Over-the-Air Statistical Estimation of Sparse Models

Chuan-Zheng Lee*, Leighton Pate Barnes†, Wenhao Zhan† and Ayfer Özgür*
*Department of Electrical Engineering, Stanford University, California
czlee@alumni.stanford.edu, aozgur@stanford.edu
†Department of Electrical and Computer Engineering, Princeton University, New Jersey
lb6858@princeton.edu, wz3993@princeton.edu

***Abstract*—We propose schemes for minimax statistical estimation of sparse parameter or observation vectors over a Gaussian multiple-access channel (MAC) under squared error loss, using techniques from statistics, compressed sensing and wireless communication. These "analog" schemes exploit the superposition inherent in the Gaussian MAC, using compressed sensing to reduce the number of channel uses needed. For the sparse Gaussian location and sparse product Bernoulli models, we derive expressions for risk in terms of the numbers of nodes, parameters, channel uses and nonzero entries (sparsity). We show that they offer exponential improvements over existing lower bounds for risk in "digital" schemes that assume nodes to transmit bits errorlessly at the Shannon capacity. This shows that analog schemes that design estimation and communication jointly can efficiently exploit the inherent sparsity in high-dimensional models and observations, and provide drastic improvements over digital schemes that separate source and channel coding in this context.**

I. Introduction

Consider the problem of estimating the parameters $\theta \in \mathbb{R}^d$ of a distribution p_θ from samples drawn from it. This fundamental problem is well studied in statistics, going back to the early 1800s [1], [2]. Yet two recent trends in data science bring new challenges to this framework.

The first trend is that data are increasingly being generated "at the edge", by many different users or sensors, who must send their observations to a central server. This has spawned burgeoning topics such as edge computing and federated learning, which aim to aggregate data from many nodes for learning, inference and estimation. This decentralization would be inconsequential to the estimation problem if the samples could be relayed to the server as they are, but as wireless engineers know well, communication is resource-constrained and costly. This makes communication a new and important bottleneck in the quest to better design these systems.

The second trend is that modern machine learning data and models are becoming increasingly high-dimensional. For example, the number of parameters d in popularly used neural network models has been growing into the millions.

The combination of these trends can pose prohibitive costs on communication resources. Consider a federated learning setting, in which a central machine learning model is trained from data split among a large number of mobile clients. This is typically done by running *distributed* stochastic gradient descent (SGD), a flavor of SGD where, in each iteration of the algorithm, clients compute gradients with respect to their local data and communicate them to the central server for aggregation. The dimension of these gradients scales with the number of model parameters, and when that is in the millions, communicating the gradients becomes impractical. Fortunately, some recent works [3], [4] have observed that these gradients can in practice be treated as sparse vectors. In many problems with large ambient dimension, such sparsity presents an opportunity to develop cost-efficient solutions.

To study these trends, we can examine the impact of communication constraints and dimensionality d on the estimation error of the parameter $\theta \in \mathbb{R}^d$. Recent works [5]–[10] have modeled this by imposing a digital data rate constraint: each node is assumed to have some finite number of bits to represent its sample U_i, which are communicated errorlessly to the server. In this abstraction, the questions of communication and estimation are in effect separated—physical constraints in the channel impose themselves on the estimation strategy only as a bit limit. These works provided schemes and information-theoretic lower bounds on estimation error in terms of such "digital" constraints, including for sparse regimes [4]–[7].

However, we have recently shown [11], [12] that, in the Gaussian MAC, "analog" schemes can drastically outperform such digital schemes. In the schemes we proposed for the Gaussian location and product Bernoulli models, nodes simply transmit a scaled uncoded version of the samples. The superposition in the MAC in effect carries out the averaging that would, in classical statistics, be used to estimate the mean. This scheme jointly designs the estimation and communication protocols, so would be preempted by the aforementioned digital abstraction. That preemption has a hefty price: judged by worst-case risk under squared error loss for the same physical resources, our analog schemes yielded an exponential improvement even over the digital *lower bounds* in [10].

Even so, a major drawback of the schemes in [11] is that they require at least as many channel uses as there are parameters. With the trend towards having millions of parameters, this would be prohibitive in many applications, including distributed SGD, a mainstay of federated and distributed learning. At the same time, digital schemes have been shown to be able to efficiently exploit sparse model structure to reduce communication requirements [5], [6]. This left a gap: If we're starved of the ability to dedicate at least one channel use per parameter, is it still possible to exploit the additive nature of the Gaussian MAC *and* the structure of sparse models to conduct parameter estimation? Can such analog schemes still trounce their digital counterparts?

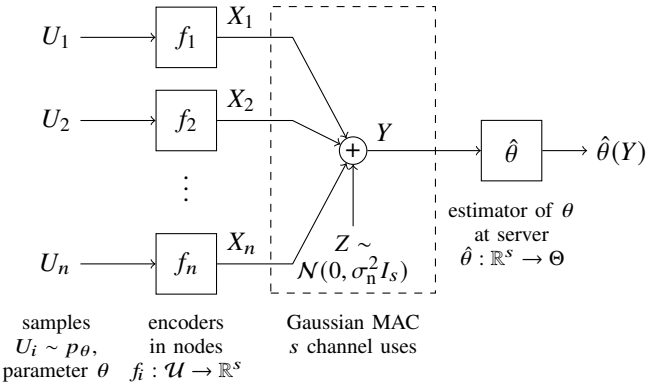

Fig. 1. System diagram

In this paper, we answer these questions in the affirmative. In distributed SGD, local gradients can often be sparsified using a number of techniques, such as communicating only the top k gradients [3], [4]. We leverage this observation to treat samples as coming from distributions with high-dimensional but k-sparse parameters or observations. This allows us to use the compressed sensing framework of [13], [14] to reduce the dimension of the transmitted vector, while still harnessing the over-the-air superposition in the Gaussian MAC. We propose schemes for the sparse Gaussian location and sparse product Bernoulli models, and show that they offer drastically lower estimation error than the digital lower bounds for the same models in [4], [7].

Our results provide a theoretical basis for a growing body of literature suggesting that analog, joint estimation-communication schemes can be markedly advantageous over digital schemes that separate source and channel coding. These include studies in source coding for sensor networks [15]–[17], experimental comparisons of analog and digital versions of distributed SGD [18], [19] and transmission of model parameters [20], as well as our own findings in [11], [12]. Other authors have progressed analog over-the-air computation methods [21], [22]. Note that in the dense models of [11], [12], the superposition in the Gaussian MAC can be easily harnessed by uncoded transmission, while the current paper develops more sophisticated encoding techniques, based on compressed sensing, that simultaneously make use of the additive nature of the channel and the sparsity of the model.

The rest of this paper is structured as follows. In Section II, we describe our models of interest. In Section III, we summarize relevant concepts from compressed sensing. Our main results come in Sections IV and V, and a discussion on their significance follows in Section VI. Proofs are in Section VII.

II. Problem formulation

Consider the Gaussian multiple-access channel (MAC) depicted in Fig. 1. In each of s channel uses, indexed $t = 1, \dots, s$, the n senders each transmit a symbol $X_{it} \in \mathbb{R}$ to a single receiver, which observes 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 set aside questions of phase synchronization, power control and fading to focus on the additive nature of MACs.

Our system seeks to estimate a parameter vector θ, which belongs to a parameter space $\Theta \subseteq \mathbb{R}^d$ and determines the distribution p_θ from which each sender $i = 1, \dots, n$ draws a sample U_i. Each sender thus chooses a vector $X_i = (X_{i1}, \dots, X_{is})$ to transmit over the s channel uses, which is a function of its sample U_i. We similarly denote the concatenation of received symbols as $Y = (Y_1, \dots, Y_s)$, and the receiver's estimate $\hat{\theta}$ must be a function of Y. We study worst-case risk under squared error loss, $\sup_{\theta \in \Theta} \mathbb{E}_\theta[\|\hat{\theta}(Y) - \theta\|_2^2]$, where $\mathbb{E}_\theta[\cdot]$ denotes the expectation under p_θ.

The senders must abide by a power constraint, and since the distribution p_θ is not known, they must obey it for the entire parameter space. That is, we require that

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

In our earlier works [11], [12] we considered the case where $s \ge d$. In this paper, we turn our attention to the case where there are fewer channel uses than parameters, $s < d$, and we focus on the sparse analogs of the models we considered in [11], [12]. In particular, we focus on two different models that allow us to investigate the impact of sparsity in the parameter space and the sample space respectively.

Our first model of interest is the **sparse Gaussian location model**, in which θ is the k-sparse mean of a Gaussian distribution of known variance σ^2. That is, $p_\theta = \mathcal{N}(\theta, \sigma^2 I_d)$ and θ belongs to the parameter space

$$\Theta_{\mathrm{SG}} \triangleq \left\{\theta \in [-B, B]^d : \|\theta\|_0 \le k\right\} \tag{3}$$

for some known $B > 0$, and some integer $0 < k < d$ representing the sparsity level. Here $\|\theta\|_0$ denotes the number of nonzero elements in θ. We say that a vector θ is *k-sparse* if it contains at most k nonzero elements, *i.e.*, if $\|\theta\|_0 \le k$. Note that this corresponds to a model where the parameter vector θ is sparse but the observed samples $U_i \sim p_\theta$ are dense.

Our second model of interest is the **sparse product Bernoulli model**, in which the elements of θ are parameters of independent Bernoulli variables, $p_\theta = \prod_{j=1}^d \mathrm{Bernoulli}(\theta_j)$, and θ belongs to

$$\Theta_{\mathrm{SB}} \triangleq \left\{\theta \in [0, 1]^d : \|\theta\|_1 \le k\right\}. \tag{4}$$

Note that θ here is not itself sparse, but rather, its elements sum to $k < d$, so the samples $U_i \sim p_\theta$ will be on average k-sparse. This models the scenario where observations rather than parameters of the model have a sparse structure. Note, however, that the sparsity is subtle: observation vectors are only sparse on average, and the assumed sparsity structure does not directly reduce the dimensionality of the problem from d to k.

III. Compressed sensing: preliminaries

The schemes we propose use a method from compressed sensing to reduce the number of channel uses needed to communicate the samples. The problem of interest in compressed

sensing is as follows. We wish to recover a k-sparse signal $x \in \mathbb{R}^d$ from only $s \ll d$ measurements

$$y_t = \langle x, a_t \rangle, \ t = 1, \dots, s, \quad \text{or} \quad y = Ax,$$

for a known, and possibly designed, $A \in \mathbb{R}^{s \times d}$. A series of works in the 2000s [13] showed that, under a certain condition on A, x can be recovered exactly by solving the convex program

$$\text{minimize} \quad \|\tilde{x}\|_1 \quad \text{subject to} \quad A\tilde{x} = y. \tag{5}$$

The condition that A must satisfy is called a *restricted isometry hypothesis*, and is defined as follows. Let $A_{\mathcal{K}}, \mathcal{K} \subseteq \{1, \dots, d\}$ be the $n \times |\mathcal{K}|$ submatrix of A whose columns are extracted from A according to the indices in $\mathcal{K}$. The *k-restricted isometry constant* δ_k of A is the smallest value such that, for all index subsets $\mathcal{K}$ with $|\mathcal{K}| \le k$, for all vectors $u \in \mathbb{R}^{|\mathcal{K}|}$

$$(1 - \delta_k)\|u\|_2^2 \le \|A_{\mathcal{K}} u\|_2^2 \le (1 + \delta_k)\|u\|_2^2. \tag{6}$$

Intuitively, when δ_k is small, (6) requires that submatrices of A with up to k columns behave approximately like an orthonormal system. A *restricted isometry hypothesis* is then a condition imposing an upper bound on a linear combination of restricted isometry constants of A, for example,

$$\delta_k + \delta_{2k} + \delta_{3k} < 1. \tag{7}$$

The work [13] showed that if A satisfies (7), then solving (5) recovers any x that is k-sparse.

Subsequent work studied *perturbed* sparse measurements, where $y = Ax + e$ for some perturbation e. Here, [14] showed that under another restricted isometry hypothesis, the signal can be recovered with error linear in the perturbation. We recite the main result therein below.

Theorem 1 (Theorem 1 of [14])**.** *Let $A \in \mathbb{R}^{s \times d}$ have restricted isometry constants satisfying $\delta_{3k} + 3\delta_{4k} < 2$, and for some perturbation level ε, let $x^\sharp$ be the solution to the convex program*

$$\textit{minimize} \quad \|\tilde{x}\|_1 \qquad \textit{subject to} \quad \|A\tilde{x} - y\|_2 \le \varepsilon. \tag{8}$$

Then for any k-sparse signal x and any perturbation e with $\|e\|_2 \le \varepsilon$, $x^\sharp$ obeys

$$\|x^\sharp - x\|_2 \le C_k \cdot \varepsilon, \tag{9}$$

where the constant C_k may only depend on δ_{4k}.

It can be shown that if $\delta_{4k} \le 0.4$ then $C_k \le 26$ in the above theorem. (This also implies $\delta_{3k} + 3\delta_{4k} < 2$, as δ_k is increasing in k.) A number of ways to generate matrices that satisfy this condition are well known in the compressed sensing literature. The most notable is that if the entries of $A \in \mathbb{R}^{s \times d}$ are i.i.d. Gaussian with zero mean and variance $1/s$, then A satisfies the condition with high probability when

$$s \ge C \cdot k \log \frac{d}{k}, \tag{10}$$

for a sufficiently large constant C. We refer the reader to [13, Sec. III] for details. In our schemes, for any k and d, this assures us of the existence of matrices $A \in \mathbb{R}^{s \times d}$, with s chosen as in (10), that satisfy (6) with $\delta_k < 0.4$ and Theorem 1 with $C_k \le 26$.

Our strategies for our two models of interest, explained in the next two sections, are to think of sampling variance and/or channel noise as "perturbations" in the sense defined in Theorem 1.

IV. Gaussian location model

A. Estimation scheme

For the sparse Gaussian location model, we propose the following scheme. Take any matrix $A \in \mathbb{R}^{s \times d}$ satisfying the restricted isometry hypothesis with $\delta_{4k} \le 0.4$. As discussed above, this requires s to satisfy (10); we will choose it to be satisfied with equality. Each sender i transmits $X_i \triangleq \alpha A U_i$ over the Gaussian MAC, where α is a scaling factor chosen as

$$\alpha \triangleq \frac{1}{1 + \delta_k} \sqrt{\frac{sP}{d(B^2 + \sigma^2)} \cdot \frac{1}{\lceil d/k \rceil}}. \tag{11}$$

The receiver observes the noisy superposition Y, given by (1). The receiver computes its estimate $\hat{\theta}$ as follows:

1) Compute $\theta^\sharp$ as the solution to the convex program

$$\underset{\vartheta}{\text{minimize}} \ \|\vartheta\|_1 \quad \text{subject to} \quad \left\| A\vartheta - \frac{1}{\alpha n} Y \right\|_2^2 \le \varepsilon^2, \tag{12}$$

where ε^2 is chosen such that

$$\varepsilon^2 \triangleq \frac{2s\sigma^2}{n}(1 + \delta_k)^2 \left\lceil \frac{d}{k} \right\rceil \left[1 + \frac{\sigma_{\mathrm{n}}^2}{nP} \frac{d}{s} \left(1 + \frac{B^2}{\sigma^2} \right) \right], \tag{13}$$

2) Compute $\hat{\theta}$ to be $\theta^\sharp$ moved into $[-B, B]^d$ by clamping wayward entries,

$$\hat{\theta}_j = \begin{cases} \theta_j^\sharp, & \text{if } \left|\theta_j^\sharp\right| \le B, \\ -B, & \text{if } \theta_j^\sharp < -B, \\ B, & \text{if } \theta_j^\sharp > B, \end{cases} \qquad j = 1, \dots, d, \tag{14}$$

where $\hat{\theta}_j$ is the θth entry of $\hat{\theta}$, and similarly for $\theta_j^\sharp$.

We verify in Proposition 1 that with the choice of α in (11), we satisfy the power constraint. The proof is in Section VII-A.

Proposition 1. *If we choose α as in* (11)*, then X_i satisfies the power constraint* (2)*.*

B. Main result on squared error risk

We are now in a position to present our first main result, an upper bound on worst-case risk for sparse estimation under the scheme proposed in Section IV-A.

Theorem 2. *For any k-sparse vector $\theta \in \Theta_{\mathrm{SG}}$, the estimate $\hat{\theta}$ provided by the scheme proposed in Section IV-A satisfies*

$$\mathbb{E}_\theta\left[\|\hat{\theta} - \theta\|_2^2\right] \le C_1 \sigma^2 \frac{d}{n} \log\left(\frac{d}{k}\right) \left[1 + \frac{\sigma_{\mathrm{n}}^2}{nP} \frac{d}{s} \left(1 + \frac{B^2}{\sigma^2} \right) \right] + 4B^2 d \cdot e^{-0.15s}, \tag{15}$$

where C_1 is an absolute (explicit) constant that does not depend on the problem parameters.

The proof is in Section VII-A. Note that the second term in (15) decays exponentially in s so when k and d are large the error will be dictated by the first term. We further discuss the implications of this result in Section VI.

V. Product Bernoulli model

A. Estimation scheme

For the sparse product Bernoulli model, we will work with a matrix that satisfies the restricted isometry hypothesis for sparsity level $2nk$ instead of k, *i.e.*, $A \in \mathbb{R}^{s\times d}$ with $\delta_{8nk} < 0.4$. As discussed in Section III, such a matrix exists if

$$s \geq C \cdot nk \log\left(\frac{d}{nk}\right).$$

Each sender i transmits $X_i \triangleq \alpha A U_i$, where α is chosen as

$$\alpha \triangleq \frac{\sqrt{sP}}{(1+\delta_k)\sqrt{4k+1}}, \tag{16}$$

The receiver observes Y as in (1) and computes $\hat{\theta}$ as follows:

1) Compute $\theta^\sharp$ as the solution to the convex program

$$\text{minimize } \|\vartheta\|_1 \quad \text{subject to } \left\|A\vartheta - \frac{1}{\alpha n}Y\right\|_2^2 \leq \varepsilon^2, \tag{17}$$

where ε^2 is chosen as

$$\varepsilon^2 \triangleq \frac{2s\sigma_{\mathrm{n}}^2}{\alpha^2 n^2}. \tag{18}$$

2) Compute $\hat{\theta}$ to be $\theta^\sharp$ moved into $[0,1]^d$ by clamping wayward entries,

$$\hat{\theta}_j = \begin{cases} \theta_j^\sharp, & \text{if } 0 \leq \theta_j^\sharp \leq 1, \\ 0, & \text{if } \theta_j^\sharp < 0, \\ 1, & \text{if } \theta_j^\sharp > 1, \end{cases} \qquad j = 1, \ldots, d, \tag{19}$$

where $\hat{\theta}_j$ is the θth entry of $\hat{\theta}$, and similarly for $\theta_j^\sharp$.

Proposition 2 verifies the power constraint. We omit the proof, but discuss it briefly in Section VII-B.

Proposition 2. *If we choose α as in* (16)*, then X_i satisfies the power constraint* (2)*.*

B. Main result on squared error risk

We have now completed the preparations to present our second main result.

Theorem 3. *For any vector $\theta \in \Theta_{\mathrm{SB}}$, the estimate $\hat{\theta}$ provided by the scheme proposed in Section V-A satisfies*

$$\mathbb{E}_\theta\left[\|\hat{\theta}-\theta\|_2^2\right] \leq \frac{k}{n}\left(C_2\frac{\sigma_{\mathrm{n}}^2}{nP}+2\right) + 2d\left(e^{-0.38nk}+e^{-0.15s}\right), \tag{20}$$

where C_2 is an absolute (explicit) constant that does not depend on the problem parameters.

We omit the proof, but discuss it briefly in Section VII-B. We again note that the second term in (20) decays exponentially in nk and s, so at large k and n, the error will be dictated by the first term. We discuss further in the next section.

VI. Comparison to digital lower bounds

Having characterized the performance of analog estimation schemes for the sparse Gaussian and Bernoulli models, we now compare their performance to digital approaches for sparse models that have recently been studied in related literature.

In *digital schemes*, rather than specify the physical-layer channel as we did in Section II, we simply assume that each sender can errorlessly transmit a message of up to m bits to the receiver. Recent work [4]–[10], [23] has used this idea to study the impact of communication constraints on distributed parameter estimation. This allows those works to abstract away the physical layer and focus effectively on the trade-off between available channel capacity and estimation error. However, this abstraction enforces a separation between source and channel coding. Owing to this, our analog schemes, which instead design estimation and communication jointly, beat even these digital lower bounds when controlling for physical resources (transmission power and number of channel uses).

To compute digital lower bounds, we draw on the results of [7] and [4], which provided information-theoretic lower bounds on estimation error for digital schemes under our two models of interest. In applying them, we assume that senders can transmit at the Shannon capacity of the Gaussian MAC. If the MAC is used s times, with rates allocated equally among senders, this would allow each sender to errorlessly communicate

$$m = \frac{s}{2n}\log_2\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right) \quad \text{bits} \tag{21}$$

to the receiver. Note that this assumption is arguably unduly optimistic about the digital schemes—at finite block lengths, senders cannot actually reach the Shannon capacity.

Proposition 3. *For the sparse Gaussian location model, for all protocols in which nodes independently send bits to the server at the Shannon capacity for s channel uses, if $k \leq \frac{d}{2}$ and $\min\{\frac{s}{2d^2}\log_2(1+\frac{nP}{\sigma_{\mathrm{n}}^2}), n\} \geq \sqrt{k\log\frac{d}{k}}$, the risk must satisfy*

$$\sup_{\theta\in\Theta_{\mathrm{SG}}} \mathbb{E}_\theta\|\hat{\theta}-\theta\|_2^2 \geq C_3\sigma^2\frac{dk\log(\frac{d}{k})}{s\log_2\left(1+\frac{nP}{\sigma_{\mathrm{n}}^2}\right)} \tag{22}$$

where C_3 is a constant independent of problem parameters.

Proof. Apply Theorem 7 of [7], using (21). □

Compare this to (15) of Theorem 2. A summary focusing on the asymptotic scaling of d, n and k is in the first row of Table I. As required by (10), we choose $s \propto k\log\frac{d}{k}$ in both the analog case (Theorem 2) and the digital case (Proposition 3) to allow a clean comparison. Here, we find that the risk of any

TABLE I
COMPARISON FOR GAUSSIAN LOCATION MODEL

	analog achievability	digital lower bound
sparse model large n $s \propto k \log \frac{d}{k}$	$O\left(\frac{d \log \frac{d}{k}}{n}\right)$	$\Omega\left(\frac{d}{\log n}\right)$
dense model [11] $s \geq d$	$O\left(\frac{d}{n}\right)$	$\Omega\left(\frac{d^2}{s \log n}\right)$

TABLE II
COMPARISON FOR PRODUCT BERNOULLI MODEL

	analog achievability	digital lower bound
sparse model $s \propto k \log \frac{d}{k}$	$O\left(\frac{k}{n}\right)$	$\Omega\left(\frac{k}{\log n}\right)$
dense model [11] $s \geq d$	$O\left(\frac{d}{n}\right)$	$\Omega\left(\frac{d^2}{s \log n}\right)$

digital scheme can decrease at best with $\Omega(d/\log n)$, while our proposed analog scheme decreases with $O(d \log(\frac{d}{k})/n)$. This implies that, controlling for physical resources, analog schemes can bring about exponentially smaller estimation error than digital schemes while using only $s \propto k \log \frac{d}{k}$ channel uses to communicate d-dimensional observation vectors with $s \ll d$. In the second row of the table, we recall our results for dense models from [11] for comparison.

We find similarly for the sparse product Bernoulli model.

Proposition 4. *For the sparse product Bernoulli model, for all protocols in which nodes independently send bits to the server at the Shannon capacity for s channel uses, if $\frac{s}{2} \log_2(1+\frac{nP}{\sigma_n^2}) \geq d \log \frac{d}{k}$ and $k \leq \frac{d}{2}$, then the risk must satisfy*

$$\sup_{\theta \in \Theta_{\mathrm{SB}}} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \geq C_4 \frac{k^2 \log \frac{d}{k}}{s \log_2\left(1 + \frac{nP}{\sigma_n^2}\right)}, \tag{23}$$

where C_4 is a constant independent of problem parameters.

Proof. Apply Theorem 2 of [4], using (21). □

A comparison summary is in Table II. The first row compares Theorem 3 and Proposition 4 asymptotically in d, n and k, again choosing $s \propto k \log \frac{d}{k}$ from (10). Note that, differently from the sparse Gaussian model, in Table II the estimation error is governed by the sparsity parameter k in place of the ambient dimension d in the numerator. The contrast between the analog and digital schemes appears in n: digital schemes can at best improve with $\Omega(k/\log n)$, while our proposed analog scheme improves with $O(k/n)$. This is similar to our result for the dense model (second row), and again shows an exponential advantage of analog schemes over any digital scheme.

These results demonstrate that, in the sparse model estimation, leveraging the physical layer in a joint estimation-communication protocol can be drastically advantageous over digital schemes separating source and channel coding. In the present case, this advantage comes about because we leverage the additive nature of the Gaussian MAC. This corroborates our findings for the dense analogs of these models [11], as well as other works comparing analog and digital approaches that also take advantage of sparsity [18], [19].

VII. PROOFS

In all proofs, we use $\mathbb{E}[\cdot]$ in place of $\mathbb{E}_\theta[\cdot]$ for brevity.

A. Sparse Gaussian location model

Proof of Proposition 1. The samples themselves satisfy

$$\mathbb{E}\left[\|U_i\|_2^2\right] = \mathbb{E}\left[\|\theta + W_i\|_2^2\right] = \|\theta\|_2^2 + d\sigma^2 \leq d(B^2 + \sigma^2). \tag{24}$$

Then the power of the transmitted symbols is at most

$$\begin{aligned}
\mathbb{E}\left[\|X_i\|_2^2\right] &= \mathbb{E}\left[\|\alpha A U_i\|_2^2\right] \\
&\leq \alpha^2 \mathbb{E}\left[(1+\delta_k)^2 \lceil d/k \rceil \|U_i\|_2^2\right] \qquad \text{(Lemma 1)} \\
&\leq \alpha^2 (1+\delta_k)^2 \lceil d/k \rceil d(B^2 + \sigma^2). \qquad \text{(by (24))}
\end{aligned}$$

Hence (11) guarantees that $\frac{1}{s}\mathbb{E}\left[\|X_i\|_2^2\right] \leq P$. □

Proof of Theorem 2. We begin by defining the perturbation

$$R \triangleq \frac{1}{n}\left(A \sum_i W_i + \frac{1}{\alpha} Z\right), \tag{25}$$

so that $\frac{1}{\alpha n} Y = A\theta + R$. Note that R, being a linear combination of zero-mean Gaussian vectors, is a zero-mean Gaussian vector, and its variance is

$$\Sigma_R \triangleq \mathbf{var}\left[\frac{1}{n}\left(A \sum_i W_i + \frac{1}{\alpha} Z\right)\right] = \frac{\sigma^2}{n} A A^{\mathsf{T}} + \frac{\sigma_n^2}{\alpha^2 n^2} I_s. \tag{26}$$

We decompose the mean squared error as

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\theta} - \theta\|_2^2\right] &= \mathbb{E}\left[\|\hat{\theta} - \theta\|_2^2 \middle| \|R\|_2 \leq \varepsilon\right] \Pr\{\|R\|_2 \leq \varepsilon\} \\
&\quad + \mathbb{E}\left[\|\hat{\theta} - \theta\|_2^2 \middle| \|R\|_2 > \varepsilon\right] \Pr\{\|R\|_2 > \varepsilon\} \\
&\leq \mathbb{E}\left[\|\hat{\theta} - \theta\|_2^2 \middle| \|R\|_2 \leq \varepsilon\right] + 4B^2 d \cdot \Pr\{\|R\|_2 > \varepsilon\}, \qquad (27)
\end{aligned}$$

where in the last step we used the fact that all $\vartheta, \vartheta' \in [-B, B]^d$, $\|\vartheta - \vartheta'\|_2 \leq 2B\sqrt{d}$ (and that $\Pr\{\|R\|_2 \leq \varepsilon\} \leq 1$).

To control the first term, representing the error under small perturbations, first note that from the definition of $\hat{\theta}$ in (14) and because $\theta \in [-B, B]^d$, we have $\|\hat{\theta} - \theta\|_2 \leq \|\theta^\sharp - \theta\|_2$. Then by Theorem 1, the solution to (12) satisfies

$$\mathbb{E}\left[\|\theta^\sharp - \theta\|_2^2 \middle| \|R\|_2 \leq \varepsilon\right] \leq C_k^2 \cdot \varepsilon^2. \tag{28}$$

As for the second term, the contribution from large perturbations, we find $\Pr\{\|R\|_2 > \varepsilon\}$ using the tail bound on $\|R\|_2$ from Lemma 2. Let $\lambda_{\max}(\cdot)$ denote the largest eigenvalue of its argument. Then by Weyl's inequality applied to (26),

$$\lambda_{\max}(\Sigma_R) \leq \frac{\sigma^2}{n} \lambda_{\max}(A A^{\mathsf{T}}) + \frac{\sigma_n^2}{\alpha^2 n^2}. \tag{29}$$

Now, by Lemma 1,

$$\lambda_{\max}(A A^{\mathsf{T}}) = \max_x \frac{\|Ax\|^2}{\|x\|^2} \leq (1+\delta_k)^2 \left\lceil \frac{d}{k} \right\rceil. \tag{30}$$

Substituting α from (11), and (30), into (29) yields a choice of ε^2 (compare to (13))

$$bs\lambda_{\max}(\Sigma_R) \le \frac{bs\sigma^2}{n}(1+\delta_k)^2\left\lceil\frac{d}{k}\right\rceil\left[1+\frac{\sigma_{\mathrm{n}}^2}{nP}\frac{d}{s}\left(1+\frac{B^2}{\sigma^2}\right)\right] \triangleq \varepsilon^2.$$

Now, $\|R\|_2^2 \ge \varepsilon^2 \implies \|R\|_2^2 \ge bs\lambda_{\max}(\Sigma_R)$, so from Lemma 2,

$$\Pr\{\|R\|_2 > \varepsilon\} \le \Pr\left\{\|R\|_2^2 \ge bs\lambda_{\max}(\Sigma_R)\right\} \le (be^{1-b})^{\frac{s}{2}}. \tag{31}$$

Substituting (28) and (31) into (27) we have

$$\mathbb{E}\left[\|\hat{\theta}-\theta\|_2^2\right] \le C_k^2\cdot\varepsilon^2 + 4B^2 d(be^{1-b})^{\frac{s}{2}}. \tag{32}$$

Finally, choosing $s = C\cdot k\log\frac{d}{k}$ to satisfy (10) with equality, and choosing $b=2$, yields the result (15). □

Remark. The second term vanishes exponentially in s, leaving just the first term as significant. If we wanted to the second term to vanish in n, we could instead choose $b = 2c\log n + 1$, and it would vanish with $(n^c/\sqrt{2c\log n+1})^{-s}$, at the cost of a $\log n$ factor in the ε^2 term.

B. Sparse product Bernoulli model

We omit the proofs of Proposition 2 and Theorem 3 for space constraints. They roughly mirror the analogous proofs for the sparse Gaussian location model (Proposition 1 and Theorem 2 respectively), with a couple of key differences.

In Proposition 2, since U_i is typically sparse, it is possible to divide the expectation $\mathbb{E}\left[\|AU_i\|_2^2\right]$ into cases corresponding to k-, $2k$-, $3k$-sparsity and so on, and apply Lemma 1 to each case separately. The expectation over these cases then allows $\mathbb{E}\left[\|AU_i\|_2^2\right]$ to be linear in k, rather than depend on d.

In Theorem 3, the main difference from the Gaussian model is that θ is not sparse, so does not permit a direct application of Theorem 1. Instead, we use the "classical" estimate $\hat{\theta}_{\mathrm{cl}} \triangleq \frac{1}{n}\sum_{i=1}^n U_i$, which is typically $2nk$-sparse (with the probability of failure in this mode bounded by the Chernoff bound for the new $e^{-0.38nk}$ term). We apply Theorem 1 to $\mathbb{E}\left[\|\theta^\sharp - \hat{\theta}_{\mathrm{cl}}\|_2^2\right]$ instead, and we bound $\mathbb{E}\left[\|\hat{\theta}_{\mathrm{cl}}-\theta\|_2^2\right]$ using classical techniques.

C. Auxiliary lemmas

We omit the proof of these two auxiliary lemmas for brevity. The first lemma is used to compute the scaling factor α in both the sparse Gaussian location and sparse product Bernoulli models.

Lemma 1. *If $A \in \mathbb{R}^{s\times d}$ has restricted isometry constant δ_k, and the vector $x\in\mathbb{R}^d$ has at most κ nonzero entries, then*

$$\|Ax\|_2 \le (1+\delta_k)\sqrt{\lceil\kappa/k\rceil}\cdot\|x\|_2. \tag{33}$$

In particular, all *vectors $x\in\mathbb{R}^d$ satisfy* (33) *with $\kappa = d$.*

This next lemma provides a tail bound on a correlated Gaussian random vector, using of the Chernoff bound.

Lemma 2. *Let $X\sim\mathcal{N}(0,\Sigma)$ where Σ is a positive semidefinite $n\times n$ matrix, and let $\lambda_{\max}$ be the largest eigenvalue of Σ. Then*

$$\Pr\left\{\|X\|_2^2 \ge bn\lambda_{\max}\right\} \le (be^{1-b})^{\frac{n}{2}}.$$

Acknowledgement

This work was supported in part by NSF grant NeTS-1817205.

References

[1] C. F. Gauss, *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*. SIAM (in Latin), 1995, first published 1820, translated by G. W. Stewart.

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

[3] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *Int. Conf. Learn. Representations*, 2018.

[4] L. P. Barnes, H. A. Inan, B. Isik, and A. Özgür, "rTop-k: A statistical estimation approach to distributed SGD," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 3, pp. 897–907, 2020.

[5] A. Garg, T. Ma, and H. Nguyen, "On communication cost of distributed statistical estimation and dimensionality," in *Advances Neural Inf. Process. Syst.*, 2014, pp. 2726–2734.

[6] 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 *Proc. 48th Annu. ACM Symp. Theory Comput.*, 2016, pp. 1011–1020.

[7] Y. Han, A. Özgür, and T. Weissman, "Geometric lower bounds for distributed parameter estimation under communication constraints," 2020, arXiv:1802.08417, updated version of paper published in *Proc. 31st Conf. Learn. Theory*, vol. 75, pp. 3163–3188, 2018.

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

[9] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints: Lower bounds from chi-square contraction," in *Proc. 32nd Conf. Learn. Theory*, vol. 99, 2019, pp. 3–17.

[10] L. P. Barnes, Y. Han, and A. Ozgur, "Lower bounds for learning distributions under communication constraints via Fisher information," *J. Mach. Learn. Res.*, vol. 21, no. 236, 2020.

[11] C.-Z. Lee, L. P. Barnes, and A. Özgür, "Over-the-air statistical estimation," in *IEEE Global Commun. Conf.*, 2020.

[12] ——, "Lower bounds for over-the-air statistical estimation," in *IEEE Int. Symp. Inf. Theory*, 2021.

[13] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[14] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. on Pure and Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

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

[16] M. Gastpar and M. Vetterli, "Source-channel communication in sensor networks," in *Inf. Process. in Sensor Netw.* Berlin, Heidelberg: Springer, 2003, pp. 162–177.

[17] W. U. Bajwa, J. D. Haupt, A. M. Sayeed, and R. D. Nowak, "Joint source–channel communication for distributed estimation in sensor networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3629–3653, 2007.

[18] M. Mohammadi Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Trans. Signal Process.*, vol. 68, pp. 2155–2169, 2020.

[19] ——, "Federated learning over wireless fading channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3546–3557, 2020.

[20] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, 2020.

[21] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2022–2035, 2020.

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

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