with density f with respect to the Lebesgue measure. We will assume that f is Hölder continuous with smoothness parameter $s \in (0,1]$ and constant L, i.e.,

$$|f(x) - f(y)| \leq L|x - y|^s, \qquad \forall x, y \in [0,1].$$

Let $\mathcal{H}_L^s([0,1])$ be the space of all such densities. We are interested in characterizing the minimax risk

$$\inf_{(\Pi, \hat{f})} \sup_{f \in \mathcal{H}_L^s([0,1])} \mathbb{E}\|f - \hat{f}\|_2^2$$

where the estimators $\hat{f}$ are functions of the transcript Y. We have the following theorem.

Theorem 4 *For any blackboard communication protocol* $\Pi \in \Pi_{\mathsf{BB}}$ *and estimator* $\hat{f}(Y)$,

$$\sup_{f \in \mathcal{H}_L^s([0,1])} \mathbb{E}\|f - \hat{f}\|_2^2 \geq c \max\{n^{-\frac{2s}{2s+1}}, (n2^k)^{-\frac{s}{s+1}}\} .$$

Moreover, this rate is achieved by an independent protocol $\Pi^\star \in \Pi_{\mathsf{Ind}}$ *so that*

$$\inf_{\hat{f}.\Pi \in \Pi_{\mathsf{Ind}}} \sup_{f \in \mathcal{H}_L^s([0,1])} \mathbb{E}\|f - \hat{f}\|_2^2 \leq C \max\{n^{-\frac{2s}{2s+1}}, (n2^k)^{-\frac{s}{s+1}}\},$$

where c, C *are constants that depend only on* s, L.

Proof We start with the lower bound. Fix a bandwidth $h = 1/d$ for some integer d, and consider a parametric subset of the densities in $\mathcal{H}^s_L([0,1])$ that are of the form

$$f_P(x) = 1 + \sum_{i=1}^{d} \frac{p_i - h}{h} g\left(\frac{x - x_i}{h}\right)$$

where $P = (p_1, \ldots, p_d)$, $x_i = (i-1)h$, and g is a smooth bump function that vanishes outside $[0,1]$ and has $\int g(x)dx = 1$. The function f_P is in $\mathcal{H}^s_L([0,1])$ provided that $\max_{i\in[d]} |p_i - h| \leq c_0 h^{s+1}$ for some constant c_0. Let

$$\mathcal{P} = \left\{ P \,:\, \sum_i p_i = 1, p_i \geq 0, |p_i - h| \leq c_0 h^{s+1} \right\} .$$

For $P \in \mathcal{P}$ and any estimator $\hat{f}$ define $\hat{P} = (\hat{p}_1 \ldots, \hat{p}_d)$ by $\hat{p}_i = \int_{x_i}^{x_{i+1}} \hat{f}(x)dx$. By Cauchy-Schwarz, we then have

$$\begin{aligned}
\mathbb{E}\|f_P - \hat{f}\|_2^2 &= \mathbb{E}\left[\int_0^1 (f_P(x) - f(x))^2 dx\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d} \int_{x_i}^{x_{i-1}} (f_P(x) - \hat{f}(x))^2 dx\right] \\
&\geq d\, \mathbb{E}\left[\sum_{i=1}^{d} \left(\int_{x_i}^{x_{i-1}} (f_P(x) - \hat{f}(x))dx\right)^2\right] \\
&= d\, \mathbb{E}\|P - \hat{P}\|_2^2 \,. \qquad (7)
\end{aligned}$$

By the proofs of Theorem 1 and Corollary 7, the quantity in (7) can be upper bounded provided that h is not too small. In particular we can pick $h^{s+1} = (n \min\{2^k, d\})^{-\frac{1}{2}}$ so that

$$\sup_{P\in\mathcal{P}} \mathbb{E}\|f_P - \hat{f}\|_2^2 \geq c \max\{n^{-\frac{2s}{2s+1}}, (n2^k)^{-\frac{s}{s+1}}\}$$

as desired.

For the achievability side note that

$$\mathbb{E}\|f - \hat{f}\|_2^2 \leq 2\mathbb{E}\|f - f_h\|_2^2 + 2\mathbb{E}\|f_h - \hat{f}\|_2^2$$

and f_h can be chosen to be a piece-wise constant function of the form

$$f_h(x) = \sum_{i=1}^{d} \frac{p_i}{h} 1(x \in [x_i, x_{i+1})),$$

which satisfies

$$\|f - f_h\|_2^2 \leq C_0 h^{2s}$$

for a constant C_0 that depends only on L. Choosing

$$\hat{f}(x) = \sum_{i=1}^{d} \frac{\hat{p}_i}{h} 1(x \in [x_i, x_{i+1}))$$

for some $\hat{P} = (\hat{p}_1, \ldots, \hat{p}_d)$ we get

$$\mathbb{E}\|f - \hat{f}\|_2^2 \leq 2C_0 h^{2s} + 2d\mathbb{E}\|P - \hat{P}\|_2^2 \, .$$

We use the following procedure for defining the estimator $\hat{P}$ along with the communication protocol. See also Han et al. (2018b).

(i) Without loss of generality we assume $d \geq 2^k$. Divide the indices $i = 1, \ldots, d$ into $m = \frac{d}{2^k - 1}$ groups of $2^k - 1$ indices each. We assume m is an integer for simplicity.

(ii) Each sample X_j gets associated with one group of indices such that there are $\frac{n}{m}$ samples associated with each group.

(iii) Each sample X_j is examined by its node, and if it matches one of the $2^k - 1$ different outcomes associated with the node's group, then the sample is communicated exactly using a unique k-bit string. If the sample is not one of the $2^k - 1$ outcomes associated with the node's group then it is ignored and we can use the 2^kth unique k-bit string to signal this outcome.

(iv) The estimate of each symbol probability $\hat{p}_i$ is then the empirical frequency of that symbol being communicated to the centralized estimator, normalized by the effective sample size $\frac{n}{m}$ instead of the original sample size n.

This estimator achieves

$$\mathbb{E}\|P - \hat{P}\|_2^2 = \sum_{i=1}^{d} \frac{m}{n} p_i(1 - p_i) \leq \frac{m}{n} \leq C_1 \frac{d}{n \min\{2^k, d\}} \, ,$$

and therefore

$$\mathbb{E}\|f - \hat{f}\|_2^2 \leq 2C_0 h^{2s} + 2C_1 \frac{d^2}{n \min\{2^k, d\}} \, .$$

Optimizing over h by setting $h^{2(s+1)} = (n \min\{2^k, d\})^{-1}$ gives the final result. ■

Acknowledgments

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.

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 $\|\mathbb{E}[S_\theta(X)|m]\|_2$ 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 $0 \leq b_m(x) \leq 1$ for all $x \in \mathcal{X}$. We use $\langle \cdot, \cdot \rangle$ to denote the usual inner product. 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 the unit vector

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

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,

$$\|\mathbb{E}[S_\theta(X)|m]\|_2 = \frac{1}{t} \mathbb{E}[\langle u_1, S_\theta(X) \rangle b_m(X)] \; . \tag{8}$$

A.1 Proof of Theorem 1

To finish the proof of Theorem 1, note that the upper bound $\mathsf{Tr}(I_M(\theta)) \leq \mathsf{Tr}(I_X(\theta))$ follows easily from the data processing inequality for Fisher information Zamir (1998). Using (8) and the Cauchy-Schwarz inequality,

$$\begin{aligned} t\|\mathbb{E}[S_\theta(X)|m]\|_2^2 &= \frac{1}{t} \left(\mathbb{E}[\langle u_1, S_\theta(X) \rangle b_m(X)] \right)^2 \\ &\leq \frac{1}{t} \mathbb{E}[\langle u_1, S_\theta(X) \rangle^2] \mathbb{E}[b_m(X)^2] \\ &\leq \frac{1}{t} \mathbb{E}[\langle u_1, S_\theta(X) \rangle^2] \mathbb{E}[b_m(X)] \\ &= \mathbb{E}[\langle u_1, S_\theta(X) \rangle^2] \; . \end{aligned}$$

So if $\mathsf{Var}\langle u_1, S_\theta(X) \rangle \leq I_0$, then because score functions have zero mean,

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

Therefore by Lemma 2,

$$\mathsf{Tr}(I_M(\theta)) \leq 2^k I_0 \; .$$

A.2 Proof of Theorem 2

Turning to Theorem 2, we now assume that for some $p \geq 1$ and any unit vector $u \in \mathbb{R}^d$, the random vector $\langle u, S_\theta(X)\rangle$ has finite squared Ψ_p norm at most I_0. 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 $\|\langle u_1, S_\theta(X)\rangle\|_{\Psi_p}^2 \leq I_0$, and the convexity of $x \mapsto \exp(|x|^p)$ gives

$$\begin{aligned}
2 &\geq \mathbb{E}[\exp(|\langle u_1, S_\theta(X)\rangle|^p / I_0^{p/2})] \\
&\geq \mathbb{E}[b_m(X) \exp(|\langle u_1, S_\theta(X)\rangle|^p / I_0^{p/2})] \\
&= t\mathbb{E}[\exp(|\langle u_1, S_\theta(X)\rangle|^p / I_0^{p/2} | m] \\
&\geq t \exp\left(|\mathbb{E}[\langle u_1, S_\theta(X)\rangle | m]|^p / I_0^{p/2}\right)
\end{aligned}$$

so that

$$\mathbb{E}[\langle u_1, S_\theta(X)\rangle | m] \leq \sqrt{I_0} \left(\log\left(\frac{2}{t}\right)\right)^{\frac{1}{p}} .$$

Therefore by (8),

$$\|\mathbb{E}[S_\theta(X)|m]\|_2 \leq \sqrt{I_0} \left(\log\left(\frac{2}{t}\right)\right)^{\frac{1}{p}} . \tag{9}$$

By Lemma 2,

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

and therefore by (9),

$$\mathsf{Tr}(I_M(\theta)) \leq I_0 \sum_m p(m|\theta) \left(\log\left(\frac{2}{p(m|\theta)}\right)\right)^{\frac{2}{p}} .$$

To bound this expression, let ϕ be the upper concave envelope of $x \mapsto x\left(\log \frac{2}{x}\right)^{\frac{2}{p}}$ on $[0, 1]$. We have

$$\begin{aligned}
\mathsf{Tr}(I_M(\theta)) &\leq I_0 2^k \sum_m \frac{1}{2^k} \phi(p(m|\theta)) \\
&\leq I_0 2^k \phi\left(\sum_m \frac{1}{2^k} p(m|\theta)\right) \\
&= I_0 2^k \phi\left(\frac{1}{2^k}\right) .
\end{aligned} \tag{10}$$

It can be easily checked that $\phi(x) = x\left(\log \frac{2}{x}\right)^{\frac{2}{p}}$ for $0 < x \leq 1/2$, and therefore

$$\begin{aligned}
\mathsf{Tr}(I_M(\theta)) &\leq I_0 \left(\log 2^{k+1}\right)^{\frac{2}{p}} \\
&\leq I_0 (k+1)^{\frac{2}{p}} \\
&\leq 4 I_0 k^{\frac{2}{p}} .
\end{aligned}$$

Appendix B. Proof of Corollaries 1 through 4

In this appendix we provide proofs of Corollaries 1, 2, 3, and 4.

B.1 Proof of Corollary 1

The score function for the Gaussian location model is

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

so that $S_\theta(X) \sim \mathcal{N}\left(0, \frac{1}{\sigma^2} I_d\right)$. Therefore $\langle u, S_\theta(X)\rangle$ is a mean-zero Gaussian with variance $1/\sigma^2$ for any unit vector $u \in \mathbb{R}^d$. Hence, the Ψ_2 norm of the projected score function vector is

$$\|\langle u, S_\theta(X)\rangle\|_{\Psi_2} = \frac{1}{\sigma}\sqrt{\frac{8}{3}}$$

since it can be checked that

$$\int \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{\sigma^2 x^2}{2}} e^{\left(\frac{x}{c}\right)^2} dx = 2$$

when $c = \frac{1}{\sigma}\sqrt{\frac{8}{3}}$. By Theorem 2,

$$\mathsf{Tr}(I_M(\theta)) \le \min\left\{\frac{d}{\sigma^2}, C\frac{k}{\sigma^2}\right\}$$

where

$$C = \frac{32}{3} \, .$$

B.2 Proof of Corollary 2

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 a shifted, scaled verson of a chi-squared distributed random variable χ_1^2 with one degree of freedom. The Ψ_1 norm of each component $S_{\theta_i}(X)$ can be bounded by

$$\|S_{\theta_i}(X)\|_{\Psi_1} \le \frac{2}{\sigma_{\min}^2} \, . \tag{11}$$

This follows since the pdf of $Y = S_{\theta_i}(X)$ is

$$f_Y(y) = \frac{2\theta_i}{\sqrt{2\pi}} \frac{\exp\left(-(\theta_i y + \frac{1}{2})\right)}{\sqrt{2\theta_i y + 1}}$$

for $y \geq -\frac{1}{2\theta_i}$, and we have

$$\begin{aligned}\int_{-\frac{1}{2\theta_i}}^{\infty} \frac{2\theta_i}{\sqrt{2\pi}} \frac{\exp\left(-(\theta_i y + \frac{1}{2})\right)}{\sqrt{2\theta_i y + 1}} \exp\left|\frac{y}{K}\right| dy &= \sqrt{\frac{\theta_i}{\pi}} \int_0^{\infty} \frac{1}{\sqrt{y}} \exp\left(-\theta_i y + \left|\frac{y - \frac{1}{2\theta_i}}{K}\right|\right) dy \\ &\leq \sqrt{\frac{\theta_i}{\pi}} \int_0^{\infty} \frac{1}{\sqrt{y}} \exp\left(-\theta_i y + \left(\frac{y + \frac{1}{2\theta_i}}{K}\right)\right) dy \\ &= \sqrt{\frac{\theta_i}{\pi}} \exp\left(\frac{1}{2\theta_i K}\right) \int_0^{\infty} \frac{1}{\sqrt{y}} \exp\left(-\left(\theta_i - \frac{1}{K}\right) y\right) dy \, .\end{aligned}$$

By picking $K = \frac{2}{\theta_i}$ and using the identity

$$\int_0^{\infty} \frac{1}{\sqrt{y}} e^{-cy} dy = \sqrt{\pi / c}$$

for $c > 0$, we get

$$\mathbb{E}\left[\exp\left|\frac{Y}{K}\right|\right] \leq \sqrt{2} e^{\frac{1}{4}} \leq 2 \, .$$

This proves (11). We next turn our attention to bounding

$$\|\langle u, S_\theta(X)\rangle\|_{\Psi_1}$$

for any unit vector $u = (v_1, \ldots, v_d) \in \mathbb{R}^d$. To this end note that Markov's inequality implies

$$\mathbb{P}(|Y| \geq t) = \mathbb{P}\left(e^{\frac{|Y|}{K}} \geq e^{\frac{t}{K}}\right) \leq 2e^{-\frac{t}{K}}, \tag{12}$$

and we have the following bound on the moments of Y:

$$\begin{aligned}\mathbb{E}[|Y|^p] &= \int_0^{\infty} \mathbb{P}(|Y| \geq t) p t^{p-1} dt \\ &\leq \int_0^{\infty} 2e^{-\frac{t}{K}} p t^{p-1} dt \\ &\leq 2p \int_0^{\infty} e^{-\frac{t}{K}} t^{p-1} dt = 2K^p \cdot p! \, . \end{aligned} \tag{13}$$

This leads to a bound on the moment generating function for Y as follows:

$$\begin{aligned}\mathbb{E}\left[e^{\frac{u_i Y}{K}}\right] &\leq 1 + \mathbb{E}\left[\frac{u_i Y}{K}\right] + \sum_{p=2}^{\infty} \mathbb{E}\left[\frac{1}{p!}\left(\frac{u_i Y}{K}\right)^p\right] \\ &\leq 1 + \sum_{p=2}^{\infty} 2|u_i|^p = 1 + \frac{2u_i^2}{1 - |u_i|} \leq e^{\frac{2u_i^2}{1-|u_i|}} \end{aligned} \tag{14}$$

for $|u_i| < 1$. Applying (14) to the moment generating function for $\langle u, S_\theta(X)\rangle$ gives

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sum_{i=1}^{d} \frac{u_i S_{\theta_i}(X)}{K}\right)\right] &= \exp\left(\frac{u_{i_0} S_{\theta_{i_0}}(X)}{K}\right) \prod_{i \neq i_0} \exp\left(\frac{u_i S_{\theta_i}(X)}{K}\right) \\ &\leq 2e^{2(2+\sqrt{2})}, \end{aligned} \tag{15}$$

where i_0 is the only index such that $|u_{i_0}| > \frac{1}{\sqrt{2}}$ (if one exists). Thus

$$\begin{aligned}
&\mathbb{E}\left[\exp\left|\sum_{i=1}^{d}\frac{u_i S_{\theta_i}(X)}{K}\right|\right] \\
&\leq \mathbb{E}\left[\exp\left(\sum_{i=1}^{d}\frac{u_i S_{\theta_i}(X)}{K}\right)\right] + \mathbb{E}\left[\exp\left(\sum_{i=1}^{d}\frac{-u_i S_{\theta_i}(X)}{K}\right)\right] \\
&\leq 4e^{2(2+\sqrt{2})}, \qquad (16)
\end{aligned}$$

and by the concavity of $x \mapsto x^\alpha$ for $x \geq 0$ and $\alpha \in [0,1]$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left|\sum_{i=1}^{d}\frac{u_i S_{\theta_i}(X)}{K/\alpha}\right|\right] &\leq \left(\mathbb{E}\left[\exp\left|\sum_{i=1}^{d}\frac{u_i S_{\theta_i}(X)}{K}\right|\right]\right)^\alpha \\
&\leq (4e^{2(2+\sqrt{2})})^\alpha = 2, \qquad (17)
\end{aligned}$$

where $\alpha = \frac{\log 2}{\log 4 + 2(2+\sqrt{2})}$. Hence, we see that

$$\|\langle u, S_\theta(X)\rangle\|_{\Psi_1} \leq \frac{2}{\sigma_{\min}^2 \alpha},$$

and using Theorem 2,

$$\mathsf{Tr}(I_M(\theta)) \leq \frac{16(\log 4 + 2(2+\sqrt{2}))^2}{(\log 2)^2}\left(\frac{k}{\sigma_{\min}^2}\right)^2 .$$

For the other bound note that $\mathsf{Var}(S_{\theta_i}(X)) = \frac{1}{2\theta_i^2}$ so that

$$\mathsf{Tr}(I_X(\theta)) \leq \frac{d}{2\sigma_{\min}^4} .$$

B.3 Proof of Corollary 3

Note that

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

and

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

for $i = 1, \ldots, d$. Recall that by assumption we are restricting our attention to $\frac{1}{4d} \leq \theta_i \leq \frac{1}{2d}$ for $i = 1, \ldots, d$. Then for any unit vector $u = (v_1, \ldots, v_d)$,

$$\begin{aligned}
\mathsf{Var}(\langle u, S_\theta(X)\rangle) &= \sum_{x=1}^{d+1} \theta_x \left(\sum_{i=1}^{d} u_i S_{\theta_i}(x) \right)^2 \\
&= \theta_{d+1} \frac{1}{\theta_{d+1}^2} \left(\sum_{i=1}^{d} u_i \right)^2 + \sum_{x=1}^{d} \theta_x \left(\sum_{i=1}^{d} u_i S_{\theta_i}(x) \right)^2 \\
&\leq 2d + \sum_{x=1}^{d} \theta_x u_x^2 \frac{1}{\theta_x^2} \leq 6d \, .
\end{aligned}$$

Finally by Theorem 1,

$$\mathsf{Tr}(I_M(\theta)) \leq 6 \min\{d^2, d2^k\} \, .$$

B.4 Proof of Corollary 4

With the product Bernoulli model

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

so that in the case $\Theta = [1/2 - \varepsilon, 1/2 + \varepsilon]^d$,

$$\mathbb{E}\left[\exp\left[\left(\frac{S_{\theta_i}(X)}{K} \right)^2 \right] \right] \leq \exp\left(\frac{1}{(\frac{1}{2} - \varepsilon)^2 K^2} \right),$$

and

$$\|S_{\theta_i}(X)\|_{\Psi_2} \leq \frac{1}{(\frac{1}{2} - \varepsilon)\sqrt{\log 2}} \, .$$

By the rotation invariance of the Ψ_2 norm Vershynin (2010), this implies

$$\|\langle u, S_{\theta_i}(X)\rangle\|_{\Psi_2} \leq \frac{C}{(\frac{1}{2} - \varepsilon)\sqrt{\log 2}}$$

for some absolute constant C. Thus by Theorem 2,

$$\mathsf{Tr}(I_M(\theta)) \leq 4 \left(\frac{C}{(\frac{1}{2} - \varepsilon)\sqrt{\log 2}} \right)^2 k \, .$$

For the other bound,

$$\mathsf{Tr}(I_X(\theta)) \leq d \left(\frac{1}{\frac{1}{2} - \varepsilon} \right)^2 \, .$$

Now consider the sparse case $\Theta = [(\frac{1}{2} - \varepsilon)\frac{1}{d}, (\frac{1}{2} + \varepsilon)\frac{1}{d}]^d$. We have

$$\begin{aligned}\mathsf{Var}\left(S_{\theta_i}(X)\right) &= \frac{1}{\theta_i^2}\theta_i + \left(\frac{-1}{1-\theta_i}\right)^2 (1-\theta_i) \\ &= \frac{1}{\theta_i} + \frac{1}{1-\theta_i} \\ &\leq \frac{2d}{\frac{1}{2} - \varepsilon}\,,\end{aligned}$$

and therefore by independence,

$$\mathsf{Var}\left(\langle u, S_\theta(X)\rangle\right) \leq \frac{2d}{\frac{1}{2} - \varepsilon}\,.$$

Thus by Theorem 1

$$\mathsf{Tr}(I_M(\theta)) \leq \frac{2d}{\frac{1}{2} - \varepsilon} \min\{d, 2^k\}\,.$$

Appendix C. Proof of Theorem 3 with Blackboard Model

In order to lower bound the minimax risk under the blackboard model, we will proceed by writing down the Fisher information from the transcript Y that is described in Section 3. Let $b_{v,y}(x_{l_v}) = b_v(x_{l_v})$ if the path $\tau(y)$ 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 $p_{i,y}(x_i) = \prod_{v \in \tau(y)\,:\,l_v = i} b_{v,y}(x_i)$. The score for component θ_i is therefore

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

To get the above display, integration and differentiation have to be interchanged just like in Lemma 1. The Fisher information from Y for estimating the component θ_i is then

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

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

$$\begin{aligned}\sum_y \mathbb{P}(Y=y)\frac{\mathbb{E}[S_{\theta_i}(X_j)p_{j,y}(X_j)]\mathbb{E}[S_{\theta_i}(X_k)p_{k,y}(X_k)]}{\mathbb{E}[p_{j,y}(X_j)]\mathbb{E}[p_{k,y}(X_k)]} \\ = \mathbb{E}\left[S_{\theta_i}(X_j)S_{\theta_i}(X_k)\sum_y\prod_{l=1}^n p_{l,y}(X_l)\right] \\ = \mathbb{E}\left[S_{\theta_i}(X_j)S_{\theta_i}(X_k)\right] = 0\,. \end{aligned} \tag{18}$$

The step in (18) 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 $\sum_y\prod_{l=1}^n p_{l,y}(x_l)=1$.

A related nontrivial identity, that

$$\sum_y\prod_{i\neq j}\mathbb{E}[p_{i,y}(X)] = 2^k \tag{19}$$

for each $j\in[n]$, will be required later in the proof. For a tree-based proof of this fact see Han et al. (2018b).

Returning to the Fisher information from Y we have that

$$\sum_{i=1}^d \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\mathbb{P}(Y=y)\right)^2\right] = \sum_y \mathbb{P}(Y=y)\sum_j\left(\frac{\mathbb{E}[S_{\theta_i}(X_j)p_{j,y}(X_j)]}{\mathbb{E}[p_{j,y}(X_j)]}\right)^2\,. \tag{20}$$

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

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

then we can simplify (20) as

$$\sum_{i=1}^d \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\mathbb{P}(Y=y)\right)^2\right] = \sum_y \mathbb{P}(Y=y)\sum_j\|\mathbb{E}_{j,y}[S_\theta(X_j)]\|_2^2\,. \tag{21}$$

(i) Suppose that $\|\langle u,S_\theta(X)\rangle\|_{\Psi_p}^2 \leq I_0$ holds for any unit vector $u\in\mathbb{R}^d$. Then letting

$$u = \frac{\mathbb{E}_{j,y}[S_\theta(X)]}{\|\mathbb{E}_{j,y}[S_\theta(X)]\|_2},$$

by the convexity of $x\mapsto\exp(|x|^p)$ we have

$$\begin{aligned} 2 &\geq \mathbb{E}[\exp(|\langle u,S_\theta(X)\rangle|^p/I_0^{p/2})] \\ &\geq \mathbb{E}[p_{j,y}(X)\exp(|\langle u,S_\theta(X)\rangle|^p/I_0^{p/2})] \\ &= \mathbb{E}[p_{j,y}(X)]\mathbb{E}_{j,y}[\exp(|\langle u,S_\theta(X)\rangle|^p/I_0^{p/2})] \\ &\geq \mathbb{E}[p_{j,y}(X)]\exp\left(\left|\mathbb{E}_{j,y}[\langle u,S_\theta(X)\rangle]\right|^p/I_0^{p/2}\right) \\ &\geq \mathbb{E}[p_{j,y}(X)]\exp\left(\mathbb{E}_{j,y}[\langle u,S_\theta(X)\rangle]^p/I_0^{p/2}\right), \end{aligned}$$

and thus

$$\|\mathbb{E}_{j,y}[S_\theta(X_j)]\|_2 \le I_0^{1/2} \left(\log \frac{2}{\mathbb{E}[p_{j,y}(X)]}\right)^{\frac{1}{p}} .$$

Continuing from (21),

$$\begin{aligned}
&\sum_{i=1}^d \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log \mathbb{P}(Y=y)\right)^2\right] \\
&\le I_0 \sum_y \mathbb{P}(Y=y) \sum_j \left(\log \frac{2}{\mathbb{E}[p_{j,y}(X)]}\right)^{\frac{2}{p}} \\
&= I_0 \sum_{y,j} \left(\prod_{i\neq j} \mathbb{E}[p_{i,y}(X)]\right) \mathbb{E}[p_{j,y}(X)] \left(\log \frac{2}{\mathbb{E}[p_{j,y}(X)]}\right)^{\frac{2}{p}} \qquad (22)
\end{aligned}$$

Finally, by upper bounding (22) with the upper concave envelope ϕ of $x \mapsto x\left(\log \frac{2}{x}\right)^{\frac{2}{p}}$ on $[0,1]$, and then combining (19) with Jensen's inequality:

$$\begin{aligned}
\sum_{i=1}^d \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log \mathbb{P}(Y=y)\right)^2\right] &\le I_0 \sum_{y,j} \left(\prod_{i\neq j} \mathbb{E}[p_{i,y}(X)]\right) \phi\left(\mathbb{E}[p_{j,y}(X)]\right) \\
&\le I_0 \sum_j 2^k \phi\left(\sum_y \frac{1}{2^k}\mathbb{P}(Y=y)\right) \\
&= I_0 n (k+1)^{\frac{2}{p}} \\
&\le 4 I_0 n k^{\frac{2}{p}} .
\end{aligned}$$

(ii) Now suppose we instead just have the finite variance condition that

$$\mathsf{Var}(\langle u, S_\theta(X)\rangle) \le I_0$$

for any unit vector $u \in \mathbb{R}^d$. Picking again

$$u = \frac{\mathbb{E}_{j,y}[S_\theta(X)]}{\|\mathbb{E}_{j,y}[S_\theta(X)]\|_2} ,$$

the Cauchy-Schwarz inequality implies

$$\begin{aligned}
\mathbb{E}[p_{j,y}(X)]\|\mathbb{E}_{j,y}[S_\theta(X)]\|_2^2 &= \frac{1}{\mathbb{E}[p_{j,y}(X)]}\left(\mathbb{E}[\langle u, S_\theta(X)\rangle p_{j,y}(X)]\right)^2 \\
&\le \frac{1}{\mathbb{E}[p_{j,y}(X)]}\mathbb{E}[\langle u, S_\theta(X)\rangle^2]\mathbb{E}[p_{j,y}(X)^2] \\
&\le \frac{1}{\mathbb{E}[p_{j,y}(X)]}\mathbb{E}[\langle u, S_\theta(X)\rangle^2]\mathbb{E}[p_{j,y}(X)] \\
&= \mathbb{E}[\langle u, S_\theta(X)\rangle^2] \le I_0 .
\end{aligned}$$

Together with (21) this gives

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

The last equality follows from (19).

In both the sub-Gaussian (i) and finite variance (ii) cases, we apply the van Trees inequality just as in the proof for the independent or sequential models to arrive at the final result.

References

Jayadev Acharya, Clement Canonne, and Himanshu Tyagi. Inference under information constraints ii: Communication constraints and shared randomness. *arXiv preprint, arXiv:1905.08302*, 2019a.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Inference under information constraints: Lower bounds from chi-square contraction. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99, pages 3–17. PMLR, 2019b.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *Proceedings of Machine Learning Research*, volume 89, pages 1120–1129. PMLR, 2019c.

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 IEEE Allerton Conference on Communication, Control, and Computing*, 2018.

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Fisher information for distributed estimation under a blackboard communication protocol. *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2019.

A. A. Borovkov. *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

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. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, page 1011–1020, 2016.

Ilias Diakonikolas, Elena Grigorescu, Jerry Li, Abhiram Natarajan, Krzysztof Onak, and Ludwig Schmidt. Communication-efficient distributed learning of discrete probability distributions. *Advances in Neural Information Processing Systems*, pages 6394–6404, 2017.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Proceedings of the 32nd Conference On Learning Theory*, volume 99, pages 1161–1191. PMLR, 2019.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pages 429–438, 2013.

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

Richard D. Gill and Boris Y. Levit. Applications of the van trees inequality: a Bayesian cramér-rao bound. *Bernoulli*, 1(1/2):059–079, 1995.

Yanjun Han, Pritam Mukherjee, Ayfer Özgür, and Tsachy Weissman. Distributed statistical estimation of high-dimensional and nonparametric distributions. In *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, 2018a.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Proceedings of Machine Learning Research*, volume 75, pages 1–26. PMLR, 2018b.

Shun ichi Amari. On optimal data compression in multiterminal statistical inference. *IEEE Transactions on Information Theory*, 57(9):5577–5587, 2011.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 2436–2444. PMLR, 2016.

E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.

Wai-Man Lam and Amy R. Reibman. Design of quantizers for decentralized estimation systems. *IEEE Transactions on Communications*, 41(11):1602–1605, 1993.

Arkadi Nemirovski. Topics in non-parametric. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

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.

Alejandro Ribeiro and Georgios B. Giannakis. Non-parametric distributed quantization-estimation using wireless sensor networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

Parvathinathan Venkitasubramaniam, Gökhan Mergen, Lang Tong, and Ananthram Swami. Quantization for distributed estimation in large scale sensor networks. In *International Conference on Intelligent Sensing and Information Processing*, pages 121–127, 2005.

Parvathinathan Venkitasubramaniam, Lang Tong, and Ananthram Swami. Minimax quantization for distributed maximum likelihood estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006.

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

Shaowei Wang, Liusheng Huang, Pengzhan Wang, Yiwen Nie, Hongli Xu, Wei Yang, Xiang-Yang Li, and Chunming Qiao. Mutual information optimally local private discrete distribution estimation. *arXiv preprint, arXiv:1607.08025*, 2016.

Aolin Xu and Maxim Raginsky. Information-theoretic lower bounds on bayes risk in decentralized estimation. *IEEE Transactions on Information Theory*, 63(3):1580–1600, 2017.

Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018.

Ram Zamir. A proof of the fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory*, 44(3):1246–1250, 1998.

Yuchen Zhang, John Duchi, Micheal I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013.