Lower Bounds for Over-the-Air Statistical Estimation

Chuan-Zheng Lee, Leighton Pate Barnes and Ayfer Özgür
Department of Electrical Engineering
Stanford University
Stanford, California
Email: {czlee, lpb, aozgur}@stanford.edu

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518157

***Abstract*—We study lower bounds for minimax statistical estimation over a Gaussian multiple-access channel (MAC) under squared error loss, using techniques from both statistical estimation and information theory. We characterize these bounds in terms of the number of nodes n and the dimension of the parameter space d, showing that the risk must be $\Omega(d/n \log n)$. This is within a $\log n$ factor of previous analog achievability results. While lower bounds for minimax statistical estimation have been previously studied under quantization constraints that abstract the physical layer as noiseless bit pipes, to our knowledge our paper provides the first lower bounds for statistical estimation over noisy multi-user channels. This adds to a body of works showing how analog schemes that consider the physical layer jointly with the estimation scheme, can outperform digital schemes that separate the two with an abstraction layer.**

I. Introduction

One interesting facet of the modern data explosion is not so much its quantity, but that increasingly, data is being generated "at the edge": by countless sensors and other devices, away from the central servers that churn through it. The study of machine learning has therefore in recent years been paying increasing attention to techniques that make inferences by combining data from many nodes. The key differentiating feature that this introduces is the need to communicate the data from edge devices to the central server, often over noisy bandwidth-limited networks. As a result, learning and estimation in networks has received significant interest in the recent years.

One approach to modelling bandwidth limitations is to consider each node to be limited in how many bits it can send to the central server. That is, the observed samples are first encoded using a fixed, finite number of bits, and these bits are then communicated errorlessly using a reliable channel coding scheme over the underlying noisy network. This amounts to introducing a digital interface separating source coding (to represent samples) from channel coding, and we refer to it in this paper as the "digital" approach. A number of recent works [1]–[6] studied this framework, and derived lower bounds on the estimation error in terms of the bit budget for each sample.

On the other hand, enforcing a digital interface between communication and statistical estimation may lead to suboptimal performance for the end-to-end system. In our earlier work [7], we introduced an "analog" framework for distributed minimax estimation over a Gaussian multiple-access channel (MAC) (see Fig. 1). This framework removed the digital interface between source and channel coding, permitting observed samples to be mapped to the MAC input by any encoding function that satisfies the power constraint at the transmitters. In this framework, we proposed analog schemes for the Gaussian and Bernoulli location estimation models, where nodes simply transmit scaled but uncoded samples over the MAC, leveraging its additive nature to average samples over the air. By comparing the performance of these schemes to the aforementioned digital lower bounds, with the bit budget set to the Shannon capacity of the Gaussian MAC for the same power constraint, we showed that, judged by minimax risk under squared error loss, these analog schemes yield an exponential improvement over even the *lower bounds* for digital schemes presented in [6].

These results reinforced similar gains that have been observed in source coding for sensor networks [8], [9], as well as a number of studies modifying common machine learning algorithms, such as gradient descent, to account for the wireless physical layer. These latter works comparing analog and digital approaches experimentally include [10], [11], which studied stochastic gradient descent, and [12], which studied transmission of model parameters. On a similar tune, several further works have continued to progress analog superposition-based methods in over-the-air learning [13]–[18].

While our own earlier work demonstrated the value of such "analog" transmission-estimation schemes, it didn't offer any fundamental lower bounds against which to assess the analog schemes therein proposed. In this paper, we fill this gap by deriving a lower bound for risk under squared error loss for estimation of sub-Gaussian models over a Gaussian MAC. Because we don't impose a separation between transmission and estimation, but instead analyze estimation over a physical multi-user channel directly, our bounds differ from the digital lower bounds developed in the previous literature both in terms of their final scaling as well as the utilized techniques. Moreover, these bounds are within a logarithmic factor of our achievability results in [7], and they are the first lower bounds of which we are aware for analog estimation over a multi-user channel.

The rest of this paper is structured as follows. In Section II, we set out our problem of interest, and in Section III we introduce some statistical definitions that will be key in our

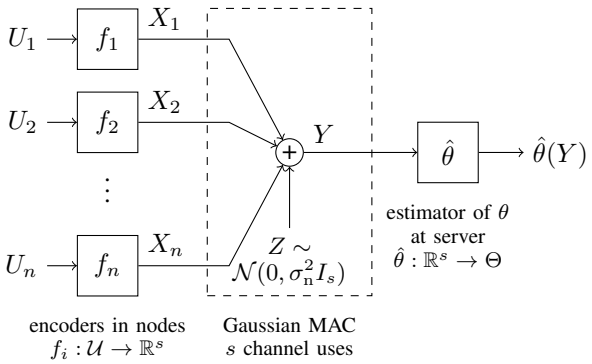

Fig. 1. System diagram

results. We present our results in Section IV, including our main result (IV-A), bounds for specific cases (IV-B), where this fits in with earlier results (IV-C) and a more general version of our main result (IV-D). We provide proofs in Section V.

II. Problem statement

We study the same Gaussian multiple-access channel as in our previous work [7], a system diagram of which is in Fig. 1. In each channel use $t = 1, \dots, s$, each of n senders transmits a symbol $X_{it} \in \mathbb{R}$ to a single receiver, which receives a noisy superposition

$$Y_t = X_{1t} + X_{2t} + \cdots + X_{nt} + Z_t, \tag{1}$$

where $Z_t \sim \mathcal{N}(0, \sigma_n^2)$ is the noise in the tth channel use. We denote $X_i = (X_{i1}, \dots, X_{is})$ and $Y = (Y_1, \dots, Y_s)$. The goal of our system is to estimate the parameter θ, belonging to a parameter space $\Theta \subseteq \mathbb{R}^d$, of the distribution p_θ from which i.i.d. samples $U_1, \dots, U_n \in \mathcal{U}$ are drawn, with U_i observed at sender i. To do this, each sender i encodes its sample using a function $f_i : \mathcal{U} \to \mathbb{R}^s$ to produce $X_i = f_i(U_i)$, and the receiver, which knows the encoding functions, uses an estimator $\hat{\theta}(Y)$. We refer to a combination of encoding functions $\mathbf{f} \triangleq (f_1, \dots, f_n)$ and an estimator function $\hat{\theta} : \mathbb{R}^s \to \Theta$ as an *estimation scheme*.

Senders are subject to a power constraint, and as the distribution p_θ is not known, they must respect it for the entire parameter space. That is, we require that

$$\frac{1}{s}\mathbb{E}_\theta\left[\|f_i(U_i)\|_2^2\right] \le P, \quad \text{for all } i \in \{1, \dots, n\}, \theta \in \Theta, \tag{2}$$

where we denote as $\mathbb{E}_\theta[\cdot]$ the expectation under p_θ. We study worst-case risk under squared error loss, $\sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta}(Y) - \theta\|_2^2$.

In this paper, we provide results in terms of the properties of the parametric model p_θ and the parameter space Θ. We will also apply this result to the two specific cases we studied in our previous work [7]. The first of these is the Gaussian location model, in which $p_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$, $\mathcal{U} = \mathbb{R}^d$ and $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le B\sqrt{d}\}$ for some known $B > 0$. The second is the product Bernoulli parameter model, in which $p_\theta = \prod_{j=1}^d \text{Bernoulli}(\theta_j)$, with $\mathcal{U} = \{0,1\}^d$ and $\Theta = [0,1]^d$.

Finally, although our main focus is the Gaussian MAC, we will also briefly consider how our result generalizes to other MACs, in terms of their sum capacity.

III. Preliminaries

We begin by introducing some useful quantities. If $U \sim p_\theta$, where p_θ is a member of a family of probability distributions parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$ and differentiable in θ, the *score function* is defined as the gradient of the log-likelihood function,

$$\begin{aligned} S_\theta(u) &\triangleq \nabla_\theta \log p_\theta(u) \\ &= \left(\frac{\partial}{\partial \theta_1} \log p_\theta(u), \dots, \frac{\partial}{\partial \theta_d} \log p_\theta(u)\right). \end{aligned}$$

Where we have many samples $U_1, \dots, U_n \sim p_\theta$, we may denote the score function of the finite sequence as

$$S_\theta(u_1, \dots, u_n) \triangleq \nabla_\theta \log p_\theta(u_1, \dots, u_n).$$

Note that both $S_\theta(u)$ and $S_\theta(u_1, \dots, u_n)$ have the same number of elements as θ, independent of the size of the argument passed into $S_\theta(\cdot)$. It is a well-known property of the score function that $\mathbb{E}[S_\theta(U)] = 0$.

The *Fisher information* is then defined as the $d \times d$ matrix

$$I_U(\theta) \triangleq \mathbb{E}[S_\theta(U) S_\theta(U)^\mathsf{T}]$$

which makes its trace equal to

$$\mathbf{tr}(I_U(\theta)) = \sum_{j=1}^d \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_j} \log p_\theta(u)\right)^2\right].$$

We say that a zero-mean random variable X is *sub-Gaussian with parameter* ρ if

$$\mathbb{E}\left[\exp(\lambda X)\right] \le \exp\left(\frac{\lambda^2 \rho^2}{2}\right) \quad \text{for all } \lambda \in \mathbb{R}.$$

Recall that if a zero-mean random variable X is bounded within $[a, b]$ with probability 1 then it is sub-Gaussian with parameter $(b-a)/2$. Also, if $X_1, \dots, X_n$ are independent and sub-Gaussian with parameters $\rho_1, \dots, \rho_n$, then their sum $X_1 + \cdots + X_n$ is sub-Gaussian with parameter $\sqrt{\rho_1^2 + \cdots + \rho_n^2}$.

Our lower bounds require the regularity conditions described in [19], which we recite here:

(i) $\sqrt{p_\theta(u_1, \dots, u_n)}$ is continuously differentiable with respect to each component θ_j at almost all $(u_1, \dots, u_n) \in \mathcal{U}^n$ (with respect to some measure dominating $\{p_\theta : \theta \in \Theta\}$).

(ii) The Fisher information for each component θ_j, $\mathbb{E}([\frac{\partial}{\partial \theta_j} \log p_\theta(U_1, \dots, U_n)]^2)$, exists and is a continuous function of θ_j.

(iii) The conditional density $p(y|u_1, \dots, u_n)$ is square integrable in the sense that for almost all $y \in \mathbb{R}^s$ for each θ, $\int p(y|u_1, \dots, u_n)^2 \, dp_\theta(u_1, \dots, u_n) < \infty$.

It is easily verified that (i) and (ii) are satisfied in both the Gaussian location and product Bernoulli parameter models. As for (iii), this follows from the bounded conditional density $p(y|x_1, \dots, x_n)$ of the Gaussian MAC; details are in the relevant proofs.

IV. Results

A. Main result: Lower bound on worst-case risk

The main result of this paper is a lower bound on the squared error risk for any estimation scheme in the setting described in Section II when the parametric model has a sub-Gaussian score.

Theorem 1. *Suppose that $[-B,B]^d \subset \Theta$, and that the samples $(U_i)_{i=1}^n$ are i.i.d. and satisfy conditions (i) and (ii), and that $\langle v, S_\theta(U_i)\rangle$ is sub-Gaussian with parameter ρ for all unit vectors $v \in \mathbb{R}^d$. Then in a Gaussian multiple-access channel with s channel uses, the worst-case risk under squared error loss of any estimation scheme $(\mathbf{f}, \hat{\theta})$ must satisfy*

$$\sup_{\theta\in\Theta} \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 \geq \frac{d}{n} \cdot \frac{1}{\frac{s}{d}\rho^2 \log\left(1 + \frac{nP}{\sigma_{\mathrm{n}}^2}\right) + \frac{\pi^2}{nB^2}}. \quad (3)$$

The proof for this builds on a relationship between Fisher information and mutual information established by [19], and a Bayesian Cramer-Rao type bound to relate minimax risk and Fisher information. We provide the proof in Section V-B.

B. Bounds for specific problem instances

In the case of the Gaussian location model, we can characterize this bound in terms of the sample variance σ^2.

Corollary 1. *In the Gaussian location model with s channel uses, the worst-case risk under squared error loss of any estimation scheme $(\mathbf{f}, \hat{\theta})$ must satisfy*

$$\sup_{\theta\in\Theta} \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 \geq \frac{d\sigma^2}{n} \cdot \frac{1}{\frac{s}{d}\log\left(1 + \frac{nP}{\sigma_{\mathrm{n}}^2}\right) + \frac{\sigma^2}{B^2}\frac{\pi^2}{n}}. \quad (4)$$

We can also derive a result for product Bernoulli models where elements of θ are close to $\frac{1}{2}$, *i.e.*, where the samples are dense.

Corollary 2. *Consider the relatively dense product Bernoulli model, where $U_1, \ldots, U_n \sim \prod_{j=1}^d \mathrm{Bernoulli}(\theta_j)$, with $\Theta = [\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon]^d$, $\varepsilon \in (0, \frac{1}{2})$, with s channel uses. The worst-case risk under squared error loss of any estimation scheme $(\mathbf{f}, \hat{\theta})$ must satisfy*

$$\sup_{\theta\in\Theta} \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 \geq \frac{d}{n} \cdot \frac{1}{\frac{s}{d}\frac{1}{(\frac{1}{2}-2\varepsilon^2)^2}\log\left(1 + \frac{nP}{\sigma_{\mathrm{n}}^2}\right) + \frac{\pi^2}{n\varepsilon^2}}. \quad (5)$$

The proofs of the above two corollaries, which both follow from Theorem 1, are in Section V-C.

C. Comparison to prior results

In our previous work [7], we considered over-the-air estimation schemes using scaled encoding functions for both models of interest, and analyzed their performance. We recite these achievability results in Proposition 1.

Proposition 1. *In the Gaussian location model, if $s \geq d$, there exists an estimation scheme $(\mathbf{f}, \hat{\theta})$ achieving the worst-case risk*

$$\sup_{\theta\in\Theta} \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 = \frac{d\sigma^2}{n}\left[1 + \frac{\sigma_{\mathrm{n}}^2}{\lfloor s/d \rfloor nP}\left(1 + \frac{B^2}{\sigma^2}\right)\right]. \quad (6)$$

In the product Bernoulli parameter model, if $s \geq d$ and $\sigma_{\mathrm{n}}^2 \leq n^{3/2}P$, there exists a scheme $(\mathbf{f}, \hat{\theta})$ achieving

$$\sup_{\theta\in\Theta} \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 = \frac{d}{4(\sqrt{n}+1)^2}\left(1 + \frac{\sigma_{\mathrm{n}}^2}{\lfloor s/d \rfloor nP}\right). \quad (7)$$

We also applied the result in [6] to find a lower bound for schemes that abstract out the physical layer, considering instead bits transmitted errorlessly at the Shannon capacity. We recite the resulting bounds in Proposition 2.

Proposition 2. *Consider all schemes in which senders send bits to the receiver at the Shannon capacity for s channel uses. In the Gaussian location model, for $B^2 \min\{\frac{s}{2d}\log_2\left(1 + nP/\sigma_{\mathrm{n}}^2\right), n\} \geq \sigma^2$, the risk associated with any such scheme is at least*

$$\sup_{\|\theta\|_2 \leq B\sqrt{d}} \mathbb{E}_\theta\|\hat{\theta} - \theta\|_2^2 \geq C\sigma^2 \max\left\{\frac{2d^2}{s\log_2\left(1 + \frac{nP}{\sigma_{\mathrm{n}}^2}\right)}, \frac{d}{n}\right\}, \quad (8)$$

with a universal constant C.

In the product Bernoulli model, for $\min\{\frac{s}{2d}\log_2\left(1 + nP/\sigma_{\mathrm{n}}^2\right), n\} \geq 1$, the risk associated with any such scheme is at least

$$\sup_{\theta\in[0,1]^d} \mathbb{E}_\theta\|\hat{\theta} - \theta\|_2^2 \geq C\max\left\{\frac{2d^2}{s\log_2\left(1 + \frac{nP}{\sigma_{\mathrm{n}}^2}\right)}, \frac{d}{n}\right\}, \quad (9)$$

with a universal constant C.

Table I summarizes behavior in s, d and n for our new lower bounds and the above prior results. The last row shows the case where s grows with d, for ease of comparison. Although the Gaussian location and product Bernoulli cases look different, their asymptotic behavior in s, d and n is the same. Our new lower bound is within a $\log n$ factor of the scheme we proposed in [7]. Note that the digital lower bound establishes that the minimax risk of any digital scheme can decrease at most logarithmically in the number of samples, while the risk of analog schemes decreases linearly in n. This shows that analog over-the-air estimation schemes can provide an exponential improvement in sample complexity for the Gaussian and Bernoulli location models.

D. Lower bound for a general multiple-access channel

The result of Theorem 1 can be generalized to other multiple-access channels, in terms of the total capacity of the network, that is, the maximum achievable sum of all rates in the network,

$$C_{\mathrm{total}} = \max_{\prod_i p_i(x_{it})} I(X_{1t}, \ldots, Y_{nt}; Y_t), \quad (10)$$

TABLE I
COMPARISON OF RESULTS

	analog achievability	analog lower bound	digital lower bound
Gaussian location	$\frac{d\sigma^2}{n}\left[1+\frac{\sigma_{\mathrm{n}}^2}{\lfloor s/d\rfloor nP}\left(1+\frac{B^2}{\sigma^2}\right)\right]$	$\frac{d\sigma^2}{n}\cdot\frac{1}{\frac{s}{d}\log\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right)+\frac{\sigma^2}{B^2}\frac{\pi^2}{n}}$	$C\frac{d\sigma^2}{n}\max\left\{\frac{2dn}{s\log_2\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right)},1\right\}$
product Bernoulli	$\frac{d}{4(\sqrt{n}+1)^2}\left(1+\frac{\sigma_{\mathrm{n}}^2}{\lfloor s/d\rfloor nP}\right)$	$\frac{d}{n}\cdot\frac{1}{\frac{s}{d}\frac{1}{(\frac{1}{2}-2\varepsilon^2)^2}\log\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right)+\frac{\pi^2}{n\varepsilon^2}}$.	$C\max\left\{\frac{2d^2}{s\log_2\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right)},\frac{d}{n}\right\}$
$s\geq d$ (both models)	$O\left(\frac{d}{n}\right)$	$\Omega\left(\frac{d^2}{s\cdot n\log n}\right)$	$\Omega\left(\frac{d^2}{s\log n}\right)$
$s\propto d$ (both models)	$O\left(\frac{d}{n}\right)$	$\Omega\left(\frac{d}{n\log n}\right)$	$\Omega\left(\frac{d}{\log n}\right)$

where the maximum is over all product distributions for $(X_{1t},\dots,X_{nt})$. We will refer to this quantity as the "sum capacity", recalling that it does not fully describe the capacity region of the network.

Theorem 2. *Suppose that the samples $(U_i)_{i=1}^n$ are i.i.d. and satisfy conditions (i) and (ii), and that $\langle v, S_\theta(U_i)\rangle$ is sub-Gaussian with parameter ρ for all unit vectors $v\in\mathbb{R}^d$. Consider any discrete memoryless multiple-access channel that is constrained by the sum capacity C_{total} (per channel use), whose conditional density $p(y|x_1,\dots,x_n)$ is bounded. The worst-case risk of any estimation scheme $(\mathbf{f},\hat{\theta})$ must satisfy*

$$\sup_{\theta\in\Theta}\mathbb{E}\|\hat{\theta}(Y)-\theta\|^2\geq\frac{d}{n}\cdot\frac{1}{2\frac{s}{d}\rho^2C_{\mathrm{total}}+\frac{\pi^2}{nB^2}}. \tag{11}$$

Proof. Follow the proof of Theorem 1, but replace the right-hand side of (19) with sC_{total} (*i.e.*, C_{total} for s channel uses). Note that the stipulation that $p(y|x_1,\dots,x_n)$ be bounded (by some different finite M) ensures that (16), and hence (iii), is satisfied. □

V. PROOFS

A. Prior results

We first present two results that are key to our main theorem. These results characterize the worst-case risk in terms of the trace of the Fisher information matrix, and in turn in terms of mutual information. First, Equation 8 of [6] tells us the following, which we list as a lemma here.

Lemma 1. *Suppose $[-B,B]^d\subset\Theta$. For any estimator $\hat{\theta}(Y_1,\dots,Y_n)$, the worst-case squared error risk must satisfy*

$$\sup_{\theta\in\Theta}\mathbb{E}\|\hat{\theta}(Y)-\theta\|^2\geq\frac{d^2}{\sup_{\theta\in\Theta}\mathbf{tr}(I_Y(\theta))+\frac{d\pi^2}{B^2}}. \tag{12}$$

We will also lean on the following theorem, due to [19].

Theorem 3. *Let $X\sim p_\theta$ and let Y be the output of a channel characterized by $p(y|x)$. Suppose that $\langle u, S_\theta(X)\rangle$ is sub-Gaussian with parameter N for any unit vector $u\in\mathbb{R}^d$. Under regularity conditions (i)–(iii),*

$$\mathbf{tr}(I_Y(\theta))\leq 2N^2I_\theta(X;Y),$$

where $I_\theta(X;Y)$ is the mutual information between X and Y when $X\sim p_\theta$.

B. Proof of main result

We first provide an upper bound for the mutual information between the channel input and output. For brevity we omit the proof, which can be derived using standard results in information theory.

Proposition 3. *Consider the Gaussian multiple-access channel with s channel uses, $Y=X_1+\cdots+X_n+Z$, $Z\sim\mathcal{N}(0,\sigma_{\mathrm{n}}^2I_s)$, with a power constraint $\frac{1}{s}\mathbb{E}[\|X_i\|^2]\leq P$. If $X_1,\dots,X_n$ are independent, then the mutual information between its input $(X_1,\dots,X_n)$ and its output Y is bounded by*

$$I(X_1,\dots,X_n;Y)\leq\frac{s}{2}\log\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right). \tag{13}$$

We now have all of the ingredients necessary to prove Theorem 1, which uses the data processing inequality to chain the above results together.

Proof of Theorem 1. Recall that $X_i=f(U_i), i=1,\dots,n$ and Y is the output of the channel $p_{Y|X}(y|x_1,\dots,x_n)$ with inputs $X_1,\dots,X_n$. The conditional distribution of Y given U can be expressed in terms of the channel's conditional distribution,

$$p_{Y|U}(y|u_1,\dots,u_n)=p_{Y|X}(y|f(u_1),\dots,f(u_n)). \tag{14}$$

That is, we have a channel from U to Y. (Note that this doesn't require invertibility in f, since it is in the condition, and $p_{Y|X}$ is defined by assumption.) To verify that this "channel" satisfies regularity condition (iii), note that $p_{Y|X}$ is bounded,

$$p_{Y|X}(y|x_1,\dots,x_n)\leq\frac{1}{\sqrt{(2\pi\sigma_{\mathrm{n}}^2)^n}}\triangleq M, \tag{15}$$

so chaining (14) and (15) verifies that

$$\int p_{Y|U}(y|u_1,\dots,u_n)^2\, dp_U(u_1,\dots,u_n) \le \int M^2\, dp_U(u_1,\dots,u_n) = M^2 < \infty. \quad (16)$$

We therefore satisfy the requirements to invoke Theorem 3, so long as we can establish that $\langle v, S_\theta(U_1,\dots,U_n)\rangle$ is sub-Gaussian for all unit vectors $v \in \mathbb{R}^d$. Note that since $U_1,\dots,U_n$ are independent,

$$S_\theta(U_1,\dots,U_n) = \sum_{i=1}^n \nabla_\theta \log p_\theta(U_i) = \sum_{i=1}^n S_\theta(U_i).$$

Then for every unit vector $v \in \mathbb{R}^d$,

$$\langle v, S_\theta(U_1,\dots,U_n)\rangle = \left\langle v, \sum_{i=1}^n S_\theta(U_i)\right\rangle = \sum_{i=1}^n \langle v, S_\theta(U_i)\rangle.$$

This is a sum of n independent sub-Gaussian random variables each with parameter ρ, and is therefore sub-Gaussian with parameter $\sqrt{n}\rho$. Theorem 3 thus gives

$$\mathbf{tr}(I_Y(\theta)) \le 2n\rho^2 I_\theta(U_1,\dots,U_n;Y). \quad (17)$$

Since $(U_1,\dots,U_n) \to (X_1,\dots,X_n) \to Y$ form a Markov chain, the data processing inequality implies that

$$I_\theta(U_1,\dots,U_n;Y) \le I_\theta(X_1,\dots,X_n;Y). \quad (18)$$

Now, $U_1,\dots,U_n$ are independent (by definition), and each $X_i, i = 1,\dots,n$ is a function of the corresponding U_i. Therefore, $X_1,\dots,X_n$ are also independent, and from Proposition 3, we have

$$I(X_1,\dots,X_n;Y) \le \frac{s}{2}\log\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right). \quad (19)$$

Putting (17), (18) and (19) together yields

$$\mathbf{tr}(I_Y(\theta)) \le n\rho^2 s \log\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right). \quad (20)$$

Substituting this expression into the result given by Lemma 1 then yields the result. □

C. Specific problem instances

To find lower bounds for the Gaussian location and product Bernoulli parameter models, we compute the sub-Gaussian parameters of their score functions and apply our main result.

Proof of Corollary 1. The score function for a single sample U_i is

$$S_\theta(u_i) = \nabla_\theta\left[\frac{(u_i-\theta)^\mathsf{T}(u_i-\theta)}{2\sigma^2} - \log 2\pi\sigma\right] = \frac{1}{\sigma^2}(u_i-\theta).$$

Then, with $U_i \sim \mathcal{N}(\theta,\sigma^2 I_d)$, the score function of each sample $S_\theta(U_i)$ is Gaussian with mean zero and covariance $\frac{1}{\sigma^2}I_d$. It follows that for any unit vector v and each sample U_i, $\langle v, S_\theta(U_i)\rangle$ is Gaussian with zero mean and variance $v^\mathsf{T}\frac{1}{\sigma^2}I_d v = \frac{1}{\sigma^2}v^\mathsf{T}v = \frac{1}{\sigma^2}$. This is sub-Gaussian with parameter $\frac{1}{\sigma}$, enabling an application of Theorem 1. □

Proof of Corollary 2. We can compute the score function of each sample in the product Bernoulli model to be $S_\theta(u_i) = (S_{\theta_1}(u_i),\dots,S_{\theta_j}(u_i))$, where

$$S_{\theta_j}(u_i) = \begin{cases} \frac{1}{\theta_j}, & \text{if } u_{ij}=1 \\ -\frac{1}{1-\theta_j}, & \text{if } u_{ij}=0. \end{cases}$$

Then $S_{\theta_j}(U_i)$ is bounded, and therefore is sub-Gaussian with parameter

$$\frac{1}{2}\left[\frac{1}{\theta_j}+\frac{1}{1-\theta_j}\right] = \frac{1}{2\theta_j(1-\theta_j)} \le \frac{1}{\frac{1}{2}-2\varepsilon^2},$$

where the last step uses the fact that $\theta \in \Theta = [\frac{1}{2}-\varepsilon, \frac{1}{2}+\varepsilon]^d$.

Being the sum of n independent sub-Gaussians, for all unit vectors $v \in \mathbb{R}^d$, $\langle v, S_\theta(U_i)\rangle$ is sub-Gaussian with parameter

$$\sqrt{\sum_{j=1}^d v_j^2 \frac{1}{(\frac{1}{2}-2\varepsilon^2)^2}} = \frac{1}{\frac{1}{2}-2\varepsilon^2}. \quad (21)$$

This gives us a value for ρ to use in Theorem 1.

For B, we may reparameterize the parameter space to $\Theta' = [-\varepsilon,\varepsilon]$ (so that the Bernoulli component means are $\theta = \theta' + \frac{1}{2}$). We can then apply Theorem 1 to arrive at Corollary 2. □

ACKNOWLEDGEMENT

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

REFERENCES

[1] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems*, 2013, pp. 2328–2336.

[2] A. Garg, T. Ma, and H. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances in Neural Information Processing Systems*, 2014, p. 2726–2734.

[3] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. 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*. ACM, 2016, p. 1011–1020.

[4] I. Diakonikolas, E. Grigorescu, J. Li, A. Natarajan, K. Onak, and L. Schmidt, "Communication-efficient distributed learning of discrete probability distributions," *Advances in Neural Information Processing Systems*, pp. 6394–6404, 2017.

[5] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints: Lower bounds from chi-square contraction," in *Proceedings of the 32nd Conference on Learning Theory*, vol. 99. Phoenix, USA: PMLR, 25–28 Jun 2019, pp. 3–17.

[6] L. P. Barnes, Y. Han, and A. Ozgur, "Lower bounds for learning distributions under communication constraints via fisher information," *Journal of Machine Learning Research*, vol. 21, no. 236, pp. 1–30, 2020.

[7] C.-Z. Lee, L. P. Barnes, and A. Özgür, "Over-the-air statistical estimation," in *IEEE Global Communications Conference (GLOBECOM)*, 2020.

[8] M. Gastpar, "Uncoded transmission is exactly optimal for a simple gaussian "sensor" network," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5247–5251, 2008.

[9] M. Gastpar and M. Vetterli, "Source-channel communication in sensor networks," in *Information Processing in Sensor Networks*, F. Zhao and L. Guibas, Eds. Berlin, Heidelberg: Springer, 2003, pp. 162–177.
[10] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.
[11] ——, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.
[12] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2020.
[13] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2897–2911, 2020.
[14] Y. Du and K. Huang, "Fast analog transmission for high-mobility wireless data acquisition in edge learning," *IEEE Wireless Communications Letters*, vol. 8, no. 2, pp. 468–471, 2019.
[15] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," 2020, arXiv:2008.11141.
[16] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "Cotaf: Convergent over-the-air federated learning," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.
[17] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.
[18] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 2604–2609.
[19] L. P. Barnes and A. Özgür, "Fisher information and mutual information constraints," 2021, arXiv:2102.05802.