

Fundamental Limits of Distributed Optimization over Multiple Access Channel

Shubham K Jha

Abstract

We consider distributed optimization over a d -dimensional space, where K remote clients send coded gradient estimates over an *additive Gaussian Multiple Access Channel (MAC)* with noise variance σ_z^2 . Furthermore, the codewords from the clients must satisfy the average power constraint P , resulting in a signal-to-noise ratio (SNR) of KP/σ_z^2 . In this paper, we study the fundamental limits imposed by MAC on the convergence rate of any distributed optimization algorithm and design optimal communication schemes to achieve these limits. Our first result is a lower bound for the convergence rate, showing that communicating over a MAC imposes a slowdown of $\sqrt{d/\frac{1}{2}\log(1 + \text{SNR})}$ on any protocol compared to the centralized setting. Next, we design a computationally tractable digital communication scheme that matches the lower bound to a logarithmic factor in K when combined with a projected stochastic gradient descent algorithm. At the heart of our communication scheme is carefully combining several compression and modulation ideas such as quantizing along random bases, *Wyner-Ziv compression*, *modulo-lattice decoding*, and *amplitude shift keying*. We also show that analog schemes, which are popular due to their ease of implementation, can give close to optimal convergence rates at low SNR but experience a slowdown of roughly \sqrt{d} at high SNR.

I. INTRODUCTION

In over-the-air distributed optimization [2], [3], the server wants to minimize an unknown function by getting gradient updates from remote clients. In this setting, the clients must communicate their gradient updates over-the-air, namely through a wireless communication channel, to the server. Due to its applications in federated learning [4], many interesting schemes have been

This work was supported by Prime Minister's Research Fellowship (PMRF), Ministry of Education (MoE), India. A preliminary version of this work [1] has appeared in IEEE Information Theory Workshop (ITW), Saint-Malo, France, 2023.

The author is with Robert Bosch Center for Cyber-Physical Systems, Indian Institute of Science, Bangalore. Email: shubhamkj@iisc.ac.in.

recently proposed for this problem [5]–[12]. However, a clear understanding of the fundamental limits of over-the-air distributed optimization is not present. In this paper, we close this gap by characterizing the fundamental limits imposed on first-order distributed optimization due to over-the-air gradient communication. We also design computationally tractable over-the-air optimization protocols which are almost optimal.

We consider the setting where a server wants to minimize an unknown smooth convex function with domain in \mathbb{R}^d by making gradient queries to K clients. Each of the K clients can generate gradient estimates within a bounded Euclidean distance σ of the true gradient. The clients can communicate their gradient estimates over an *additive Gaussian Multiple Access Channel* (MAC) with variance σ_z^2 . Furthermore, each client’s communication must also satisfy a power constraint of P , which results in a signal-to-noise ratio (SNR) of KP/σ_z^2 . We establish an information-theoretic lower bound on the convergence rate of any over-the-air optimization protocol. Our lower bound shows that there is $\left(\sqrt{\frac{d}{\min(\frac{1}{2}\log(1+\text{SNR}), d)}}\right)$ factor slowdown in convergence rate of any over-the-air optimization protocol when compared to that of centralized setting. Next, we design a digital, computationally tractable communication scheme that, combined with the standard *projected stochastic gradient descent* (PSGD) algorithm, almost matches this lower bound.

We elaborate on several key ideas in our communication scheme. In this scheme, we divide the clients into two halves and send the gradients updates from the first half of the clients to form a preliminary estimate. We then employ *Wyner-Ziv* compression to send gradient updates from the second half of clients. This first step is crucial in getting close-to-optimal dependence on the parameter σ in the convergence rate. We also employ quantizing along random bases to get optimal dependence on the dimension d in the convergence rate. Finally, to send a d -dimensional gradient update in a minimum number of channel uses, we use *lattice encoding* and a *modulo lattice decoder*, and *amplitude shift keying* (ASK) modulation.

We also derive tight lower and upper bounds on the performance of analog schemes. Our bounds show that analog schemes are close to the optimal performing schemes at low SNR, but they are highly suboptimal at high SNR and have a slowdown of \sqrt{d} as SNR tends to infinity. Table I provides a concise summary of all our results.

Our work is closely related to [13] and [14]. [13], too, studies fundamental limits of over-the-air optimization, but they do so in the single client setting and when the communication channel is the more straightforward additive Gaussian noise channel. The application of distributed

TABLE I: Convergence rates of our proposed schemes for large K , N , and for $\frac{1}{2} \log(1 + \text{SNR})$ less than d .

Lower Bound (General)	Proposed Scheme (General)	Lower Bound (Analog)	Proposed Scheme (Analog)
$\frac{D\sigma}{\sqrt{KN}} \cdot \sqrt{\frac{d}{\frac{1}{2} \log(1 + \text{SNR})}}$	$\frac{D\sqrt{B}\sigma}{\sqrt{KN}} \cdot \sqrt{\frac{d(\log K + \log \log N)}{\frac{1}{2} \log(1 + \text{SNR})}}$	$\frac{D\sigma}{\sqrt{KN}} \cdot \sqrt{\frac{d}{\text{SNR}}}$	$\frac{DB}{\sqrt{KN}} \cdot \sqrt{\frac{d}{\text{SNR}}}$
(Theorem III.3)	(Theorem IV.3)	(Theorem V.2)	(Theorem V.3)

optimization considered in [14, Section 5] is similar to ours. However, in their setup, the K remote clients can perfectly communicate any update up to r bits. While the more complicated channel considered in this paper prohibits the direct application of schemes from these papers, we build on the ideas proposed in these two papers to come up with our almost optimal scheme.

In a slightly different direction, distributed optimization with compressed gradient estimates has also been extensively studied in recent years (see, for instance, [15]–[33]). Here gradient compression is employed to mitigate the slowdown in convergence when full gradients are communicated.

The rest of the paper is organized as follows. We setup the problem in the next section and provide basic preliminaries and lower bounds in Section III. Section IV and V contain our proposed schemes and the associated results. All the proofs are given in Section VI. Section VII contains the experiments, followed by the concluding remarks in Section VIII.

II. SETUP

Consider the following distributed optimization problem. A *server* wants to minimize an unknown convex function $f: \mathcal{X} \rightarrow \mathbb{R}$ over its domain $\mathcal{X} \subset \mathbb{R}^d$ using gradient updates from K remote clients. At each iteration, the server queries the clients for gradient estimates of the unknown function. On receiving the query, each of the K clients generates a stochastic gradient estimate of the function at the queried point, encodes it, and transmits it over a MAC. The output of this channel is available to the server, which it first decodes and then uses it to update the query point for the next iteration using a first-order optimization algorithm (such as Stochastic Gradient Descent). This setting models practical distributed optimization scenarios arising in federated learning and is of independent theoretical interest.

Our goal is twofold: 1) To understand the fundamental limits imposed by communicating gradients over a MAC on the convergence rate; 2) To design the encoding algorithms at the clients, and the decoding and optimization algorithm at the server to come close to the aforementioned fundamental limit.

A. Functions and gradient estimates

a) *Convex and smooth function family:* We assume that the server wants to minimize an unknown function f which is convex and L -smooth functions. That is, for all¹ $x, y \in \mathcal{X}$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y), \quad (1)$$

$$f(y) - f(x) \leq \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|^2. \quad (2)$$

b) *Stochastic gradient estimates:* We assume that client $C_k, k \in [K]$, outputs a noisy gradient $\hat{g}_k(x)$ at a query point $x \in \mathcal{X}$ which satisfies the following standard conditions:

$$\mathbb{E} [\hat{g}_k(x)|x] = \nabla f(x), \quad (\text{unbiasedness}) \quad (3)$$

$$\mathbb{E} [\|\hat{g}_k(x) - \nabla f(x)\|^2|x] \leq \sigma^2, \quad (\text{bounded deviation}) \quad (4)$$

$$\|\hat{g}_k(x)\|^2 \leq B^2. \quad (\text{almost surely bounded}) \quad (5)$$

Denote by \mathcal{O} the set of tuple (f, \mathcal{C}) of functions and clients satisfying the conditions (1), (2), (3), (4) and (5).

B. Communication schemes and the multiple access channel

For the t th query x_t made by the server, clients C_1, \dots, C_K generate gradient estimates $\hat{g}_{1,t}, \dots, \hat{g}_{K,t}$, respectively. In our setting, these gradient estimates are not directly available to the server. These are first encoded by the clients for error correction and then transmitted over MAC, and only the output of the channel is available to the server. For all the clients, we consider encoders of length ℓ with average power less than P . That is, the encoder $\varphi_k: \mathbb{R}^d \times \mathcal{U} \rightarrow \mathbb{R}^\ell$ used by client C_k satisfies the power constraint

$$\mathbb{E} [\|\varphi_k(\hat{g}_{k,t}, U)\|^2] \leq \ell P, \quad \forall k \in [K], \quad (6)$$

¹ $\|\cdot\|$ refers to the standard euclidean norm.

where U is string of public randomness available to all the k clients' encoders and the server's decoder, and \mathcal{U} is the space of such random strings. For notational convenience, we will drop the argument U of φ_k in the rest of the paper.

The encoded codewords $\{\varphi_k(\hat{g}_{k,t})\}_{k=1}^K$ are sent over MAC using ℓ channel uses. The server receives the channel output $Y_t \in \mathbb{R}^\ell$ given by

$$Y_t(j) = \sum_{k=1}^K \varphi_k(\hat{g}_{k,t})(j) + Z_t(j), \quad \forall j \in [\ell], \quad (7)$$

where $Z_t(j)$ is Gaussian distributed with mean 0 and variance σ_z^2 . We denote the *signal-to-noise ratio* by $\text{SNR} := \frac{KP}{\sigma_z^2}$.

The decoder $\psi: \mathbb{R}^\ell \times \mathcal{U} \rightarrow \mathbb{R}^d$ at the server projects back the ℓ -length channel output to a vector in \mathbb{R}^d , which the optimization algorithm uses to update the query point.

Any tuple of mappings $(\varphi_1, \dots, \varphi_k, \psi)$, is said to be a (d, ℓ, P, K) -communication scheme if φ_k , $k \in [K]$, and ψ are described as above. Denote by \mathcal{Q}_ℓ the set of all possible (d, ℓ, P, K) -communication schemes.

C. Over-the-air optimization over MAC

We now describe the optimization algorithm π interacting with the tuple $(\varphi_1, \dots, \varphi_k, \psi) \in \mathcal{Q}_\ell$. At iteration t , the optimization algorithm uses all the previous query points, $\{x_{t'}\}_{t'=1}^{t-1}$, and the decoded gradient estimates, $\{\psi(Y_{t'})\}_{t'=1}^{t-1}$, to decide on the query point $x_t \in \mathcal{X}$. The server then queries the clients at the point x_t , resulting in a gradient estimate $\psi(Y_t)$. This continues for T iterations, after which the algorithm outputs a point $x_T \in \mathcal{X}$.

Denote by $\Pi_{T,\ell}$ the set of all optimization algorithms π making T queries to the clients and interacting with a (d, ℓ, P, K) -communication scheme.

For an optimization algorithm $\pi \in \Pi_{T,\ell}$ and a communication scheme $Q \in \mathcal{Q}_\ell$, we call the tuple (π, Q) an *over-the-air optimization protocol*. For a tuple of function and clients $(f, \mathcal{C}) \in \mathcal{O}$, we measure the performance of any over-the-air optimization protocol (π, Q) by the convergence error

$$\mathcal{E}(f, \mathcal{C}, \pi, Q) := \mathbb{E}[f(\bar{x}_T)] - \min_{x \in \mathcal{X}} f(x).$$

We will study this error when the total number of channel uses, $T\ell$, is restricted to be at most N . We can use communication schemes of arbitrary length ℓ . Note, however, that to increase the length ℓ , we must decrease the number of queries T , since the total number of channel uses

is limited to N . Conversely, to increase the number of queries, we must decrease the length of the communication schemes. Let $\Lambda(N) := \{(\pi, Q) : \pi \in \Pi_{T,\ell}, Q \in \mathcal{Q}_\ell, T\ell \leq N\}$ be the set of over-the-air optimization protocols with at most N channel uses. The smallest worst-case error possible over all such protocols is given by

$$\mathcal{E}^*(N, K, \text{SNR}, \mathcal{X}) := \inf_{(\pi, Q) \in \Lambda(N)} \sup_{(f, \mathcal{C}) \in \mathcal{O}} \mathcal{E}(f, \mathcal{C}, \pi, Q).$$

Let $\mathbb{X} := \{\mathcal{X} : \sup_{x, y \in \mathcal{X}} \|x - y\| \leq D\}$. In this paper, we will characterize² $\mathcal{E}^*(N, K, \text{SNR}) := \sup_{\mathcal{X} \in \mathbb{X}} \mathcal{E}^*(N, K, \text{SNR}, \mathcal{X})$.

III. PRELIMINARIES AND AN INFORMATION THEORETIC LOWER BOUND

A. A benchmark from prior results

We recall the results for the centralized case, which we can model by setting $\text{SNR} = \infty$. In this case, clients can perfectly communicate the gradient estimates in only one channel use. Denote by $\mathcal{E}^*(N, K, \infty)$ the smallest worst-case optimization in this case. A direct application of [34, Theorem 6.3] leads to the following upper bound on $\mathcal{E}^*(N, K, \infty)$ which serves as a basic benchmark for our results in this paper.

Theorem III.1. $\mathcal{E}^*(N, K, \infty) \leq \frac{\sqrt{2}D\sigma}{\sqrt{KN}} + \frac{LD^2}{2N}.$

B. A general convergence bound

Throughout the paper, we will use projected stochastic gradient descent (PSGD) as the first-order optimization algorithm π ; the overall over-the-air optimization protocol is described in Algorithm 1. PSGD proceeds as stochastic gradient descent with the additional projection step where it projects the updates back to domain \mathcal{X} using the map $\Gamma_{\mathcal{X}}(y) := \min_{x \in \mathcal{X}} \|x - y\|$, $\forall y \in \mathbb{R}^d$. The convergence rate of Algorithm 1 is controlled by the square root of worst-case root mean square error (RMSE) $\alpha(Q)$ and the worst-case bias $\beta(Q)$ of the gradient estimates decoded by the server. They are defined as follows:

$$\alpha(Q) := \sup_{\substack{\forall x, k \in [K], \hat{g}_k \in \mathbb{R}^d: \\ \mathbb{E} \|\hat{g}_k - \nabla f(x)\|^2 \leq \sigma^2}} \sqrt{\mathbb{E} [\|\psi(Y) - \nabla f(x)\|^2]},$$

²While our upper bound techniques can handle an arbitrary, fixed \mathcal{X} , the supremum over \mathbb{X} is to ensure that the lower bounds are independent of the geometry of set \mathcal{X} .

1: **for** $t = 0$ to $T - 1$ **do**
 2: $x_{t+1} = \Gamma_{\mathcal{X}}(x_t - \eta_t \psi(Y_t))$
 3: **Output** $\bar{x}_T = \frac{1}{T} \sum_{t=1}^T x_t$

Algorithm 1: PSGD for over-the-air optimization

$$\beta(Q) := \sup_{\substack{\forall x, k \in [K], \hat{g}_k \in \mathbb{R}^d: \\ \mathbb{E} \|\hat{g}_k - \nabla f(x)\|^2 \leq \sigma^2}} \|\mathbb{E}[\psi(Y)] - \nabla f(x)\|,$$

where for $i \in [d]$, $Y(i)$ satisfies (7) and the expectation is taken over all the randomness in the set up. We now recall a lemma from [14] that upper bounds the convergence rate of Algorithm 1 in terms of $\alpha(Q)$ and $\beta(Q)$.

Lemma III.2 ([14, Lemma II.2]). *Let π be the PSGD algorithm making T queries to the clients and Q be any communication scheme in \mathcal{Q}_ℓ . Moreover, the over-the-air optimization protocol uses the MAC channel $N = T \cdot \ell$ times. Then, we have $\sup_{(f, \mathcal{O}) \in \mathcal{O}} \mathcal{E}(f, \mathcal{C}, \pi, Q)$*

$$\leq \frac{\sqrt{2}D\alpha(Q)}{\sqrt{N/\ell}} + \beta(Q) \left(D + \frac{DB}{\alpha(Q)\sqrt{2N/\ell}} \right) + \frac{LD^2}{2N/\ell}.$$

with the learning rate $\eta_t = \min \left\{ \frac{1}{L}, \frac{D}{\alpha(Q)\sqrt{2T}} \right\}, \forall t \in [T]$.

As a result, it is enough to control the RMSE α and bias β of the communication scheme Q to upper bound the overall convergence rate of the corresponding OTA optimization protocol.

Remark 1. We remark here that for the noiseless case, i.e., $\text{SNR} = \infty$, the choices for φ and ψ are *identity* and *averaging* functions, respectively. Specifically, $\varphi(\hat{g}_{k,t}) = \hat{g}_{k,t}$ and $\psi(Y) = Y/K$ with $Y = \sum_k \hat{g}_{k,t}$. Further, due to no constraint on power, the d coordinates can be sent in just one channel use, $\ell = 1$. That gives $\alpha(Q) = \frac{\sigma}{\sqrt{K}}$ and $\beta(Q) = 0$ retrieving the result in Theorem III.1. Therefore, in a nutshell, the primary goal is to design an “efficient” distributed mean estimator under MAC constraints in the sense that its performance parameters α, β , and ℓ are close to these ideal values.

C. Lower bound for over-the-air optimization

We now present an information-theoretic lower bound for any over-the-air optimization protocol. We note that [13] shows a similar lower bound in the single client setting. We build on

their proof and extend the result to the more general setting of K clients. The key step involves showing that over-the-air optimization over parallel independent additive Gaussian noise channel is much easier than over MAC and then proceeding as in [13].

Theorem III.3. *For some universal constant $c \in (0, 1)$ and $N \geq \frac{d}{K \log(1+\text{SNR})}$, we have*

$$\mathcal{E}^*(N, K, \text{SNR}) \geq \frac{cD\sigma}{\sqrt{KN}} \sqrt{\frac{d}{\min\{d, \frac{1}{2} \log(1 + \text{SNR})\}}}.$$

Our lower bound states that, except for very high values of SNR, any over-the-air optimization protocol will experience a slowdown by a factor of $\sqrt{\frac{d}{\frac{1}{2} \log(1+\text{SNR})}}$ over the convergence rate of centralized setting.

IV. A DIGITAL COMMUNICATION SCHEME FOR OVER-THE-AIR OPTIMIZATION

In this section, we present our main result: a digital communication scheme that, combined with PSGD, will almost match the lower bound in Theorem III.3. Our scheme below is “universal” in the sense that the clients don’t require the knowledge of σ for the transmission of gradient estimates. As pointed out in Remark 1, our focus should be on designing an efficient distributed mean estimator.

A. Warm-up scheme: UQ-OTA

For ease of presentation, we first present a warm-up OTA optimization protocol, UQ-OTA, based on uniform quantization. We will build on the components described below to present our final digital scheme. Throughout the description of our schemes, we omit the subscript t for convenience.

a) Uniform quantization.: Each client C_k first divides the gradient estimate \hat{g}_k by the number of clients K to form \tilde{g}_k and quantizes it using an unbiased v -level coordinate-wise uniform quantizer v -CUQ. The v -CUQ takes i th coordinate $\tilde{g}_k(i) \in [-\frac{B}{K}, \frac{B}{K}]$ as input and outputs $z_{k,i} \in \{0, \dots, v-1\}$ as per the following rule:

$$z_{k,i} = \begin{cases} \left\lceil \frac{(v-1)(K\tilde{g}_k(i)+B)}{2B} \right\rceil, & \text{w.p. } \frac{\tilde{g}_k(i) - \lfloor \frac{\tilde{g}_k(i)K(v-1)}{2B} \rfloor}{2B/(K(v-1))} \\ \left\lfloor \frac{(v-1)(K\tilde{g}_k(i)+B)}{2B} \right\rfloor, & \text{w.p. } \frac{\lceil \frac{\tilde{g}_k(i)K(v-1)}{2B} \rceil - \tilde{g}_k(i)}{2B/(K(v-1))} \end{cases}.$$

That is, the quantizer first finds the two consecutive quantization points containing $\tilde{g}_k(i)$ and declares exactly one of the corresponding indices stochastically. The probability distribution is

chosen in such a way that the output $z_{k,i}$ suffices to form an unbiased estimate of $\tilde{g}_k(i)$. Define $\mathbf{Z}_k := \{z_{k,i} : i \in [d]\}$ as the quantized output for client k . We now process the quantized output for transmission over MAC.

b) *Lattice encoding and ASK modulation using $\mathcal{M}(\mathbb{Q}_k, v, p)$* .: Client C_k sends \mathbf{Z}_k over MAC by first encoding them onto a one-dimensional lattice and further modulating them onto an ASK code. We describe below the entire procedure and refer to it by $\mathcal{M}(\mathbb{Q}_k, v, p)$ with parameters v and p to be specified later.

For an integer³ $p \leq d$, we first partition the set of coordinates $[d]$ equally into blocks of size p . That way, we have d/p blocks. For $j \in [d/p]$, denote by \mathcal{B}_j the j th block is given by $\mathcal{B}_j = \{(j-1)p + 1, \dots, (j-1)p + p\}$. For each \mathcal{B}_j , the corresponding quantized values are mapped onto a one-dimensional lattice Λ_w generated by a set of basis $\{w^0, \dots, w^{p-1}\}$ for some positive integer w and is given by

$$\Lambda_w = \{q_1 \cdot w^0 + \dots + q_p \cdot w^{p-1} : 0 \leq q_1, \dots, q_p \leq v-1\}.$$

Denote by $\tau_{k,j}$ the lattice point corresponding to block \mathcal{B}_j for the client C_k is given by

$$\tau_{k,j} = Z_k(\mathcal{B}_j(1)) + Z_k(\mathcal{B}_j(2)) \cdot w + \dots + Z_k(\mathcal{B}_j(p)) \cdot w^{p-1},$$

with $w = K(v-1) + 1$. Note that this choice of w ensures successful recovery of the sum of client updates at the server.

Definition IV.1. A code is an *Amplitude Shift Keying (ASK) code* satisfying the average power constraint (6) if the range \mathcal{A} of the encoder mapping is given by

$$\mathcal{A} := \left\{ -\sqrt{P} + (i-1) \cdot \frac{2\sqrt{P}}{r-1} : i \in [r] \right\},$$

for some $r \in \mathbb{N}$. Note that this is a code of length 1. Note that this is a code of length 1.

To satisfy the power constraints of MAC, we then modulate each $\tau_{k,j}$ to $[-\sqrt{P}, \sqrt{P}]$ using an ASK code. Since each $\tau_{k,j}$ takes values in $\{0, \dots, \frac{w^p-1}{K}\}$, we set the size of ASK code $r = \frac{w^p-1}{K} + 1$ to establish one-to-one correspondence. Consequently, the encoded value is given by $\varphi_k(j) = \mathcal{A}(\tau_{k,j} + 1), \forall j \in [d/p], \forall k \in [K]$. The transmission takes place over MAC as in (7).

³For simplicity, we assume p divides d .

c) *Lattice decoding at server* $\mathcal{L}(Y, v, p)$.: On the server side, our goal will be to compute an unbiased estimate of sum $\sum_k \tilde{g}_k$ from Y . Note that the natural aggregation property of the MAC channel makes it somewhat easier to recover this sum instead of individual Z_k s. This further implies recovering the sum of quantized outputs $\sum_k Z_k$ which suffices to form the unbiased estimate of $\sum_k \tilde{g}_k$.

Towards that, each coordinate $Y(j)$, $j \in [\ell]$, is first fed into a coordinate-wise MD decoder to locate the nearest ASK codeword $\hat{Y}(j)$ in $\{-K\sqrt{P} + (i-1) \cdot \frac{2\sqrt{P}}{r-1} : i \in [r]\}$. Using the one-to-one correspondence, the decoded point $\hat{Y}(j)$ is then mapped back to the lattice⁴ Λ_w . Denote by $\hat{\tau}_j$ the decoded lattice point can be expressed as

$$\hat{\tau}_j = \lambda(\mathcal{B}_j(1)) \cdot w^0 + \dots + \lambda(\mathcal{B}_j(p)) \cdot w^{p-1},$$

for some vector $\lambda \in \{0, \dots, K(v-1)\}^d$. Therefore, to recover the desired sum, the server uses a *modulo-lattice decoder* for each \mathcal{B}_j , $j \in [d/p]$, that successively outputs the corresponding coordinates of λ . In particular, $\forall i \in [p]$,

$$\lambda(\mathcal{B}_j(i)) = \frac{\hat{\tau}_j - \lambda(\mathcal{B}_j(1)) \cdot \dots - \lambda(\mathcal{B}_j(i-1))w^{i-2}}{w^{i-1}} \pmod{w},$$

where for positive integers a, b, m and t such that $a = m \cdot b + t$ and $0 \leq m \leq b-1$, the mod-operation is defined as $a \pmod{b} = t$. Note that such recovery is feasible with the current choice of w as each coordinate of $\sum_{k \in [K]} Z_k$ is at most $w-1$. The vector λ obtained above is finally used to form $\psi(Y)$ to be used in Algorithm 1 as

$$\psi(Y) = -B + \frac{2B}{K(v-1)} \cdot \lambda. \quad (8)$$

Theorem IV.2. *Let π be the optimization algorithm described in Algorithm 1, where $\psi(Y)$ is obtained in (8) with $v = \sqrt{d} + 1$. Then, for a universal constant $c_1 > 0$ and integers p, K such that $d \geq p \geq 1$ and $K \geq B^2/\sigma^2$, we have*

$$\sup_{(f, \mathcal{C}) \in \mathcal{O}} \mathcal{E}(f, \mathcal{C}, \pi, Q) \leq \frac{c_1 DB}{\sqrt{KN}} \sqrt{\frac{d}{p}} + \frac{LD^2 d}{2Np},$$

where $p = \left\lfloor \frac{\log\left(1 + \sqrt{\frac{2K \text{SNR}}{\ln(KN^{1.5})}}}\right)}{\log(Kd)} \right\rfloor$.

Remark 2. We remark that both the encoding and decoding complexity of UQ-OTA is $O(d)$.

⁴Note that under perfect decoding, this yields the sum of transmitted lattice points $\sum_k \tau_{k,j}$.

Remark 3. We remark that the size of \mathcal{A} used for MAC transmission grows with the operating SNR as $r \approx \min(\sqrt{d} + 1, \lfloor 1 + \sqrt{2\text{SNR}/(K \ln(KN^{1.5}))} \rfloor)$.

Remark 4. At large values of $\text{SNR} \geq 2(2^d - 1)$, $p \approx \frac{\frac{1}{2} \log(1+\text{SNR})}{\log(Kd)}$. We remark that UQ-OTA incur a $(B/\sigma)\sqrt{\log K + \log d + \log \log N}$ factor slowdown in convergence rate compared to that of centralized setting, which can still cause a slowdown for high-dimensional settings and large values of B compared to σ .

B. Wyner-Ziv digital scheme: WZ-OTA

We are now ready to present our main digital scheme WZ-OTA which significantly improves over the performance of UQ-OTA and is almost optimal.

In this scheme, we partition the clients \mathcal{C} equally into two sets \mathcal{C}_1 and \mathcal{C}_2 . In each iteration t , the clients in \mathcal{C}_1 construct the side information at the server, and the remaining clients in \mathcal{C}_2 exploit this information to form a Wyner-Ziv estimate of $\nabla f(x_t)$ at the server.

a) Side information construction: The clients in \mathcal{C}_1 use the previously described UQ-OTA communication scheme to form a preliminary estimate (8) at the server. This requires $\ell = d/p$ channel uses. Note that the clients in \mathcal{C}_2 send 0 during these transmissions.

The server divides this preliminary estimate by $K/2$ to form S and then rotates it by a random matrix \mathbf{R} to form the side information $\mathbf{R}S$. Here $\mathbf{R} = 1/\sqrt{d}\mathbf{H}\mathbf{D}'$ where \mathbf{H} is the Walsh-Hadamard⁵ matrix [35], and \mathbf{D}' is a random diagonal matrix with non-zero entries generated uniformly and independently from $\{-1, +1\}$.

b) The Wyner-Ziv estimate: The clients in \mathcal{C}_2 use a Wyner-Ziv estimator boosted DAQ from [14] to construct the final estimate, while those in \mathcal{C}_1 transmit 0 in all channel uses. The boosted DAQ uses the idea of correlated sampling between the input and the side information to reduce quantization error. Specifically, for an input $|x| \leq M$ at the encoder and a corresponding side information $|y| \leq M$ at the decoder, the boosted DAQ estimate is given by

$$\hat{X} = (2M/I) \sum_{i \in [I]} (\mathbb{1}_{\{U_i \leq x\}} - \mathbb{1}_{\{U_i \leq y\}}) + y, \quad (9)$$

where each $U_i \sim \text{unif}[-M, M]$ is a uniform random variable. Note that \hat{X} is an unbiased estimate of x with MSE at most $2M|x - y|/I$.

⁵Without loss of generality, we assume d is a power of 2. If not, we can zero-pad the gradient estimates and make the resulting dimension power of 2; this only adds a constant multiplicative factor to our upper bounds.

In our setting, each client $C_k \in \mathcal{C}_2$ first pre-processes its noisy estimate as $\tilde{g}_k = \frac{2\hat{g}_k}{K}$ and uses shared randomness to draw I uniform random vectors $U_{k,i} \in [-M, M]^d, i \in [I]$, independently. The choice of M and I are crucial for our scheme and will be specified later. Using shared randomness again, each \tilde{g}_k is rotated using the same random matrix \mathbf{R} used earlier. Each coordinate of this rotated vector is then quantized to an element in $\{0, \dots, I\}$ as

$$\mathbf{Q}_k(j) = \sum_{i \in [I]} \mathbb{1}_{\{U_{k,i}(j) \leq \mathbf{R}\tilde{g}_k(j)\}}, \forall j \in [d].$$

As an aside, it is instructive to note that under the event $\mathcal{V}_j = \{|\mathbf{R}S(j)| \leq M, |\mathbf{R}\tilde{g}_k(j)| \leq M\}$, $\mathbf{Q}_k(j)$ suffices to form an unbiased estimate of $\mathbf{R}\tilde{g}_k(j)$ using boosted DAQ (see (9)). Coming back to our scheme, each client k transmits the quantized vector \mathbf{Q}_k over the MAC channel by first using the lattice encoder and then using ASK modulation. The entire operation is described by the function $\mathcal{M}(\mathbf{Q}_k, v', p')$ (see Section IV-A) with $v' = I + 1$ and p' to be specified shortly. Note that there are $\ell = d/p'$ channel uses per iteration.

At the server, the channel output $Y \in \mathbb{R}^{d/p'}$ is passed through $\mathcal{L}(Y, v', p')$ to obtain λ . Following the boosted DAQ estimator (9), the final output $\psi(Y)$ is given by

$$\psi(Y) = \frac{2M}{I} \mathbf{R}^{-1} \sum_{j \in [d]} (\lambda(j) - \omega(j)) \cdot e_j + \frac{K}{2} S, \quad (10)$$

where each $\omega(j) = \sum_{k \in \mathcal{C}_2} \sum_{i \in [I]} \mathbb{1}_{\{U_{k,i}(j) \leq \mathbf{R}S(j)\}}$ can be realized at the server using shared randomness and the available side-information. We next characterise the performance of WZ-OTA.

Theorem IV.3. *Let c_2, c_3 be positive universal constants and π be the optimization algorithm described in Algorithm 1, where $\psi(Y)$ is obtained using (10) with $v = 7, M = \frac{c_2 B}{K\sqrt{d}} \sqrt{\ln(K^{1.5}N)}$ and $I = c_2 \sqrt{\ln(K^{1.5}N)}$. Then, for integers p, p' and K such that $K \geq B^2 d / \sigma^2$ and $d \geq p, p' \geq 1$, we have*

$$\sup_{(f, \mathcal{C}) \in \mathcal{O}} \mathcal{E}(f, \mathcal{C}, \pi, Q) \leq \frac{c_3 D \sqrt{B\sigma}}{\sqrt{KN}} \sqrt{\frac{d}{q} + \frac{LD^2 d}{2Nq}},$$

where $\frac{1}{q} = \frac{1}{p} + \frac{1}{p'}$ with $p = \left\lfloor \frac{\log\left(1 + \sqrt{\frac{K \text{SNR}}{2 \ln(K^{1.5}N)}}}\right)}{\log K} \right\rfloor$ and $p' = \left\lfloor \frac{\log\left(1 + \sqrt{\frac{K \text{SNR}}{2 \ln(K^{1.5}N)}}}\right)}{\log K + \log \log N} \right\rfloor$.

Remark 5. For large K, N , we remark that the WZ-OTA combined with PSGD is off only by a factor of $\sqrt{(B/\sigma)(\log K + \log \log N)}$ from our lower bound. In comparison, from Theorem IV.2, UQ-OTA combined with PSGD is off by a factor $(B/\sigma) \sqrt{(\log K + \log d + \log \log N)}$. Quantization along random bases and Wyner-ziv compression allows WZ-OTA to improve by factors $\log d$ and $\sqrt{B/\sigma}$ over UQ-OTA.

Remark 6. We remark that the random rotation step using the Walsh-Hadamard matrix can be performed in nearly linear-time and, in particular, requires $O(d \log d)$ operations. Since this is the most expensive step in WZ-OTA, the encoding and decoding complexity of WZ-OTA is $O(d \log d)$.

V. PERFORMANCE OF ANALOG SCHEMES

Definition V.1. A communication scheme is an *analog scheme* if the encoder mapping φ is linear, i.e., $\varphi(x) = \mathbf{A}x$ for $\mathbf{A} \in \mathbb{R}^{\ell \times d}$ and $\ell \leq d$. We allow random entries for \mathbf{A} as long as the randomness is independent of x . For the class of (d, ℓ, P, K) -communication schemes restricted to using such analog schemes, we denote by $\mathcal{E}_{\text{analog}}^*(N, K, \text{SNR})$ the corresponding min-max optimization error. Clearly, $\mathcal{E}_{\text{analog}}^*(N, K, \text{SNR}) \geq \mathcal{E}^*(N, K, \text{SNR})$.

We begin by proving a lower bound for analog communication schemes.

Theorem V.2. For some universal constant $c \in (0, 1)$, and $N \geq \frac{d}{K}(\sigma^2 + \frac{\sigma^2}{\text{SNR}})$, we have

$$\mathcal{E}_{\text{analog}}^*(N, K, \text{SNR}) \geq \frac{cD}{\sqrt{KN}} \sqrt{d\sigma^2 + \frac{d\sigma^2}{\text{SNR}}}.$$

The following lower bound also uses affine functions as difficult functions and builds on a class of Gaussian oracles proposed, recently, towards proving a similar result in [13].

For our upper bound, we use the well-known *scaled transmission* scheme from [8]. In this scheme, the gradient estimates are scaled-down by \sqrt{dP}/B by every client $C_k \in \mathcal{C}$ to satisfy the power constraint in (6), sent coordinate-by-coordinate over d channel uses, and then scaled-up by B/\sqrt{dP} and averaged at the server before using it in a gradient descent procedure. It is not difficult to see the following upper bound.

Theorem V.3. Let π be the PSGD optimization algorithm and Q be the scaled transmission communication scheme described above. Then, we have

$$\sup_{(f, \mathcal{C}) \in \mathcal{O}} \mathcal{E}(f, \mathcal{C}, \pi, Q) \leq \frac{\sqrt{2}D}{\sqrt{KN}} \sqrt{d\sigma^2 + \frac{dB^2}{\text{SNR}}} + \frac{dLD^2}{2N}.$$

Remark 7. For $\text{SNR} \geq B^2/\sigma^2$, Theorem V.2 shows that compared to the centralized setting discussed in Theorem III.1, analog schemes will have a slowdown of \sqrt{d} . However, for small values of SNR, an analog communication scheme combined with PSGD gives close optimal performance. It matches the lower bound in Theorem III.3 up to a factor of B/σ . This observation follows by noting that $\log(1 + \text{SNR}) \approx \text{SNR}$ for small values of SNR.

VI. PROOFS

Difficult functions for lower bound

For our lower bounds, we use affine functions as difficult functions which are 0-smooth and are admissible in the class of L -smooth functions. These functions are considered in the same spirit as in showing the lower bounds for convex, smooth optimization under communication constraints [14]. We consider the domain $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \leq D/(2\sqrt{d})\}$, and consider the following class of functions on \mathcal{X} : For $v \in \{-1, 1\}^d$, let

$$f_v(x) := \frac{2\sigma\delta}{\sqrt{d}} \sum_{i=1}^d \left| x(i) - \frac{v(i)D}{2\sqrt{d}} \right|, \quad \forall x \in \mathcal{X},$$

and x_v^* be its minimizer. Note that the gradient of f_v at $x \in \mathcal{X}$ is independent of x , i.e., $-2\sigma\delta v/\sqrt{d}$. For any such f_v , each coordinate of d -dimensional noisy gradient $\hat{g}_{k,t}(i)$ takes $-\sigma/\sqrt{d}$ or σ/\sqrt{d} independently with probabilities $(1+2\delta v(i))/2$ and $(1-2\delta v(i))/2$, respectively. The parameter $\delta > 0$ is to be chosen later. Note that the above construction satisfies the set of assumptions in (3), (4) and (5).

A. Proof of Theorem III.3

Draw $V \sim \text{unif}\{-1, 1\}^d$. With respect to the associated random function f_V , each client C_i chooses a quantizer φ to generate output $\varphi(\hat{g}_{k,t})$. Denote by $Y^T = (Y_1, \dots, Y_T)$ the vector of ℓ -dimensional MAC channel outputs observed at the server. We follow a standard approach to reduce the underlying optimization problem to a multiple hypothesis problem of estimating V from output Y^T . Denote by $\mathbf{p}_{+i}^{Y^T}$ and $\mathbf{p}_{-i}^{Y^T}$ the distribution of Y^T given $V(i) = +1$ and $V(i) = -1$, respectively. We can relate the expected gap to optimality to the average total variational distance between $\mathbf{p}_{+i}^{Y^T}$ and $\mathbf{p}_{-i}^{Y^T}$ using the techniques from [36, Lemma 3, 4], which in turn builds on [37], [38]. In particular,

$$\begin{aligned} \mathbb{E}[f_V(\bar{x}_T) - f_V(x_V^*)] &= \sum_{i=1}^d \mathbb{E} \left[\frac{2\sigma\delta}{\sqrt{d}} \left| \bar{x}_T(i) - \frac{V(i)D}{2\sqrt{d}} \right| \right] \\ &\geq \frac{D\sigma\delta}{3d} \sum_{i=1}^d \mathbb{P} \left(\frac{2\sigma\delta}{\sqrt{d}} \left| x(i) - \frac{V(i)D}{2\sqrt{d}} \right| \geq \frac{D\sigma\delta}{3d} \right) \\ &\geq \frac{D\sigma\delta}{6} \left[1 - \frac{1}{d} \sum_{i=1}^d \mathbf{d}_{\text{TV}} \left(\mathbf{p}_{+i}^{Y^T}, \mathbf{p}_{-i}^{Y^T} \right) \right], \end{aligned}$$

where the first inequality is Markov's inequality and the second is the standard lower bound on probability of error in binary hypothesis testing under uniform prior. The average total-variational distance term can be further bounded using the convenient ‘‘plug-and-play’’ bound from [38, Theorem 2]

$$\begin{aligned} \left(\frac{1}{d} \sum_{i=1}^d d_{\text{TV}} \left(\mathbf{p}_{+i}^{Y^T}, \mathbf{p}_{-i}^{Y^T} \right) \right)^2 &\leq \frac{284T\delta^2}{d} \max_{v \in \{-1,1\}^d} \max_{\varphi \in \mathcal{Q}} I(\hat{g}_{1,1}, \dots, \hat{g}_{K,1} \wedge Y_1) \\ &\leq \frac{284T\delta^2}{d} \cdot \min \left(\max_{v \in \{-1,1\}^d} \max_{\varphi \in \mathcal{Q}} I(\varphi(\hat{g}_{1,1}), \dots, \varphi(\hat{g}_{K,1}) \wedge Y_1), Kd\ell \right), \end{aligned}$$

where the first inequality holds for $\delta \in (0, 1/6)$, and the second inequality uses $I(\hat{g}_{1,1}, \dots, \hat{g}_{K,1} \wedge Y_1) \leq H(I(\hat{g}_{1,1}, \dots, \hat{g}_{K,1} \wedge Y_1)) \leq Kd\ell$ and data-processing for the second. Further, we consider an auxiliary generative model for obtaining the MAC output Y_1 . Specifically, we assume K parallel Gaussian noise outputs given by

$$Y_{k,1}(i) = \varphi(\hat{g}_{k,1})(i) + Z_{k,1}(i), \quad i \in [\ell], k \in [K],$$

where $Z_{k,1} \sim \mathcal{N}(0, \sigma_z^2/K\mathbf{I}_\ell)$. Note that the output Y_1 and the sum of $\sum_{k \in [K]} Y_{k,1}$ are statistically equivalent. Thus, we can write

$$\begin{aligned} I(\varphi(\hat{g}_{1,1}), \dots, \varphi(\hat{g}_{K,1}) \wedge Y_1) &= I(\varphi(\hat{g}_{1,1}), \dots, \varphi(\hat{g}_{K,1}) \wedge Y_{1,1} + \dots + Y_{K,1}) \\ &\leq I(\varphi(\hat{g}_{1,1}), \dots, \varphi(\hat{g}_{K,1}) \wedge Y_{1,1}, \dots, Y_{K,1}) \\ &\leq \frac{K\ell}{2} \log(1 + \text{SNR}). \end{aligned}$$

Combining, we have for some universal constant c' ,

$$\mathbb{E}[f_V(\bar{x}_T) - f_V(x_V^*)] \geq \frac{D\sigma\delta}{3} \left[1 - \sqrt{\frac{c'KN\delta^2 \min(d, \frac{1}{2} \log(1 + \text{SNR}))}{d}} \right].$$

Setting $\delta = \sqrt{d/(2c' \min(2d, \log(1 + \text{SNR}))KN)}$, we finally get for some universal constant $c \in (0, 1)$.

$$\mathbb{E}[f_V(\bar{x}_T) - f_V(x_V^*)] \geq \frac{cD\sigma}{\sqrt{KN}} \sqrt{\frac{d}{\min(d, \frac{1}{2} \log(1 + \text{SNR}))}},$$

where we need $N \geq 18d/(c'''K \log(1 + \text{SNR}))$ in order to enforce $\delta \leq 1/6$. The proof is completed by noting that $\mathcal{E}^*(N, K, \text{SNR}) \geq \mathbb{E}[f_V(\bar{x}_T) - f_V(x_V^*)]$.

B. A general recipe for upper bounds

We now provide the general recipe to prove our upper bounds. For the minimum-distance decoder, denote by A_N the event where all the ASK constellation points sent in N channel uses are decoded correctly by the algorithm and by A_N^c its complement, i.e., $A_N^c := \cup_{t=1}^N \{|Z_t(1)| \geq 2\sqrt{P}/(r-1)\}$, where $Z_t(1)$ is defined in (7). We have

$$\begin{aligned} \mathbb{E}[f(\bar{x}_T) - f(x^*)] &= \mathbb{E}[(f(\bar{x}_T) - f(x^*)) | A_N] \cdot \mathbb{P}(A_N) + \mathbb{E}[(f(\bar{x}_T) - f(x^*)) | A_N^c] \cdot \mathbb{P}(A_N^c) \\ &\leq \mathbb{E}[(f(\bar{x}_T) - f(x^*)) | A_N] + DB \cdot \mathbb{P}(A_N^c). \end{aligned}$$

Using Chernoff's bound: $\mathbb{P}(A_N^c) \leq N \exp\left(-\frac{2K\text{SNR}}{(w^p-1)^2}\right) \leq \frac{1}{K\sqrt{N}}$, where the last line follows by setting $p = \lceil \log_w \left(\sqrt{\frac{2K\text{SNR}}{\ln(KN^{1.5})}} + 1\right) \rceil$. The first term under perfect decoding is bounded by calculating the performance measures $\alpha(Q)$ and $\beta(Q)$ for different schemes, and the proof is completed using Lemma III.2.

C. Proof of Theorem IV.2

Note that the number of channel uses $\ell = d/p$ per iteration. Under perfect decoding, $\lambda = \sum_k Q_k$ implying $\beta(Q) = 0$. Using conditional expectation, we also have

$$\mathbb{E} \left[\left\| \psi(Y_t) - \sum_{k \in [K]} \tilde{g}_{k,t} \right\|^2 \right] \leq \frac{4B^2d}{K(v-1)^2}.$$

Further, from (4) we have $\mathbb{E} \left[\left\| \sum_{k \in [K]} \tilde{g}_{k,t} - \nabla f(x_t) \right\|^2 \right] \leq \frac{\sigma^2}{K}$. At last, using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, setting $v = \sqrt{d} + 1$, and the fact $\sigma \leq B$, we get $\alpha^2(Q) \leq \frac{10B^2}{K}$.

D. Proof of Theorem IV.3

Based on the proof for [39, Lemma 4.1], we begin by a lemma capturing the performance of boosted DAQ estimator.

Lemma VI.1. *Given that $x, y \in [-M, M]^d$. For the boosted DAQ estimate \hat{X} in (9), we have*

$$\mathbb{E}[\hat{X}] = x \text{ and } \mathbb{E} \left[\|\hat{X} - x\|^2 \right] \leq (2M/I)\sqrt{d}\|x - y\|.$$

We have $\lambda(j) = \sum_{k \in \mathcal{C}_2} \sum_{i \in [I]} \mathbb{1}_{\{U_{k,i}(j) \leq \mathbf{R}\tilde{g}_k(j)\}}$ under perfect decoding, which implies

$$\mathbf{R}\psi = (2M/I) \sum_{j \in [d]} \sum_{k \in \mathcal{C}_2} \sum_{i \in [I]} \mathbb{1}_{\{U_{k,i}(j) \leq \mathbf{R}\tilde{g}_k(j)\}} e_j - \mathbb{1}_{\{U_{k,i}(j) \leq \mathbf{R}S(j)\}} e_j + (K/2)\mathbf{R}S.$$

Similar to the proof of Theorem IV.2, we first bound $\mathbb{E} [\|\psi(Y) - \sum_{k \in \mathcal{C}_2} \tilde{g}_k\|^2]$ using conditional expectation. Define $\psi_k := (2M/I)\mathbf{R}^{-1} \sum_{j \in [d]} \sum_{i \in [I]} (\mathbb{1}_{\{U_{k,i}(j) \leq \mathbf{R}\tilde{g}_k(j)\}} - \mathbb{1}_{\{U_{k,i}(j) \leq \mathbf{R}S(j)\}})e_j + S$, such that $\psi = \sum_k \psi_k$. We have

$$\begin{aligned}
& \mathbb{E} \left[\mathbb{E} \left[\left\| \sum_{k \in \mathcal{C}_2} \psi_k - \sum_{k \in \mathcal{C}_2} \tilde{g}_k \right\|^2 \middle| \mathbf{R} \right] \right] \\
&= \sum_{k \in \mathcal{C}_2} \mathbb{E} [\mathbb{E} [\|\psi_k - \tilde{g}_k\|^2 | \mathbf{R}]] + \sum_{k \neq k'} \mathbb{E} [\mathbb{E} [\langle \psi_k - \tilde{g}_k, \psi_{k'} - \tilde{g}_{k'} \rangle | \mathbf{R}]] \\
&= \sum_{k \in \mathcal{C}_2} \mathbb{E} [\|\psi_k - \tilde{g}_k\|^2] + \sum_{k \neq k'} \mathbb{E} [\langle \mathbb{E} [\psi_k - \tilde{g}_k | \mathbf{R}], \mathbb{E} [\psi_{k'} - \tilde{g}_{k'} | \mathbf{R}] \rangle] \\
&= \sum_{k \in \mathcal{C}_2} \mathbb{E} [\|\psi_k - \tilde{g}_k\|^2] + \mathbb{E} \left[\left(\sum_{k \in \mathcal{C}_2} \|\mathbb{E} [\psi_k - \tilde{g}_k | \mathbf{R}]\| \right)^2 \right] - \mathbb{E} \left[\sum_{k \in \mathcal{C}_2} \|\mathbb{E} [\psi_k - \tilde{g}_k | \mathbf{R}]\|^2 \right] \\
&\leq \sum_{k \in \mathcal{C}_2} \mathbb{E} [\|\psi_k - \tilde{g}_k\|^2] + \frac{K}{2} \sum_{k \in \mathcal{C}_2} \mathbb{E} [\|\mathbb{E} [\psi_k - \tilde{g}_k | \mathbf{R}]\|^2], \tag{11}
\end{aligned}$$

where the second equality uses the fact that $\psi_k - \tilde{g}_k$ and $\psi_{k'} - \tilde{g}_{k'}$ are independent given \mathbf{R} , and the last line uses Jensen's inequality. Now consider an event $\mathcal{V}_j = \{|\mathbf{R}S(j)| \leq M, |\mathbf{R}\tilde{g}_k(j)| \leq M\}$ and use unitary property of \mathbf{R} to write the first term in (11) as

$$\sum_{k \in \mathcal{C}_2} \mathbb{E} [\|\mathbf{R}\psi_k - \mathbf{R}\tilde{g}_k\|^2] \tag{12}$$

$$= \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} [(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j}] + \mathbb{E} [(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c}] \tag{13}$$

The first term in (12) is bounded as

$$\begin{aligned}
\sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} [(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j}] &\leq \sum_{k \in \mathcal{C}_2} \mathbb{E} [\|\mathbf{R}\psi_k - \mathbf{R}\tilde{g}_k\|^2] \\
&= \sum_{k \in \mathcal{C}_2} \mathbb{E} [\mathbb{E} [\|\mathbf{R}\psi_k - \mathbf{R}\tilde{g}_k\|^2 | S, \hat{g}_k]] \\
&\leq \frac{2M\sqrt{d}}{I} \sum_{k \in \mathcal{C}_2} \mathbb{E} [\|\mathbf{R}\tilde{g}_k - \mathbf{R}S\|] \\
&\leq \frac{2M\sqrt{d}}{I} \sum_{k \in \mathcal{C}_2} \sqrt{\mathbb{E} [\|\mathbf{R}\tilde{g}_k - \mathbf{R}S\|^2]} \\
&\leq \frac{4M\sqrt{d}}{KI} \sqrt{\frac{4\sigma^2}{K} + \frac{32B^2d}{K(v-1)^2} + \sigma^2} \\
&\leq \frac{4\sqrt{3}M\sqrt{d}\sigma}{I},
\end{aligned}$$

where the first inequality uses $\mathbb{1}_{\mathcal{V}_j} \leq 1$, the second inequality is due to Lemma VI.1, the third inequality is Jensen's inequality, the fourth one uses the fact that $(K/2)S$ is the output of UQ – OTA along with the proof of Theorem (IV.2), and the last inequality holds for parameters $v = 7, K \geq B^2d/\sigma^2$. For the second term in (12), each summand is bounded as

$$\begin{aligned} \mathbb{E} \left[(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \right] &\leq 8M^2 \cdot \mathbb{P}(\mathcal{V}_j^c) + 2\mathbb{E} \left[(\mathbf{R}(S - \tilde{g}_k)(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \right] \\ &= 8M^2 \cdot \mathbb{P}(\mathcal{V}_j^c) + 2\mathbb{E} \left[(\mathbf{R}(S - \tilde{g}_k)(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \cdot \mathbb{1}_{\mathcal{V}_j'} \right] \\ &\quad + 2\mathbb{E} \left[(\mathbf{R}(S - \tilde{g}_k)(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \cdot \mathbb{1}_{\mathcal{V}_j'^c} \right] \\ &\leq 8M^2 \cdot \mathbb{P}(\mathcal{V}_j^c) + 2M^2 \cdot \mathbb{P}(\mathcal{V}_j^c) + 2\mathbb{E} \left[(\mathbf{R}(S - \tilde{g}_k)(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j'^c} \right] \end{aligned}$$

where the first inequality uses the definition of ψ_k along with $(a + b)^2 \leq 2(a^2 + b^2)$, and the second equality follows from considering a new event $\mathcal{V}_j' = \{|\mathbf{R}(S - \tilde{g}_k)(j)| \leq M\}$ and $\mathbb{1}_{\mathcal{V}_j'^c} \leq 1$. Note that for first two terms above $\mathbb{P}(\mathcal{V}_j^c) \leq 4e^{-\frac{dK^2M^2}{8B^2}}$. For the third term above, we note that $\mathbf{R}(S - \tilde{g}_k)(j)$ is sub-Gaussian with variance factor $\frac{16B^2}{dK^2}$ (see for instance, [21, Lemma V.8]) and use the following concentration result.

Lemma VI.2. [39, Lemma 8.1] *For a sub-Gaussian random Z with variance factor σ^2 and every $t \geq 0$, we have*

$$\mathbb{E} \left[Z^2 \mathbb{1}_{\{|Z|>t\}} \right] \leq 2(2\sigma^2 + t^2)e^{-t^2/2\sigma^2}.$$

Using $e^{-\frac{dK^2M^2}{8B^2}} \leq e^{-\frac{dK^2M^2}{32B^2}}$ and the Lemma VI.2, we now have

$$\begin{aligned} \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} \left[(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \right] &\leq \left(\frac{64B^2}{K} + 22M^2dK \right) e^{-\frac{dK^2M^2}{32B^2}} \\ &\leq 26\delta^2, \end{aligned} \tag{14}$$

where for the last inequality, we choose $M^2 = \frac{c_2B^2}{K^2d} \ln \left(\frac{c_2B}{\sqrt{K}\delta} \right)$, for $\delta \in \left(0, \frac{c_2B}{\sqrt{K}} \right)$. Further, the second term in (11) is 0 under \mathcal{V} and can be bounded under \mathcal{V}^c using the Jensen's inequality as

$$\begin{aligned} \sum_{k \in \mathcal{C}_2} \mathbb{E} \left[\|\mathbb{E}[\psi_k - \tilde{g}_k | \mathbf{R}]\|^2 \right] &= \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} \left[(\mathbb{E}[\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j) | \mathbf{R}])^2 \right] \\ &= \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} \left[(\mathbb{E}[\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j) | \mathbf{R}])^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \right] \\ &\leq \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} \left[\mathbb{E} \left[(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 | \mathbf{R} \right] \cdot \mathbb{1}_{\mathcal{V}_j^c} \right] \\ &= \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} \left[(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \right] \end{aligned}$$

$$\leq 26\delta^2,$$

where the first equality uses the unitary property of \mathbf{R} , the second uses the fact that boosted DAQ yields an unbiased estimate (see Lemma VI.1), the first inequality follows from Jensen's inequality, and the last line uses the bound in (14). Finally, we set $\delta = \frac{c_2 B}{K^2 N}$. Combining all above in (11) and using (4), we get

$$\begin{aligned} \alpha^2(Q) &= \frac{4\sqrt{3}M\sigma\sqrt{d}}{I} + 26\delta^2 + 13K\delta^2 + \frac{2\sigma^2}{K} \\ &\leq \frac{4\sqrt{3}B\sigma}{K} + \frac{3\sigma^2}{K}, \end{aligned}$$

where the last line holds whenever $K \geq B^2 d / \sigma^2$ and by choosing $I = \left\lceil \sqrt{c_2 \ln\left(\frac{c_2 B}{\sqrt{K}\delta}\right)} \right\rceil$. Also,

$$\begin{aligned} \beta^2(Q) &= \|\mathbb{E}[\psi(Y)] - \sum_{k \in \mathcal{C}_2} \tilde{g}_k\|^2 \\ &\leq \left(\sum_{k \in \mathcal{C}_2} \|\mathbb{E}[\mathbf{R}\psi_k - \mathbf{R}\tilde{g}_k]\| \right)^2 \\ &\leq K \sum_{k \in \mathcal{C}_2} \|\mathbb{E}[\mathbf{R}\psi_k - \mathbf{R}\tilde{g}_k]\|^2 \\ &= K \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} \left[(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j)) \cdot \mathbb{1}_{\mathcal{V}_j^c} \right]^2 \\ &\leq K \sum_{k \in \mathcal{C}_2} \sum_{j \in [d]} \mathbb{E} \left[(\mathbf{R}\psi_k(j) - \mathbf{R}\tilde{g}_k(j))^2 \cdot \mathbb{1}_{\mathcal{V}_j^c} \right] \\ &\leq 9K\delta^2 \\ &= \frac{9c_5 B^2}{K^3 N^2}, \end{aligned}$$

where the first identity follows from (3), the first inequality is triangle inequality, the second inequality uses Cauchy-Schwarz, the second identity uses unbiased nature of estimate under \mathcal{V}_j , the third inequality is Jensen's, and the last inequality uses the bound in (14). Since the final output is obtained in $d/p + d/p'$ channel uses per iteration, where $p = \left\lceil \log_w \left(1 + \sqrt{\frac{K \text{SNR}}{2 \ln(KN^{1.5})}} \right) \right\rceil$ and $p' = \left\lceil \log_{w'} \left(1 + \sqrt{\frac{K \text{SNR}}{2 \ln(KN^{1.5})}} \right) \right\rceil$ with $w = 3K + 1, w' = KI/2 + 1$. Using Lemma III.2, the proof is completed.

VII. EXPERIMENTS

We demonstrate the performance of our proposed digital schemes UQ-OTA and WZ-OTA for the following mean estimation task under MAC constraints.

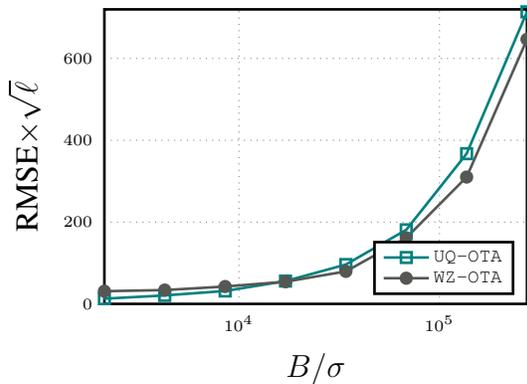


Fig. 1: Comparison of UQ-OTA and WZ-OTA at SNR = 50dB and $d = 32$.

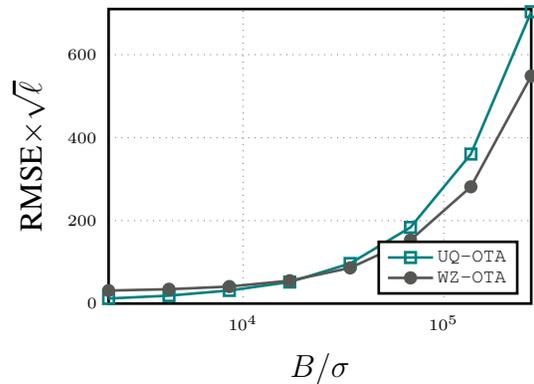


Fig. 2: Comparison of UQ-OTA and WZ-OTA at SNR = 75dB and $d = 32$.

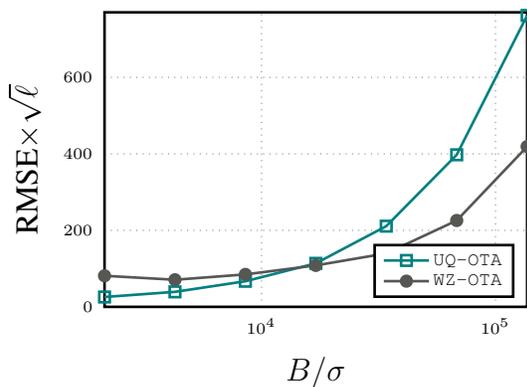


Fig. 3: Comparison of UQ-OTA and WZ-OTA at SNR = 100dB and $d = 64$.

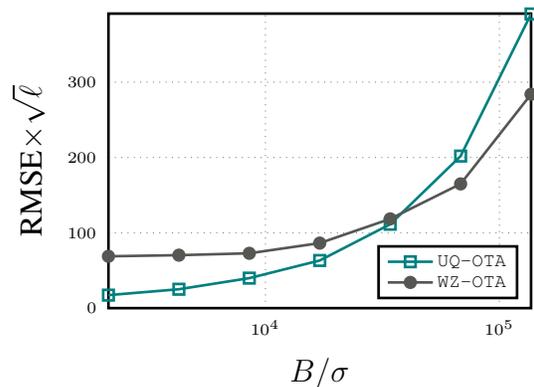


Fig. 4: Comparison of UQ-OTA and WZ-OTA at SNR = 180dB and $d = 64$.

Each client C_k has a d -dimensional vector \hat{g}_k , defined as $\hat{g}_k = \mu + U_k^c$, where $\mu \in [-1, 1]^d$ is a constant mean vector, and U_k^c is a random vector whose coordinates are independently drawn from a uniform random variable denoted as $\text{unif}(-\sigma', \sigma')$. Note that $\mathbb{E}[\hat{g}_k] = \mu$ and $\mathbb{E}[\|\hat{g}_k - \mu\|^2] = \frac{d\sigma'^2}{3}$. The goal is to recover the sample average $\bar{g} = \frac{1}{K} \sum_{k \in K} \hat{g}_k$ which is an unbiased estimate for the mean vector μ .

We compare the two proposed digital over-the-air schemes UQ-OTA and WZ-OTA for estimating \bar{g} . We evaluate the performance of our proposed schemes by RMSE between the μ and the estimated sample average vector $\hat{\bar{g}}$ formed by the server. We then plot a combined error metric which is the product of RMSE and the square root of number of MAC channel transmissions. Such a metric, in essence with the Lemma III.2, is a crucial term contributing to the overall

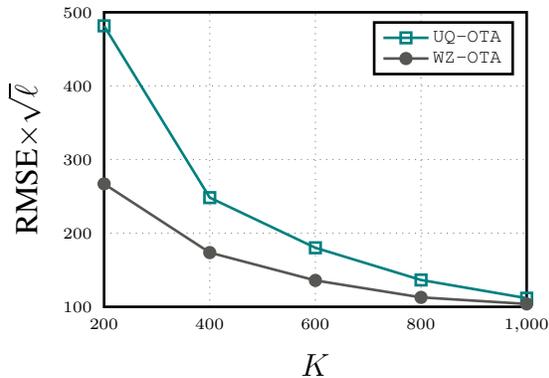


Fig. 5: Comparison of UQ-OTA and WZ-OTA at SNR = 100dB, $B/\sigma = 1.36 \times 10^5$, and $d = 64$.

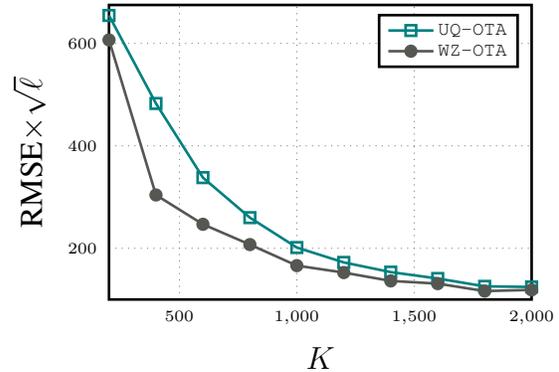


Fig. 6: Comparison of UQ-OTA and WZ-OTA at SNR = 180dB, $B/\sigma = 1.36 \times 10^5$, and $d = 64$.

performance analysis. Our codes are implemented in Python language and are available online on GitHub [40].

We fix the number of clients $K = 500$ and conduct the experiments for dimensions $d = 32$ at SNR = 50dB and 75dB, and for dimensions $d = 64$ at SNR = 100dB and 180dB. We fix the value of $\sigma' = 0.05196$ which gives a valid choice for σ to be $0.03\sqrt{d}$. For all of these experiments, we vary the B/σ for different choices of B . All the experiments are averaged over 20 runs for statistical consistency.

In Figures 1, 2, 3, and 4, we observe that the WZ-OTA outperforms the UQ-OTA for large values of B/σ . However, for lower values of B/σ , UQ-OTA performs better than WZ-OTA. This observation is in accordance with the Remark 5.

In other direction, we note that for both the schemes, there is decrease in the error performance for the same B/σ and dimension d as the SNR increases. This is because of better channel decoding with increasing operating SNR.

Next, we demonstrate the error performance of both the digital schemes for increasing number of clients. Specifically, we fix the values of ratio $B/\sigma = 1.36 \times 10^5$ and the dimension $d = 64$. In Figure 5, we show the error performances of our schemes at SNR = 100dB and for the number of clients $K = 200, 400, 600, 800$, and 1000. In Figure 6, we show the same at SNR = 180dB and for the number of clients $K = 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800$, and 2000. As

can be seen from both the figures, the error performance decreases with increase in the number of clients.

VIII. CONCLUSION

We provide an almost complete characterization of the min-max convergence rate of over-the-air distributed optimization. Our bounds show that a simple analog coding scheme is optimal at low values of SNR, but they can be far from optimal at high values of SNR (Remark 7). This observation mirrors the observation made by [13], albeit in the single client setting. Furthermore, we design an explicit digital communication scheme based on lattice coding to match our lower bound for all values of SNR. We hope our work inspires other explicit communication schemes for similar distributed optimization problems. Our upper bound matches our lower bound up to a nominal $\sqrt{\log K + \log \log N}$ factor (Theorem IV.3). Further closing the gap between our upper and lower bound would lead to new communication schemes or lower bound techniques for distributed optimization and is an exciting research direction.

ACKNOWLEDGEMENT

The author would like to thank Himanshu Tyagi, Prathamesh Mayekar, and Naina Nagpal for many useful discussions and help with problem formulation and proof ideas.

REFERENCES

- [1] S. K. Jha and P. Mayekar, “Fundamental limits of distributed optimization over multiple access channel,” in *2023 IEEE Information Theory Workshop (ITW)*, pp. 406–411, 2023.
- [2] M. M. Amiri and D. Gündüz, “Over-the-Air Machine Learning at the Wireless Edge,” in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2019.
- [3] M. M. Amiri and D. Gündüz, “Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air,” in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1432–1436, 2019.
- [4] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.
- [5] A. Sahin and R. Yang, “A survey on over-the-air computation,” <https://arxiv.org/abs/2210.11350>, Nov 2022.
- [6] W.-T. Chang and R. Tandon, “Communication Efficient Federated Learning over Multiple Access Channels,” <https://arxiv.org/abs/2001.08737>, 2020.
- [7] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, “COTAF: Convergent Over-the-Air Federated Learning,” in *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2020.
- [8] 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.

- [9] 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.
- [10] J. Zhang, N. Li, and M. Dedeoglu, “Federated Learning over Wireless Networks: A Band-limited Coordinated Descent Approach,” <https://arxiv.org/abs/2102.07972>, 2021.
- [11] G. Zhu, Y. Du, D. Gündüz, and K. Huang, “One-Bit Over-the-Air Aggregation for Communication-Efficient Federated Edge Learning: Design and Convergence Analysis,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 2120–2135, 2021.
- [12] R. Saha, S. Rini, M. Rao, and A. Goldsmith, “Decentralized optimization over noisy, rate-constrained networks: How we agree by talking about how we disagree,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5055–5059, IEEE, 2021.
- [13] S. K. Jha, P. Mayekar, and H. Tyagi, “Fundamental limits of over-the-air optimization: Are analog schemes optimal?,” *IEEE Journal on Selected Areas in Information Theory*, vol. 3, no. 2, pp. 217–228, 2022.
- [14] P. Mayekar, S. K. Jha, A. T. Suresh, and H. Tyagi, “Wyner-ziv estimators for distributed mean estimation with side information and optimization,” <https://arxiv.org/abs/2011.12160.v2>, 2022.
- [15] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, “QSGD: Communication-efficient SGD via gradient quantization and encoding,” *Advances in Neural Information Processing Systems*, pp. 1709–1720, 2017.
- [16] V. Gandikota, D. Kane, R. Kumar Maity, and A. Mazumdar, “vqsgd: Vector quantized stochastic gradient descent,” in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, pp. 2197–2205, PMLR, 2021.
- [17] D. Basu, D. Data, C. Karakus, and S. Diggavi, “Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations,” *Advances in Neural Information Processing Systems*, 2019.
- [18] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, “1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns,” *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [19] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright, “Atomo: Communication-efficient learning via atomic sparsification,” *Advances in Neural Information Processing Systems*, pp. 9850–9861, 2018.
- [20] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, “TernGrad: Ternary gradients to reduce communication in distributed deep learning,” *Advances in Neural Information Processing Systems*, pp. 1509–1519, 2017.
- [21] P. Mayekar and H. Tyagi, “RATQ: A universal fixed-length quantizer for stochastic optimization,” *IEEE Transactions on Information Theory*, 2020.
- [22] C.-Y. Lin, V. Kostina, and B. Hassibi, “Differentially Quantized Gradient Descent,” in *IEEE International Symposium on Information Theory (ISIT)*, 2021.
- [23] P. Mayekar and H. Tyagi, “Limits on gradient compression for stochastic optimization,” *Proceedings of the IEEE International Symposium of Information Theory (ISIT’ 20)*, 2020.
- [24] D. Jhunjunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, “Adaptive quantization of model updates for communication-efficient federated learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3110–3114, 2021.
- [25] A. Ghosh, R. K. Maity, and A. Mazumdar, “Distributed newton can communicate less and resist byzantine workers,” *Advances in Neural Information Processing Systems*, 2020.
- [26] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, “Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization,” in *International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031, PMLR, 2020.

- [27] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, “Distributed mean estimation with limited communication,” *Proceedings of the International Conference on Machine Learning (ICML’ 17)*, vol. 70, pp. 3329–3337, 2017.
- [28] W.-N. Chen, P. Kairouz, and A. Özgür, “Breaking the communication-privacy-accuracy trilemma,” *Neural Information Processing Systems (NeurIPS)*, 2020.
- [29] J. Acharya, C. Canonne, P. Mayekar, and H. Tyagi, “Information-constrained optimization: can adaptive processing of gradients help?,” in *Advances in Neural Information Processing Systems*, vol. 34, pp. 7126–7138, 2021.
- [30] J. Acharya, C. De Sa, D. J. Foster, and K. Sridharan, “Distributed Learning with Sublinear Communication,” *International Conference on Machine Learning*, 2019.
- [31] Z. Huang, W. Yilei, K. Yi, *et al.*, “Optimal sparsity-sensitive bounds for distributed mean estimation,” *Advances in Neural Information Processing Systems*, pp. 6371–6381, 2019.
- [32] J. Konečný and P. Richtárik, “Randomized distributed mean estimation: Accuracy vs. communication,” *Frontiers in Applied Mathematics and Statistics*, vol. 4, p. 62, 2018.
- [33] P. Davies, V. Gurusamy, N. Moshrefi, S. Ashkboos, and D.-A. Alistarh, “New bounds for distributed mean estimation and variance reduction,” in *9th International Conference on Learning Representations*, 2021.
- [34] S. Bubeck, “Convex optimization: Algorithms and complexity,” *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.
- [35] K. J. Horadam, *Hadamard matrices and their applications*. Princeton university press, 2012.
- [36] J. Acharya, C. L. Canonne, P. Mayekar, and H. Tyagi, “Information-constrained optimization: can adaptive processing of gradients help?,” <https://arxiv.org/abs/2104.00979>, 2021.
- [37] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar, “Information-theoretic lower bounds on the oracle complexity of convex optimization,” *Advances in Neural Information Processing Systems*, pp. 1–9, 2009.
- [38] J. Acharya, C. L. Canonne, Z. Sun, and H. Tyagi, “Unified lower bounds for interactive high-dimensional estimation under information constraints,” <http://arxiv.org/abs/2010.06562v5>, 2020.
- [39] P. Mayekar, A. T. Suresh, and H. Tyagi, “Wyner-Ziv estimators: Efficient distributed mean estimation with side-information,” in *International Conference on Artificial Intelligence and Statistics*, pp. 3502–3510, PMLR, 2021.
- [40] S. Jha, “Over-the-air optimization.” https://github.com/shubhamjha-46/OTA_MAC/. [Online; accessed 08-Sept-2023].

APPENDIX

MATHEMATICAL DETAILS CONCERNING REMARKS 4 AND 5

a) *Sub-optimality of UQ-OTA at high SNR:* Let $x = \text{SNR}$, $y = \frac{K}{2 \ln(KN^{1.5})}$. Then, we can write

$$\begin{aligned}
 p &\geq \frac{\log(1 + \sqrt{xy})}{\log(Kd)} - 1 \\
 &= \frac{\log(1 + \sqrt{x})}{\log(Kd)} + \frac{\log\left(\frac{1}{1+\sqrt{x}} + \frac{\sqrt{xy}}{1+\sqrt{x}}\right)}{\log(Kd)} - 1 \\
 &\geq \frac{\log(1 + \sqrt{x})}{\log(Kd)} + \frac{\log(\min(1, \sqrt{y}))}{\log(Kd)} - 1 \\
 &\geq \frac{\log(1 + \sqrt{x})}{\log(Kd)} - 1
 \end{aligned}$$

$$\geq \frac{\frac{1}{2} \log(1+x)}{\log(Kd)} - 1,$$

where the first inequality holds due to the fact that $\lfloor p \rfloor \geq p-1$, the second inequality is due to the term inside the second convex combination of points 1 and \sqrt{y} , the third inequality holds for K sufficiently large satisfying $y \geq 1$, and the last one holds as $\log(1+\sqrt{x}) = \frac{1}{2} \log(1+x+2\sqrt{x}) \geq \frac{1}{2} \log(1+x)$.

b) Approximation of p, p' at large K, N and SNR: At large N , we have $w \leq w'$ which implies $p \geq p'$ and thus $q \geq p'/2$. Again, considering $x = \text{SNR}$, $y = \frac{K}{2 \ln(KN^{1.5})}$ and proceeding as earlier, we can show that $p' \approx \frac{\frac{1}{2} \log(1+\text{SNR})}{\log(KI)}$ for large $\text{SNR} \geq 2(2^d - 1)$ regime.