Fundamental limits of over-the-air optimization: Are analog schemes optimal?

Shubham K Jha* Prathamesh Mayekar† Himanshu Tyagi*†

***Abstract*—We consider convex optimization on a d dimensional space where coded gradients are sent over an additive Gaussian noise channel with variance σ^2. The codewords satisfy an average power constraint P, resulting in the signal-to-noise ratio (SNR) of P/σ^2. Many schemes have been proposed for this problem, termed over-the-air optimization, in recent years. We present lower and upper bounds for the convergence rates for over-the-air optimization. Our first result is a lower bound for the convergence rate showing that any code must slowdown the convergence rate by a factor of roughly $\sqrt{d/\log(1+\mathtt{SNR})}$. Next, we consider a popular class of schemes called *analog coding*, where a linear function of the gradient is sent. We show that a simple scaled transmission analog coding scheme results in a slowdown in convergence rate by a factor of $\sqrt{d(1+1/\mathtt{SNR})}$. This matches the previous lower bound up to constant factors for low SNR, making the scaled transmission scheme optimal at low SNR. However, we show that this slowdown is necessary for any analog coding scheme. In particular, a slowdown in convergence by a factor of $\sqrt{d}$ remains even when SNR tends to infinity, a clear shortcoming of analog coding schemes at high SNR. Remarkably, we present a simple quantize-and-modulate scheme that uses *Amplitude Shift Keying* and almost attains the optimal convergence rate at all SNRs.**

I. Introduction

Distributed optimization is a classic topic with decades of work building basic theory. The last decade has seen increased interest in this topic motivated by distributed and large scale machine learning. For instance, parallel implementation of training algorithms for deep learning models over multi-GPU has become commonplace. In another direction, over the past 5 years or so, federated learning applications that require building machine learning models for data distributed across multiple users have motivated optimization algorithms that limit communication from the users to a parameter server ($cf.$ [12]). Most recently, there has been a lot of interest in the scenario where this communication is *over-the-air*, namely the users are connected over a wireless communication channel ($cf.$ [6], [7]).

Many different optimization algorithms have been proposed using different kind of codes. However, there is no work addressing information-theoretic limits on the performance of these algorithms. In particular, it remains unclear whether simple analog schemes for communication over AWGN channel are optimal in any setting and whether there is any fundamental limitation to their performance. More broadly, do we still need sophisticated error-correcting codes to attain the optimal convergence rate for the optimization problem? In this work, we address these questions for convex optimization problems.

We establish an information-theoretic lower bound on the convergence rate for any scheme for convex stochastic optimization which shows that, for $d-$dimensional domain, there is a $\sqrt{d/\log(1+\mathtt{SNR})}$ factor slowdown in convergence rate. Furthermore, for low SNR, analog codes with stochastic gradient descent (SGD) attain this optimal rate. Next, we establish a general lower bound on the performance of analog codes and show that there is a factor $\sqrt{d(1+1/\mathtt{SNR})}$ slowdown in convergence rate when analog codes are used. Note that as SNR goes to infinity one can expect that the convergence rate should tend to the classic one. But our bound shows that for analog codes, there is at least a factor $\sqrt{d}$ slowdown even as the SNR tends to infinity, making them suboptimal at high SNR. Finally, we show that a simple quantize-and-modulate SGD scheme that uses a vector quantizer for the gradients and sends the quantized values using amplitude shift keying (ASK) is almost rate optimal.

There has been a very interesting line of work on these topics, including [6], [7], [9], [17]–[19], [21]–[23]. Most works have considered the multiparty setting, with more complicated channels than AWGN. In this paper, for simplicity, we restrict to the two-terminal setting. But our qualitative results apply to the multiparty setting as well.

Broadly, the gradient coding schemes proposed in these works can be divided into two categories: analog and digital. In more detail, in analog schemes, the coded gradients sent over the noisy channel are a linear transformation of the subgradient supplied by the oracle. Typical analog schemes include scaling, sparsification, or direct transmission of gradients over a wireless channel. For instance, authors in [7] send only top k gradient coordinates along with error feedback. In [18], the subgradient estimates are scaled-down appropriately to satisfy the power constraint. Each coordinate is then transmitted over the Gaussian channel using one channel-use per transmission. Similar scaling approaches are also presented in [19], [21]–[23]. On the other hand, digital schemes rely on gradient quantization and channel coding. For instance, authors in [9] propose to quantize the subgradients using stochastic quantization and the precision is chosen so that the transmission rate is the same as channel capacity. Then they are transmitted using any capacity-achieving code. In [23], authors perform one-bit quantization of subgradients similar

*Robert Bosch Center for Cyber-Physical Systems, Indian Institute of Science, Bangalore, India.

†Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India.

Email: {shubhamkj, prathamesh, htyagi}@iisc.ac.in

The detailed proofs of all the results in this paper are available in the extended version [11].

GLOBECOM 2021 - 2021 IEEE Global Communications Conference | 978-1-7281-8104-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/GLOBECOM46510.2021.9685770

to signSGD [8] and send them over-the-air using OFDM modulation, taking into account the frequency selective-fading and inter-symbol interference.

In summary, most of the prior work either use analog schemes or capacity achieving channel codes. Further, even works such as [23] which use a quantize-and-modulate approach like our work, do not comment on the optimality of the rate of convergence. In fact, in our proposed scheme, we use a one dimensional signal constellation and let the number of bits used to quantize grow roughly as $\log(1+\texttt{SNR})$ to get optimal convergence rate.

In a slightly different direction, the variant of distributed optimization with compressed subgradient estimates has also been studied extensively, primarily to mitigate the slowdown in convergence of distributed optimization procedures when full gradients are communicated (see, for instance, [4], [13]–[15], [20]). We build on the quantizers proposed in these works to obtain a nearly optimal convergence rate algorithm.

For our lower bounds, we follow a similar strategy as [2] (which in turn builds on [1], [3]) where optimization under communication constraints (not over-the-air) was considered. While the difficult oracles of these prior works yield our general lower bound, for deriving the limitation for analog schemes, we consider a new class of Gaussian oracles; see Section IV for more details.

The rest of the paper is organized as follows. We setup the problem in the next section and provide all our main results in Sections III. All the proofs are given in Section IV and concluding remarks are in Section V.

II. Problem Formulation and Preliminaries

A. Functions and gradient oracles

For a convex set $\mathcal{X} \subset \mathbb{R}^d$ with $\sup_{x,y\in\mathcal{X}} \|x-y\| \le D$, we consider the minimization of an unknown convex function $f:\mathcal{X}\to\mathbb{R}$ using access to a first order *oracle* O that reveals noisy subgradient estimates for any queried point. We assume that the oracle outputs $\hat{g}(x)$ when a point $x\in\mathcal{X}$ is queried satisfy the following conditions:

$$\mathbb{E}\left[\hat{g}(x)|x\right] \in \partial f(x), \quad \text{(unbiasedness)} \tag{1}$$

$$\mathbb{E}\left[\|\hat{g}(x)\|^2|x\right] \le B^2, \quad \text{(mean square bounded oracle)} \tag{2}$$

where $\partial f(x) \subset \mathbb{R}^d$ denotes the set of subgradients of f at input x. Denote by $\mathcal{O}$ the set of pairs (f,O) of functions and oracles satisfying the conditions above.

B. Codes and Gaussian channel

In our setting, the gradient estimates are not directly available to the optimization algorithm π but must be coded for error correction, sent over a noisy channel, and decoded to be used by π. We consider fixed length codes of length ℓ with average power less than P. Specifically, we consider (d,ℓ,P)-codes consisting of encoder mappings $\varphi:\mathbb{R}^d\times\mathcal{U}\to\mathbb{R}^\ell$ such that the codeword $\varphi(\hat{g},U)\in\mathbb{R}^\ell$ used to send the subgradient estimate $\hat{g}\in\mathbb{R}^d$ satisfies the average power constraint

$$\mathbb{E}\left[\|\varphi(\hat{g},U)\|^2\right] \le \ell P, \tag{3}$$

where $U\in\mathcal{U}$ denotes the public randomness used to randomize the encoder and is assumed to be available to both φ and optimization algorithm π. For convenience, we drop the argument U from the notation of φ for the rest of the paper. Denote by $\mathcal{C}_\ell$ the set of all (d,ℓ,P)-codes.

After the tth query by the algorithm, when the oracle supplies a subgradient estimate $\hat{g}_t$, the codeword $C_t=\varphi(\hat{g}_t)$ is sent over an *additive Gaussian noise* channel. That is, after the tth query to the oracle, the algorithm π observes $Y_t\in\mathbb{R}^\ell$ given by

$$Y_t(i) = C_t(i) + Z_t(i),\ 1\le i\le \ell, \tag{4}$$

where $\{Z_t(i)\}_{i\in[\ell],t\in\mathbb{N}}$ is a sequence of i.i.d. random variables with common distribution $(0,\sigma^2)$ – the Gaussian distribution with mean 0 and variance σ^2. We denote the *signal-to-noise ratio* by $\texttt{SNR} := P/\sigma^2$.

C. Over-the-air Optimization

We now describe an optimization algorithm π using (d,ℓ,P)-code φ. In any iteration t, the optimization algorithm π, upon observing the previous channel outputs $Y_1,...,Y_{t-1}\in\mathbb{R}^\ell$, queries the oracle with point[1] x_t. The oracle gives $\hat{g}_t\in\partial f(x_t)$, encodes it as $\varphi(\hat{g})$ and sends it over the Gaussian channel. The algorithm π observes the output $Y_t\in\mathbb{R}^\ell$ of the channel and moves to iteration $t+1$.

After T iterations, the algorithm outputs x_T. Denote by $\Pi_{\ell,T}$ the class of all algorithms using a (d,ℓ,P)-code and making T oracle queries.

We abbreviate the overall algorithm π with access to oracle O and using encoder φ by $\pi^{\varphi O}$. We call the tuple (π,φ) consisting of the optimization algorithm and the encoding procedure φ as an *over-the-air optimization protocol.* The convergence error of this over-the-air optimization protocol is given by

$$\mathcal{E}(f,\pi^{\varphi O}) := \mathbb{E}\left[f(x_T)\right] - \min_{x\in\mathcal{X}} f(x).$$

We want to study how the convergence error goes to zero as a function of the total number of channel uses $N=T\ell$. We are allowed to use codes with any length ℓ but note that increase in the length of encoding protocol will lead to decrease in the number of oracle queries as the number of channel uses is restricted to N. Similarly, while we are allowed to use optimization algorithm which can make as many as N queries to the oracle, increase the number of queries will lead to a smaller block length encoding protocol. Let $\Lambda(N) := \{\pi\in\Pi_{\ell,T},\varphi\in\mathcal{C}_\ell\colon \ell\cdot T\le N\}$. That is, $\Lambda(N)$ is the set of all over-the-air optimization protocols using N channel transmissions. Then, the smallest worst-case convergence error possible by using N channel transmissions is given by $\mathcal{E}^*(N,\mathcal{X}) := \inf_{(\pi,\varphi)\in\Lambda(N)}\sup_{(f,O)\in\mathcal{O}}\mathcal{E}(f,\pi^{\varphi O})$.

[1] We assume that the downlink communication channel from the algorithm to the oracle is noiseless.

Let $\mathbb{X} := \{\mathcal{X}\colon \sup_{x,y\in\mathcal{X}} \|x-y\| \le D\}$. In this paper, we will characterize the following quantity[2]:

$$\mathcal{E}^*(N) := \sup_{\mathcal{X}\in\mathbb{X}} \mathcal{E}^*(N,\mathcal{X}). \tag{5}$$

D. Special coding schemes

In addition to the general coding scheme above, we are interested in the following two special classes of simple coding schemes: Analog codes and ASK codes.

Definition II.1. A code is an *analog code* if the encoder mapping φ is linear, i.e., when $\varphi(x) = \mathbf{A}x$ for an $\ell \times d$ matrix $\mathbf{A}$, for any $\ell \le d$. We allow for random matrices $\mathbf{A}$ as long as they are independent of the observed gradient estimates. Also, we denote by $\mathcal{E}^*_{analog}(N)$ the minmax optimization error when the class of (d,ℓ,P)-encoding protocol is restricted to analog schemes (with everything else remaining the same as in (5)). Clearly, $\mathcal{E}^*_{analog}(N) \ge \mathcal{E}^*(N)$.

Definition II.2. A code is an[3] *Amplitude Shift Keying (ASK) code* satisfying the average power constraint (3) if the range of the encoder mapping is given by

$$\left\{-\sqrt{P} + \frac{(k-1)\cdot 2\sqrt{P}}{2^r-1}\colon k \in [2^r]\right\},$$

for some $r \in \mathbb{N}$. Namely, the encoder first quantizes $\hat{g}$ to r bits and then uses ASK modulation for sending the quantized subgradient estimate. Note that this is a code of length 1.

E. A benchmark from prior results

We recall results for the case $\mathtt{SNR} = \infty$, namely the classic case when gradients estimates supplied by the oracle are directly available to π, since perfect decoding is possible for every channel-use. We denote the minmax error in this case by $\mathcal{E}^*_{classic}(N)$. In this standard setup for first-order convex optimization, prior work gives a complete characterization of the minmax error $\mathcal{E}^*_{classic}(N)$; see, for instance, [16]. We summarize these well-known results below.

Theorem II.3. *For absolute constants $c_1 \ge c_0 > 0$, we have*

$$\frac{c_0 DB}{\sqrt{N}} \le \mathcal{E}^*_{classic}(N) \le \frac{c_1 DB}{\sqrt{N}}.$$

Thus, the $1/\sqrt{N}$ convergence rate that SGD provides for convex functions is optimal up to constant factors, with dependence on the dimension d coming only through the parameters D and B. This convergence rate will serve as a basic benchmark for our results in this paper.

[2]Our goal behind considering the minmax cost in (5) is to ensure that the lower bounds are independent of the geometry of set $\mathcal{X}$. But our upper bound techniques can handle an arbitrary, fixed $\mathcal{X}$ as well.

[3]For simplicity, we have considered AWGN channel for transmission. In many practical communication systems, a two-dimensional signal space is available through the in-phase and quadrature-phase components. For these systems, our results for ASK code continue to hold with a QAM or QPSK constellation based code.

III. Main Results

A. Lower Bound for over-the-air optimization

We begin by proving a lower bound for over-the-air optimization. The proof of the lower bound uses recent results in information-constrained optimization given in [2], which in turn builds on the results of [1], [3]. As is usual in other lower bounds in stochastic optimization, our lower bound holds for a sufficiently large N.

Theorem III.1. *For some universal constant[4] $c \in (0,1)$ and $N \ge \frac{d}{\log(1+\mathtt{SNR})}$, we have[5]*

$$\mathcal{E}^*(N) \ge \frac{cDB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\min\{d, 1/2\log(1+\mathtt{SNR})\}}}.$$

Our lower bound states that there is slowdown by a factor of $\sqrt{d/\log(1+\mathtt{SNR})}$ over the classic convergence rate and that no over-the-air optimization scheme can achieve the classic convergence rate unless the $\mathtt{SNR}$ is sufficiently high.

B. Performance and limitations of analog schemes

Next, we show that a simple analog coding scheme attains the optimal convergence rate at low $\mathtt{SNR}$. Specifically, we consider the scheme from [18] where the subgradient estimate is scaled-down appropriately to satisfy the power constraint in (3), sent coordinate-by-coordinate over d channel-uses, and then scaled-up before using it in a gradient descent procedure. We call this analog code the *scaled transmission* analog code. Throughout the paper our first-order optimization algorithm remains projected subgradient descent algorithm (PSGD), with different codes and associated decoding schemes to get back the transmitted subgradient estimate.

Theorem III.2. *The over-the-air optimization procedure (π,φ) comprising the scaled transmission analog code and PSGD satisfies*

$$\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f,\pi^{\varphi O}) \le \frac{cDB}{\sqrt{N}} \cdot \sqrt{d + \frac{d}{\mathtt{SNR}}},$$

where c is a universal constant.

Since $\sqrt{d+(d/\mathtt{SNR})} \le \sqrt{2d/\mathtt{SNR}} \le \sqrt{3d/\log(1+\mathtt{SNR})}$ for a sufficiently small $\mathtt{SNR}$, we get the following corollary in view of Theorem III.1 and the result above.

Corollary III.3. *For $\mathtt{SNR} \in (0,1)$ and $N \ge \frac{d}{\log(1+\mathtt{SNR})}$, we have*

$$\mathcal{E}^*_{analog}(N) = \Theta\left(\frac{DB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\log(1+\mathtt{SNR})}}\right).$$

Interestingly, our next result shows that the scaled transmission scheme is the optimal analog coding scheme up to constant factors. In particular, while analog codes are optimal for low $\mathtt{SNR}$, they can be far from optimal at high $\mathtt{SNR}$.

[4]The universal constants differ in different theorem statements.

[5]$\log(\cdot)$ and $\ln(\cdot)$ denote logarithms to the base 2 and base e, respectively.

Theorem III.4. *For some universal constant $c \in (0,1)$ and $N \geq d(1+1/\mathtt{SNR})$, we have*

$$\mathcal{E}^*_{analog}(N) \geq \frac{cDB}{\sqrt{N}} \cdot \sqrt{d + \frac{d}{\mathtt{SNR}}}.$$

Theorem III.4 shows that in comparison to Theorem III.1 analog schemes can lead to a slowdown of $\sqrt{d}$ for high values of SNR. Even when SNR goes to infinity, we can't get the classic, dimension-free convergence rate back. Note that the upper bound in Theorem III.2 matches the lower bound of Theorem III.4 for large SNR, establishing that the scaled transmission analog code of [18] is optimal among analog coding schemes even at high SNR. We remark that the convergence analysis in [18] required additional smoothness assumptions and is not valid for our setting.

Remark 1. While our definition of analog schemes does not include the top-k [5] analog coding schemes, we can also derive a lower bound for such schemes. Even for such analog schemes, similar lower bound as above holds and the convergence rate does not match the classic convergence rate at high SNR. We defer the details to the extended version [11].

C. Optimality of ASK

We now present a code that almost attains the convergence rate in the lower bound of Theorem III.1. Our encoder φ quantizes the noisy subgradient estimates by using a *gain-shape* quantizer. That is, the encoder separately quantizes the norm of the subgradient, its *gain*, and the normalized vector obtained after dividing the subgradient by its norm, its *shape*. The quantized gain and shape are sent over two different channel-uses, both using ASK. We note that this scheme is not strictly an ASK code since we use the channel twice. However, this is just a technicality and can be avoided by a more tedious analysis. To clearly present our ideas, we first present an ASK code which works in a slightly more idealized setting, captured by the following assumptions for the quantized subgradient:

1) (Perfect gain quantization) We assume that the norm of subgradient vector can be perfectly sent to the algorithm i.e., without any induced noise. Further, we don't account for the channel-uses in sending the norm.
2) (An ideal shape quantizer) There exists an ideal shape[6] quantizer which quantizes the shape of the vector to a means square error of d/r and where the quantized output is an unbiased estimate of the input.

Recall that our optimization algorithm is PSGD with an appropriate decoding rule to decode the noisy codewords sent over the channel.

Theorem III.5. *Under Assumptions 1-2 above, there exists an over-the-air optimization procedure (π, φ) with an ASK code φ for which we have*

$$\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{\varphi O}) \leq \frac{2DB}{\sqrt{N}} \sqrt{\frac{d}{\min\{d, \log\left(\sqrt{4\mathtt{SNR}/\ln N} + 1\right)\}}}.$$

Furthermore, the ASK code quantizes the subgradient vector to $r = \log\left(\sqrt{4\mathtt{SNR}/\ln N} + 1\right)$ bits.

Remark 2 (Resolution grows with SNR). We remark that the number of bits r used to express the subgradients in our algorithm grows with SNR as $r = \log\left(\sqrt{4\mathtt{SNR}/\ln N} + 1\right)$ bits, namely the resolution must grow logarithmically with SNR.

We now state our complete result, without making ideal assumptions. This time the gain is sent in one channel-use by scaling it appropriately to satisfy the power constraint and the shape is quantized using RATQ [15]. As noted above, this is not formally an ASK code since we sent the quantized gain over a separate channel use, but is the same in essence.

Theorem III.6. *For d, SNR, and N satisfying*[7] $\ln^*(d/3) \leq 7$ *and* $\log\left(\sqrt{4\mathtt{SNR}/\ln N} + 1\right) \geq 6$, *we have*

$$\mathcal{E}^*(N) \leq \frac{2DB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\min\{d, r/48\}}},$$

where $r = \log\left(\sqrt{4\mathtt{SNR}/\ln N} + 1\right)$. Furthermore, this bound is attained by using an over-the-air optimization procedure consisting of PSGD as the optimization algorithm and an ASK-like encoding procedure.[8]

IV. Proofs

We first prove our lower bounds, before coming to the algorithms and upper bounds. For brevity, we only present the main ideas and defer the further details to [cite].

A. The general recipe for proving lower bounds

We follow the recipe of [2] to prove our lower bounds. The difficult functions we construct are the same as in previous lower bounds for convex functions such as [3]. 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

$$f_v(x) := \frac{2B\delta}{\sqrt{d}} \sum_{i=1}^{d} \left| x(i) - \frac{v(i)D}{2\sqrt{d}} \right|, \quad \forall\, x \in \mathcal{X}. \tag{6}$$

Note that the gradient $g_v(x)$ of f_v at $x \in \mathcal{X}$ is equal to $-2B\delta v/\sqrt{d}$, i.e., it is independent of x. We will fix our noisy subgradient oracle O_v later. For any O_v, let $\hat{g}_t$ denote the output of the gradient oracle in iteration t. We will consider a noisy oracle which outputs $\hat{g}_t$ that are i.i.d. from a distribution p_v with mean $-2B\delta v/\sqrt{d}$.

For a given code φ of length ℓ, let $C_t = \varphi(g_t)$, $t = 1, ..., T$. Let $V \sim Unif\{-1,1\}^d$ and $Y^T = (Y_1, \ldots, Y_T)$ denote output of the AWGN channel when the inputs are $C^T =$

[6] We call this an ideal quantizer because it would achieve the lower bound for stochastic optimization [15], where the gradients are quantized to r-bits.

[7] $\ln^* a$ denotes the smallest number of ln operations on a required to make it less than 1. Also, we remark that the bound on $\ln^*(d/3)$ is just for simplicity, and a similar convergence bound can be shown for any $d > 0$.

[8] In particular, the encoding procedure uses two channel-uses where in the first channel-use, an ASK code is used to transmit the shape of the subgradient vector, which is quantized to r bits, and in the second channel-use, the gain of the subgradient vector is transmitted after an appropriate scaling.

$(C_1, \ldots, C_T)$. The following lower bound can be established by using results from[9] [2, Lemma 3, 4]:

$$\mathbb{E}\left[f_V(x_T) - f_V(x_V^*)\right] \geq \frac{DB\delta}{6}\left[1 - \sqrt{\frac{2}{d}\sum_{i=1}^{d} I(V(i) \wedge Y^T)}\right], \tag{7}$$

By the definition of $\mathcal{E}^*(N)$, we have

$$\mathcal{E}^*(N) \geq \mathbb{E}\left[f_V(x_T) - f_V(x_V^*)\right]. \tag{8}$$

Thus, it only remains to bound the mutual-information term. Note that this bound holds for any oracle O_v; we choose difficult oracles satisfying (1) and (2) to derive our lower bounds.

B. Proof of Theorem III.1

a) A difficult gradient oracle: For each f_v in (6), consider a gradient oracle O_v which outputs $\hat{g}_t$ with independent coordinates, each taking values $-B/\sqrt{d}$ or $B/\sqrt{d}$ with probabilities $(1 + 2\delta v(i))/2$ and $(1 - 2\delta v(i))/2$, respectively. The parameter $\delta > 0$ is to be chosen suitably later. Note that $\hat{g}_t$ are product Bernoulli distributed vectors with mean $-2B\delta v/\sqrt{d}$.

b) Bounding the mutual-information: The following strong data processing inequality was derived in [1] for $I\big(V(i) \wedge Y^T\big)$ when the observations are product Bernoulli vectors:

$$\sum_{i\in[d]} I\big(V(i) \wedge Y^T\big) \leq c'\delta^2 \max_{v\in\{-1,1\}^d} \max_{\varphi\in\mathcal{C}_\ell} I\big(\hat{g}^T \wedge Y^T\big),$$

where c' is some constant. Using the well-known formula for AWGN capacity (see [10]), we can show using the data processing inequality that

$$\sum_{i\in[d]} I\big(V(i) \wedge Y^T\big) \leq c'\delta^2 N \min\{d, 1/2\log(1 + \texttt{SNR})\}.$$

The proof is completed by combining this bound with (7) and (8), and maximizing the right-side of (7) by setting $\delta = \sqrt{d/(4c'N\min\{2d, \log(1 + \texttt{SNR})\})}$.

C. Proof of Theorem III.4

Consider the encoder $\varphi(\hat{g}_t) = \mathbf{A}\hat{g}_t$ corresponding to an analog coding scheme for the functions in (6).

a) Gaussian oracle: For every f_v and any query point x_t, consider a Gaussian oracle that outputs: $\hat{g}(x_t) = -2B\delta v/\sqrt{d} + G$, where $G \sim \mathcal{N}(0, B^2/d\mathbf{I}_d)$. For matrix $\mathbf{A} \in \mathbb{R}^{\ell\times d}$, the subgradients are encoded as $\varphi(\hat{g}(x_t)) = \mathbf{A}\hat{g}(x_t)$ and sent over the Gaussian channel.

b) Bounding the mutual-information: We proceed as in the previous lower bound proof and first note that $\sum_{i=1}^{d} I(V(i) \wedge Y^T) \leq I(V \wedge Y^T)$ since $V(i)$ are i.i.d.. Further, since $Y_1, .., Y_T$ are i.i.d. conditioned on V, we have $I(V \wedge Y^T) \leq TI(V \wedge Y_1)$. Thus, it suffices to bound the mutual-information $I(V \wedge Y_1)$ which we do in the following lemma. Recall that the outputs $C_t = (-2B\delta/\sqrt{d})\mathbf{A}V + \mathbf{A}G$, where G denotes the Gaussian noise of the oracle, satisfies the power constraint $\sum_{t=1}^{T}\mathbb{E}\left[\|C_t\|_2^2\right] \leq T\ell P$, which implies that $\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)B^2/d \leq \ell P/(1 + 4\delta^2)$. Further, $Y_t = C_t + Z_t$, where Z_t is the channel noise in ℓ uses.

Lemma IV.1. *For $\mathbf{A}, G, V$ and Y_t defined above, if $\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)B^2/d \leq \ell P/(1 + 4\delta^2)$, then $\forall t \in [T]$, $I(V \wedge Y_t) \leq (2\log e)\cdot\ell\delta^2(1 + 1/\texttt{SNR})^{-1}$.*

The proof uses the fact that Gaussian maximizes entropy, along with Jensen's inequality; refer to [11] for more details. Combining the previous bound with (7) and maximizing the right-side with $\delta = \sqrt{(1 + 1/\texttt{SNR})d(\log e)(16N)}$, the proof is completed using (8).

D. A general convergence bound for over-the-air optimization

For an $\ell-$length coding scheme $\varphi : \mathbb{R}^d \to \mathbb{R}^\ell$, recall that the overall output of the channel Y_t after the tth query is given by (4). Our proposed schemes in sections IV-E and IV-F below involve projecting back this channel output in $\mathbb{R}^\ell$ to $\mathbb{R}^d$. In particular, as a part of the optimization algorithm π, Y_t is passed through a *decoder mapping* $\psi : \mathbb{R}^\ell \to \mathbb{R}^d$ which gives back a $d-$dimensional vector to be used by the first-order optimization algorithm.

We use PSGD as the first-order optimization algorithm; the overall over-the-air optimization procedure is described in Algorithm 1. PSGD proceeds as SGD, with the additional projection step where it projects the updates back to domain $\mathcal{X}$ using the map $\Gamma_{\mathcal{X}}(y) := \min_{x\in\mathcal{X}}\|x - y\|$, $\forall\, y \in \mathbb{R}^d$.

1: **for** $t = 0$ to $T - 1$ **do**
2: Observe Y_t given by (4)
3: $x_{t+1} = \Gamma_{\mathcal{X}}\left(x_t - \eta_t\psi(Y_t)\right)$
4: Output $\frac{1}{T}\cdot\sum_{t=1}^{T} x_t$

Algorithm 1: Over-the-air PSGD with encoder φ, decoder ψ

We now describe a convergence bound for over-the-air optimization described in Algorithm 1. In our formulation, the decoder ψ is a part of the optimization protocol π. However, for concreteness, with a slight abuse of notation we now denote the overall over-the-air optimization protocol using the tuple (π, φ, ψ). The performance of (π, φ, ψ) is controlled by the worst-case L_2-norm $\alpha(\pi, \varphi, \psi)$ and the worst-case bias $\beta(\pi, \varphi, \psi)$ of the subgradient obtained after processing the received vector, defined below:

$$\alpha(\pi, \varphi, \psi) := \sup_{\hat{g}\in\mathbb{R}^d : \mathbb{E}\left[\|\hat{g}\|^2\right]\leq B^2} \sqrt{\mathbb{E}\left[\|\psi(Y)\|^2\right]},$$

$$\beta(\pi, \varphi, \psi) := \sup_{\hat{g}\in\mathbb{R}^d : \mathbb{E}\left[\|\hat{g}\|^2\right]\leq B^2} \|\mathbb{E}\left[(\hat{g} - \varphi(Y)\right]\|,$$

where for all $i \in [d]$, $Y(i)$ satisfies (4). The next result is only a minor modification of the standard PSGD proof and is very similar to [15, Theorem 2.4].

Lemma IV.2. *For the above PSGD equipped over-the-air optimization protocol (π, φ, ψ) with N channel uses, we have*

$$\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{\varphi O}) \leq D\left(\frac{\alpha(\pi, \varphi, \psi)}{\sqrt{N/\ell}} + \beta(\pi, \varphi, \psi)\right),$$

[9]Note that the result in [2] is for a more general class of adaptive channels.

provided that the learning rate η_t is set to $\frac{D}{\alpha(\pi,\varphi,\psi)\sqrt{N/\ell}}$ for all iterations $t \in [N/\ell]$.

This general convergence bound will establish our upper bound proofs below.

E. Proof of Theorem III.2

a) Downscale the power: The subgradient vector is multiplied by $\sqrt{Pd}/B$ to meet the power constraints and sent using d channel-uses,one channel-use per coordinate. Thus, our encoded output is $\varphi(\hat{g}(x_t)) = \sqrt{Pd}/B \cdot \hat{g}(x_t)$.

b) Upscale the power: The optimization algorithm π observes Y_t given by (4) and re-scales it back by a factor $B/\sqrt{Pd}$. Thus, the decoding ψ rule at the algorithm's end is given by $\psi(Y_t) = B/\sqrt{Pd}Y_t$. It is easy to see that $\mathbb{E}[\psi(Y_t)|x_t] = \mathbb{E}[\hat{g}(x_t)|x_t]$ implying $\beta(\pi,\varphi,\psi) = 0$. Also, using the independence of zero mean noise Z_t and $\hat{g}(x_t)$, $\mathbb{E}\left[\|\psi(Y_t)\|^2|x_t\right] = \mathbb{E}\left[\|\hat{g}(x_t)\|^2 + B^2/(Pd)\|Z_t\|^2\right]$ which can be bounded by $B^2 + B^2\sigma^2/P$. That implies $\alpha(\pi,\varphi,\psi) \leq B\sqrt{(1+1/\mathtt{SNR})}$ and the proof is completed using Lemma IV.2.

F. Proof of Theorem III.5

Since ASK code is of length 1, we can have N queries in N channel-uses. For the minimum-distance decoder ψ, 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 as its complement, i.e., $A_N^c := \cup_{t=1}^N\{|Z_t| \geq 2\sqrt{P}/(2^r-1)\}$, where Z_t is defined in (4). By the assumption about the ideal quantizer, under the event A_N which depends only on the channel noise, Lemma IV.2 with $\alpha(\pi,\varphi,\psi) = \sqrt{d/r}$ gives $\mathbb{E}[(f(x_T) - f(x^*))\mathbb{1}_{A_N}] \leq \frac{DB}{\sqrt{N}}\sqrt{\frac{d}{r}}$. Further, due to Gaussianity, $\mathbb{P}(A_N^c) \leq N\exp(-\frac{2\mathtt{SNR}}{(2^r-1)^2})$ and that $\mathbb{E}\left[(f(x_T) - f(x^*))\mathbb{1}_{A_N^c}\right] \leq DB \cdot \mathbb{P}(A_N^c)$, the proof is completed upon setting $r = \log\left(\sqrt{\frac{4\mathtt{SNR}}{\ln N}} + 1\right)$ (which gives $\mathbb{P}(A_N^c) \leq \frac{1}{\sqrt{N}}$). □

V. Concluding remarks

We showed the optimality of analog schemes at low SNR in Corollary III.3. However, Theorem III.4 shows that there is a $\sqrt{d}$ factor bottleneck that analog codes can't overcome, no matter how high the SNR is. Finally, we show in Theorem III.6 that ASK codes almost attain the optimal convergence rate at all SNRs. It is important to note that more sophisticated coding schemes can still help improving the small $\log\log N$ and $\ln^* d$ factors seen in the performance of ASK codes.

In another direction, it is important to consider multiparty algorithms and multiterminal communication over Gaussian additive MAC channel. While the limitations for analog schemes apply to that setting as well, we may need to use lattice codes to extend our ASK coding scheme to a MAC. This is an interesting direction for future work.

References

[1] 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.

[2] 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.

[3] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright, "Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization," *IEEE Transactions on Information Theory*, vol. 5, no. 58, pp. 3235–3249, 2012.

[4] 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*, 2017.

[5] D. Alistarh, T. Hoefler, M. Johansson, S. Khirirat, N. Konstantinov, and C. Renggli, "The convergence of sparsified gradient methods," *Advances in Neural Information Processing Systems*, 2018.

[6] 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)*, 2019.

[7] ——, "Over-the-Air Machine Learning at the Wireless Edge," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019, pp. 1–5.

[8] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed Optimisation for Non-Convex Problems," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 560–569.

[9] W.-T. Chang and R. Tandon, "Communication Efficient Federated Learning over Multiple Access Channels," *https://arxiv.org/abs/2001.08737*, 2020.

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

[11] S. K. Jha, P. Mayekar, and H. Tyagi, "Fundamental limits of over-the-air optimization: Are analog schemes optimal?" *https://arxiv.org/abs/2109.05222*, 2021.

[12] 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.

[13] 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*. PMLR, 2021.

[14] P. Mayekar and H. Tyagi, "Limits on gradient compression for stochastic optimization," *Proceedings of the IEEE International Symposium of Information Theory (ISIT' 20)*, 2020.

[15] ——, "RATQ: A universal fixed-length quantizer for stochastic optimization," *IEEE Transactions on Information Theory*, 2020.

[16] A. Nemirovsky, "Information-based complexity of convex programming," 1995, Available Online http://www2.isye.gatech.edu/ne-mirovs/Lec_EMCO.pdf.

[17] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[18] 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.

[19] T. Sery, N. Shlezinger, K. Cohen, and Y. C. Eldar, "COTAF: Convergent Over-the-Air Federated Learning," in *IEEE Global Communications Conference (GLOBECOM)*, 2020, pp. 1–6.

[20] 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)*, 2017.

[21] 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.

[22] 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.

[23] 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.