Over-the-Air Statistical Estimation

Chuan-Zheng Lee, Leighton Pate Barnes, Ayfer Özgür
Department of Electrical Engineering
Stanford University
Stanford, California

***Abstract*—We study minimax statistical estimation over a Gaussian multiple-access channel (MAC) under squared error loss, in a framework combining statistical estimation and wireless communication. We develop "analog" joint estimation-communication schemes that leverage the additive nature of the Gaussian MAC and characterize their minimax risk in terms of the number of nodes n, the dimension of the parameter space d and the signal-to-noise ratio of the MAC, for two estimation tasks: Gaussian location and product Bernoulli model. We then compare this risk to existing lower bounds for risk in digital schemes, in which nodes transmit bits noiselessly at the Shannon capacity. We show that, by leveraging the summation inherent in the Gaussian MAC, our analog schemes in both cases outperform these lower bounds, scaling with $O(d/n)$ rather than $\Omega(d/\log n)$. This suggests that in over-the-air statistical estimation, drastic improvements in estimation error can be obtained by using analog schemes that work in tandem with the physical layer, rather than digital schemes using a physical-layer abstraction.**

I. Introduction

In recent years, the use of machine learning has rocketed, the ubiquity of wireless devices has touched ever more applications, and the amount of data generated by sensors everywhere has exploded. The combination of these trends brings new opportunities to combine data from many sources to do estimation, inference and learning. The key features that contrast this setting with more traditional statistical estimation and learning are that data is generated at the edge, and that it needs to be communicated to a central server, often over wireless channels, to solve the desired statistical problems. This leads to new challenges that lie in the intersection of two decades-old disciplines: statistics and wireless communication.

Over the recent years, there has been significant interest in problems in this intersection. In particular, recent works in the machine learning literature [1]–[6] have studied the impact of communication constraints on distributed statistical estimation and testing. These works simplify the communication problem by assuming that capacity constraints in the physical layer dictate the number of bits available to represent each sample. Each observed sample is then quantized into a fixed given number of bits, which are assumed to be communicated to the central receiver without any errors.

In this paper, we seek to study the joint estimation communication problems from first principles. We formulate a model for minimax optimal parameter estimation over the Gaussian multiple-access channel (MAC) and study two canonical estimation models, Gaussian and Bernoulli location models. We develop "analog" transmission-estimation schemes over this channel, where each sensor simply scales and transmits its sample to the receiver in an uncoded fashion, which allows to leverage the additive nature of the Gaussian MAC to average the statistical samples over the air. We characterize the performance of our schemes and show that their resultant estimation error is exponentially smaller than the digital schemes in the aforementioned literature when both approaches use the same amount of physical resources. Similar gains have been observed in source coding for sensor networks [7], [8], as well as experimentally in recent works [9], which build on the additive nature of the Gaussian MAC for gradient accumulation in federated learning-type settings.

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 then summarize our main results in Section III. We compare them to existing lower bounds for digital schemes in Section IV, showing that analog schemes leveraging the superposition nature of the wireless channel can significantly outperform digital schemes. We provide proofs of our results in Sections V and VI.

II. Problem formulation

We study statistical estimation over a Gaussian multiple-access channel. In each use t of this channel, n senders transmit their respective symbols $X_{1t}, \dots, X_{nt} \in \mathbb{R}$ to a single receiver, which receives a noisy superposition Y_t,

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

where $Z_t \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2)$ is the noise in the tth channel use. We assume an average power constraint P on each sender. That is, if a task takes s channel uses, we require that

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

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

This system has the following estimation task: Each of the n senders has an i.i.d. sample U_i, $i = 1, \dots, n$, 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 $Y \triangleq (Y_1, \dots, Y_s)$.

To complete this task, each sender i chooses $X_i \triangleq (X_{i1}, \dots, X_{is}) = f_i(U_i)$ using a function f_i chosen in advance

and known to the receiver, and the receiver 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 $(\mathbf{f}_\mathrm{M}, \hat{\theta}_\mathrm{M})$ 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, $R(\theta; f, \hat{\theta}) \triangleq R(\theta; \mathbf{f}, \hat{\theta})$.

In this paper, we will be concerned with two cases of this 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 $\Theta = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le B\sqrt{d}\}$ for some known $B > 0$. That is, the goal of the receiver is to estimate the unknown 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 $p_\theta = \prod_{i=1}^d \text{Bernoulli}(\theta)$, with $\mathcal{U} = \{0, 1\}^d$ and $\Theta = [0, 1]^d$. The goal of the receiver is to estimate the unknown 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.* [10].

III. Main results

We develop linear estimation schemes for two estimation models, Gaussian and Bernoulli mean estimation, as described in the previous section. An important characteristic of these schemes is that they are "analog" in the sense that senders simply scale and transmit their samples to the receiver in an uncoded fashion. This is in contrast to a digital approach where each sample is encoded with a finite number of bits, which are then reliably communicated to the receiver using channel coding techniques. The analog nature of the transmissions allow us to make use of the additive nature of the Gaussian MAC to combine and average the statistical samples over the air. The following theorems characterize the risk of these analog schemes. In the next section, we compare their performance to and quantify their gain over digital approaches that separate estimation and channel coding.

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 $s \ge d$, 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 V.

Where $\lfloor s/d\rfloor$ 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* t*th 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*

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

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

$$\sup_{\theta} R(\theta; f_{\mathrm{M}}, \hat{\theta}_{\mathrm{M}}) = \begin{cases} \dfrac{d}{4(\sqrt{n}+1)^2}\left(1 + \dfrac{\sigma_{\mathrm{n}}^2}{nP}\right), & \text{if } \sigma_{\mathrm{n}}^2 \le n^{3/2}P, \\ \dfrac{d}{4} \cdot \dfrac{1}{1 + n \cdot \frac{nP}{\sigma_{\mathrm{n}}^2}}, & \text{if } \sigma_{\mathrm{n}}^2 \ge n^{3/2}P. \end{cases} \tag{10}$$

We can also similarly extend this using a repetition code.

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

$$\sup_{\theta} R(\theta; f_{\mathrm{R}}, \hat{\theta}_{\mathrm{R}}) = \begin{cases} \dfrac{d}{4(\sqrt{n}+1)^2}\left(1 + \dfrac{\sigma_{\mathrm{n}}^2}{\lfloor s/d \rfloor nP}\right), & \text{if } \sigma_{\mathrm{n}}^2 \le n^{3/2}P, \\ \dfrac{d}{4} \cdot \dfrac{1}{1 + n \cdot \frac{\lfloor s/d \rfloor nP}{\sigma_{\mathrm{n}}^2}}, & \text{if } \sigma_{\mathrm{n}}^2 \ge n^{3/2}P. \end{cases} \tag{11}$$

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.

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 $\Theta \subset \mathbb{R}^d$,

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

where $\mathcal{F}_k^{\mathrm{D}}$ 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 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 [11]

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

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

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

to the receiver. We then substitute this expression into the lower bounds developed in [6] for the Gaussian location model. 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.

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* $B^2 \min\{\frac{s}{2d} \log_2 \left(1 + nP/\sigma_{\mathrm{n}}^2\right), n\} \ge \sigma^2$*, the risk associated with any such scheme is at least*

$$\sup_{\|\theta\|_2 \le B\sqrt{d}} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \ge C\sigma^2 \max\left\{\frac{2d^2}{s \log_2\left(1 + \frac{nP}{\sigma_{\mathrm{n}}^2}\right)}, \frac{d}{n}\right\}. \tag{14}$$

Proof. Apply Corollary 5 from [6], using (13). Note that $[-B, B]^d \subset \Theta \triangleq \{\theta : \|\theta\|_2 \le B\sqrt{d}\}$, as required by their lower bound result. □

Compare this to (7) from Corollary 1. Note that this corollary 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 $O(d/n)$. (The effect of s vanishes as it converges to the classical noiseless case.) This implies that, when $s \ge d$, the analog schemes can lead to an exponentially smaller estimation error 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.

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 $\min\{\frac{s}{2d}\log_2\left(1 + nP/\sigma_n^2\right), n\} \geq 1$, the risk associated with any such scheme is at least*

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

Proof. Apply Corollary 8 from [6], using (13). □

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 decreases as $O(d/n)$. At sufficiently low SNR nP/σ_n^2, the analog scheme appears to achieve $O(d^2/sn^2)$ (compared to $\Omega(d/n)$), but since increasing n also increases the SNR, 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 provide drastic gains in estimation performance. Similar gains have been observed in asymptotic lossy source coding for Gaussian sensor networks in [7], [8], 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 [9].

V. 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*

$$\hat{\theta}_M(Y) = \frac{1}{\alpha n}Y - \frac{1}{\alpha}\beta, \quad (16)$$

which yields risk

$$\mathbb{E}_\theta\|\hat{\theta}_M(Y) - \theta\|_2^2 = \frac{d}{n}\left(\sigma^2 + \frac{\sigma_n^2}{n\alpha^2}\right). \quad (17)$$

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 $\boldsymbol{\theta}$ 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*

$$\hat{\theta}_{\mu,b^2}(Y) = \mu + \frac{\alpha n b^2}{\alpha^2 n^2 b^2 + \alpha^2 n\sigma^2 + \sigma_n^2}(Y - \alpha n\mu + n\beta), \quad (18)$$

and the Bayes risk is

$$\mathbb{E}\|\hat{\theta}_{\mu,b^2}(Y) - \theta\|^2 = \frac{d(\alpha^2 n\sigma^2 + \sigma_n^2)}{\alpha^2 n^2 + \frac{\alpha^2 n\sigma^2 + \sigma_n^2}{b^2}}. \quad (19)$$

Proof. Under squared error loss, the Bayes estimator for $\mathcal{N}(\mu, b^2 I_d)$ is (by well-known theorem, *e.g.* [12, Cor. 4.1.2(a)]) $\hat{\theta}_{\mu,b^2}(y) = \mathbb{E}(\boldsymbol{\theta}|y)$, which we will evaluate. The relevant covariance matrices are

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

Then the Bayes estimate is given by

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

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

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

□

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

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

The minimax estimator is then

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

□

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

$$\begin{aligned}\mathbb{E}_\theta\left[\|X_i\|_2^2\right] &= \alpha^2(\|\theta\|^2 + d\sigma^2) + 2\alpha\theta^\top\beta + \|\beta\|^2 \\ &= \|\alpha\theta + \beta\|_2^2 + \alpha^2 d\sigma^2. \end{aligned} \tag{20}$$

We thus solve

$$\begin{aligned} &\text{maximize} \quad \alpha \\ &\text{subject to} \quad \|\alpha\theta + \beta\|_2^2 + \alpha^2 d\sigma^2 \le dP \quad \forall\theta : \|\theta\| \le \sqrt{d}B. \end{aligned} \tag{21}$$

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

$$\alpha = \sqrt{\frac{P}{B^2 + \sigma^2}}, \quad \beta = 0, \tag{22}$$

and verify that it also satisfies the constraints of, and is therefore also a solution to, (21). □

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. □

This then yields the achievability result of Corollary 1.

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

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 $s/d \to \infty$.

VI. Bernoulli parameter 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

$$f_C(u) = \begin{cases} -C, & \text{if } u = 0 \\ C, & \text{if } u = 1. \end{cases} \tag{23}$$

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*

$$f_M(u) = \begin{cases} -\sqrt{P}, & \text{if } u = 0 \\ \sqrt{P}, & \text{if } u = 1, \end{cases} \tag{24}$$

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

$$\sup_\theta R(\theta; f_M, \hat{\theta}_M) = \begin{cases} \frac{1}{4(\sqrt{n}+1)^2}\left(1 + \frac{\sigma_n^2}{nP}\right), & \text{if } \sigma_n^2 \le n^{3/2}P, \\ \frac{1}{4}\cdot\frac{1}{1+n\cdot nP/\sigma_n^2}, & \text{if } \sigma_n^2 \ge n^{3/2}P. \end{cases} \tag{25}$$

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 (23), and the receiver uses the affine estimator $\hat{\theta}_{\alpha,\beta}(Y) = \alpha Y + \beta$, then the risk is*

$$\begin{aligned} R(\theta; f_C, \hat{\theta}_{\alpha,\beta}) = \alpha^2\left[4nC^2\theta(1-\theta) + \sigma_n^2\right] \\ + \left[\alpha nC(2\theta - 1) + \beta - \theta\right]^2. \end{aligned} \tag{26}$$

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

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

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 (23), 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*

$$\alpha_M = \begin{cases} \frac{1}{2\sqrt{n}C(\sqrt{n}+1)}, & \text{if } \sigma_n^2 \le n^{3/2}C^2, \\ \frac{nC}{2(\sigma_n^2 + n^2C^2)}, & \text{if } \sigma_n^2 \ge n^{3/2}C^2. \end{cases} \tag{27}$$

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

$$\sup_\theta R(\theta; f_M, \hat{\theta}_M) = \begin{cases} \frac{1}{4(\sqrt{n}+1)^2}\left(1 + \frac{\sigma_n^2}{nC^2}\right), & \text{if } \sigma_n^2 \le n^{3/2}C^2, \\ \frac{1}{4}\cdot\frac{1}{1+n\cdot nC^2/\sigma_n^2}, & \text{if } \sigma_n^2 \ge n^{3/2}C^2. \end{cases} \tag{28}$$

Proof. Define $\alpha_{\rm lo} = \frac{1}{2\sqrt{n}C(\sqrt{n}+1)}$ and $\alpha_{\rm hi} = \frac{nC}{2(\sigma_{\rm n}^2+n^2C^2)}$. Note that then $\alpha_{\rm M} = \min\{\alpha_{\rm lo}, \alpha_{\rm hi}\}$. For convenience, and with some abuse of notation, let $R(\theta; f_C, \alpha, \beta)$ refer to the expression in (26). We will repeatedly use the facts that:

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

These can all be verified by appropriate substitutions into (26). 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 $\sup_\theta R(\theta; \alpha_{\rm M}, \beta_{\rm M})$. We divide into three cases.

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

$$\begin{aligned} R(\tfrac{1}{2}; f_C, \alpha, \beta) &= \alpha^2(nC^2 + \sigma_{\rm n}^2) + (\beta - \tfrac{1}{2})^2 \\ &> \alpha_{\rm lo}^2(nC^2 + \sigma_{\rm n}^2) = R(\tfrac{1}{2}; f_C, \alpha_{\rm lo}, \beta_{\rm M}). \end{aligned}$$

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

$$\begin{aligned} R(\tfrac{1}{2}; f_C, \alpha_{\rm lo}, \beta_{\rm M}) &\overset{\text{(a)}}{=} R(0; f_C, \alpha_{\rm lo}, \beta_{\rm M}) \\ &\overset{\text{(c)}}{\geq} R(0; f_C, \alpha_{\rm hi}, \beta_{\rm M}) \overset{\text{(b)}}{=} \sup_\theta R(\theta; f_C, \alpha_{\rm M}, \beta_{\rm M}), \end{aligned}$$

where labeled steps refer to corresponding facts above.

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

$$\begin{aligned} R(0; f_C, \alpha, \beta) &= \alpha^2\sigma_{\rm n}^2 + (\beta - \alpha nC)^2 \\ &\geq \alpha^2\sigma_{\rm n}^2 + (\tfrac{1}{2} - \alpha nC)^2 = R(0; f_C, \alpha, \beta_{\rm M}), \end{aligned}$$

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

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

If $\alpha_{\rm M} = \alpha_{\rm hi}$, then we have

$$R(0; f_C, \alpha, \beta_{\rm M}) \overset{\text{(c)}}{\geq} R(0; f_C, \alpha_{\rm hi}, \beta_{\rm M}) \overset{\text{(b)}}{=} \sup_\theta R(\theta; f_C, \alpha_{\rm M}, \beta_{\rm M}).$$

Case 3: $\alpha < \alpha_{\rm lo}$ and $\beta \leq \frac{1}{2}$. Take $\theta = 1$ and argue similarly to case 2 that $R(1; f_C, \alpha, \beta_{\rm M}) > \sup_\theta R(\theta; f_C, \alpha_{\rm M}, \beta_{\rm M})$. □

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 $\hat\theta(Y)$. □

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 (23). The minimax affine estimator for such encoding functions is found in Proposition 5. From (28), 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), $C = \pm\sqrt{P}$. □

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. □

Finally, Corollary 2 follows using Lemma 2 again.

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

Acknowledgement

This work was supported in part by NSF award CCF-1704624.

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 Neural Inf. Process. Syst.*, 2013, pp. 2328–2336.

[2] 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.

[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," in *Advances Neural Inf. Process. Syst.*, 2017, pp. 6394–6404.

[5] 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.

[6] L. P. Barnes, Y. Han, and A. Özgür, "Learning distributions from their samples under communication constraints," 2019, arXiv:1902.02890.

[7] M. Gastpar, "Uncoded transmission is exactly optimal for a simple gaussian "sensor" network," vol. 54, no. 11, pp. 5247–5251, 2008.

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

[9] M. Mohammadi Amiri and D. Gunduz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," 2020, early access.

[10] L. P. Barnes, H. A. Inan, B. Isik, and A. Ozgur, "rTop-k: A statistical estimation approach to distributed SGD," 2020, arXiv:2005.10761.

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

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