Limits on Gradient Compression for Stochastic Optimization

Prathamesh Mayekar† Himanshu Tyagi†

January 27, 2020

Abstract

We consider stochastic optimization over ℓ_p spaces using access to a first-order oracle. We ask: What is the minimum precision required for oracle outputs to retain the unrestricted convergence rates? We characterize this precision for every $p \geq 1$ by deriving information theoretic lower bounds and by providing quantizers that (almost) achieve these lower bounds. Our quantizers are new and easy to implement. In particular, our results are exact for $p = 2$ and $p = \infty$, showing the minimum precision needed in these settings are $\Theta(d)$ and $\Theta(\log d)$, respectively. The latter result is surprising since recovering the gradient vector will require $\Omega(d)$ bits.

1 Introduction

We consider the classic problem of minimizing an unknown convex functions which is Lipschitz continuous in the ℓ_p norm. Our setting is that of first-order stochastic optimization, where we are given oracle access to noisy, unbiased estimates of the subgradients which have their ℓ_q norm bounded, where q is the Hölder conjugate of p. Motivated by recent works on gradient descent using quantized gradient updates (*cf.* [2, 4–6, 8, 11–14, 18–22]), we seek to determine the minimum number of bits r to which the gradients can be expressed while retaining the standard, unrestricted gradient rates.

We show that for $p \in [1, 2)$ and $p \geq 2$, respectively, roughly d and $d^{2/p} \log(d^{1-2/p} + 1)$ bits are necessary and sufficient for retaining the standard convergence rates. These bounds are tight upto an $O(\log d)$ factor, in general, but are exact for $p = 2$ and $p = \infty$. Prior work has only considered the problem for the Euclidean case, and not for general ℓ_p geometry. Further, even for the Euclidean case, the best known bounds are from our recent work [14] where the bounds are shown to be tight only upto a mild $O(\log \log \log \ln^* d)$ factor. The results in this paper get rid of this nagging factor and establish tight bounds.

We use different quantizers for $p \geq 2$ and $p \in [1, 2)$. In the $p \geq 2$ range, we use a quantizer we call $SimQ^+$. $SimQ^+$, in turn, uses multiple repetitions of another quantizer we call $SimQ$ which expresses a vector as a convex combination of corner points of an ℓ_1 ball. It is $SimQ$ that yields an $O(\log d)$ bit quantizer for optimization over ℓ_∞. Also, $SimQ^+$ yields the exact upper bound in the ℓ_2 case. In the $[1, 2)$ range, we divide the vector into two parts with small and large coordinates. We use a uniform quantizer for the first part and RATQ of [14] for the second part.

†Department of Electrical Communication Engineering, Indian Institute of Science, India. Email: {prathamesh, htyagi}@iisc.ac.in.

The main observation in our analysis for upper bound is that the role of quantizer in optimization is not to express the gradient with small error. It suffices to have an unbiased estimate with appropriately bounded norms. Our lower bounds are based on those in [3,14]. But interestingly we show that bounds that are useless in the classic setting become useful under precision constraints.

We remark that while our quantizers are related to the ones used in prior works, our main contribution is to show that our specific design choices yield optimal precision. For instance, the quantizers in [11] expresses the input as a convex combination of set of points, similar to *SimQ*. In fact, one of the quantizers in [11] uses similar set of points as that of SimQ with a different scaling. However, the quantizers in [11] are designed keeping in mind other objectives and they fall short of attaining the optimal precision guarantees of *SimQ* and $SimQ^+$.

Also, stochastic optimization over ℓ_p spaces using a biased first-order oracle which is constructed by using only statistical query access to the underlying data was considered in [10]. Here, the authors characterize the number of statistical queries needed to optimize a function upto a given accuracy. This differs from our objective where we assume access to a noisy, unbiased first-order oracle and seek to compress the oracle output to the minimum number of bits, without losing the uncompressed convergence guarantees.

Notation: Throughout the paper q will denote the Hölder conjugate of p (that is, $q = p/(p-1)$). $a \vee b$ and $a \wedge b$ denote the $\max\{a, b\}$ and $\min\{a, b\}$, respectively. We denote by $\log(\cdot)$ logarithm to the base 2 and by $\ln(\cdot)$ logarithm to the base e. $\ln^*(a)$ denotes the number of times ln must be iteratively applied to a before the result is less than or equal to 1. $\{e_1, ..., e_d\}$ denotes the standard basis. We keep notation consistent with our earlier work [14].

2 Preliminaries

2.1 The Setup

We extend the formulation of [14] for Euclidean space to general ℓ_p spaces. Formally, we consider the problem of minimizing an unknown convex function $f : \mathcal{X} \to \mathbb{R}$ using oracle access to noisy subgradients of the function (*cf.* [7,16]). We assume that the function f is convex over the compact, convex domain $\mathcal{X}$ such that $\sup_{x,y\in\mathcal{X}} \|x - y\|_p \le D$; we denote the set of all such sets $\mathcal{X}$ by $\mathbb{X}$. For a query point $x \in \mathcal{X}$, the oracle outputs random estimates of the subgradient $\hat{g}(x)$ which for all $x \in \mathcal{X}$ satisfy[1]

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

$$P(\|\hat{g}(x)\|_q^2 \le B^2|x) = 1, \tag{2}$$

where q is the Hölder conjugate of p and $\partial f(x)$ denotes the set of subgradients of f at x.

Definition 2.1 (Almost surely bounded oracles)**.** A first order oracle which upon a query x outputs only the subgradient estimate $\hat{g}(x)$ satisfying the assumptions (1) and (2) is termed an almost surely bounded oracle. We denote the class of convex functions and oracles satisfying assumptions (1) and (2) by $\mathcal{O}_{0,p}$.

Remark 1 (Mean square bounded oracles). We remark that in the classic literature a more general oracle model has been considered (*cf.* [16], [15]). Specifically, (2) is replaced by $\mathbb{E}\left[\|\hat{g}(x)\|_q^2|x\right] \le B^2$.

[1]Assumptions (1) and (2) imply that any function f in $\mathcal{O}_{0,p}$ is Lipschitz continuous in the ℓ_p norm over the domain $\mathcal{X}$.

In this paper we restrict ourselves to the almost surely bounded oracles. Nonetheless, using the general recipe in [14] for converting quantizers for the almost surely bounded oracles to mean square bounded ones, we can design quantizers for the latter model as well.

In our setting, the outputs of the oracle are passed through a quantizer. An *r-bit quantizer* consists of randomized mappings $(Q^{\mathsf{e}}, Q^{\mathsf{d}})$ with the encoder mapping $Q^{\mathsf{e}} : \mathbb{R}^d \to \{0,1\}^r$ and the decoder mapping $Q^{\mathsf{d}} : \{0,1\}^r \to \mathbb{R}^d$. The overall quantizer is given by the composition mapping $Q = Q^{\mathsf{d}} \circ Q^{\mathsf{e}}$. Let $\mathcal{Q}_r$ be the set of all such r-bit quantizers.

For an oracle $(f, O) \in \mathcal{O}_{0,p}$ and an r-bit quantizer Q, let $QO = Q \circ O$ denote the composition oracle that outputs $Q(\hat{g}(x))$ for each query x. Let π be an algorithm with at most T iterations with oracle access to QO. We will call such an algorithm an optimization protocol. Denote by Π_T the set of all such optimization protocols with T iterations.

Remark 2 (Memoryless, fixed length quantizers). We note that the quantizers in $\mathcal{Q}_r$ are memoryless as well as fixed length quantizers. That is, each new subgradient estimate at time t will be quantized without using any information from the previous updates to a fixed length code of r bits.

Denoting the combined optimization protocol with its oracle QO by π^{QO} and the associated output as $x^*(\pi^{QO})$, we measure the performance of such an optimization protocol for a given (f, O) using the metric $\mathcal{E}_0(f, \pi^{QO}, p)$ defined as $\mathcal{E}_0(f, \pi^{QO}, p) := \mathbb{E}\left[f(x^*(\pi^{QO})) - \min_{x \in \mathcal{X}} f(x)\right]$.

Before proceeding, we recall the results for the case $r = \infty$. These bounds will serve as a basic benchmark for our problem.

Theorem 2.2. *There exist absolute constants c_0 and c_1 where $c_1 \geq c_0 > 0$ such that the following hold:*

1. *For $p \geq 2$,*
$$\frac{c_1 d^{1/2-1/p} DB}{\sqrt{T}} \geq \mathcal{E}_0^*(T, \infty, p) \geq \frac{c_0 d^{1/2-1/p} DB}{\sqrt{T}};$$

2. *for[2] $2 > p \geq 1$,*
$$\frac{c_1 DB\sqrt{\log d}}{\sqrt{T}} \geq \mathcal{E}_0^*(T, \infty, p) \geq \frac{c_0 DB}{\sqrt{T}}.$$

The lower bounds and the upper bounds can be found, for instance, in [3, Theorem 1] and [3, Appendix C].

Remark 3. An optimal achievable scheme for $p \in [1, 2)$ is the stochastic mirror descent with the mirror maps $\|x\|_{p'}^2/(p'-1)$, where p' is chosen appropriately for a given p. These algorithms require only that the expected squared ℓ_q norm of the gradient estimates are bounded.

Remark 4. An optimal achievable scheme for p greater than 2 is simply projected subgradient descent(PSGD). To see this, note that PSGD gives a guarantee of $D'B'/\sqrt{T}$ (*cf.* [15]), where D' is the ℓ_2 diameter and B' is the bound on $\mathbb{E}\left[\|\hat{g}\|_2^2\right]$. Using the monotonicity of ℓ_q norms in q, for $q \geq 2$ we have $\mathbb{E}\left[\|\hat{g}\|_2^2\right] \leq \mathbb{E}\left[\|\hat{g}\|_q^2\right] \leq B^2$. Also, the ℓ_2 diameter of a unit ℓ_p ball is $d^{1/2-1/p}$. It follows that PSGD attains the upper bounds in Theorem 2.2.

[2]For certain range of p closer to 2 the $\sqrt{\log d}$ factor can be removed; for simplicity, we state the slightly weaker result.

The fundamental quantity of interest in this work is $r^*(T,p)$, the minimum precision to achieve the optimization accuracy $\mathcal{U}(T,p)$, the benchmark above from the classic setting. Specifically, we define

$$r^*(T,p) := \inf\{r \in \mathbb{N} : \mathcal{E}_0^*(T,r,p) \le \mathcal{U}(T,p)\}, \tag{3}$$

where[3]

$$\mathcal{E}_0^*(T,r,p) := \sup_{\mathcal{X}\in\mathbb{X}} \inf_{\pi\in\Pi_T} \inf_{Q\in\mathcal{Q}_r} \sup_{(f,O)\in\mathcal{O}_{0,p}} \mathcal{E}(f,\pi^{QO},p), \tag{4}$$

$$\mathcal{U}(T,p) := \frac{4c_1 d^{1/2-1/p} DB}{\sqrt{T}}, \quad \forall p \in [2,\infty],$$

$$\mathcal{U}(T,p) := \frac{4c_1 \sqrt{\log d} DB}{\sqrt{T}}, \quad \forall p \in [1,2).$$

2.2 A Basic Convergence Bound for Quantized Gradients

While our lower bounds hold for any kind of quantizers, but we attain this bounds using unbiased quantizers. For such an unbiased quantizer Q, we characterize the performance in terms of a parameter $\alpha_0(Q)$ defined earlier in [14]. We define

$$\alpha_0(Q;p) := \sup_{Y\in\mathbb{R}^d : \|Y\|_q^2 \le B^2 \text{ a.s.}} \sqrt{\mathbb{E}\left[\|Q(Y)\|_2^2\right]}, \quad p \in [2,\infty],$$

$$\alpha_0(Q;p) := \sup_{Y\in\mathbb{R}^d : \|Y\|_q^2 \le B^2 \text{ a.s.}} \sqrt{\mathbb{E}\left[\|Q(Y)\|_q^2\right]}, \quad p \in [1,2).$$

Note that for all $p \ge 1$, the composed oracle QO satisfies assumption (1). Moreover, in view of Remarks 3 and 4, we have the following convergence guarantees for first-order stochastic optimization using gradients quantized by Q.

Theorem 2.3. *Consider a quantizer Q for the gradients. There exists algorithms $\pi \in \Pi_T$ which when used with oracle outputs quantized by Q performs as follows.*

1. *For $p \ge 2$,*
$$\frac{c_1 d^{1/2-1/p} D\alpha_0(Q;p)}{\sqrt{T}} \ge \sup_{(f,O)\in\mathcal{O}_{0,p}} \mathcal{E}(f,\pi^{QO},p);$$
2. *for $2 > p \ge 1$,*
$$\frac{c_1 \sqrt{\log d} D\alpha_0(Q;p)}{\sqrt{T}} \ge \sup_{(f,O)\in\mathcal{O}_{0,p}} \mathcal{E}(f,\pi^{QO},p).$$

In the rest of the paper, we design unbiased, fixed length quantizers which have $\alpha_0(\cdot;p)$ of the same order as B. Then, using Theorem 2.3 the quantized updates give the same convergence guarantees as that of the classical case, which leads to upper bounds for $r^*(T,p)$. Further, we derive lower bounds for $r^*(T,p)$ to prove optimality of our quantizers.

[3]We take the sup over all $X \in \mathbb{X}$ in the definition of $\mathcal{E}_0^*(T,r,p)$, as it is unclear even in the infinite precision case if we can get matching upper bounds and lower bounds for any convex set in $\mathbb{X}$.

An interesting insight offered by the result above, which is perhaps simple in hindsight, is that even when dealing with ℓ_p oracles for $p \geq 2$, we only need to be concerned about the expected ℓ_2 norm of the quantizers output. It is this insight that leads to the realization that $SimQ^+$ is optimal for these settings.

3 Main Result: Characterization of $r^*(T, p)$

The main result of our paper is the almost complete characterization of $r^*(T, p)$. We divide the result into cases $p \in [1, 2)$ and $p \geq 2$; as mentioned earlier, we use different quantizers for these two cases.

Theorem 3.1. *For stochastic optimization using T accesses to a first-order oracle, the following bounds for $r^*(T, p)$ hold.*

1. *For $p \geq 2$, we have*

$$d^{2/p} \log \left(2e \cdot d^{1-2/p} + 2e\right) \geq r^*(T, p) \geq \left(\frac{c_0}{4c_1} \cdot d^{1/p}\right)^2 \vee 2 \log \left(\frac{c_0}{4c_1} \cdot d^{1/2}\right).$$

2. *For $2 > p \geq 1$, we have*

$$d \left(\left\lceil \log(2\sqrt{2}{\Delta_1}^{1/q} + 2)\right\rceil + 3\right) + \Delta_2 \geq \; r^*(T, p) \geq \left(\frac{c_0}{4c_1\sqrt{\log d}}\right)^2 \cdot d,$$

where $\Delta_1 = \left\lceil \log\left(2 + \sqrt{18 + 6 \ln \Delta_2} \cdot d^{1/2-1/q}\right)\right\rceil$ *and* $\Delta_2 = \lceil \log(1 + \ln^*(d/3)) \rceil$.

Note that for $p \geq 2$ the upper bounds and lower bounds for $r^*(T, p)$ are off by nominal factor of $\log(d^{1-2/p} + 1)$. Also, for $p \in [1, 2)$ the bounds are roughly off by $O(\log d \cdot \log(\log d^{1/2-1/q})^{1/q})$ (ignoring the $\log^* d$ terms).

We present the quantizers achieving these upper bounds, and the proof of the upper bounds, in the next two sections. For $p \geq 2$, we use a quantizer $SimQ$ and its extension $SimQ^+$, presented in Section 4. For $p \in [1, 2)$, we use a combination of uniform quantization and the quantizer RATQ from [14], presented in Section 5.

Our lower bound proof is based on small modification of existing proofs. But an interesting element is that constructions that yield trivial bounds for convergence rate yield tight bounds when information constraints such as constraints on gradient precision are placed. The proof is deferred to Section 6.

We highlight the most interesting features of the result above in separate remarks below.

Remark 5 ($r^*(T, p)$ is independent of T). Theorem 3.1 shows that $r^*(T, p)$ is a function only of p and d, and is independent of T. The number of queries T is a proxy for the desired optimization accuracy. Therefore, the fact that $r^*(T, p)$ is independent of such a parameter is interesting. We note, however, that for oracle models with milder assumptions, such as mean square bounded oracles, this may not hold. In fact, the results of [14] suggest that for mean square bounded oracles $r^*(T, 2)$ is dependent on T.

Remark 6 (Optimality for $p = \infty$). Our bounds match for $p = \infty$, namely our quantizer $SimQ$ offers optimal convergence rate with gradient updates at the least precision. A surprising observation is that this precision is merely $O(\log d)$, much smaller than $O(d)$ bits needed to recover the gradient vector under any reasonable loss function.

Remark 7 (Optimality for $p = 2$). The lower bound for the case when $p = 2$ already appeared in [14], giving $r^*(T, 2) = \Omega(d)$. Both [5] and [20] give variable-length quantization schemes to achieve this lower bound, but the worst-case precision can be order-wise greater than d. Our recent work [14] proposed a quantizer termed RATQ that was within a small factor of $O(\log\log\log\ln^* d)$ of this lower bound. Current work removes this nagging factor using a different quantizer $SimQ^+$.

Remark 8 (Fixed precision). The quantizer RATQ in [14] remains optimal upto a factor of $O(\sqrt{\log\ln^* d})$ for the more general problem of characterizing $\mathcal{E}_0^*(T, r, 2)$ for any precision r less than d bits. In this setting of small precision, the performance of $SimQ^+$ is much worse.

4 Our quantizers for $p \geq 2$

We present our quantizer $SimQ$ and its extension $SimQ^+$. The former is seen to be optimal for $p = \infty$ while the latter for $p = 2$.

4.1 An optimal quantizer for $p = \infty$

Require: Input $Y \in \mathbb{R}^d$, Parameter B

$i^* = \begin{cases} i & w.p. \quad |Y(i)|/B \\ 0 & w.p. \quad 1 - \|Y\|_1/B \end{cases}$

if $i^* \in [d]$ **then**

$j^* = sign(Y(i^*))$

else

$j^* = 1$

Output: $Q^{\mathtt{e}}_{\mathtt{SimQ}}(Y; B) = i^* \cdot j^*$

Figure 1: Encoder $Q^{\mathtt{e}}_{\mathtt{SimQ}}(Y; B)$ for $SimQ$

Require: Input $i' \in \{-d, -(d-1), \cdots 0, \cdots, d\}$

if $i' \neq 0$ **then**

$Z = Bsign(i')e_{|i'|}$

else

$Z = 0$

Output: $Q^{\mathtt{d}}_{\mathtt{SimQ}}(i'; B) = Z$

Algorithm 2: Decoder $Q^{\mathtt{d}}_{\mathtt{SimQ}}(i'; B)$ for $SimQ$

Simplex Quantizer (SimQ) Our first quantizer $SimQ$ is described in Algorithms 1 and 2. For $p = \infty$, our quantizer's input vector Y is an unbiased estimate of the subgradient of the function at the point queried and satisfies $\|Y\|_1 \leq B$. $SimQ$ takes such a Y as an input and produces an output vector which, too, satisfies both these properties. The main idea behind $SimQ$ is the fact that any point inside the unit ℓ_1 ball can be represented as a convex combination of at the most $2d$ points: $\{e_i, -e_i : i \in [d], j \in \{-1, 1\}\}$. With this observation, we use an ℓ_1 sampling procedure to obtain an unbiased estimate $i^*.j^*$ of the vector. At the decoder, upon observing $i' = i^*.j^*$, the decoder simply declares $Bj^*e_{i^*}$.

Theorem 4.1. *Let Q be the quantizer SimQ described in Algorithms 1, 2. Then, for Y such that $\|Y\|_1 \leq B$ a.s., $Q(Y)$ can be represented in $\log(2d+1)$ bits, $\mathbb{E}[Q(Y)|Y] = Y$, and $\alpha_0(Q,\infty) \leq B$.*

Proof. Since $i^* \in [d]$ and $j^* \in \{-1, 1\}$, we can represent the output of the encoder of *SimQ* using $\log(2d+1)$ bits. Next, denoting the quantizer *SimQ* by Q, note that

$$\mathbb{E}[Q(Y)|Y] = \sum_{i=1}^{d} B \cdot sign(Y(i)) \cdot e_i \cdot \frac{|Y(i)|}{B} = Y,$$

namely *SimQ* is unbiased. To complete the proof, note that $\|Q(Y)\|_2^2 \leq B^2$ *a.s.*. □

Theorem 4.1 along with Theorem 2.3 establishes Theorem 3.1 for $p = \infty$.

4.2 Our Quantizer for $p \in [2, \infty)$

For this case, we need to quantize inputs that are bounded in ℓ_q norm with $q \in (1, 2]$ so that the quantized output is unbiased and has small expected ℓ_2 norm square; we will use $SimQ^+$ to do this.

SimQ$^+$ The quantizer $SimQ^+$ outputs the average of k independent repetitions of the *SimQ* quantizer for a given input vector. The input vectors Y satisfy $\|Y\|_1 \leq Bd^{1/p}$. Therefore, we use *SimQ* with parameter $Bd^{1/p}$ instead of B. The repetitions help reduce the error to compensate for the extra loss factor. Specifically, the output of $SimQ^+$ denoted by $Q(Y)$ is given by

$$Q(Y) = \frac{1}{k} \cdot \sum_{i=1}^{k} Q^i_{\texttt{SimQ}}(Y; Bd^{1/p}), \tag{5}$$

where $Q^i_{\texttt{SimQ}}$ are independent iterations of *SimQ*.

The next component of $SimQ^+$ is how the encoder of $SimQ^+$ expresses the output of these k copies of *SimQ* to attain compression. If represented naively, this will require $O(d^{2/p} \log d)$. But we can do much better since we only need the average value of these entries. For that, we can simply report the *type* of this vector – the frequency of each index in the k length sequence. The signs of the input coordinates for the non-zero entries can be sent separately.

Note that there are $d+1$ indices overall, as *SimQ* can pick any index from $\{0, \ldots d\}$. Therefore, the total number of types is $\binom{d+k}{k}$, which can at the most be $(\frac{ed+ek}{k})^k$ bits. Hence, the precision needed to represent the type is at the most $k \log e + k \log(\frac{d}{k} + 1)$.

The type of the input can be used to determine a set $\mathcal{I}_0$ of non-zero indices that appear at least once. There are at most k such entries. Therefore, we can use a binary vector of length k to store the signs for these entries. We use this representation in $SimQ^+$, with the indices in $\mathcal{I}_0$ represented in the vector in increasing order.

Theorem 4.2. *For a $p \in [2, \infty)$, let Q be the quantizer $SimQ^+$ described in (5). Then, for Y such that $\|Y\|_q \leq B$ a.s., $Q(Y)$ can be represented in $k \log e + k \log(\frac{d}{k} + 1) + k$ bits, $\mathbb{E}[Q(Y)|Y] = Y$, and $\alpha_0(Q; p) \leq \sqrt{\frac{B^2 d^{2/p}}{k} + B^2}$.*

Proof. We already saw how to represent the output of the k copies of $SimQ$ using $k\log e + k\log(\frac{d}{k}+1)+k$ bits. For bounding $\alpha_0(Q;p)$, note from (5) that $SimQ^+$ is an unbiased quantizer since $SimQ$ is unbiased. Further, denoting by $Q_i(Y)$ the output $Q^i_{\mathtt{SimQ}}(Y; Bd^{1/p})$, we get

$$\begin{aligned}
\mathbb{E}\left[\|Q(Y)\|_2^2\right] &= \mathbb{E}\left[\|Q(Y)-Y\|_2^2\right] + \mathbb{E}\left[\|Y\|_2^2\right] \\
&= \frac{1}{k^2}\sum_{i=1}^{k}\mathbb{E}\left[\mathbb{E}\left[\|Q_i(Y)-Y\|_2^2|Y\right]\right] + \mathbb{E}\left[\|Y\|_2^2\right] \\
&= \frac{\mathbb{E}\left[\|Q_1(Y)-Y\|_2^2\right]}{k} + \mathbb{E}\left[\|Y\|_2^2\right] \\
&\le \frac{d^{2/p}B^2}{k} + B^2,
\end{aligned}$$

where the first identity uses the fact that $Q(Y)$ is an unbiased estimate of Y; the second uses the fact that $Q_i(Y)-Y$ are zero-mean, independent random variables when conditioned on Y; the third uses the fact that $Q_i(Y)-Y$ are identically distributed; and the final inequality is by the performance of $SimQ$. □

The proof of upper bound for $p\in[2,\infty)$ in Theorem 3.1 is completed by setting $k = d^{2/p}$ and using Theorems 4.2 and 2.3.

5 Our Quantizers for $p \in [1, 2)$

For p in $[1,2)$, the oracle yields unbiased subgradient estimates Y such that $\|Y\|_q \le B$ almost surely. Our goal is to quantize such Ys in an unbiased manner and ensure that $\mathbb{E}\left[\|Q(Y)\|_q^2\right]$ is $O(B^2)$. It can be seen that a simple unbiased uniform quantizer will achieve this using $d(\log d)^{1/2-1/q}$. However, our goal here is to get a result that is stronger than this baseline performance. To that end, we split the input vector Y in two parts Y_1 and Y_2 with the first part having ℓ_∞ norm less than c and the second part having less than d/Δ_1 nonzero coordinates. We use an "ℓ_∞ ball quantizer" (a uniform quantizer) for Y_1 and an "ℓ_2 ball quantizer" for Y_2.

Specifically, set $c := \frac{B\Delta_1^{1/q}}{d^{1/q}}$, where Δ_1 is that in Theorem 3.1. Then, define

$$Y_1 := \sum_{i=1}^{d} Y(i)\mathbb{1}_{\{|Y(i)|\le c\}}e_i, \quad Y_2 := \sum_{i=1}^{d} Y(i)\mathbb{1}_{\{|Y(i)|> c\}}e_i. \tag{6}$$

Clearly, $\|Y_1\|_\infty \le c$. Further, since $\|Y\|_q \le B$, the number of nonzero coordinates in Y_2 can be at the most $B^q/c^q = d/\Delta_1$. For quantizing Y_1, we use the coordinate-wise uniform quantizer (CUQ) described below.

CUQ We note that CUQ is an unbiased, randomized uniform quantizer which has appeared in multiple works recently (*cf.* [20], [14]). We follow the treatment in [14]. CUQ has two parameters: M which describes the complete dynamic range $[-M, M]$ of the quantizer; k which describes the number of levels to which $[-M, M]$ is quantized to uniformly. In order to quantize Y_1 in (6), we set

$$M = c, \quad \log(k+1) = \left\lceil \log(2\sqrt{2}\Delta_1^{1/q}+2)\right\rceil. \tag{7}$$

Lemma 5.1. *Let $Q_{\mathtt{u}}$ be the quantizer CUQ with parameters M and k set as in* (7)*. Then, for Y such that $\|Y\|_q \le B$ a.s. and Y_1 as that in* (6)*, $Q_{\mathtt{u}}(Y_1)$ can be represented in $d\left\lceil \log(2\sqrt{2}\Delta_1^{1/q}+2)\right\rceil$ bits, $\mathbb{E}\left[Q_{\mathtt{u}}(Y_1) \mid |Y\right] = Y_1$, and $\mathbb{E}\left[\|Q_{\mathtt{u}}(Y_1)\|_q^2\right] \le 3B^2$.*

Proof. CUQ requires a precision of $d\log(k+1)$, which coincides with the statement above for our choice of k. To see unbiasedness, note that CUQ is an unbiased quantizer as long as all the coordinates of the input do not exceed M. Since we have set $M = c$ and $\|Y_1\|_\infty = c$, this property holds. Finally, to show that $\mathbb{E}\left[\|Q_{\mathtt{u}}(Y_1)\|_q^2\right] \le 3B^2$, note that $\mathbb{E}\left[\|Q_{\mathtt{u}}(Y_1)\|_q^2\right] \le 2\mathbb{E}\left[\|Q_{\mathtt{u}}(Y_1) - Y_1\|_q^2\right] + 2\mathbb{E}\left[\|Y_1\|_q^2\right]$. Also,

$$\begin{aligned}\mathbb{E}\left[\|Q_{\mathtt{u}}(Y_1) - Y_1\|_q^2\right] &\le \mathbb{E}\left[\sum_{i\in[d]} |Q_{\mathtt{u}}(Y_1(i)) - Y_1(i)|^q\right]^{2/q}\\ &\le B^2,\end{aligned}$$

where the first inequality uses Jensen's inequality as $(\cdot)^{2/q}$ is a concave function and second follows from fact that for M set as in (7) we have that $|Q_{\mathtt{u}}(Y_1)(i) - Y_1(i)| \le \frac{2M}{(k-1)}$ a.s., $\forall i \in [d]$, by the description of CUQ. □

In order to quantize Y_2, we indicate the coordinates with non-zero entries. This takes less than d bits. Then, we quantize the restriction Y_2' of Y_2 to these nonzero entries. Recall that the dimension of Y_2' is less than $d' := d/\Delta_1$. Also, the ℓ_2 norm of Y_2' is less than $\|Y\|_2 \le \|Y\|_q d^{1/2-1/q} \le Bd^{1/2-1/q} =: B'$.

We need a quantizer Q such that $\mathbb{E}\left[\|Q(Y_2')\|_q^2\right]$ is $O(B^2)$. As seen in the proof of Lemma 5.1, one way to do this is to ensure $\mathbb{E}\left[\|Q(Y_2') - Y'\|_q^2\right]$ is $O(B^2)$, which, in turn, can be ensured if $\mathbb{E}\left[\|Q(Y_2') - Y_2'\|_2^2\right]$ is $O(B^2)$. To achieve this, we can use an unbiased quantizer for the unit ℓ_2 ball in $\mathbb{R}^d$, which can quantize it to an MSE of $O(1/d^{1-2/q})$ using $O(d\log(d^{1/2-1/q})$ bits. We note that $SimQ^+$, while optimal for the stochastic optimization use-case, does not yield the required scaling of bits in MSE. A candidate quantizer is RATQ of [14], which is, in fact, close to information theoretically optimal.

We note that RATQ is a quantizer used to quantize random vectors in $\mathbb{R}^d$ which have Euclidean norm almost surely bounded by B. The optimal parameters for RATQ are set in terms of B and d. Since Y_2' has the Euclidean norm almost surely bounded by $B' = Bd^{1/2-1/q}$ and its dimension is $d' = d/\Delta_1$, we can set the parameters of RATQ in terms of B' and d'. We set

$$\begin{aligned}&m = \frac{3B'^2}{d'}, \quad m_0 = \frac{2B'^2}{d'}\cdot \ln s, \quad \log h = \lceil \log(1+\ln^*(d'/3))\rceil,\\ &s = \log h, \quad \log(k+1) = \Delta_1. \end{aligned}\tag{8}$$

Lemma 5.2. *Let $Q_{\mathtt{at},R}$ be the quantizer RATQ with parameters set as* (8)*. Then, for Y such that $\|Y\|_q \le B$ a.s. and Y_2' the restriction of Y_2 in* (6)*, $Q_{\mathtt{at},R}(Y_2')$ can be represented in $2d + \Delta_2$ bits, $\mathbb{E}\left[Q_{\mathtt{at},R}(Y_2') \mid |Y\right] = Y_2'$, and $\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2')\|_q^2\right] \le 3B^2$.*

Proof. First, we note that the output of RATQ can be represented in $\lceil d'/s\rceil(\log h) + d\log(k+1)$ bits, which, in this case, is less than

$$\frac{d}{\Delta_1 \log h}\cdot(\log h) + \log h + \left(\frac{d}{\Delta_1}\log(k+1)\right) \le 2d + \Delta_2.$$

For unbiasedness, note that for our choice of m, m_0, h, RATQ is always an unbiased quantizer of the input. Finally, for showing $\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2')\|_q^2\right] \le 3B^2$, we note that

$$\begin{aligned}\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2')\|_q^2\right] &\le 2\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2') - Y_2'\|_q^2\right] + 2\mathbb{E}\left[\|Y_2'\|_q^2\right] \\ &\le 2\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2') - Y_2'\|_q^2\right] + 2B^2 \\ &\le 2\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2') - Y_2'\|_2^2\right] + 2B^2.\end{aligned}$$

The proof will be complete upon showing that $\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2') - Y_2'\|_2^2\right] \le B^2/2$, towards which we apply [14, Lemma 6.3] to get

$$\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y_2') - Y_2'\|_2^2\right] \le B^2 d^{1-2/q} \cdot \frac{9 + 3\ln s}{(k-1)^2},$$

and substituting our choice of k. □

The overall quantizer Q of input vector Y is the sum of quantized outputs of Y_1 and Y_2. By Lemmas 5.1 and 5.2, the quantized output of Y can be represented in $d\left(\left\lceil \log(2\sqrt{2}{\Delta_1}^{1/q} + 2)\right\rceil + 3\right) + \Delta_2$ bits[4]. Furthermore, $\alpha_0(Q;p) \le \sqrt{12}B$. These facts along with Theorem 2.3 prove the upper bound in Theorem 3.1 for $p \in [1, 2)$.

6 Proof of lower bounds in Theorem 3.1

We derive the lower bounds on the maxmin error $\mathcal{E}_0^*(T, r, p)$ defined in (4). The lower bounds of Theorem 3.1 will follow by upper-bounding $\mathcal{E}_0^*(T, r, p)$ by $\mathcal{U}(T, p)$. These information theoretic lower bounds generalize the ones derived for the Euclidean case in [14].

In our proof below, we use the reduction of convex optimization to mean estimation given in [3]. We use two different types of oracle constructions. The first one is based on Bernoulli product distribution and yields the lower bound using a strong data processing inequality from [9]. The second one is Paninski's construction [17] and uses a strong data processing inequality type bound (*cf.* the chi-square contraction bound in [1]). Interestingly, for $p > 2$ we need to use both the oracle constructions, whereas for $p \in [1, 2)$ Bernoulli product construction is sufficient. Heuristically, for $p > 2$, Paninski's construction captures the difficulty for optimization but does not pose much additional difficulty for quantized oracles. On the other hand, the Bernoulli product construction is not the bottleneck for optimization but poses a significant challenge if the oracle is quantized.

Theorem 6.1. *For $p \in [2, \infty]$, we have*

$$\mathcal{E}_0^*(T, r, p) \ge \left(\frac{c_0 DB \cdot d^{1/2-1/p}}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge 2^r}}\right) \vee \left(\frac{c_0 DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}\right).$$

Proof.

[4] This accounts for the communication needed to send the nonzero indices of Y_2, too.

First Lower Bound For simplicity, we assume $\mathcal{X} = \{x : \|x\|_\infty \leq D/(2d^{1/p})\}$. Let $\mathcal{V} \subset \{-1, 1\}^d$ be the maximal $d/4$-packing in Hamming distance, namely it is a collection of vectors such that any two vectors $\alpha, \alpha' \in V$, $d_H(\alpha, \alpha') \geq d/4$. As is well-known, there exists such a packing of cardinality $2^{c_2 d}$, where c_2 is a constant. Consider convex functions f_α, $\alpha \in \mathcal{V}$, with domain $\mathcal{X}$ and satisfying assumptions (1) and (2) given below:

$$f_\alpha(x) := \frac{2B\delta}{d} \sum_{i=1}^{d} \alpha(i) x(i).$$

Note that the gradient of $f_\alpha(x)$ is given by $2B\delta\alpha/d$ for each $x \in \mathcal{X}$. For each f_α, consider the corresponding gradient oracles O_α which outputs $e_i \cdot B$ and $-e_i \cdot B$ with probabilities $(1+2\delta\alpha(i))/2d$ and $(1-2\delta\alpha(i))/2d$, respectively. We denote the distribution of output of oracle O_α by P_α.

Let V be distributed uniformly over $\mathcal{V}$. Consider the multiple hypothesis testing problem of determining V by observing samples from $Q^{\mathsf{e}}(Y)$ with Y distributed as P_V. Consider an optimization algorithm that outputs x_T after T iterations. Then, we have

$$\begin{aligned}
\mathbb{E}\left[f_\alpha(x_T) - f_\alpha(x^*)\right] \geq & \frac{DB\delta}{4d^{1/p}} P\left(f_\alpha(x_T) - f_\alpha(x^*) \geq \frac{DB\delta}{4d^{1/p}}\right) \\
= & \frac{DB\delta}{4d^{1/p}} P\left(\frac{B\delta}{d}\alpha^T(x_T - x^*) \geq \frac{DB\delta}{8d^{1/p}}\right) \\
= & \frac{DB\delta}{4d^{1/p}} P\left(\frac{B\delta}{d}\|x_T - x^*\|_1 \geq \frac{DB\delta}{8d^{1/p}}\right) \\
= & \frac{DB\delta}{4d^{1/p}} P\left(\|(2d^{1/p}/D)x_T + \alpha\|_1 \geq \frac{d}{4}\right),
\end{aligned}$$

where the second identity holds since $sign(\alpha(i)) = sign(x_T - x^*)$ and the final identity is obtained by noting that the optimal value x^* for f_α is $-(D/2d^{1/p})\alpha$. Note that all $\alpha, \alpha' \in \mathcal{V}$ satisfy $\|\alpha - \alpha'\|_1 \geq d/2$. Consider the following test for the aforementioned hypothesis testing problem. We execute the optimization protocol using oracle O_V and declare the unique $\alpha \in V$ such that $\|(2d^{1/p}/D)x_T + \alpha\|_1 < d/4$. The probability of error for this test is bounded above by $P\left(\|(2d^{1/p}/D)x_T + \alpha\|_1 \geq \frac{d}{4}\right)$, whereby the previous bound and Fano's inequality give

$$\mathbb{E}\left[f_\alpha(x_T) - f_\alpha(x^*)\right] \geq \frac{DB\delta}{4d^{1/p}}\left(1 - \frac{TI(V; Q(Y)) + 1}{\log|\mathcal{V}|}\right).$$

For a quantizer Q with precision r, using, for instance, the chi-square contraction bound from [1], we have $I(V; Q(Y)) \leq \delta^2 \min\{2^r, d\}/d$. Therefore,

$$\max_\alpha \mathcal{E}_0(f, \pi^{QO}) \geq \frac{DB\delta}{4d^{1/p}}\left(1 - \frac{T\delta^2(\min\{2^r, d\}/d)}{c_2 d}\right).$$

The proof is completed by maximizing the right-side over δ.

Second Lower Bound We consider a slightly different family of convex functions still parametrized by $\alpha \in V$, with everything else remaining the same as the first bound. Consider convex functions

f_α, $\alpha \in \mathcal{V}$, with domain $\mathcal{X}$ and satisfying assumptions (1) and (2) given below:

$$f_\alpha(x) := \frac{2B\delta}{d^{1/q}} \sum_{i=1}^{d} \alpha(i)x(i).$$

Note that the gradient of $f_\alpha(x)$ is given by $2B\delta\alpha/d^{1/q}$ for each $x \in \mathcal{X}$. For each f_α, consider the corresponding gradient oracles O_α which outputs independent values for each coordinate, with the value of ith coordinate taking values $B/d^{1/q}$ and $-B/d^{1/q}$ with probabilities $(1 + 2\delta\alpha(i))/2$ and $(1 - 2\delta\alpha(i))/2$, respectively. We denote the distribution of output of oracle O_α by P_α.

By similar argument as in the first bound, we have

$$\begin{aligned}
\mathbb{E}\left[f_\alpha(x_T) - f_\alpha(x^*)\right] \geq \frac{DB\delta}{4} P\left(f_\alpha(x_T) - f_\alpha(x^*) \geq \frac{DB\delta}{4}\right) \\
= \frac{DB\delta}{4} P\left(\frac{B\delta}{d^{1/q}} \alpha^T(x_T - x^*) \geq \frac{DB\delta}{8}\right) \\
= \frac{DB\delta}{4} P\left(\frac{B\delta}{d^{1/q}} \|x_T - x^*\|_1 \geq \frac{DB\delta}{8}\right) \\
= \frac{DB\delta}{4} P\left(\|(2d^{1/p}/D)x_T + \alpha\|_1 \geq \frac{d}{4}\right), \\
= \frac{DB\delta}{4} \left(1 - \frac{TI(V; Q(Y)) + 1}{\log|\mathcal{V}|}\right)
\end{aligned}$$

For a quantizer Q with precision r, using the strong data processing inequality bound from [9, Proposition 2], we have $I(V; Q(Y)) \leq 360\delta^2 \min\{r, d\}$. Therefore,

$$\max_\alpha \mathcal{E}_0(f, \pi^{QO}) \geq \frac{DB\delta}{4}\left(1 - \frac{1}{c_2 d} - \frac{360T\delta^2 \min\{r, d\}}{c_2 d}\right).$$

The proof is completed by maximizing the right-side over δ.

□

Next, noting that the second lower bound in Theorem 6.1 also holds for $p \in [1, 2)$, we obtain the following lower bound.

Theorem 6.2. *For $p \in [1, 2)$, we have*

$$\mathcal{E}_0^*(T, r, p) \geq \frac{c_0 DB}{\sqrt{T}} \cdot \sqrt{\frac{d}{d \wedge r}}.$$

7 Comments on general tradeoff and mean square bounded oracles

We close with the remark that an almost complete characterization of $\mathcal{E}_0^*(T, r, p)$, for any r, p can be obtained using our quantizers and the ideas developed in this paper. In fact, our lower bounds on $r^*(T, p)$ are derived via lower bounds on $\mathcal{E}_0^*(T, r, p)$ which hold for all r, p (these can be found in the extended version). For upper bounds when $p \in [1, 2)$, note that the parameter k of $SimQ^+$

gives us a nice lever to operate under any precision constraint $r \geq \log d$. It turns out that such a quantizer leads to upper bounds which are off by at the most by a $\sqrt{\log d}$ factor. For upper bounds in the case of $p \in [1, 2)$, note that classical sampling techniques such as uniform sampling without replacement maybe used to sample a subset of coordinates and quantize them using the quantizers described here. Even in this case the upper bound and lower bounds are off by a nominal factor of $\sqrt{\log d \cdot \log\log d/q}$. However, we believe that removing these remaining factors can lead to new quantizers, and is of research interest.

For mean square bounded oracles mentioned in Remark 1, the bias in the quantized oracle output is nearly inevitable. In our previous work [14], we proposed appropriate *gain-shape* quantizers for quantizing the oracle output in the Euclidean setup, which resulted in lesser bias over standard quantizers. This idea is valid for the general ℓ_p setup; in particular, we can use a gain quantizer to quantize the ℓ_q norm of the oracle output and a shape quantizer to quantize the oracle output vector normalized by the ℓ_q norm, the shape of the oracle output vector. Note that the shape vector has has ℓ_q norm 1, which allows us to use the quantizers developed in this paper to quantize the shape. The gain is a scalar random variable which has its second moment bounded by B^2. To quantize such a random variable, we can use the quantizer proposed in [14] termed *Adaptive Geometric Uniform Quantizer* (AGUQ). Clearly, the lower bounds for almost surely bounded oracles remain valid for mean square bounded oracles as well. Additionally, we can also derive lower bounds for a specific class of quantizers, such as those derived in [14], which help in capturing the reduction in the convergence rate due to mean square bounded noise. However, we do not have matching bounds, even for the Euclidean case.

Acknowledgement

The authors would like to thank Arya Mazumdar and Anand Theertha Suresh for useful discussions.

Prathamesh Mayekar is supported by a PhD fellowship from Wipro Limited. Himanshu Tyagi is supported by a grant from Robert Bosch Center for Cyberphysical Systems (RBCCPS), Indian Institute of Science, Bangalore and the grant EMR/2016/002569 from the Department of Science and Technology (DST), India.

References

[1] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under Information Constraints I: Lower Bounds from Chi-Square Contraction," *arXiv:1812.11476 (a shorter version appeared at COLT 2019)*, 2018.

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

[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] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," *Advances in Neural Information Processing Systems*, pp. 7564–7575, 2018.

[5] 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.

[6] D. Basu, D. Data, C. Karakus, and S. Diggavi, “Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations,” *arXiv:1906.02367*, 2019.

[7] S. Bubeck, “Convex optimization: Algorithms and complexity,” *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[8] C. M. De Sa, C. Zhang, K. Olukotun, and C. Ré, “Taming the wild: A unified analysis of Hogwild-style algorithms,” *Advances in Neural Information Processing Systems*, pp. 2674–2682, 2015.

[9] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and Y. Zhang, “Optimality guarantees for distributed statistical estimation,” *arXiv:1405.0782*, 2014.

[10] V. Feldman, C. Guzmán, and S. Vempala, “Statistical query algorithms for mean vector estimation and stochastic convex optimization,” *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA’17)*, pp. 1265–1277, 2017.

[11] V. Gandikota, R. K. Maity, and A. Mazumdar, “vqSGD: Vector quantized stochastic gradient descent,” *arXiv preprint arXiv:1911.07971*, 2019.

[12] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, “Deep learning with limited numerical precision,” *Proceedings of the International Conference on Machine Learning (ICML’ 15)*, pp. 1737–1746, 2015.

[13] S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi, “Error feedback fixes signsgd and other gradient compression schemes,” *arXiv:1901.09847*, 2019.

[14] P. Mayekar and H. Tyagi, “RATQ: A universal fixed-length quantizer for stochastic optimization,” *arXiv:1908.08200 (a shorter version to appear at AISTATS 2020)*, 2019.

[15] A. Nemirovski, “Information-based complexity of convex programming,” 1995, Available Online http://www2.isye.gatech.edu/ne-mirovs/Lec_EMCO.pdf.

[16] A. S. Nemirovsky and D. B. Yudin, “Problem complexity and method efficiency in optimization.” *Wiley series in Discrete Mathematics and Optimization*, 1983.

[17] L. Paninski, “A coincidence-based test for uniformity given very sparsely sampled discrete data,” *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750–4755, 2008.

[18] A. Ramezani-Kebrya, F. Faghri, and D. M. Roy, “Nuqsgd: Improved communication efficiency for data-parallel sgd via nonuniform quantization,” *arXiv preprint arXiv:1908.06077*, 2019.

[19] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, “Sparsified SGD with memory,” *Advances in Neural Information Processing Systems*, pp. 4447–4458, 2018.

[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’ 17)*, vol. 70, pp. 3329–3337, 2017.

[21] 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.

[22] 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.