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

Shubham K Jha [*] Prathamesh Mayekar[†] Himanshu Tyagi[*] [†]

April 4, 2022

Abstract

We consider over-the-air 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. We derive 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 $\sqrt{d/\log(1+\texttt{SNR})}$. 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 $\sqrt{d(1+1/\texttt{SNR})}$. 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 $\sqrt{d}$ for analog coding remains even when SNR tends to infinity. Remarkably, we present a simple quantize-and-modulate scheme that uses *Amplitude Shift Keying* and almost attains the optimal convergence rate at all SNRs.

[*]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

An abridged version of this paper will appear in the proceedings of IEEE Global Communications Conference (GLOBECOM), 2021.

Contents

1 Introduction **3**

2 Problem formulation and preliminaries **4**
- 2.1 Functions and gradient oracles . . . 4
- 2.2 Codes and Gaussian channel . . . 4
- 2.3 Over-the-air Optimization . . . 5
- 2.4 Special coding schemes . . . 6
- 2.5 A benchmark from prior results . . . 6

3 Main Results **6**
- 3.1 Lower Bound for over-the-air optimization . . . 6
- 3.2 Performance and limitations of analog schemes . . . 7
- 3.3 Optimality of ASK . . . 8

4 Proofs **9**
- 4.1 Summary of the method used for proving lower bounds . . . 10
- 4.2 Proof of Theorem 3.1 . . . 10
- 4.3 Proof of Theorem 3.4 . . . 11
- 4.4 A general convergence bound for over-the-air optimization . . . 12
- 4.5 Proof of Theorem 3.2 . . . 13
  - 4.5.1 The scaled transmission analog scheme . . . 13
  - 4.5.2 The sampled version of scaled transmission analog scheme . . . 13
- 4.6 Proof of Theorem 3.6 . . . 14
- 4.7 Proof of Theorem 3.7 . . . 15

5 Experiments **19**

6 Concluding remarks **20**

A Mathematical details concerning Remark 2 **25**

1 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 ($cf.$ [1]). 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 ($cf.$ [2, 3]).

Many different optimization algorithms have been proposed using different kinds 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 $\sqrt{\dfrac{d}{\log(1+\texttt{SNR})}}$ factor slowdown in convergence rate. Furthermore, for low $\texttt{SNR}$, 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 $\sqrt{d(1+\frac{1}{\texttt{SNR}})}$ slowdown in convergence rate when analog codes are used. Note that as $\texttt{SNR}$ 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 $\texttt{SNR}$ tends to infinity, making them suboptimal at high $\texttt{SNR}$. 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 [2–18]. 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 [2] send only top k gradient coordinates along with error feedback. In [10], 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 [9, 11, 13, 14, 17]. On the other hand, digital schemes rely on gradient quantization and channel coding. For instance, authors in [7] 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 [14], authors perform one-bit quantization of subgradients similar to signSGD [19] 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 uses analog schemes or capacity-achieving channel codes. Further, even works such as [14] 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 for quantization 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, [20–29, 29–36]).

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 [37] (which in turn builds on [38–40]) 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 4 for more details.

The rest of the paper is organized as follows. We set up the problem in the next section and provide all our main results in Sections 3. All the proofs are given in Section 4 and concluding remarks are in Section 6.

2 Problem formulation and preliminaries

2.1 Functions and gradient oracles

For a convex set $\mathcal{X} \subset \mathbb{R}^d$ with $\sup_{x,y\in\mathcal{X}} \|x-y\| \leq D$, 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:

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

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

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.

2.2 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

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

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

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

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 $\mathtt{SNR} := \frac{P}{\sigma^2}$.

2.3 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 an increase in the length of encoding protocol will lead to a decrease in the number of oracle queries as the number of channel-uses is restricted to N. Similarly, while we are allowed to use an optimization algorithm that can make as many as N queries to the oracle, increasing the number of queries will lead to a smaller block length encoding protocol. Let $\Lambda(N) := \{\pi \in \Pi_{\ell,T}, \varphi \in \mathcal{C}_\ell \colon \ell \cdot T \leq N\}$. 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 $\mathcal{E}^*(N, \mathcal{X}) := \inf_{(\pi,\varphi)\in\Lambda(N)} \sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{\varphi O})$. Let $\mathbb{X} := \{\mathcal{X} \colon \sup_{x,y\in\mathcal{X}} \|x - y\| \leq D\}$. In this paper, we will characterize the following quantity[2]:

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

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

[2]Our goal behind considering the min-max 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.

2.4 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 2.1. 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 $\ell \leq d$. 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 min-max 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, $\mathcal{E}^*_{analog}(N) \geq \mathcal{E}^*(N)$.

Definition 2.2. 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

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

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.

2.5 A benchmark from prior results

We recall results for the case $\texttt{SNR} = \infty$, 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 min-max 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 min-max error $\mathcal{E}^*_{classic}(N)$; see, for instance, [41]. We summarize these well-known results below.

Theorem 2.3. *For absolute constants $c_1 \geq c_0 > 0$, we have*

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

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.

3 Main Results

3.1 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 [37], which in turn builds on the results of [38,39]. As is usual in other lower bounds in stochastic optimization, our lower bound holds for a sufficiently large N.

[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.

Theorem 3.1. *For some universal constant*[4] $c \in (0,1)$ *and* $N \geq \frac{d}{\log(1+\texttt{SNR})}$, *we have*[5]

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

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

3.2 Performance and limitations of analog schemes

Next, we show that a simple analog coding scheme attains the optimal convergence rate at low SNR. Specifically, we consider the scheme from [10] 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 3.2. *The over-the-air optimization procedure* (π, φ) *comprising the scaled transmission analog code and PSGD satisfies*

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

where c *is a universal constant.*

Since $\sqrt{d + (d/\texttt{SNR})} \leq \sqrt{2d/\texttt{SNR}} \leq \sqrt{3d/\log(1+\texttt{SNR})}$ for a sufficiently small SNR, we get the following corollary in view of Theorem 3.1 and the result above.

Corollary 3.3. *There exist universal constants* c_1, c_2 *such that for* $\texttt{SNR} \in (0,1)$ *(i.e., low* SNR*) and* $N \geq \frac{d}{\log(1+\texttt{SNR})}$, *we have*

$$\frac{c_1 DB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\log(1+\texttt{SNR})}} \leq \mathcal{E}^*_{analog}(N) \leq \frac{c_2 DB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\log(1+\texttt{SNR})}}.$$

Remark 1. We remark that a slightly different analog coding scheme can also guarantee the same performance as the scaled transmission analog code given in Theorem 3.2 and performs better in our experiments presented in Section 5. In this scheme, the noisy subgradient estimate is first randomly rotated by a random matrix, and then only a few of its coordinates are used for the gradient descent procedure, which, in turn, are sampled randomly. Both the random matrix and random coordinate sampling are generated using shared randomness between the encoder and the algorithm. Notice that such an algorithm needs only few channel-uses per descent step instead of scaled transmission analog code that uses d channel-uses per descent step. We provide a detailed description and analysis of this scheme in Section 4.5.

[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.

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 $\mathtt{SNR}$.

Theorem 3.4. *For some universal constant $c \in (0,1)$ and $N \geq d(1 + 1/\mathtt{SNR})$, we have*

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

Note that for small $\mathtt{SNR}$, we have $1 + \frac{1}{\mathtt{SNR}} \approx \frac{1}{\log(1+\mathtt{SNR})}$, and thus, Theorem 3.4 shows that analog codes are optimal at low $\mathtt{SNR}$. Theorem 3.4 also shows that in comparison to Theorem 3.1 analog schemes can lead to a slowdown of $\sqrt{d}$ for high values of $\mathtt{SNR}$. Even when $\mathtt{SNR}$ goes to infinity, we can't get the classic, dimension-free convergence rate back. Note that the upper bound in Theorem 3.2 matches the lower bound of Theorem 3.4 for large $\mathtt{SNR}$, establishing that the scaled transmission analog code of [10] is optimal among analog coding schemes even at high $\mathtt{SNR}$. We remark that the convergence analysis in [10] required additional smoothness assumptions and is not valid for our setting.

Remark 2. While our definition of analog schemes does not include the top-k (see, for instance, [42] and the references therein) 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 $\mathtt{SNR}$. We defer the details to the Appendix A.

3.3 Optimality of ASK

We now present a code that almost attains the convergence rate in the lower bound of Theorem 3.1. Our encoder φ quantizes the noisy subgradient estimates by using a *gain-shape* quantizer (*cf.*[43]).

Definition 3.5 (Gain-shape quantizer)**.** A Quantizer Q is defined to be a gain-shape quantizer if it has the following form

$$Q(Y) = Q_g(\|Y\|_2) \cdot Q_s(Y/\|Y\|_2),$$

where Q_g is any $\mathbb{R} \to \mathbb{R}$ quantizer and Q_s is any $\mathbb{R}^d \to \mathbb{R}^d$ 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 using ASK code. 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 an ideal 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 mean square error of d/r and where the quantized output is an unbiased estimate of the input.

[6]We call this an ideal quantizer because such a quantizer would achieve the lower bound for stochastic optimization in [28], where the gradients are quantized to r-bits.

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

Theorem 3.6. *Under Assumptions 1-2 above, there exists an over-the-air optimization procedure (π, φ) with an ASK code φ for which we have*

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

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

Remark 3 (Resolution grows with $\mathtt{SNR}$). We remark that the number of bits r used to express the subgradients in our algorithm grows with $\mathtt{SNR}$ as $r = \log\left(\sqrt{\frac{4\mathtt{SNR}}{\ln N}}+1\right)$ bits, namely the resolution must grow logarithmically with $\mathtt{SNR}$.

We now state our complete result, without making ideal assumptions. This time the gain is sent in one channel-use by simply scaling the gain value appropriately to satisfy the power constraint, which is similar to the scaled transmission analog code from Theorem 3.2. For quantizing the shape, our scheme uses the quantizer RATQ from [28]. Again note above, this scheme is not formally an ASK code since we send the gain over a separate channel. Nonetheless, they are similar, in essence, to ASK codes as only the transmission of gain, a scalar, is not accounted for in the ASK code.

Theorem 3.7. *For d, $\mathtt{SNR}$, and N satisfying*[7] $\ln^*(d/3) \le 7$ *and* $\log\left(\sqrt{\frac{4\mathtt{SNR}}{\ln N}}+1\right) \ge 6$*, we have*

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

where $r = \log\left(\sqrt{\frac{4\mathtt{SNR}}{\ln N}}+1\right)$. 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]

4 Proofs

We first prove our lower bounds before coming to the algorithms and upper bounds.

[7] $\ln^* a$ denotes the smallest number of ln operations on a required to make it less than 1. Also, we remark that we can prove a similar convergence bound without any upper bound on $\ln^*(d/3)$; we only make this assumption to simplify the upper bound expression.

[8] In particular, the encoding procedure uses two channel-uses for transmitting the subgradient estimate. In the first channel-use, an ASK code is used to transmit the shape of the subgradient vector, which is quantized to $r = \log\left(\sqrt{\frac{4\mathtt{SNR}}{\ln N}}+1\right)$ bits. In the second channel-use, the gain of the subgradient vector is transmitted after scaling it appropriately to satisfy the power constraint.

4.1 Summary of the method used for proving lower bounds

We follow the recipe of [37] to prove our lower bounds. The difficult functions we construct are the same as in previous lower bounds for convex functions such as [38]. We consider the domain $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\|_\infty \le D/(2\sqrt{d}\}$, 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 $C^T = (C_1, \ldots, C_T)$. The following lower bound can be established by using results from[9] [37, Lemma 3, 4]:

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

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

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

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.

4.2 Proof of Theorem 3.1

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 $-2B\delta v/\sqrt{d}$.

Bounding the mutual-information The following strong data processing inequality was derived in [39] for $I\big(V(i) \wedge Y^T\big)$ when the observations are product Bernoulli vectors:

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

where c' is some constant. Using the well-known formula for AWGN capacity (see [44]), we can show using the data processing inequality that

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

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

[9] Note that the result in [37] is for a more general class of adaptive channels.

4.3 Proof of Theorem 3.4

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

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.

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 4.1. *For $\mathbf{A}, G, V$ and Y_t defined above, if $\frac{\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)B^2}{d} \leq \frac{\ell P}{1+4\delta^2}$, then $\forall t \in [T]$,*

$$I(V \wedge Y_t) \leq (2\log e) \cdot \ell\delta^2 (1 + 1/\mathtt{SNR})^{-1}.$$

Proof. Since $C_t = (-2B\delta/\sqrt{d})\mathbf{A}V + \mathbf{A}G$, we have $\mathbb{E}\left[C_t C_t^\top\right] = \left(\frac{B^2(1+4\delta^2)}{d}\right)\mathbf{A}\mathbf{A}^\top$ which implies

$$\mathbb{E}\left[\|C_t\|^2\right] = \mathrm{Tr}(\mathbb{E}\left[C_t C_t^\top\right]) = \left(\frac{B^2(1+4\delta^2)}{d}\right)\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top) \leq \ell P. \tag{9}$$

As Z_t is independent of C_t, we also have

$$\mathbb{E}\left[Y_t Y_t^\top\right] = \left(\frac{B^2(1+4\delta^2)}{d}\right)\mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I}_\ell. \tag{10}$$

By definition of mutual-information and the fact that Gaussian maximizes the entropy,

$$I(V \wedge Y_t) = h(Y_t) - h(Y_t|V) \leq \frac{1}{2}\log\frac{\det\left(\frac{B^2(1+4\delta^2)}{d}\mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I}_\ell\right)}{\det\left(\frac{B^2}{d}\mathbf{A}\mathbf{A}^\top + \sigma^2\mathbf{I}_\ell\right)}.$$

Let $\lambda_1, \ldots, \lambda_\ell$ be the eigen values of $\mathbf{A}\mathbf{A}^\top$. Then, right-side can be further bounded as

$$\begin{aligned}
I(V \wedge Y_t) &\leq \frac{1}{2}\sum_{i=1}^{\ell}\log\frac{\frac{B^2(1+4\delta^2)}{d}\lambda_i + \sigma^2}{\frac{B^2}{d}\lambda_i + \sigma^2} \\
&\leq \frac{\ell}{2}\log\frac{\frac{B^2(1+4\delta^2)}{\ell d}\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top) + \sigma^2}{\frac{B^2}{\ell d}\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top) + \sigma^2} \\
&\leq \frac{(2\log e)\cdot\ell\delta^2}{1 + \ell d\sigma^2/(B^2\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top))} \\
&\leq \frac{(2\log e)\cdot\ell\delta^2}{1 + \mathtt{SNR}^{-1}},
\end{aligned}$$

where the first inequality is the Hadamard inequality; the second is Jensen's inequality; the third one uses $\log(1+x) \leq x \log e$; and the last inequality follows from (9). $\square$

Combining the previous bound with (7) and maximizing the right-side using $\delta = \sqrt{(1+\frac{1}{\texttt{SNR}})\frac{d}{((\log e)\cdot 16N)}}$, the proof is completed using (8).

4.4 A general convergence bound for over-the-air optimization

For an ℓ-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 4.5 and 4.6 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: $x_{t+1} = \Gamma_{\mathcal{X}}\left(x_t - \eta\psi(Y_t)\right)$
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 derive 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:

$$\alpha(\pi, \varphi, \psi) := \sup_{\hat{g}\in\mathbb{R}^d : \mathbb{E}\left[\|\hat{g}\|^2\right]\leq B^2} \sqrt{\mathbb{E}\left[\|\psi(Y)\|^2\right]}, \tag{11}$$

$$\beta(\pi, \varphi, \psi) := \sup_{\hat{g}\in\mathbb{R}^d : \mathbb{E}\left[\|\hat{g}\|^2\right]\leq B^2} \|\mathbb{E}\left[(\hat{g} - \psi(Y)\right]\|, \tag{12}$$

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 [28, Theorem 2.4].

Lemma 4.2. *For the above PSGD equipped over-the-air optimization protocol (π, φ, ψ) with N channel-uses, we have*

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

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 be used in our upper bound proofs below.

4.5 Proof of Theorem 3.2

4.5.1 The scaled transmission analog scheme

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 $\varphi(\hat{g}(x_t)) = \sqrt{Pd}/B \cdot \hat{g}(x_t)$.

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 $\mathbb{E}\left[\psi(Y_t)|x_t\right] = \mathbb{E}\left[\hat{g}(x_t)|x_t\right]$ 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 $\alpha(\pi, \varphi, \psi) \leq B\sqrt{(1 + 1/\texttt{SNR})}$ and the proof is completed using Lemma 4.2.

4.5.2 The sampled version of scaled transmission analog scheme

Rotate randomly At each iteration t, the subgradient vector is rotated by multiplying it with a random matrix

$$\mathbf{R} := \frac{1}{\sqrt{d}}\mathbf{H}\mathbf{D},$$

where $\mathbf{H}$ is a $(d \times d)-$ Walsh-Hadamard matrix [45][10] and $\mathbf{D}$ is diagonal matrix with each non-zero entry generated uniformly from $\{+1, -1\}$. The diagonal matrix is generated via public randomness between the encoder and the algorithm, and can therefore be used for decoding at the algorithms end.

From [28, Lemma 5.8], each coordinate $\mathbf{R}\hat{g}_t(i)$ of the rotated subgradient $\mathbf{R}\hat{g}_t$ satisfies

$$\mathbb{E}\left[\mathbf{R}\hat{g}_t(i)^2\right] \leq \frac{B^2}{d}, \quad \forall i \in [d].$$

Subsampling Using shared randomness between the encoder and the decoder, a set $S \subseteq [d]$ is sampled uniformly over all subsets of $[d]$ of cardinality ℓ. The rotated subgradient vector is sampled at S and is denoted as

$$\tilde{g}_{\mathbf{R},S,t} := \sum_{i=1}^{d} \mathbf{R}\hat{g}_t(i)\mathbb{1}_{\{i\in S\}} \cdot e_i.$$

Downscale the power The subsampled vector $\tilde{g}_{\mathbf{R},S,t}$ is multiplied by $\sqrt{Pd}/B$ to meet the power constraints and sent using ℓ channel-uses, one channel-use per coordinate. Thus, the encoded output is $\varphi(\hat{g}(x_t)) = \sqrt{Pd}/B \cdot \tilde{g}_{\mathbf{R},S,t}$.

Upscale the power The optimization algorithm π observes Y_t given by (4) and re-scales it back by a factor $\frac{dB}{\ell\sqrt{Pd}} \cdot \mathbf{R}^{-1}$. Thus, the decoding ψ rule at the algorithm's end is given by $\psi(Y_t) =$

[10] We assume that d is a power of 2.

$\frac{dB}{\ell\sqrt{Pd}} \cdot \mathbf{R}^{-1} Y_t$. It is easy to see that

$$\begin{aligned}
\mathbb{E}\left[\psi(Y_t)|x_t\right] &= \mathbb{E}\left[\frac{d}{\ell}\cdot \mathbf{R}^{-1}\left(\sum_{i=1}^{d}\mathbf{R}\hat{g}_t(i)\mathbb{1}_{\{i\in S\}}\cdot e_i\right)|x_t\right]\\
&= \mathbb{E}\left[\hat{g}_t\mathbb{E}\left[\frac{d}{\ell}\mathbb{1}_{\{i\in S\}}\right]|x_t\right]\\
&= \mathbb{E}\left[\hat{g}(x_t)|x_t\right],
\end{aligned}$$

implying $\beta(\pi,\varphi,\psi)=0$. Also, using the independence of AWGN noise $Z_t \sim \mathcal{N}(0,\sigma^2\mathbf{I}_\ell)$ and $\hat{g}(x_t)$, we get

$$\begin{aligned}
\mathbb{E}\left[\|\psi(Y_t)\|^2|x_t\right] &= \frac{d^2}{\ell^2}\mathbb{E}\left[\|\hat{g}(x_t)\|^2\mathbb{1}_{\{i\in S\}}|x_t\right] + \frac{dB^2}{P\ell^2}\|Z_t\|^2\\
&\leq \frac{B^2 d}{\ell}\left(1+\frac{\sigma^2}{P}\right),
\end{aligned}$$

which implies $\alpha(\pi,\varphi,\psi) \leq B\sqrt{\frac{d}{\ell}\left(1+\frac{1}{\texttt{SNR}}\right)}$, and the proof is complete using Lemma 4.2.

4.6 Proof of Theorem 3.6

Since an 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.,

$$A_N^c := \bigcup_{t=1}^{N}\left\{|Z_t| \geq \frac{2\sqrt{P}}{(2^r-1)}\right\},$$

where Z_t is defined in (4). By the assumptions about an ideal quantizer (*c.f.* Section 3.3), under the event A_N, which depends only on the channel noise, Lemma 4.2 with $\alpha(\pi,\varphi,\psi)=\sqrt{d/r}$ gives

$$\mathbb{E}\left[(f(x_T)-f(x^*))\mathbb{1}_{A_N}\right] \leq \frac{DB}{\sqrt{N}}\cdot\sqrt{\frac{d}{r}}.$$

Further, due to Gaussian[11] noise, we have $\mathbb{P}\left(A_N^c\right) \leq N\exp(-\frac{2P}{\sigma^2(2^r-1)^2}) = N\exp(-\frac{2\texttt{SNR}}{(2^r-1)^2})$. Setting $r=\log\left(\sqrt{\frac{4\texttt{SNR}}{\ln N}}+1\right)$, we have $\mathbb{P}\left(A_N^c\right)\leq \frac{1}{\sqrt{N}}$, which leads to

$$\begin{aligned}
\mathbb{E}\left[(f(x_T)-f(x^*))\right] &= \mathbb{E}\left[(f(x_T)-f(x^*))\mathbb{1}_{A_N}\right] + \mathbb{E}\left[(f(x_T)-f(x^*))\mathbb{1}_{A_N^c}\right]\\
&\leq \frac{DB}{\sqrt{N}}\cdot\sqrt{\frac{d}{r}} + \frac{DB}{\sqrt{N}}\\
&\leq 2\frac{DB}{\sqrt{N}}\cdot\sqrt{\frac{d}{\min\{d,r\}}}.
\end{aligned}$$

[11] In fact, the proof requires noise to be only sub-Gaussian, a weaker assumption than being Gaussian.

4.7 Proof of Theorem 3.7

For communication, we consider an ASK code in $[-\sqrt{P}, \sqrt{P}]$ with the following 2^r-constellation points

$$\left\{-\sqrt{P} + \frac{(i-1)\cdot 2\sqrt{P}}{2^r - 1} : i \in [2^r]\right\}, \tag{13}$$

for some $r \in \mathbb{N}$.

We separately send the *gain* $\|\hat{g}_t\| \in \mathbb{R}$ and *shape* $\frac{\hat{g}_t}{\|\hat{g}_t\|} \in \mathbb{R}^d$ of subgradient $\hat{g}_t$. The encoder φ is a tuple which consists of separate gain and shape encoders φ_g and φ_s, i.e., $\varphi = (\varphi_g, \varphi_s)$. The internal randomness used in these gain and shape encoders will be independent, which will result in the output of these encoders being conditionally independent given any subgradient estimate $\hat{g}_t$. Similarly, the decoding mechanism is also a tuple consisting of two separate decoders ψ_g and ψ_s that are used to decode the transmitted gain and shape values, respectively. The final decoded output of ψ is taken to be the product of the decoded outputs of ψ_g and ψ_s.

1. Communicating the gain. Recall that the gain sub-encoders above need to satisfy the average power constraint (9) from Section 2.2.

The gain $\|\hat{g}_t\|$ is multiplied by $\sqrt{P}/B$ to meet the power constraints and sent in one channel-use. Thus, our encoded output is $\varphi_g(\|\hat{g}_t\|) = (\sqrt{P}/B)\|\hat{g}_t\|$. The optimization algorithm π observes the channel output $Y_{g,t}$ given by

$$Y_{g,t} = \varphi_g(\|\hat{g}_t\|) + Z_{g,t},$$

where $Z_{g,t} \sim \mathcal{N}(0, \sigma^2)$ denotes the Gaussian noise, and re-scales it back by a factor $B/\sqrt{P}$, i.e.,

$$\psi_g(Y_{g,t}) = \|\hat{g}_t\| + B/\sqrt{P} \cdot Z_{g,t}.$$

We evaluate the performance measures $\alpha(\pi, \varphi_g, \psi_g)$ and $\beta(\pi, \varphi_g, \psi_g)$, viewing the gain $\|\hat{g}_t\|$ as a 1-dimensional subgradient. Similar to (11), (12) we have,

$$\alpha(\pi, \varphi_g, \psi_g) := \sup_{\|\hat{g}_t\| \in \mathbb{R} : \mathbb{E}\left[\|\hat{g}_t\|^2\right] \le B^2} \sqrt{\mathbb{E}\left[\|\psi_g(Y_{g,t})\|^2\right]},$$
$$\beta(\pi, \varphi_g, \psi_g) := \sup_{\|\hat{g}_t\| \in \mathbb{R} : \mathbb{E}\left[\|\hat{g}_t\|^2\right] \le B^2} \|\mathbb{E}\left[\|\hat{g}_t\| - \psi_g(Y_{g,t})\right]\|.$$

Specifically, it is easy to see that

$$\alpha(\pi, \varphi_g, \psi_g) = \sqrt{B^2 + B^2/\texttt{SNR}}, \quad \beta(\pi, \varphi_g, \psi_g) = 0. \tag{14}$$

2. Quantizing the shape. We denote the shape $\hat{g}_t/\|\hat{g}_t\|$ by $\hat{g}_{t,\texttt{shape}}$. In every iteration t, L_2-norm of $\hat{g}_{t,\texttt{shape}}$ is almost surely bounded by 1. Accordingly, to quantize the shape, we are interested in quantizers for almost surely bounded oracles. We use a subsampled version of RATQ [28, Section 3.5] to quanitze the shape, as this quantizer is almost optimal for communication-constrained optimization with almost surely bounded oracles. The encoder φ_s is composed of four components: *rotation*, *subsampling*, *tetra-iterated adaptive quantization* and *mapping to ASK code*, which we describe below.

Rotation. Assuming that d is a power of 2, the subgradient shape $\hat{g}_{t,shape}$ is rotated by multiplying it with a random matrix

$$\mathbf{R} := \frac{1}{\sqrt{d}}\mathbf{H}\mathbf{D},$$

where $\mathbf{H}$ is a $(d \times d)-$ Walsh-Hadamard matrix [45] and $\mathbf{D}$ is a diagonal matrix with diagonal entries generated uniformly from $\{+1, -1\}$. The diagonal matrix is generated via public randomness between the encoder and the algorithm, and can therefore be used for decoding at the algorithms end.

Note that since $\mathbf{R}$ is a unitary matrix, the norm remains unaltered even after rotation, i.e., $\|\mathbf{R}\hat{g}_{t,\mathtt{shape}}\| = \|\hat{g}_{t,\mathtt{shape}}\| = 1$ a.s..

Subsampling. Using shared randomness between the encoder and the decoder, a set $U \in [d]$ is sampled uniformly over all subsets of $[d]$ of cardinality μd. The rotated shape vector is sampled at U and is denoted by

$$\tilde{g}_{\mathbf{R},U,t} := \{\mathbf{R}\hat{g}_{t,\mathtt{shape}}(i)\}_{i \in U}.$$

We now quantize every coordinate of $\tilde{g}_{\mathbf{R},U,t} \in \mathbb{R}^{\mu d}$ using the following.

Tetra-iterated Adaptive Quantization. Consider a sequence of intervals $\{[-M_i, M_i]\}_{i \in [h_s]}$ where $M_1, \ldots, M_{h_s}$ grows using[12] tetra-iteration:

$$M_1^2 = \frac{3}{d}, M_i^2 = \frac{3}{d} \cdot e^{*(i-1)},\ i \in [h_s],$$

where parameter h_s satisfies $\log h_s = \lceil [\log(1 + \ln^*(d/3))\rceil$. We choose these values such that the largest interval must contain $\|\tilde{g}_{\mathbf{R},U,t}\|_\infty$, i.e., $1 \leq M_{h_s}$.

For each coordinate $\tilde{g}_{\mathbf{R},U,t}(i)$, the quantizer first identifies the smallest index $j \in [h_s]$ such that $|\tilde{g}_{\mathbf{R},U,t}(i)| \leq M_j$ and then represent $\tilde{g}_{\mathbf{R},U,t}(i)$ using a uniform k_s-level shape quantizer Q_{M_j,k_s} in interval $[-M_j, M_j]$. The k_s levels of shape quantizer are given by

$$B_{M_j,k_s}(l) := -M_j + (l-1) \cdot \frac{2M_j}{k_s - 1},\ l \in [k_s].$$

These levels partition $[-M_j, M_j]$ into $k_s - 1$ sub-intervals $\{[B_{M_j}(l), B_{M_j}(l+1)]\}_{l \in [k_s - 1]}$. The uniform quantizer locates a sub-interval that contains $\tilde{g}_{\mathbf{R},U,t}(i)$, say $[B_{M_j}(l^*), B_{M_j}(l^*+1)]$ for some $l^* \in [k_s - 1]$, and outputs

$$Q_{M_j,k_s}(\tilde{g}_{\mathbf{R},U,t}(i)) = \begin{cases} B_{M_j}(l^*), & \text{w.p. } \frac{B_{M_j}(l^*+1) - \tilde{g}_{\mathbf{R},U,t}(i)}{B_{M_j}(l^*+1) - B_{M_j}(l)} \\ B_{M_j}(l^*+1), & \text{w.p. } \frac{\tilde{g}_{\mathbf{R},U,t}(i) - B_{M_j}(l^*)}{B_{M_j}(l^*+1) - B_{M_j}(l)} \end{cases}.$$

This is done for all μd coordinates and we represent the output as $Q(\tilde{g}_{\mathbf{R},U,t})$ given by

$$Q(\tilde{g}_{\mathbf{R},U,t}) := (Q_{M_{j_{(i)}}}(\tilde{g}_{\mathbf{R},U,t}(i)) : 1 \leq i \leq \mu d),$$

where $j_{(i)}$ corresponds to the index identified for ith coordinate $\tilde{g}_{\mathbf{R},U,t}(i)$.

Note that there is no overflow because of the choice of h_s and the quantized output $Q(\tilde{g}_{\mathbf{R},U,t})$ can be represented using precision of at most $\mu d \cdot (\log(k_s) + \log(h_s))$ bits. We denote this binary representation by $[Q(\tilde{g}_{\mathbf{R},U,t})]_2$.

[12] The ith tetra-iteration of e^{*i} is defined as: $e^{*1} = e, e^{*i} := e^{e^{*(i-1)}}$.

Mapping to ASK code. Using the ASK code in (13), when $r = \mu d \cdot \log(h_s k_s)$, there exists a one-to-one mapping between $[Q(\hat{g}_{\mathbf{R},U,t})]_2$ and ASK code, say $\zeta_s : \{0,1\}^r \to [-\sqrt{P}, \sqrt{P}]$. We therefore send the codeword $\varphi_s(\hat{g}_{t,\mathtt{shape}}) = \zeta([Q(\tilde{g}_{\mathbf{R},U,t})]_2)$ in one channel-use as

$$Y_{s,t} = \varphi_s(\hat{g}_{t,\mathtt{shape}}) + Z_{s,t}, \tag{15}$$

where $Z_{s,t} \sim \mathcal{N}(0, \sigma^2)$ denotes the Gaussian noise. Note that the power constraint is always satisfied.

At the algorithm's end, the decoder ψ_s primarily makes use of three components, namely the minimum-distance decoder, inverse mapping ζ_s^{-1}, and inverse rotation, and performs the following steps:

1. The channel output $Y_{s,t}$ is fed into a minimum-distance decoder that locates the nearest possible ASK codeword in $[-\sqrt{P}, \sqrt{P}]$ which further is fed into ζ_s^{-1} retrieving an r-bit sequence.
2. The recovered r-bit sequence is split into blocks of size $\mu d \cdot \log(k_s)$ and $\mu d \cdot \log(h_s)$, each of which gets further split into sub-blocks of sizes $\log(k_s)$ and $\log(h_s)$, respectively. These sub-blocks can uniquely identify the quantization intervals and the corresponding quantization levels for all sampled coordinates in U. We denote that by $\hat{Q}(\tilde{g}_{\mathbf{R},U,t}) \in \mathbb{R}^d$. Note that all the remaining unsampled coordinates are decoded to be 0, i.e.,
$$\hat{Q}(\tilde{g}_{\mathbf{R},U,t})(i) = 0, \quad \forall i \notin U.$$
3. The last step is to multiply $\hat{Q}(\tilde{g}_{\mathbf{R},U,t})$ by $\frac{1}{\mu}$ and perform inverse rotation to get the decoded output $\psi_s(Y_{s,t}) = \frac{1}{\mu}\mathbf{R}^{-1}\hat{Q}(\tilde{g}_{\mathbf{R},U,t})$.

Under perfect minimum-distance decoding, we have $\hat{Q}(\tilde{g}_{\mathbf{R},U,t}) = Q(\tilde{g}_{\mathbf{R},U,t})$. Again, similar to (11), (12) we have,

$$\alpha(\pi, \varphi_s, \psi_s) := \sup_{\hat{g}_t \in \mathbb{R}^d : \|\hat{g}_t\|^2 \le 1} \sqrt{\mathbb{E}\left[\|\psi_s(Y_{s,t})\|^2\right]},$$
$$\beta(\pi, \varphi_s, \psi_s) := \sup_{\hat{g}_t \in \mathbb{R}^d : \|\hat{g}_t\|^2 \le 1} \|\mathbb{E}\left[\|\hat{g}_t\| - \psi_s(Y_{s,t})\right]\|,$$

where $Y_{s,t}$ is defined in (15). Since the shape vector is almost-surely bounded by 1, note that this time, the performance measures are defined over the class of almost-surely bounded oracles.

Following the proof of [28, Theorem 3.7], we can derive lemma below.

Lemma 4.3. *For φ_s, ψ_s as defined above and under perfect minimum-distance decoding event, we have*

$$\alpha(\pi, \varphi_s, \psi_s) \le \sqrt{\frac{1}{\mu}\left(\frac{9}{(k_s-1)^2} + 1\right)} \text{ and } \beta(\pi, \varphi_s, \psi_s) = 0.$$

3. Combining the gain and the shape quantizers. The final decoded output is taken to be product of outputs from gain decoder ψ_g and shape decoder ψ_s, i.e.,

$$\psi(Y_{g,t}, Y_{s,t}) = \psi_g(Y_{g,t}) \cdot \psi_s(Y_{s,t}).$$

The lemma below is again adapted from [28, Theorem 4.2] and can be proved in a similar way.

Lemma 4.4. *For the decoded output $\psi(Y_{g,t}, Y_{s,t})$ defined above, we have*

$$\alpha(\pi, \varphi, \psi) \le \alpha(\pi, \varphi_g, \psi_g) \cdot \alpha(\pi, \varphi_s, \psi_s) \text{ and}$$
$$\beta(\pi, \varphi, \psi) \le \beta(\pi, \varphi_g, \psi_g).$$

4. Analysis. Since communicating gain and shape for each query requires 2 channel-uses, we can have atmost $N/2$ queries. 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., $A_N^c := \cup_{t=1}^N \{|Z_{s,t}| \ge 2\sqrt{P}/(2^r - 1)\}$, where $Z_{s,t}$ is defined in (15). We have

$$\begin{aligned}\mathbb{E}\left[f(x_N) - f(x^*)\right] &= \mathbb{E}\left[(f(x_N) - f(x^*)) \mid A_N\right] \cdot \mathbb{P}\left(A_N\right) + \mathbb{E}\left[(f(x_T) - f(x^*)) \mid A_N^c\right] \cdot \mathbb{P}\left(A_N^c\right) \\ &\le \mathbb{E}\left[(f(x_N) - f(x^*)) \mid A_N\right] + DB \cdot \mathbb{P}\left(A_N^c\right). \end{aligned} \tag{16}$$

As $\mathbb{P}\left(A_N^c\right) \le N \exp\left(-\frac{2\texttt{SNR}}{(2^r-1)^2}\right)$, setting $r = \log\left(\sqrt{\frac{4\texttt{SNR}}{\ln N}} + 1\right)$ gives $\mathbb{P}\left(A_N^c\right) \le \frac{1}{\sqrt{N}}$. Using Lemma 4.2, the first term on the right-side can be bounded as

$$\mathbb{E}\left[(f(x_N) - f(x^*)) \mid A_N\right] \le D\left(\frac{\alpha(\pi, \varphi, \psi)}{\sqrt{N/2}} + \beta(\pi, \varphi, \psi)\right).$$

We now analyse the overall performance measures $\alpha(\pi, \varphi, \psi)$ and $\beta(\pi, \varphi, \psi)$ of gain-shape quantizer described above.

Recall that the gain value is sent in one channel-use after appropriate scaling. The shape is quantized using RATQ and sent over the channel using the ASK code given by (13) with $r = \mu d \log(h_s k_s)$.

Using the individual performance measures from (14) and Lemma 4.3, and combining them via Lemma 4.4, we have

$$\begin{aligned} D\left(\frac{\alpha(\pi, \varphi, \psi)}{\sqrt{N/2}} + \beta(\pi, \varphi, \psi)\right) &\le D\left(\frac{B\sqrt{1 + \frac{1}{\texttt{SNR}}} \cdot \sqrt{\frac{1}{\mu}\left(\frac{9}{(k_s-1)^2} + 1\right)}}{\sqrt{N/2}}\right) \\ &= DB\sqrt{\frac{d}{N}}\sqrt{\frac{2}{\mu d}\left(1 + \frac{1}{\texttt{SNR}}\right)\left(\frac{9}{(k_s-1)^2} + 1\right)} \\ &= DB\sqrt{\frac{d}{Nr}}\sqrt{2\log(h_s k_s)\left(1 + \frac{1}{\texttt{SNR}}\right)\left(\frac{9}{(k_s-1)^2} + 1\right)}, \end{aligned}$$

where the last line uses the fact that $\mu d = \lceil r/\log(h_s k_s)\rceil$. Using the inequality above, (16) can be further bounded as

$$\mathbb{E}\left[f(x_N) - f(x^*)\right] \le \frac{DB\sqrt{d}}{\sqrt{N}}\sqrt{\frac{2\log(h_s k_s)\left(1 + \frac{1}{\texttt{SNR}}\right)\left(\frac{9}{(k_s-1)^2} + 1\right)}{\log(1 + \sqrt{4\texttt{SNR}/\ln N})}} + \frac{DB}{\sqrt{N}}.$$

At last, we use an 8-level shape quantizer for every coordinate, i.e., $k_s = 8$, and choose the number of quantization intervals h_s satisfying $\log h_s = \lceil[\log(1+\ln^*(d/3))\rceil$. For $\ln^*(d/3) \le 7$, we have $\mu d = \lceil r/6 \rceil$, which further implies that $r \ge 6$, and that

$$\begin{aligned}
\mathbb{E}\left[f(x_N) - f(x^*)\right] &\le \frac{DB}{\sqrt{N}}\sqrt{\frac{d \cdot 24\left(1+\frac{1}{\mathtt{SNR}}\right)}{\log(1+\sqrt{4\mathtt{SNR}/\ln N})}} + \frac{DB}{\sqrt{N}} \\
&\le \frac{2DB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\min\{d, \frac{\log(1+\sqrt{4\mathtt{SNR}/\ln N})}{24\left(1+\frac{1}{\mathtt{SNR}}\right)}\}}} \\
&\le \frac{2DB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\min\{d, \frac{\log(1+\sqrt{4\mathtt{SNR}/\ln N})}{48}\}}},
\end{aligned}$$

where the last line uses the inequality that $1 + \frac{1}{\mathtt{SNR}} \le 2$ for $\mathtt{SNR} > 1$, which further holds since $r \ge 6$. $\square$

5 Experiments

We evaluate the performance of our proposed analog and digital schemes (*c.f.* Sections 4.5, 4.7) which achieve the optimality of over-the-air optimization at low and high $\mathtt{SNR}$s, respectively. Our experiments validate all our claims and are described below.

We consider the task of image classification and perform experiments on MNIST dataset, which has 60000 training and 10000 test samples. In particular, the classifier for the MNIST dataset is implemented by training a 3-layer *Convolutional Neural Network* (CNN) that consists of a single convolution layer with 16 filters of dimension 3×3 each and ReLU activation function, followed by a 2×2 max-pooling; one fully connected layer with dimensions 2704×10; and a final softmax output layer, i.e., $d = 27210$. We choose the optimization algorithm to be SGD with learning rates proportional to $\mathtt{SNR}$s (as can be inferred from Lemma 4.2).

For our experiments, we consider the proposed digital scheme using ASK described in Section 4.7. Recall that the gain is always sent in one channel-use after scaling, and the shape is quantized using RATQ and then sent over the Gaussian channel using the ASK code. For RATQ, we set $h_s = 4, k_s = 8$ in tetra-iterated adaptive shape quantizer. Further, the descriptions of quantization interval ($\log h_s$ bits per dimension) and the corresponding uniform quantization point ($\log k_s$ bits per dimension) are sent separately[13] in two different channel uses, and the best values for r in ASK code are chosen proportional to operating $\mathtt{SNR}$.

On the other hand, for the proposed analog scheme, we consider the sampled version of scaled transmission scheme described in Section 4.5.2 with sampling only three coordinates, i.e., $\ell = 3$. This choice of ℓ is considered for a fair performance comparison with the digital scheme in terms of the number of channel uses. Our codes are available online [46] on GitHub.

We investigate the performance of the proposed analog and digital over-the-air schemes at various $\mathtt{SNR}$s; specifically, -30dB, 40dB, 100dB and 180dB. We plot the training loss and test

[13]Note that our proposed digital scheme (see Section 4.7 for details) uses one channel transmission for sending the shape gradient quantization. Still, in the experiments, we send it using two channel transmissions. We do this to mitigate the precision issues we run into for ASK coding at high values of r in Python.

accuracy for the image classification task at these SNRs in Figures 1, 2, 3 and 4, respectively. The choice for these SNRs are for illustrating the validity of theoretical claims, not for practical considerations.

The performance of both the analog and the digital scheme improves as SNR increases. However, the improvement is faster for the digital scheme than for the analog scheme. In more detail, Figure 1 shows that the proposed analog scheme performs much better than the proposed digital scheme at very low SNR of -30dB. As we increase the SNR, the performance gap between the digital and analog schemes gets reduced. This can be observed in Figure 2 where the performance of both the schemes at $\texttt{SNR} = 40$dB is similar. With further increase in SNR values, the proposed digital scheme surpasses the performance of the proposed analog scheme, with the gap between their performance widening with an increase in SNR, as can be seen in Figure 3 and Figure 4.

Figure 4 also shows the performance of the classic *baseline* scheme, where perfect stochastic gradient estimates are available for the optimization protocol. In other words, the gradients are passed through a Gaussian channel of zero variance. As observed in Figure 4, the proposed digital scheme is close to the baseline[14] performance.

Thus our experiments validate our theory. In particular, Figure 1 validates our theoretical claim that analog schemes are optimal at low SNR. On the other hand, Figures 2, 3, 4 validates our theoretical claim that analog schemes go further away from optimality with an increase in SNR and digital schemes need to be used for optimal convergence at high SNR.

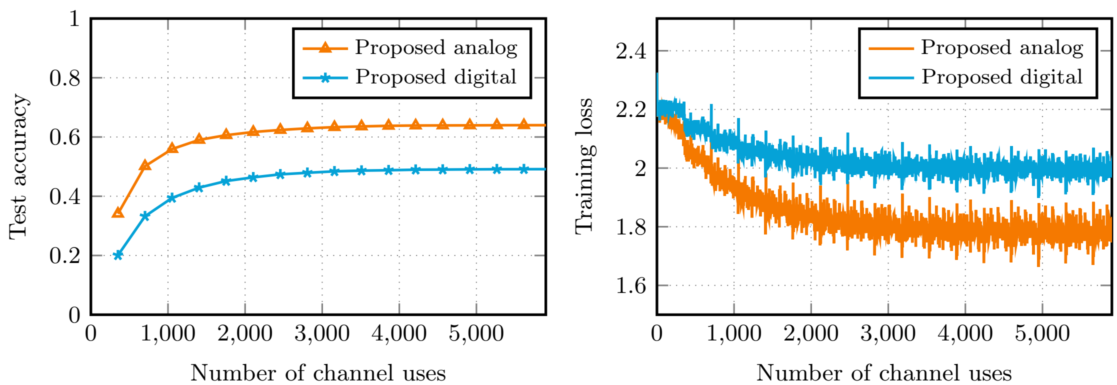

Figure 1: Comparison between proposed analog and proposed digital scheme at $\texttt{SNR} = -30$dB.

6 Concluding remarks

We showed the optimality of analog schemes at low SNR in Corollary 3.3. However, Theorem 3.4 shows that there is a $\sqrt{d}$ factor bottleneck that analog codes can't overcome, no matter how high

[14] In an ideal scenario, we expect the digital scheme to attain the baseline performance for a larger SNR value, as we increase the value of r accordingly. Unfortunately, our python code runs into precision issues for optimally tuned ASK schemes at higher SNRs (beyond 180 dB).

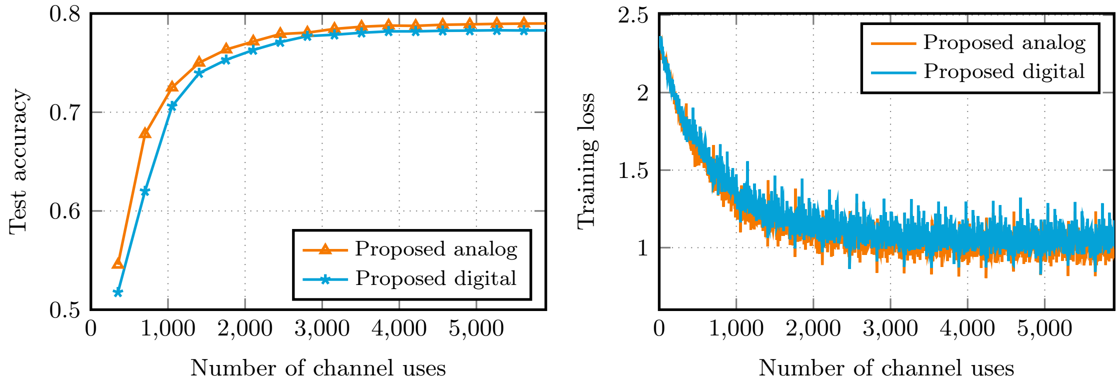

Figure 2: Comparison between proposed analog and proposed digital scheme at $\mathtt{SNR} = 40\text{dB}$.

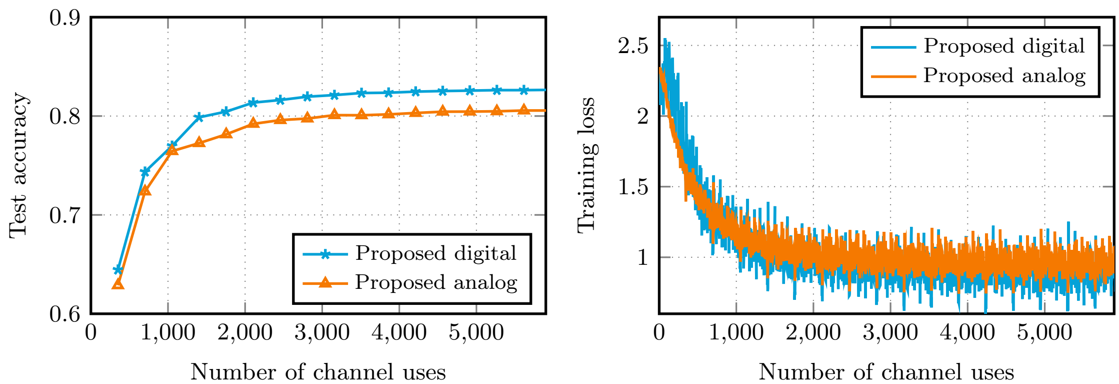

Figure 3: Comparison between proposed analog and proposed digital scheme at $\mathtt{SNR} = 100\text{dB}$.

the $\mathtt{SNR}$ is. Finally, we show in Theorem 3.7 that the proposed digital scheme using ASK codes almost attain the optimal convergence rate at all $\mathtt{SNR}$s.

It is important to note that more sophisticated coding schemes can still help in 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. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

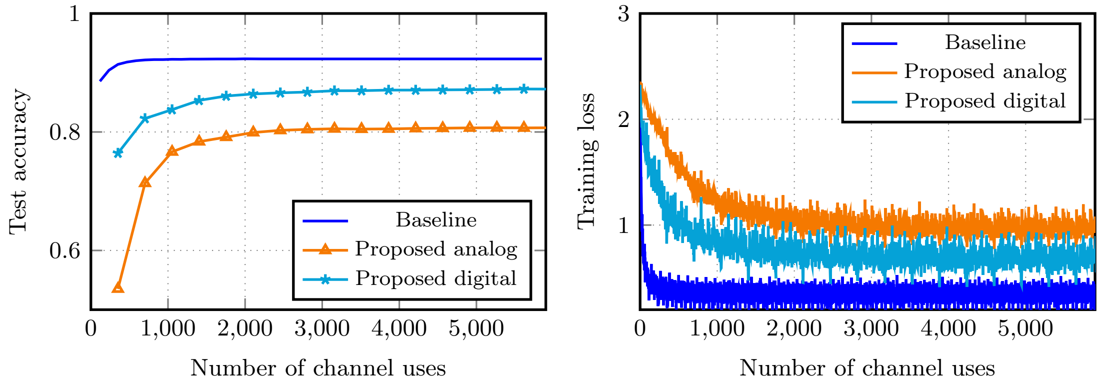

Figure 4: Comparison between proposed analog and proposed digital scheme at $\texttt{SNR} = 180$dB.

[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)*, 2019, pp. 1–5.

[3] ——, "Machine Learning at the Wireless Edge: Distributed Stochastic Gradient Descent Over-the-Air," in *IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1432–1436.

[4] ——, "Federated Learning Over Wireless Fading Channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

[5] M. M. Amiri, T. M. Duman, D. Gündüz, S. R. Kulkarni, and H. Vincent Poor, "Collaborative Machine Learning at the Wireless Edge with Blind Transmitters," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2021.

[6] M. S. H. Abad, E. Ozfatura, D. Gündüz, and O. Ercetin, "Hierarchical Federated Learning ACROSS Heterogeneous Cellular networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8866–8870.

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

[8] T. Sery and K. Cohen, "A Sequential Gradient-Based Multiple Access for Distributed Learning over Fading Channels," in *57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2019, pp. 303–307.

[9] 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.

[10] 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.

[11] 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.

[12] G. Zhu, Y. Wang, and K. Huang, "Broadband Analog Aggregation for Low-Latency Federated Edge Learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2020.

[13] 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.

[14] 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.

[15] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning," in *IEEE Conference on Computer Communications (INFOCOM)*, 2018, pp. 63–71.

[16] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A Joint Learning and Communications Framework for Federated Learning over Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2021.

[17] Y. Sun, S. Zhou, and D. Gündüz, "Energy-Aware Analog Aggregation for Federated Learning with Redundant Data," in *IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.

[18] 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)*. IEEE, 2021, pp. 5055–5059.

[19] 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.

[20] 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.

[21] V. Gandikota, D. Kane, R. K. Maity, and A. Mazumdar, "vqsgd: Vector quantized stochastic gradient descent," *arXiv preprint arXiv:1911.07971*, 2019.

[22] 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.

[23] F. Faghri, I. Tabrizian, I. Markov, D. Alistarh, D. Roy, and A. Ramezani-Kebrya, "Adaptive gradient quantization for data-parallel sgd," *Advances in Neural Information Processing Systems*, 2020.

[24] 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.

[25] 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.

[26] 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.

[27] J. Acharya, C. De Sa, D. J. Foster, and K. Sridharan, "Distributed Learning with Sublinear Communication," *arXiv:1902.11259*, 2019.

[28] P. Mayekar and H. Tyagi, "RATQ: A universal fixed-length quantizer for stochastic optimization," *IEEE Transactions on Information Theory*, 2020.

[29] C.-Y. Lin, V. Kostina, and B. Hassibi, "Achieving the fundamental convergence-communication tradeoff with differentially quantized gradient descent," *arXiv preprint arXiv:2002.02508*, 2020.

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

[31] 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.

[32] W.-N. Chen, P. Kairouz, and A. Özgür, "Breaking the communication-privacy-accuracy trilemma," *arXiv preprint arXiv:2007.11707*, 2020.

[33] 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.

[34] 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, pp. 3502–3510.

[35] D. Jhunjhunwala, A. Gadhikar, G. Joshi, and Y. C. Eldar, "Adaptive quantization of model updates for communication-efficient federated learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3110–3114.

[36] A. Ghosh, R. K. Maity, and A. Mazumdar, "Distributed newton can communicate less and resist byzantine workers," *Advances in Neural Information Processing Systems*, 2020.

[37] 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.

[38] 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.

[39] J. Acharya, C. L. Canonne, and H. Tyagi, "General lower bounds for interactive high-dimensional estimation under information constraints," *http://arxiv.org/abs/2010.06562v4*, 2020.

[40] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi, "Interactive inference under information constraints," *http://arxiv.org/abs/2007.10976*, 2021.

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

[42] D. Alistarh, T. Hoefler, M. Johansson, S. Khirirat, N. Konstantinov, and C. Renggli, "The convergence of sparsified gradient methods," *arXiv preprint arXiv:1809.10505*, 2018.

[43] A. Gersho and R. M. Gray, *Vector quantization and signal compression.* Springer Science & Business Media, 2012, vol. 159.

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

[45] K. J. Horadam, *Hadamard matrices and their applications.* Princeton university press, 2012.

[46] Available online: https://github.com/shubhamjha-46/OTA_Optimization.

A Mathematical details concerning Remark 2

Recall that in the top-k gradient coding scheme only the absolute largest k values of the gradients are used to update the query point. We begin by defining a strict generalization of top-k gradient coding schemes which we call k-coordinate sampling codes.

Definition A.1. A code is a k-coordinate sampling code if the encoder mapping φ consist of only k-coordinate values and their indices, i.e., when $\varphi(x) = (S, \{x(i)\}_{i\in S})$, where S is a subset of $[d]$ with cardinality k. Further, we allow for the set S to be dependent on x. Also, we denote by $\mathcal{E}^*_{\mathrm{k}cs}(N)$ the min-max 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, $\mathcal{E}^*_{\mathrm{k}cs}(N) \geq \mathcal{E}^*(N)$.

Lemma A.2. *For all values of* $\mathtt{SNR}$, *we have*

$$\mathcal{E}^*_{\mathrm{k}cs}(N) \geq \frac{cDB}{\sqrt{N}} \cdot \sqrt{\frac{d}{\min\{d, k\log\frac{d}{k}\}}}$$

Proof. For bounding $\mathcal{E}^*_{\mathrm{k}cs}(N)$, our function class remains the same as in (6) and the oracle remain the same as in the proof of Theorem 3.1. Now note that since gradient estimates supplied by the oracle are Bernoulli vectors, the encoder $\varphi(\cdot)$ can thought of as quantizer with precision of $\log\binom{d}{k} + k$ bits, where the first term in the addition is used to represent S and the second to represent k. Therefore, even at infinite $\mathtt{SNR}$, we have

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

where the result directly follows from [37, Theorem 5] Then, by noting that $\log\binom{d}{k} + k \leq k\log\frac{d}{k} + k(1+\log e)$ and proceeding as in proof of Theorem 3.1, the proof is complete. ☐

Thus, if we employ top-k gradient coding schemes, even at very high $\mathtt{SNR}$ values we do not attain the classic convergence rate.