Over-the-Air Statistical Estimation

Chuan-Zheng Lee, Leighton Pate Barnes, and Ayfer Özgür, *Senior Member, IEEE*

Abstract—We study schemes and lower bounds for distributed minimax statistical estimation over a Gaussian multiple-access channel (MAC) under squared error loss, in a framework combining statistical estimation and wireless communication. First, we develop "analog" joint estimation-communication schemes that exploit the superposition property of the Gaussian MAC and we characterize their risk in terms of the number of nodes and dimension of the parameter space. Then, we derive information-theoretic lower bounds on the minimax risk of any estimation scheme restricted to communicate the samples over a given number of uses of the channel and show that the risk achieved by our proposed schemes is within a logarithmic factor of these lower bounds. We compare both achievability and lower bound results to previous "digital" lower bounds, where nodes transmit errorless bits at the Shannon capacity of the MAC, showing that estimation schemes that leverage the physical layer offer a drastic reduction in estimation error over digital schemes relying on a physical-layer abstraction.

Index Terms—federated learning, over-the-air learning, statistical estimation

I. INTRODUCTION

To fully appreciate the plenitude of data fueling the modern rise of machine learning, we might pause to consider not just its volume, but its origins. While the computational lifting is often concentrated in powerful central servers, the data they rely on is largely and increasingly born "at the edge"—spawned in a myriad of devices, scattered ubiquitously, equipped with sensors and user input to collect all sorts of information from the world.

Recognizing this growing decentralization of data, there has been growing interest in the study of techniques to combine samples from many nodes to make inferences. The key distinction between this and traditional approaches to learning and estimation is an explicit consideration of the communication channels between edge devices and the central server, which are often noisy or unreliable channels, like wireless links. Two disciplines with decades-long histories, wireless networks and statistical estimation, thus find themselves with common cause.

Recent years have seen significant activity in this nascent intersection. One simple and intuitive way to study the impact of bandwidth-limitations on estimation performance is to model them as imposing constraints on the number of bits available to encode each observed sample. A number of works in the machine learning literature have taken this tack [1]–[6], providing both achievable schemes and information-theoretic lower bounds under bit constraints, and characterizing the dependence of the estimation error on the number of bits available to represent each sample. Since these analyses

The authors are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (email: czlee@stanford.edu; lpb@stanford.edu; aozgur@stanford.edu).

assume the encoded bits to be received without error, they can be interpreted as assuming that a reliable scheme is used for transmission over the underlying noisy channels. This in effect suggests an abstraction layer, separating the question of physical-layer communication from the statistical estimation problem.

Our goal in this paper is to study the problem of distributed statistical estimation over a noisy multiple-access channel from first principles. Our main contributions are as follows.

First, we introduce a model for minimax parameter estimation over a fixed number of uses of the Gaussian multiple-access channel (MAC), defining the estimation schemes of interest in this setting and providing a rigorous and tractable mathematical model for quantifying their performance and studying their optimality.

Second, we develop analog transmission-estimation schemes for two canonical estimation tasks, the Gaussian location and product Bernoulli models, in which each node scales and transmits its uncoded sample to the central server, leveraging the superposition of the Gaussian MAC to perform averaging over the air. We analyze these schemes and characterize their risk under squared error loss. When compared to information theoretic *lower* bounds for digital schemes, controlling for physical resources, we find that the worst-case risk of these analog schemes is exponentially smaller than that of any digital scheme. This suggests that analog schemes that consider estimation and physical layer transmission jointly can bring about drastic improvements over digital schemes that separate the two with an abstraction layer.

We next address the question of whether the analog schemes we develop are close to optimality. We derive a fundamental lower bound on the risk achievable by any estimation-communication scheme satisfying the physical constraints of our model. This bound uses the recent result of [7], which showed an upper bound on the Fisher information of a channel's output in terms of the channel's Shannon mutual information when the statistical model being estimated exhibits a sub-Gaussian score. We apply this result to the aforementioned two estimation tasks, and find the risk achieved by our estimation schemes to be within a logarithmic factor of the lower bound. To the best of our knowledge, this is the first information-theoretic lower bound for distributed minimax estimation over a noisy multi-user channel.

A. Related works

Our results corroborate similar gains from uncoded analog transmission that have been observed in source coding for sensor networks [8], [9]. They also reinforce a number of recent works proposing adaptations of common learning algorithms

in a wireless federated learning context. A notable example is distributed stochastic gradient descent, where leveraging the Gaussian MAC for analog averaging has been observed experimentally to far outperform digital approaches [10]–[12]. On the theoretical side, analysis of methods using analog over-the-air gradient aggregation has shown convergence rates similar to error-free channels [13]–[15], albeit without comparison to digital counterparts.

Other works with analog aggregation showing an advantage over digital methods include [16], which replaced digital with analog modulation of model parameters to improve latency with comparable accuracy, and [17], which used an analog method with MIMO antennas. On the other hand, [18] found the digital and analog schemes it studied to perform comparably in numerical experiments. Such analog methods have seen wider development efforts [19], [20], including with privacy considerations in mind [21]–[24]. This idea has also been applied to over-the-air computation more generally [25]–[29]. While the current literature supports the utility of analog schemes, none of these works provide an explicit analytical comparison from first principles between analog achievability and digital lower bounds for canonical estimation and inference tasks.

More broadly, there has been interest in federated learning over wireless networks from a range of angles. One idea to preserve bandwidth is to take advantage of gradient sparsity [10], [30], [31]. Other angles include quantization [32], [33], incentive mechanisms [34], power control [24], [35] and optimization [36]–[38].

B. Structure of paper

The rest of this paper is structured as follows. In Section II we define the problem and introduce the definition of a minimax estimation scheme in this setting. We summarize our main results for achievability in Section III and discuss how they compare to existing digital lower bounds in Section IV. We then derive our lower bounds and discuss them in Section V. Proofs of our results are in Sections VI (achievability) and VII (lower bounds). We validate our results with simulations in Section VIII and conclude in Section IX.

II. PROBLEM STATEMENT

We study statistical estimation over a Gaussian multiple-access channel, a system diagram of which is in Fig. 1. In each channel use $t = 1, \ldots, s$, each of n senders transmits a symbol $X_{1t}, \ldots, X_{nt} \in \mathbb{R}$ to the central server (which we interchangeably refer to as the *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 assume an average power constraint P on each sender,

$$\frac{1}{s} \sum_{t=1}^{s} \mathbb{E}[X_{it}^2] \le P, \quad \text{for all } i = 1, \ldots, n, \tag{2}$$

where the expectation is over whatever randomness might exist in X_{it}, which we will make more precise shortly.
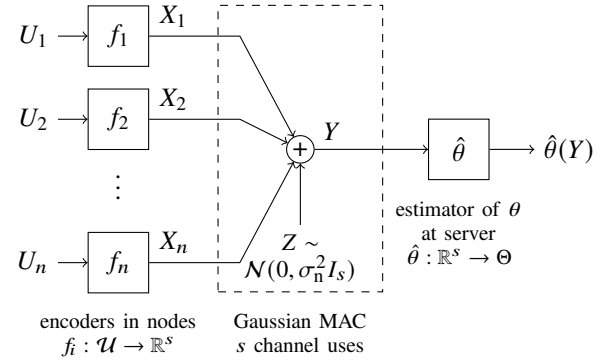

Fig. 1. System diagram

This system has the following estimation task: Each of the n senders has an i.i.d. sample U_i, $i = 1, \ldots, n$, from an unknown distribution p_θ on an alphabet $\mathcal{U}$, belonging to a parameterized family of distributions $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ with parameter space $\Theta \subseteq \mathbb{R}^d$. We use the notation $\mathbb{E}_\theta[\cdot]$ to mean expectation under the distribution p_θ. The goal of the receiver is to estimate θ given $Y \triangleq (Y_1, \ldots, Y_s)$.

To complete this task, each sender i encodes its sample using a function $f_i : \mathcal{U} \to \mathbb{R}^s$ to produce $X_i \triangleq (X_{i1}, \ldots, X_{is}) = f_i(U_i)$. The receiver, which knows the encoding functions, uses an estimator $\hat{\theta}(Y)$. We thus define how an estimation is carried out.

Definition 1. An *estimation scheme* for s channel uses is a pair $(\mathbf{f}, \hat{\theta})$ comprising n encoding functions $\mathbf{f} = (f_1, \ldots, f_n)$, where $f_i : \mathcal{U} \to \mathbb{R}^s$ is used by sender i, and an estimator function $\hat{\theta} : \mathbb{R}^s \to \Theta$ used by the receiver.

We are now in a position to elaborate on the average power constraint in (2). The distribution of X_i depends (via f) on p_θ, which is not known in advance. We therefore require that schemes respect this power constraint for every $\theta \in \Theta$, that is, that the encoding functions $\{f_i\}$ satisfy

$$\frac{1}{s} \mathbb{E}_\theta \left[\|f_i(U_i)\|_2^2 \right] \le P, \quad \text{for all } i \in \{1, \ldots, n\}, \theta \in \Theta. \tag{3}$$

To evaluate possible schemes, we study risk under squared error loss, with the goal of minimizing the squared error $\mathbb{E}_\theta \|\hat{\theta}(Y) - \theta\|_2^2$. If we fix the encoding functions $\mathbf{f}$, all that remains is to choose an estimator function $\hat{\theta}$. We can understand these estimators using the same frameworks as in classical statistics; the difference is that our estimator can access only Y, not the samples $\{U_i\}$. In particular, when $\mathbf{f}$ is fixed, we will call an estimator *minimax* if it minimizes the worst-case risk (over $\theta \in \Theta$).

In our context, it is natural to extend this idea to schemes. When referring to the risk of a scheme $(\mathbf{f}, \hat{\theta})$, we mean the risk when that scheme is used. To remind ourselves that this also depends on the encoding functions $\mathbf{f}$, we write the risk as $R(\theta; \mathbf{f}, \hat{\theta}) = \mathbb{E}_\theta \|\hat{\theta}(Y) - \theta\|_2^2$, with $\mathbf{f}$ being implicit on the right-hand side. We can then extend minimaxity to schemes.

Definition 2. Consider a class $\mathcal{S}$ of estimation schemes for s channel uses. A scheme $(\mathbf{f}_M, \hat{\theta}_M)$ is *minimax* for $\mathcal{S}$ if it minimizes the maximum risk among all those schemes in $\mathcal{S}$

that also satisfy the power constraint (3). That is, if $\mathcal{S}_P$ is the subset of $\mathcal{S}$ satisfying (3), then a scheme $(\mathbf{f}, \hat{\theta})$ is minimax if it satisfies

$$\inf_{(\mathbf{f},\hat{\theta}) \in \mathcal{S}_P} \sup_\theta R(\theta; \mathbf{f}, \hat{\theta}) = \sup_\theta R(\theta; \mathbf{f}_{\mathrm{M}}, \hat{\theta}_{\mathrm{M}}). \tag{4}$$

Where a scheme's encoding functions are the same for all nodes, $f_i = f$ for all $i = 1, \ldots, n$, we will abuse notation by writing the common encoding function f in place of the collection $\mathbf{f}$, for example, $R(\theta; f, \hat{\theta}) \triangleq R(\theta; \mathbf{f}, \hat{\theta})$.

The main achievability results in Section III and the lower bound corollaries in Section V are concerned with two cases of the general problem. The first is the **Gaussian location** model, in which $p_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$, with $\mathcal{U} = \mathbb{R}^d$ and $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\sqrt{d}\}$ for some known $B > 0$. The goal of the receiver is to estimate the mean θ of the multivariate Gaussian distribution with known covariance matrix $\sigma^2 I_d$.

The second is the **product Bernoulli parameter** model, in which $p_\theta = \prod_{j=1}^{d} \mathrm{Bernoulli}(\theta_j)$, with $\mathcal{U} = \{0, 1\}^d$ and $\Theta = [0, 1]^d$. The goal of the receiver is to estimate the mean θ of the Bernoulli distribution.

We note that the gradient aggregation problem in distributed stochastic gradient descent, a key part of federated machine learning, can be cast as a distributed parameter estimation problem of this type; see *e.g.* [39].

III. RESULTS FOR ANALOG ACHIEVABILITY

We develop linear estimation schemes for the Gaussian and Bernoulli mean estimation tasks described in Section II. A defining feature of these schemes is their *analog* nature: they simply scale and transmit their samples to the central server in an uncoded fashion. This is in contrast to a digital approach where samples are encoded with a finite number of bits, which are then reliably communicated to the server using channel coding techniques. This analog approach allows us to exploit the additive nature of the Gaussian MAC to average the statistical samples over the air. The following theorems characterize the risk of these analog schemes.

Theorem 1. *In the Gaussian location model, consider the class of all estimation schemes for d channel uses, and using a scale-and-offset encoding function common to all senders $f(u) = \alpha u + \beta$ for some $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d$ (and any estimator function). The minimax scheme is given by the choice*

$$f_{\mathrm{M}}(u) = \sqrt{\frac{P}{B^2 + \sigma^2}} u, \quad \hat{\theta}_{\mathrm{M}}(Y) = \frac{1}{n}\sqrt{\frac{B^2 + \sigma^2}{P}} Y, \tag{5}$$

and yields the minimax risk

$$\sup_\theta R(\theta; f_{\mathrm{M}}, \hat{\theta}_{\mathrm{M}}) = \frac{d\sigma^2}{n}\left[1 + \frac{\sigma_{\mathrm{n}}^2}{nP}\left(1 + \frac{B^2}{\sigma^2}\right)\right]. \tag{6}$$

By using a repetition code, Theorem 1 can be extended to cases where $s > d$.

Corollary 1. *In the Gaussian location model, if $s \geq d$, there exists a scheme $(f_{\mathrm{R}}, \hat{\theta}_{\mathrm{R}})$ achieving the worst-case risk*

$$\sup_\theta R(\theta; f_{\mathrm{R}}, \hat{\theta}_{\mathrm{R}}) = \frac{d\sigma^2}{n}\left[1 + \frac{\sigma_{\mathrm{n}}^2}{\lfloor s/d \rfloor nP}\left(1 + \frac{B^2}{\sigma^2}\right)\right]. \tag{7}$$

This scheme involves repeating the encoding function (5) $\lfloor s/d \rfloor$ times, leaving the remaining $s - d\lfloor s/d \rfloor$ channel uses unused, and averaging the corresponding repeated estimates.

The proofs of Theorem 1 and Corollary 1 are in Section VI-A.

Where s/d is not an integer, the unused channel uses could be filled with another partial repetition, giving a slight improvement on (7) but a more unwieldy expression.

For the product Bernoulli parameter model, we provide the minimax scheme among those using affine estimators.

Theorem 2. *In the product Bernoulli parameter model, consider the class of all estimation schemes for d channel uses ($s = d$), and using affine estimators. The minimax scheme in this class is the one using the encoding function defined per element*

$$[f_{\mathrm{M}}(u)]_t = \begin{cases} -\sqrt{P}, & \text{if } [u]_t = 0 \\ \sqrt{P}, & \text{if } [u]_t = 1, \end{cases} \tag{8}$$

where $[\cdot]_t$ is the tth element of its (vector) argument, and the estimator function $\hat{\theta}_{\mathrm{M}}(Y) = \alpha_{\mathrm{M}} Y + \beta_{\mathrm{M}} \mathbf{1}$, where $\beta_{\mathrm{M}} = \frac{1}{2}$ and

$$\alpha_{\mathrm{M}} = \begin{cases} \dfrac{1}{2\sqrt{nP}(\sqrt{n}+1)}, & \text{if } \sigma_{\mathrm{n}}^2 \leq n^{3/2}P, \\[3mm] \dfrac{n\sqrt{P}}{2(\sigma_{\mathrm{n}}^2 + n^2 P)}, & \text{if } \sigma_{\mathrm{n}}^2 \geq n^{3/2}P. \end{cases} \tag{9}$$

The minimax risk given by this choice of f_{M} and $(\alpha_{\mathrm{M}}, \beta_{\mathrm{M}})$ is

$$\sup_\theta R(\theta; f_{\mathrm{M}}, \hat{\theta}_{\mathrm{M}}) =$$

$$\begin{cases} \dfrac{d}{4(\sqrt{n}+1)^2}\left(1 + \dfrac{\sigma_{\mathrm{n}}^2}{nP}\right), & \text{if } \sigma_{\mathrm{n}}^2 \leq n^{3/2}P, \\[3mm] \dfrac{d}{4} \cdot \dfrac{1}{1 + n \cdot \frac{nP}{\sigma_{\mathrm{n}}^2}}, & \text{if } \sigma_{\mathrm{n}}^2 \geq n^{3/2}P. \end{cases} \tag{10}$$

We can also similarly extend this using a repetition code.

Corollary 2. *In the product Bernoulli parameter model, if $s \geq d$, there exists a scheme $(f_{\mathrm{R}}, \hat{\theta}_{\mathrm{R}})$ achieving the risk*

$$\sup_\theta R(\theta; f_{\mathrm{R}}, \hat{\theta}_{\mathrm{R}}) =$$

$$\begin{cases} \dfrac{d}{4(\sqrt{n}+1)^2}\left(1 + \dfrac{\sigma_{\mathrm{n}}^2}{\lfloor s/d \rfloor nP}\right), & \text{if } \sigma_{\mathrm{n}}^2 \leq n^{3/2}P, \\[3mm] \dfrac{d}{4} \cdot \dfrac{1}{1 + n \cdot \frac{\lfloor s/d \rfloor nP}{\sigma_{\mathrm{n}}^2}}, & \text{if } \sigma_{\mathrm{n}}^2 \geq n^{3/2}P. \end{cases} \tag{11}$$

This scheme involves repeating the encoding function (8) $\lfloor s/d \rfloor$ times, leaving the remaining $s - d\lfloor s/d \rfloor$ channel uses unused, and averaging the corresponding repeated estimates.

The proofs of Theorem 2 and Corollary 2 are in Section VI-B.

IV. COMPARISON TO DIGITAL LOWER BOUNDS

In the previous section, we characterized the performance of analog estimation schemes for the Gaussian and Bernoulli

models. In this section, we compare their performance to digital approaches that have been studied in the recent literature, and show that analog schemes can lead to drastically smaller estimation error for the same amount of physical resources, *i.e.* transmission power and number of channel uses.

In particular, recent work in machine learning [1]–[6] has studied the impact of communication constraints on distributed parameter estimation. These works abstract out the physical layer, simply assuming a constraint on the number of bits available to represent each sample. This implicitly corresponds to assuming that communication is done in a digital fashion, with channel coding used to transmit the resultant bits without any errors. For example, in [6], the authors develop information-theoretic lower bounds on the minimax squared error risk over a parameter space $\Theta \subset \mathbb{R}^d$,

$$\sup_{\theta \in \Theta, \mathbf{f} \in \mathcal{F}_k^D} R(\theta; \mathbf{f}, \hat{\theta}) = \sup_{\theta \in \Theta, \mathbf{f} \in \mathcal{F}_k^D} \mathbb{E}_\theta \|\hat{\theta}(Y) - \theta\|_2^2,$$

where $\mathcal{F}_k^D$ now is defined as the set of all possible encoding schemes $\mathbf{f} \triangleq (f_1, \ldots, f_n)$, where $f_i(u) \in \{1, 2, \ldots, 2^k\}$ for all $i = 1, \ldots n$, *i.e.* each sample U_i is quantized to k bits, which are then noiselessly communicated to the receiver. Note that these information-theoretic results lower bound the minimax risk achieved by any k-bit digital estimation scheme. In this section, we compare our results to these lower bounds, as applied to the Gaussian MAC we study in this paper.

We assume that senders can transmit errorlessly at the Shannon capacity of the channel. The capacity region of a Gaussian multiple-access channel with n users, power P and channel noise σ_n^2 is given by the region of all $(R_1, \ldots, R_n)$ satisfying [40]

$$\sum_{i \in S} R_i < \frac{1}{2} \log_2 \left(1 + \frac{|S|\bar{P}}{\sigma_n^2}\right), \qquad \forall S \subseteq \{1, \ldots, n\}. \quad (12)$$

We allocate rates equally among all the senders, in which case the inequality in which S comprises all the senders dominates. If the MAC channel is utilized s times, we assume that each sender is able to noiselessly communicate

$$k = \sup_{(R_1, \ldots, R_n)} \frac{s}{n} \sum_{i=1}^n R_i = \frac{s}{2n} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) \qquad \text{bits} \quad (13)$$

to the receiver. Note that at finite block lengths, the senders cannot communicate to the receiver at the Shannon capacity and that this optimistic assumption benefits the performance of the digital schemes.

We can then substitute (13) into the aforementioned lower bound on minimax risk in [6]. However, that result (Corollary 5 therein) has an unspecified constant, inhibiting direct comparisons to our result from Corollary 1 above. Therefore, we instead take advantage of more recent work [7] to rederive the bound without the unspecified constant, and combine this with (13) to arrive at the following digital lower bound for the Gaussian location model.

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

$$\sup_{\|\theta\|_2 \leq B\sqrt{d}} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \geq \frac{d\sigma^2}{\frac{s}{d} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) + \frac{\pi^2 \sigma^2}{B^2}}. \quad (14)$$

Proof. Use Theorem 5 (see Section VII) to rederive Theorem 2 of [6] without the unspecified constant. Then follow the modifications to Corollary 1 of [6], then Corollary 5 of [6]. Finally, substitute (13) into this modified result. Note that $[-B, B]^d \subset \Theta \triangleq \{\theta : \|\theta\|_2 \leq B\sqrt{d}\}$, as required by their lower bound result. $\square$

Compare this to (7) from Corollary 1. Note that this proposition implies that as the number of nodes and therefore the number of samples n increases, the risk of any digital scheme decreases as $\Omega(d^2/s \log n)$, whereas the risk of the scheme from Corollary 1 scales with $O(d/n)$. This implies that, when $s \geq d$, the analog schemes can lead to an exponentially smaller estimation error, or equivalently require an exponentially smaller number of samples to achieve the same estimation accuracy, as compared to digital schemes employing the same physical resources.

On the other hand, we make a brief note on the case where $s < d$. Here, an analog scheme transmitting scaled versions of samples cannot easily communicate more coordinates than it has channel uses. A natural approach would be for each node to transmit only the (scaled) first s elements of U_i. In this case, the worst-case risk would scale as $\Theta(d)$, independent of s and n, which is the maximal risk achievable even in the absence of any samples. Thus, the digital scheme achieves risk better than $\Theta(d)$ whenever $s = \omega(d/\log(1 + \frac{nP}{\sigma_n^2}))$, which can be the case when the SNR or n is large, while our analog scheme requires $s \geq d$ to be viable.

We also have a similar situation for the Bernoulli model, derived in the same way.

Proposition 2. *In the product Bernoulli model, consider all schemes in which senders send bits to the receiver at the Shannon capacity for s channel uses. For $\frac{s}{n} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) < d$, the risk associated with any such scheme is at least*

$$\sup_{\theta \in [0,1]^d} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \geq \frac{d}{\frac{s}{d} \frac{1}{(\frac{1}{2} - 2\varepsilon^2)^2} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) + \frac{\pi^2}{\varepsilon^2}} \quad (15)$$

Proof. Apply a similarly modified version of Corollary 8 from [6] (via Corollary 4 of [6] and Theorem 2 of [7]), using (13). $\square$

Note that analogously to the Gaussian case, this result implies that the risk of any digital scheme for the Bernoulli model scales as $\Omega(d^2/s \log n)$, while the risk of our analog scheme from Corollary 2 decreases as $O(d/n)$. In the low SNR case, $\sigma_n^2 \geq n^{3/2} P$, the analog scheme appears to achieve $O(d^2/sn^2)$ (compared to $\Omega(d/n)$), but since increasing n also increases the SNR given by nP/σ_n^2, this relationship will eventually give way to the high SNR regime, where $\sigma_n^2 \leq n^{3/2} P$. These results show that building on the inherent summation of transmitted signals in the Gaussian MAC to perform the averaging that classical statistical estimators would do can

TABLE I
COMPARISON OF RESULTS

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

provide drastic gains in estimation performance. Similar gains have been observed in asymptotic lossy source coding for Gaussian sensor networks in [8], [9], where one is interested in communicating an i.i.d. Gaussian source over a MAC under mean-squared error distortion, as well as for distributed stochastic gradient descent in experimental comparisons in [10].

We summarize all these results, as well as the new lower bounds we present in Section V below, in Table I.

V. RESULTS FOR ANALOG LOWER BOUNDS

So far in this paper, we have shown that the analog joint communication-estimation schemes in Section III can exponentially outperform lower bounds for digital schemes with a physical layer abstraction. Having shown this striking advantage, the next question this naturally raises is what lower bounds exist for analog schemes. That is: if these analog schemes can achieve what digital schemes could never hope for, how close are we to the fundamental limits for such analog approaches?

In this section, we present new lower bounds on worst-case squared error risk for any estimation scheme in our setting of interest where the parametric model has a sub-Gaussian score function. This work draws on the findings of [7], for which we first introduce some relevant quantities. We then apply them to the same models for which we presented achievability results in Section III. We find that these lower bounds are within a logarithmic factor of the risk achieved by the schemes presented above.

A. Preliminaries

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,

$$S_\theta(u) \triangleq \nabla_\theta \log p_\theta(u)$$
$$= \left(\frac{\partial}{\partial\theta_1}\log p_\theta(u),\ldots,\frac{\partial}{\partial\theta_d}\log p_\theta(u)\right).$$

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

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

Note that both $S_\theta(u)$ and $S_\theta(u_1,\ldots,u_n)$ 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

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

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 $X_1,\ldots,X_n$ are independent and sub-Gaussian with parameters $\rho_1,\ldots,\rho_n$, 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 [7], which we recite here:
(i) $\sqrt{p_\theta(u_1,\ldots,u_n)}$ is continuously differentiable with respect to each component θ_j at almost all $(u_1,\ldots,u_n) \in \mathcal{U}^n$ (with respect to some measure dominating $\{p_\theta : \theta \in \Theta\}$).
(ii) The Fisher information for each component θ_j, $\mathbb{E}([\frac{\partial}{\partial\theta_j}\log p_\theta(U_1,\ldots,U_n)]^2)$, exists and is a continuous function of θ_j.
(iii) The conditional density $p(y|u_1,\ldots,u_n)$ is square integrable in the sense that for almost all $y \in \mathbb{R}^s$ for each θ, $\int p(y|u_1,\ldots,u_n)^2 \, dp_\theta(u_1,\ldots,u_n) < \infty$.

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 $p(y|x_1,\ldots,x_n)$ of the Gaussian MAC; details are in the relevant proofs.

B. Main lower bound on worst-case risk

We have now covered the background necessary to state our main lower bound, which is 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 3. *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*

$$\sup_{\theta \in \Theta} R(\theta; \mathbf{f}, \hat{\theta}) \geq \frac{d}{n} \cdot \frac{1}{\frac{s}{d} \rho^2 \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) + \frac{\pi^2}{nB^2}}. \quad (16)$$

The proof for this builds on a relationship between Fisher information and Shannon information established by [7], and a Bayesian Cramer-Rao type bound to relate minimax risk and Fisher information. We provide the proof in Section VII-B.

C. Bounds for specific models

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

Corollary 3. *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*

$$\sup_{\theta \in \Theta} R(\theta; \mathbf{f}, \hat{\theta}) \geq \frac{d\sigma^2}{n} \cdot \frac{1}{\frac{s}{d} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) + \frac{\sigma^2}{B^2} \frac{\pi^2}{n}}. \quad (17)$$

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 4. *Consider the relatively dense product Bernoulli model, where $U_1, \ldots, U_n \sim \prod_{j=1}^d \text{Bernoulli}(\theta_j)$, 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*

$$\sup_{\theta \in \Theta} R(\theta; \mathbf{f}, \hat{\theta}) \geq \frac{d}{n} \cdot \frac{1}{\frac{s}{d} \frac{1}{(\frac{1}{2} - 2\varepsilon^2)^2} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) + \frac{\pi^2}{n\varepsilon^2}}. \quad (18)$$

The proofs of the above two corollaries, which both follow from Theorem 3, are in Section VII-C.

D. Lower bound for a general multiple-access channel

The result of Theorem 3 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,

$$C_{\text{total}} = \max_{\prod_i p_i(x_{it})} I(X_{1t}, \ldots, X_{nt}; Y_t), \quad (19)$$

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

Theorem 4. *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 $p(y | x_1, \ldots, x_n)$ is bounded. The worst-case risk of any estimation scheme $(\mathbf{f}, \hat{\theta})$ must satisfy*

$$\sup_{\theta \in \Theta} R(\theta; \mathbf{f}, \hat{\theta}) \geq \frac{d}{n} \cdot \frac{1}{2 \frac{s}{d} \rho^2 C_{\text{total}} + \frac{\pi^2}{nB^2}}. \quad (20)$$

Proof. Follow the proof of Theorem 3, but replace the right-hand side of (40) with sC_{total} (*i.e.*, C_{total} for s channel uses). Note that the stipulation that $p(y | x_1, \ldots, x_n)$ be bounded (by some different finite M) ensures that (37), and hence (iii), is satisfied. □

VI. PROOFS OF ANALOG ACHIEVABILITY

A. Gaussian location model

In this section, we prove our main results for the Gaussian location model, Theorem 1 and Corollary 1. In this model, the samples $U_i \sim \mathcal{N}(\theta, \sigma^2 I_d)$, where θ lies in an ℓ_2-ball in a d-dimensional space, $\{\|\theta\|_2 \leq B\sqrt{d}\}$, and the goal is to estimate θ.

Since the multiple-access channel already produces a sum, one might suspect that that an estimation scheme emulating the sample mean would be a natural candidate, given its properties in classical estimation. Indeed, it is minimax among schemes using affine encoders (and any estimator). First, we show that this estimator is minimax for a fixed affine encoder, as we state formally in the following proposition.

Proposition 3. *In the Gaussian location model, let the senders use any scale-and-offset encoding function $f(u) = \alpha u + \beta$ for some $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d$, common to all senders, and assume that this encoding function satisfies the power constraint, and that the channel is used d times (i.e., $s = d$). Then the minimax estimator is given by*

$$\hat{\theta}_M(Y) = \frac{1}{\alpha n} Y - \frac{1}{\alpha} \beta,$$

which yields risk

$$\mathbb{E}_\theta \|\hat{\theta}_M(Y) - \theta\|_2^2 = \frac{d}{n} \left(\sigma^2 + \frac{\sigma_n^2}{n\alpha^2}\right). \quad (21)$$

Remark. The estimator given by Proposition 3 is also the maximum likelihood estimator.

The proof for this follows similar lines to the classical result using a least favorable sequence of priors, with modifications for channel noise.

Lemma 1. *If θ is distributed according to the prior $\mathcal{N}(\mu, b^2 I_d)$, and all senders use the common encoding function $f(u) = \alpha u + \beta$ for some $\alpha \in \mathbb{R}, \beta \in \mathbb{R}^d$, then the Bayes estimator $\hat{\theta}_{\mu, b^2}(y)$ is given by*

$$\hat{\theta}_{\mu, b^2}(Y) = \mu + \frac{\alpha n b^2}{\alpha^2 n^2 b^2 + \alpha^2 n \sigma^2 + \sigma_n^2} (Y - \alpha n \mu + n\beta), \quad (22)$$

and the Bayes risk is

$$\mathbb{E}\|\hat{\theta}_{\mu,b^2}(Y) - \theta\|^2 = \frac{d(\alpha^2 n\sigma^2 + \sigma_n^2)}{\alpha^2 n^2 + \frac{\alpha^2 n\sigma^2 + \sigma_n^2}{b^2}}. \qquad (23)$$

Proof. Under squared error loss, the Bayes estimator for $\mathcal{N}(\mu, b^2 I_d)$ is (by well-known theorem, *e.g.* [41, Cor. 4.1.2(a)]) $\hat{\theta}_{\mu,b^2}(y) = \mathbb{E}(\theta|y)$, which we will evaluate. The relevant covariance matrices are

$$\Sigma_Y = (\alpha^2 n^2 b^2 + \alpha^2 n\sigma^2 + \sigma_n^2) I_d,$$
$$\Sigma_{Y\theta} = \mathbb{E}\left[(\alpha n W + \alpha \sum_i V_i) W^\top\right] = \alpha n b^2 I_d.$$

Then the Bayes estimate is given by

$$\hat{\theta}_{\mu,b^2}(Y) = \mathbb{E}(\theta|Y) = \mathbb{E}\theta + \Sigma_{\theta Y}\Sigma_Y^{-1}(Y - \mathbb{E}Y)$$
$$= \mu + \alpha n b^2 \cdot \frac{1}{\alpha^2 n^2 b^2 + \alpha^2 n\sigma^2 + \sigma_n^2} \cdot (Y - \alpha n\mu - n\beta),$$

and since this estimator is unbiased, the squared error is given by the trace of the conditional variance,

$$\mathbb{E}\|\hat{\theta}_{\mu,b^2} - \theta\|^2 = \mathbf{tr}\,\mathbf{var}(\theta|Y) = \mathbf{tr}(\Sigma_\theta - \Sigma_{\theta Y}\Sigma_Y^{-1}\Sigma_{Y\theta})$$
$$= db^2 - \frac{d(\alpha n b^2)^2}{\alpha^2 n^2 b^2 + \alpha^2 n\sigma^2 + \sigma_n^2}. \qquad \square$$

Proof of Proposition 3. Take the Bayes estimator from Lemma 1. Let $b^2 \to \infty$, then we have a sequence of priors $\mathcal{N}(\mu, b^2)$ yielding increasing Bayes risk converging to

$$\lim_{b\to\infty}\mathbb{E}\|\hat{\theta}_{\mu,b^2} - \theta\|^2 = \frac{d(\alpha^2 n\sigma^2 + \sigma_n^2)}{\alpha^2 n^2} = \frac{d}{n}\left(\sigma^2 + \frac{\sigma_n^2}{n\alpha^2}\right).$$

The minimax estimator is then

$$\lim_{b\to\infty}\hat{\theta}_{\mu,b^2}(Y) = \mu + \frac{1}{\alpha n}Y - \mu - \frac{1}{\alpha}\beta. \qquad \square$$

In the absence of a power constraint, the offset β has no effect—since it is known, it is easily cancelled by the receiver's estimator. Intuitively, with a power constraint, one would expect no offset to be preferable. In Theorem 1, where we find the best choice of (α, β), we find that this is indeed the case.

Proof of Theorem 1. For any given α, β, the minimax risk from Proposition 3 is decreasing in α. Therefore, we choose the largest α satisfying the power constraint (3). Note that

$$\mathbb{E}_\theta\left[\|X_i\|_2^2\right] = \alpha^2(\|\theta\|^2 + d\sigma^2) + 2\alpha\theta^\top\beta + \|\beta\|^2$$
$$= \|\alpha\theta + \beta\|_2^2 + \alpha^2 d\sigma^2. \qquad (24)$$

We thus solve

maximize $\quad \alpha$
subject to $\quad \|\alpha\theta + \beta\|_2^2 + \alpha^2 d\sigma^2 \le dP \quad \forall \theta : \|\theta\| \le \sqrt{d}B.$
$$(25)$$

If we relax the constraint to $\|\beta \pm \alpha B\mathbf{1}\|_2^2 + \alpha^2 d\sigma^2 \le dP$, we can use Lagrange multipliers to find the solution

$$\alpha = \sqrt{\frac{P}{B^2 + \sigma^2}}, \quad \beta = 0, \qquad (26)$$

and verify that it also satisfies the constraints of, and is therefore also a solution to, (25). $\qquad \square$

We now turn to the case where $s > d$. A natural extension of the scheme from Theorem 1 would be to transmit repetitions of the sample.

Lemma 2. *Let* $(\mathbf{f}, \hat{\theta})$ *be a scheme with* $\hat{\theta}(Y)$ *affine in Y and consider a scheme* $(\mathbf{f}_R, \hat{\theta}_R)$ *that repeats the encoding function m times and averages the estimates for each repetition,* $\hat{\theta}_R(Y) = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}([Y]_j)$, *where* $[Y]_j$ *is the part of Y corresponding to the jth repetition. The risk of* $(\mathbf{f}_R, \hat{\theta}_R)$ *is the same as for* $(\mathbf{f}, \hat{\theta})$, *but with* σ_n^2/m *in place* σ_n^2.

Proof. The bias of the estimator is unaffected by the repetition (and is independent of σ_n^2), and if the original estimator is written as $\hat{\theta}(Y) = AY + c$, the variance can be shown to be $\sum_i \mathbf{var}(AX_i) + \frac{1}{m}\mathbf{var}(AZ)$. Relative to the original estimator variance, this is equivalent to dividing $\sigma_n^2 I$ by m. $\qquad \square$

This then yields the achievability result of Corollary 1.

Proof of Corollary 1. Apply Lemma 2 to Theorem 1, with $m = \lfloor s/d \rfloor$ and ignoring the leftover channel uses. $\qquad \square$

Comparing this to the $s = d$ case, the repetition reduces the noise by a factor of roughly s/d, which is the expected effect of averaging a repeated transmission. The minimax risk then converges more quickly to the noiseless case as $s/d \to \infty$.

B. Product Bernoulli model

In this section, we prove our main results for the Bernoulli parameter model, Theorem 2 and Corollary 2. In this model, $U_i \sim \prod_{i=1}^{d} \text{Bernoulli}(\theta)$, and the goal is to estimate θ, which is in $[0, 1]^d$. Our calculations in this section will work with the parameterized encoding function common to all senders

$$f_C(u) = \begin{cases} -C, & \text{if } u = 0 \\ C, & \text{if } u = 1. \end{cases} \qquad (27)$$

Our analysis of the Bernoulli parameter model focuses on the scalar case, as stated in Proposition 4 below. Theorem 2 will then follow by extension to independent dimensions.

Proposition 4. *In the scalar Bernoulli parameter model ($d = 1$), consider the class of all estimation schemes using affine estimators* $\hat{\theta}_{\alpha,\beta}(Y) = \alpha Y + \beta, \alpha, \beta \in \mathbb{R}$ *(and any scalar encoding function with $s = 1$). The minimax scheme in this class is the one using the encoding function*

$$f_M(u) = \begin{cases} -\sqrt{P}, & \text{if } u = 0 \\ \sqrt{P}, & \text{if } u = 1, \end{cases} \qquad (28)$$

and the estimator $\hat{\theta}_M(Y) = \alpha_M Y + \beta_M$, *where* $\beta_M = \frac{1}{2}$ *and* α_M *is as provided in* (9). *The minimax risk given by this choice of* (α_M, β_M) *is*

$$\sup_\theta R(\theta; f_M, \hat{\theta}_M) = \begin{cases} \frac{1}{4(\sqrt{n}+1)^2}\left(1 + \frac{\sigma_n^2}{nP}\right), & \text{if } \sigma_n^2 \le n^{3/2}P, \\ \frac{1}{4}\cdot\frac{1}{1+n\cdot nP/\sigma_n^2}, & \text{if } \sigma_n^2 \ge n^{3/2}P. \end{cases} \qquad (29)$$

Our steps for proving Proposition 4 will be first to establish the minimax risk for the common encoding function f_C, then to show that a scheme using any other encoding function can

be transformed to one using f_C for some C of equal risk, and finally to show that the optimal value for C is $\sqrt{P}$. Before we continue, we compute the risk for a general affine estimator.

Lemma 3. *In the scalar Bernoulli parameter model ($d = 1$), if all senders use the encoding function f_C (27), and the receiver uses the affine estimator $\hat{\theta}_{\alpha,\beta}(Y) = \alpha Y + \beta$, then the risk is*

$$R(\theta; f_C, \hat{\theta}_{\alpha,\beta}) = \alpha^2 \left[4nC^2\theta(1-\theta) + \sigma_n^2\right] + \left[\alpha nC(2\theta - 1) + \beta - \theta\right]^2. \quad (30)$$

Proof. Recall that $Y = \sum_{i=1}^n f_C(U_i) + Z$ and that $f_C(U_i) = C$ w.p. θ and $f_C(U_i) = -C$ w.p. $1 - \theta$. The variance and bias of the estimator are then

$$\mathbf{var}_\theta[\hat{\theta}_{\alpha,\beta}(Y)] = \alpha^2 \mathbf{var}(Y) = \alpha^2 \left[4nC^2\theta(1-\theta) + \sigma_n^2\right].$$
$$\mathbf{bias}_\theta[\hat{\theta}_{\alpha,\beta}(Y)] \triangleq \mathbb{E}_\theta \hat{\theta}_{\alpha,\beta}(Y) - \theta = \alpha nC(2\theta - 1) + \beta - \theta.$$

The result then follows from combining these as

$$\mathbb{E}_\theta[\hat{\theta}_{\alpha,\beta}(Y) - \theta]_2^2 = \mathbf{var}_\theta[\hat{\theta}_{\alpha,\beta}(Y)] + (\mathbf{bias}_\theta[\hat{\theta}_{\alpha,\beta}(Y)])^2. \quad \square$$

The bulk of the work in proving Proposition 4 is in showing Proposition 5, which establishes the minimax estimator for the encoding function f_C.

Proposition 5. *In the scalar Bernoulli parameter model, let all senders use the encoding function $f_C(u)$ from (27), and consider the class of all affine estimators $\hat{\Theta}_{\text{aff}} = \{\hat{\theta}_{\alpha,\beta}(Y) = \alpha Y + \beta, \alpha, \beta \in \mathbb{R}\}$. The minimax affine estimator is given by $\hat{\theta}_M(Y) = \alpha_M Y + \beta_M$, where $\beta_M = \frac{1}{2}$ and*

$$\alpha_M = \begin{cases} \dfrac{1}{2\sqrt{n}C(\sqrt{n}+1)}, & \text{if } \sigma_n^2 \le n^{3/2}C^2, \\[2mm] \dfrac{nC}{2(\sigma_n^2 + n^2C^2)}, & \text{if } \sigma_n^2 \ge n^{3/2}C^2. \end{cases} \quad (31)$$

The minimax risk given by this choice of (α_M, β_M) is

$$\sup_\theta R(\theta; f_M, \hat{\theta}_M) = \begin{cases} \dfrac{1}{4(\sqrt{n}+1)^2}\left(1 + \dfrac{\sigma_n^2}{nC^2}\right), & \text{if } \sigma_n^2 \le n^{3/2}C^2, \\[2mm] \dfrac{1}{4} \cdot \dfrac{1}{1 + n \cdot nC^2/\sigma_n^2}, & \text{if } \sigma_n^2 \ge n^{3/2}C^2. \end{cases} \quad (32)$$

Proof. Define $\alpha_{\text{lo}} = \frac{1}{2\sqrt{n}C(\sqrt{n}+1)}$ and $\alpha_{\text{hi}} = \frac{nC}{2(\sigma_n^2 + n^2C^2)}$. Note that then $\alpha_M = \min\{\alpha_{\text{lo}}, \alpha_{\text{hi}}\}$. For convenience, and with some abuse of notation, let $R(\theta; f_C, \alpha, \beta)$ refer to the expression in (30). We will repeatedly use the facts that:

(a) $R(\theta; f_C, \alpha_{\text{lo}}, \beta_M)$ is constant with respect to θ.
(b) $R(\theta; f_C, \alpha_{\text{hi}}, \beta_M)$ is convex in θ and minimized at $\theta \in \{0, 1\}$, at which the risk is equal.
(c) $R(0; f_C, \alpha, \beta_M)$ is convex in α and minimized at $\alpha = \alpha_{\text{hi}}$.

These can all be verified by appropriate substitutions into (30). Where we invoke these facts, we will label the equality or inequality signs accordingly.

We will show that for every other choice (α, β), there exists some $\theta \in [0, 1]$ with risk exceeding $\sup_\theta R(\theta; \alpha_M, \beta_M)$. We divide into three cases.

Case 1: $\alpha > \alpha_{\text{lo}}$, or $\alpha = \alpha_{\text{lo}}$ and $\beta \ne \frac{1}{2}$. In this case, take $\theta = \frac{1}{2}$ and we have

$$R(\tfrac{1}{2}; f_C, \alpha, \beta) = \alpha^2(nC^2 + \sigma_n^2) + (\beta - \tfrac{1}{2})^2$$
$$> \alpha_{\text{lo}}^2(nC^2 + \sigma_n^2) = R(\tfrac{1}{2}; f_C, \alpha_{\text{lo}}, \beta_M).$$

Then, if $\alpha_M = \alpha_{\text{lo}}$, then by (a), the right-hand side is equal to $\sup_\theta R(\theta; f_C, \alpha_M, \beta_M)$. If $\alpha_M = \alpha_{\text{hi}}$, then note that

$$R(\tfrac{1}{2}; f_C, \alpha_{\text{lo}}, \beta_M) \overset{\text{(a)}}{=} R(0; f_C, \alpha_{\text{lo}}, \beta_M)$$
$$\overset{\text{(c)}}{\ge} R(0; f_C, \alpha_{\text{hi}}, \beta_M) \overset{\text{(b)}}{=} \sup_\theta R(\theta; f_C, \alpha_M, \beta_M),$$

where labeled steps refer to corresponding facts above.

Case 2: $\alpha < \alpha_{\text{lo}}$ and $\beta \ge \frac{1}{2}$. Take $\theta = 0$ and we have

$$R(0; f_C, \alpha, \beta) = \alpha^2\sigma_n^2 + (\beta - \alpha nC)^2$$
$$\ge \alpha^2\sigma_n^2 + (\tfrac{1}{2} - \alpha nC)^2 = R(0; f_C, \alpha, \beta_M),$$

where in the inequality we used the fact that $\alpha nC < \alpha_{\text{lo}}nC = \frac{\sqrt{n}}{2(\sqrt{n}+1)} < \frac{1}{2}$. Then, if $\alpha_M = \alpha_{\text{lo}}$, we also have $\alpha_{\text{lo}} < \alpha_{\text{hi}}$, and by fact (c), is strictly decreasing in α for all $\alpha < \alpha_{\text{lo}}$, thus

$$R(0; f_C, \alpha, \beta_M) > R(0; f_C, \alpha_{\text{lo}}, \beta_M) \overset{\text{(a)}}{=} \sup_\theta R(\theta; f_C, \alpha_M, \beta_M).$$

If $\alpha_M = \alpha_{\text{hi}}$, then we have

$$R(0; f_C, \alpha, \beta_M) \overset{\text{(c)}}{\ge} R(0; f_C, \alpha_{\text{hi}}, \beta_M) \overset{\text{(b)}}{=} \sup_\theta R(\theta; f_C, \alpha_M, \beta_M).$$

Case 3: $\alpha < \alpha_{\text{lo}}$ and $\beta \le \frac{1}{2}$. Take $\theta = 1$ and argue similarly to case 2 that $R(1; f_C, \alpha, \beta_M) > \sup_\theta R(\theta; f_C, \alpha_M, \beta_M)$. $\square$

Lemma 4. *In the scalar Bernoulli parameter model, consider the scheme $(f, \hat{\theta})$, in which all senders use the encoding function $f(0) = A$, $f(1) = B$, and the receiver uses the estimator $\hat{\theta}$. Then there exists a scheme $(f', \hat{\theta}')$ satisfying $f'(0) = -f'(1)$ and with minimax risk equal to that of $(f, \hat{\theta})$.*

Proof. Choose $C = \frac{B-A}{2}$, so that $f'(u) \triangleq f_C(u) = f(u) - \frac{A+B}{2}$. By construction, $f'(0) = -f'(1) = \frac{A-B}{2}$. Then, if Y and Y' are what the receiver observes under f and f' respectively, we have $Y' = \sum_i f'(U_i) + Z = \sum_i [f(U_i) - \frac{A+B}{2}] + Z = Y - n\frac{A+B}{2}$. We can then define $\hat{\theta}'(Y') \triangleq \hat{\theta}(Y' + n\frac{A+B}{2})$, and this will have exactly the same statistical properties as $\hat{\theta}(Y)$. $\square$

Now we may complete the proof of Proposition 4.

Proof of Proposition 4. Because Lemma 4 shows there is no sacrifice in minimax risk, it suffices to consider just schemes using encoding functions of the form f_C in (27). The minimax affine estimator for such encoding functions is found in Proposition 5. From (32), the minimax risk for f_C is strictly decreasing in C^2. Therefore, to minimize over all encoding functions f_C, we take the highest-magnitude C satisfying the power constraint (3), $C = \pm\sqrt{P}$. $\square$

The extension of this result to the product Bernoulli model is then an application of the scalar case on a per-sample basis.

Proof of Theorem 2. Because each dimension $1, \ldots, d$ is independent, each dimension can be optimized separately. Each sender transmits its jth sample $[f_C(U_i)]_j$ using the scheme from Proposition 4. The even division of power still satisfies the average power constraint (3). The minimax risk is then d times the minimax risk along one dimension. $\square$

Finally, Corollary 2 follows using Lemma 2 again.

Proof of Corollary 2. Apply Lemma 2 to Theorem 2, with $m = \lfloor s/d \rfloor$ and ignoring the leftover channel uses. $\square$

VII. PROOFS OF ANALOG LOWER BOUNDS

A. Preliminaries

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 Shannon information. First, Equation 8 of [6] tells us the following, which we list as a lemma here.

Lemma 5. *Suppose* $[-B, B]^d \subset \Theta$. *For any estimator* $\hat{\theta}(Y_1, \ldots, Y_n)$, *the worst-case squared error risk must satisfy*

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 \geq \frac{d^2}{\sup_{\theta \in \Theta} \mathbf{tr}(I_Y(\theta)) + \frac{d\pi^2}{B^2}}. \tag{33}$$

We will also lean on the following theorem, due to [7].

Theorem 5. *Suppose that* $\langle u, S_\theta(X) \rangle$ *is sub-Gaussian with parameter* ρ *for any unit vector* $u \in \mathbb{R}^d$. *Under regularity conditions (i)–(iii),* $\mathbf{tr}(I_Y(\theta)) \leq 2\rho^2 I_\theta(X; Y)$.

B. Proof of main lower bound

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

$$I(X_1, \ldots, X_n; Y) \leq \frac{s}{2} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right). \tag{34}$$

We now have all of the ingredients necessary to prove Theorem 3, which uses the data processing inequality to chain the above results together.

Proof of Theorem 3. Recall that $X_i = f(U_i), i = 1, \ldots, n$ and Y is the output of the channel $p_{Y|X}(y|x_1, \ldots, x_n)$ with inputs $X_1, \ldots, X_n$. The conditional distribution of Y given U can be expressed in terms of the channel's conditional distribution,

$$p_{Y|U}(y|u_1, \ldots, u_n) = p_{Y|X}(y|f(u_1), \ldots, f(u_n)). \tag{35}$$

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,

$$p_{Y|X}(y|x_1, \ldots, x_n) \leq \frac{1}{\sqrt{(2\pi\sigma_n^2)^n}} \triangleq M, \tag{36}$$

so chaining (35) and (36) verifies that

$$\int p_{Y|U}(y|u_1, \ldots, u_n)^2 \, dp_U(u_1, \ldots, u_n)$$
$$\leq \int M^2 \, dp_U(u_1, \ldots, u_n) = M^2 < \infty. \tag{37}$$

We therefore satisfy the requirements to invoke Theorem 5, so long as we can establish that $\langle v, S_\theta(U_1, \ldots, U_n) \rangle$ is sub-Gaussian for all unit vectors $v \in \mathbb{R}^d$. Note that since $U_1, \ldots, U_n$ are independent,

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

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

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

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 5 thus gives

$$\mathbf{tr}(I_Y(\theta)) \leq 2n\rho^2 I_\theta(U_1, \ldots, U_n; Y). \tag{38}$$

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

$$I_\theta(U_1, \ldots, U_n; Y) \leq I_\theta(X_1, \ldots, X_n; Y). \tag{39}$$

Now, $U_1, \ldots, U_n$ are independent (by definition), and each $X_i, i = 1, \ldots, n$ is a function of the corresponding U_i. Therefore, $X_1, \ldots, X_n$ are also independent, and from Proposition 6, we have

$$I(X_1, \ldots, X_n; Y) \leq \frac{s}{2} \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right). \tag{40}$$

Putting (38), (39) and (40) together yields

$$\mathbf{tr}(I_Y(\theta)) \leq n\rho^2 s \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right). \tag{41}$$

Substituting this expression into the result given by Lemma 5 then yields

$$\sup_{\theta \in \Theta} \mathbb{E}\|\hat{\theta}(Y) - \theta\|^2 \geq \frac{d^2}{n\rho^2 s \log_2 \left(1 + \frac{nP}{\sigma_n^2}\right) + \frac{d\pi^2}{B^2}}. \qquad \square$$

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 3. The score function for a single sample U_i is

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

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 $v^\top \frac{1}{\sigma^2} I_d v = \frac{1}{\sigma^2} v^\top v = \frac{1}{\sigma^2}$. This is sub-Gaussian with parameter $\frac{1}{\sigma}$, enabling an application of Theorem 3. $\square$
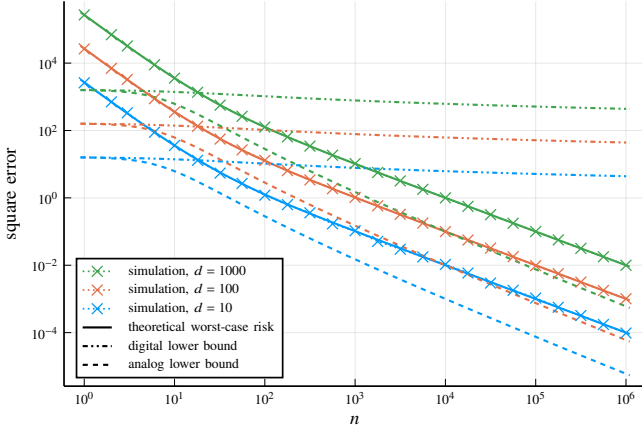
Fig. 2. Simulation of Gaussian location model with $\frac{P}{\sigma_n^2} = 0.1$, $\frac{B}{\sigma^2} = 0.4$

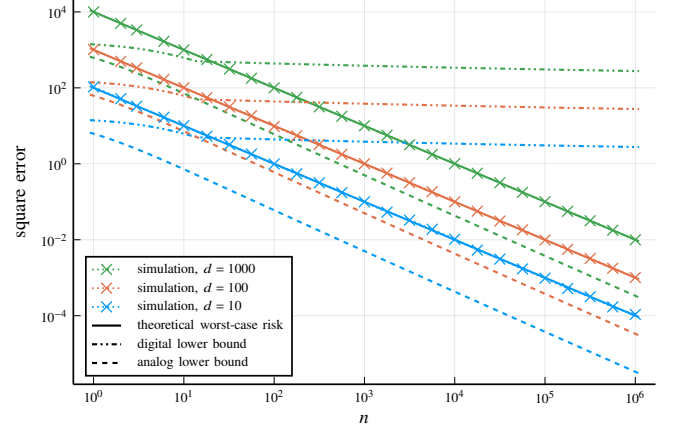
Fig. 3. Simulation of Gaussian location model with $\frac{P}{\sigma_n^2} = 1000$, $\frac{B}{\sigma^2} = 0.4$

Proof of Corollary 4. We can compute the score function of each sample in the product Bernoulli model to be $S_\theta(u_i) = (S_{\theta_1}(u_i), \ldots, S_{\theta_j}(u_i))$, 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

$$\sqrt{\sum_{j=1}^d v_j^2 \frac{1}{(\frac{1}{2} - 2\varepsilon^2)^2}} = \frac{1}{\frac{1}{2} - 2\varepsilon^2}. \tag{42}$$

This gives us a value for ρ to use in Theorem 3.

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 3 to arrive at Corollary 4. □

VIII. SIMULATIONS

We ran simulations to validate the schemes proposed in Section III for both of our models of interest, and we plot the results alongside each of our three theoretical results in Figs. 2–5. Lines in the same color relate to the same value of d, with different patterns corresponding to different results.

The points with '×' markers are the squared error of our proposed analog schemes, averaged over 100 simulations with θ drawn uniformly at random from the parameter space Θ, at various values of d and n, in the case where $s = d$. The solid lines are the theoretical results we presented in Section III, namely Theorems 1 and 2, which are the worst-case risk over parameters $\theta \in \Theta$ for the schemes we proposed (which are minimax among the classes of estimators noted in those theorems).
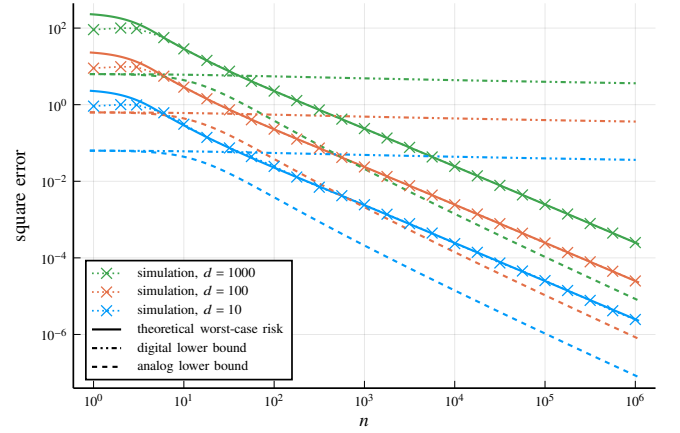

Fig. 4. Simulation of product Bernoulli model with $\frac{P}{\sigma_n^2} = 0.1$
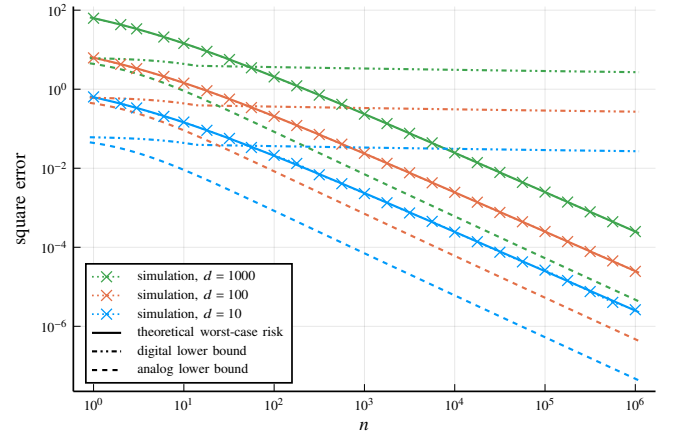

Fig. 5. Simulation of product Bernoulli model with $\frac{P}{\sigma_n^2} = 1000$

Since the theoretical results are the *worst-case* risk over the parameters $\theta \in \Theta$, we expect them to be at least as large as the *average* squared error found in simulations with randomly sampled $\theta \in \Theta$. It turns out that in most cases, the worst-case and average risk are equal, because the risk of the minimax scheme is constant with respect to θ. The one exception is the low-SNR regime ($\sigma_n^2 > n^{3/2}P$, see Theorem 2) of the product Bernoulli model estimator (Fig. 4), where the risk depends on θ, leading to the observed difference between the theoretical line and the average over randomly drawn θ at low n.

The digital lower bounds of Propositions 1 and 2 are plotted in dash-dot-dot lines. It is here that we see the marked improvement discussed in Section IV. Since the error in digital schemes scales at best with $1/\log n$, the simulated analog schemes rapidly become significantly advantageous even in moderately large values of n.

Finally, the analog lower bounds of Corollaries 3 and 4 are plotted in dashed lines, and run a $\log n$ factor from the achievability and simulation results.

Recall that in the product Bernoulli model, both lower bounds depend on ε, where $\Theta = [\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon]^d$. Our achievability results and simulations, on the other hand, are for the full parameter space $\Theta = [0, 1]^d$. Note, however, that the lower bounds for any $\varepsilon < \frac{1}{2}$, yielding $\Theta \subset [0, 1]^d$, also imply a lower bound for $[0, 1]^d$. To generate the plots in Figs. 4 and 5, we used $\varepsilon = \frac{1}{4}$.

IX. CONCLUSIONS

We introduced and studied a new model for minimax parameter estimation over the Gaussian multiple-access channel, developing estimation schemes for the Gaussian location model and product Bernoulli model. These "analog" estimation schemes directly leverage the superposition property of the Gaussian MAC, and our analysis of their risk under squared error loss showed that they exponentially outperform even lower bounds on the risk of "digital" schemes that separate the communication and estimation problems. We then derived new "analog" lower bounds for this estimation problem that are within a logarithmic factor of our achievability results. We confirmed our findings in simulations for both models. This adds theoretical insight to a growing body of literature on the advantages of analog schemes in over-the-air learning and inference, demonstrating that even fundamental limits of digital schemes can be beaten when estimation and communication are considered jointly.

This opens a number of further questions to be examined in future work. First, while the lower bounds we presented are general, our achievability results pertain specifically to the two estimation models we studied. With the additive nature of the Gaussian MAC, one would imagine that this extends to other mean estimation problems, but it raises the question of whether other estimation problems, such as distribution or empirical frequency estimation [6], [42], [43], can also harness this or other channels to achieve similarly startling gains over digital schemes.

Moreover, our estimation schemes work only when there is at least one channel use available for each parameter ($s \geq d$).

In many applications, models may have too many parameters for this to be feasible, so the setting where $s < d$ is also important to study. Developing analog schemes for this regime that outperform digital approaches remains as future work.

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] L. P. Barnes and A. Özgür, "Fisher information and mutual information constraints," 2021, arXiv:2102.05802.

[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] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Convergence of federated learning over a noisy downlink," 2020, arXiv:2008.11141.

[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] 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.

[15] ——, "Over-the-air federated learning from heterogeneous data," 2020, arXiv:2009.12787.

[16] 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.

[17] 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.

[18] J. Ahn, O. Simeone, and J. Kang, "Wireless federated distillation for distributed edge learning with heterogeneous data," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2019, pp. 1–6.

[19] 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.

[20] H. Guo, A. Liu, and V. K. N. Lau, "Analog gradient aggregation for federated learning over wireless networks: Customized design and convergence analysis," *IEEE Internet of Things Journal*, vol. 8, no. 1, pp. 197–210, 2021.

[21] 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.

[22] A. Abdi, Y. M. Saidutta, and F. Fekri, "Analog compression and communication for federated learning over wireless MAC," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.

[23] L. P. Barnes, W. N. Chen, and A. Özgür, "Fisher information under local differential privacy," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 645–659, 2020.

[24] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 170–185, 2021.

[25] X. Wu, S. Zhang, and A. Özgür, "Stac: Simultaneous transmitting and air computing in wireless data center networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 4024–4034, 2016.

[26] W. Liu, X. Zang, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimization, analysis and scaling laws," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5488–5502, 2020.

[27] J. Dong, Y. Shi, and Z. Ding, "Blind over-the-air computation and data fusion via provable wirtinger flow," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1136–1151, 2020.

[28] X. Zang, W. Liu, Y. Li, and B. Vucetic, "Over-the-air computation systems: Optimal design with sum-power constraint," *IEEE Wireless Communications Letters*, vol. 9, no. 9, pp. 1524–1528, 2020.

[29] L. Chen, N. Zhao, Y. Chen, F. R. Yu, and G. Wei, "Over-the-air computation for iot networks: Computing multiple functions with antenna arrays," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 5296–5306, 2018.

[30] H. Sun, S. Li, F. R. Yu, Q. Qi, J. Wang, and J. Liao, "Toward communication-efficient federated learning in the internet of things with edge computing," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 053–11 067, 2020.

[31] Y. S. Jeon, M. M. Amiri, J. Li, and H. V. Poor, "A compressive sensing approach for federated learning over massive mimo communication systems," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[32] W. T. Chang and R. Tandon, "MAC aware quantization for distributed gradient descent," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[33] B. Tegin and T. M. Duman, "Machine learning at wireless edge with OFDM and low resolution ADC and DAC," 2020, arXiv:2010.00350.

[34] L. U. Khan, S. R. Pandey, N. H. Tran, W. Saad, Z. Han, M. N. H. Nguyen, and C. S. Hong, "Federated learning for edge networks: Resource optimization and incentive mechanism," *IEEE Communications Magazine*, vol. 58, no. 10, pp. 88–93, 2020.

[35] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning in fading channels," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[36] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, 2019, pp. 1387–1395.

[37] M. Chen, H. Vincent Poor, W. Saad, and S. Cui, "Convergence time optimization for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2020.

[38] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2021.

[39] L. P. Barnes, H. A. Inan, B. Isik, and A. Özgür, "rTop-k: A statistical estimation approach to distributed SGD," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 897–907, 2020.

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., 2006.

[41] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. Springer-Verlag, 1998.

[42] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," in *Proceedings of the 31st Conference On Learning Theory*, vol. 75. PMLR, 6–9 Jul 2018, pp. 3163–3188.

[43] W.-N. Chen, P. Kairouz, and A. Ozgur, "Breaking the communication-privacy-accuracy trilemma," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3312–3324.