RATQ: A Universal Fixed-Length Quantizer for Stochastic Optimization

Prathamesh Mayekar[†] Himanshu Tyagi[†]

Abstract

We present Rotated Adaptive Tetra-iterated Quantizer (RATQ), a fixed-length quantizer for gradients in first order stochastic optimization. RATQ is easy to implement and involves only a Hadamard transform computation and adaptive uniform quantization with appropriately chosen dynamic ranges. For noisy gradients with almost surely bounded Euclidean norms, we establish an information theoretic lower bound for optimization accuracy using finite precision gradients and show that RATQ almost attains this lower bound.

For mean square bounded noisy gradients, we use a gain-shape quantizer which separately quantizes the Euclidean norm and uses RATQ to quantize the normalized unit norm vector. We establish lower bounds for performance of any optimization procedure and shape quantizer, when used with a uniform gain quantizer. Finally, we propose an adaptive quantizer for gain which when used with RATQ for shape quantizer outperforms uniform gain quantization and is, in fact, close to optimal.

As a by-product, we show that our fixed-length quantizer RATQ has almost the same performance as the optimal variable-length quantizers for distributed mean estimation. Also, we obtain an efficient quantizer for Gaussian vectors which attains a rate very close to the Gaussian rate-distortion function and is, in fact, universal for subgaussian input vectors.

[†]Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India. Email: {prathamesh, htyagi}@iisc.ac.in

Contents

1 Introduction

Stochastic gradient descent (SGD) and its variants are popular optimization methods for machine learning. In its basic form, SGD performs iterations $x_{t+1} = x_t - \eta \hat{g}(x_t)$, where $\hat{g}(x)$ is a noisy estimate of the subgradient of the function being optimized at x. Our focus in this work is on a distributed implementation of this algorithm where the output $\hat{g}(x)$ of the first order oracle must be quantized to a precision of r bits. This abstraction models important scenarios ranging from distributed optimization to federated learning, and is of independent theoretical interest.

We study the tradeoff between the convergence rate of first order optimization algorithms and the precision r available per subgradient update. We consider two *oracle models*: the first where the subgradient estimate's Euclidean norm is *almost surely bounded* and the second where it is *mean square bounded*. Our main contributions include new quantizers for the two oracle models and theoretical insights into the limitations imposed by heavy-tailed gradient distributions admitted under the mean square bounded oracles. A more specific description of our results and their relation to prior work is provided below.

1.1 Prior work

SGD and the oracle model abstraction for it appeared in classic works [42] and [38], respectively. We refer the reader to textbooks and monographs [13, 37, 39] for a review of the basic setup. Recently, variants of this problem with quantization or communication constraints on oracle output have received a lot of attention [3, 6, 9, 10, 16, 24, 34, 41, 46–49]. Our work is motivated by the results in [9, 47], and we elaborate on the connection.

Specifically, [9] considers a problem very similar to ours. The paper [47] considers the related problem of distributed mean estimation, but the quantizer and its analysis is directly applicable to distributed optimization. The two papers present different quantizers that encode each input using a variable number of bits. Both these quantizers are of optimal expected precision for almost surely bounded oracles. However, their worst-case (fixed-length) performance is suboptimal.

In fact, the problem of designing fixed-length quantizers for almost surely bounded oracles is closely related to designing small-size covering for the Euclidean unit ball. There has been a longstanding interest in this problem in the vector quantization and information theory literature (*cf.* [15, 19, 23, 33, 35, 51]). A closely related problem is that of Gaussian rate-distortion where we seek to quantize a random Gaussian vector to within a specified mean squared error, while using as few bits per dimension as possible (*cf.* [14, 20]). Typical fixed-length schemes for this problem draw on its duality with the channel coding problem and modify channel codes to obtain coverings; see, for instance, [36, 45, 53]. However, for our application of distributed optimization, these schemes may not be acceptable for two reasons: First the resulting complexity is still too high for hardware implementation; and second, the resulting schemes are not universal and are tied to Gaussian distributions specifically.

To the best of our knowledge, the spherical covering code construction with the least known encoding complexity is from [26, 27, 29] (see [28] for more details). However, its performance rate-distortion and computational complexity have been analyzed only for the Gaussian source; it has been shown in [28] that the computational complexity needed grows linearly in rate. In comparison, the variant of our scheme for the Gaussian source is especially simple and has only constant computational complexity per dimension.

In a slightly different direction, a seminal, but perhaps not so widely known, result of [55] provides a very simple universal quantizer for random vectors with independent and identically distributed (*iid*) coordinates, with each coordinate almost surely bounded. In this scheme, we first quantize each coordinate uniformly, separately using a "scalar-quantizer," and then apply a universal entropic compression scheme to the quantized vector. We note that the variable-length schemes proposed in [9, 47] are very similar, albeit with a specific choice of the entropic compression scheme.

All these schemes are variable-length schemes, while it is desirable to get a fixed-length scheme for the ease of both protocol and hardware implementation. We remark that indeed [47] presents an interesting randomly-rotate and quantize fixed-length scheme, but it still requires communicating $O(\log\log d)$ times more than the optimal fixed-length quantizer for the unit Euclidean ball given in [51]. To the best of our knowledge, prior to our work, the quantizer in [47] is the best known efficient fixed-length quantizer for the unit Euclidean ball.

In fact, a randomized orthogonal transform scheme similar to that in [47] appeared almost concurrently in [25] as well, where an analysis for Gaussian source is presented. However, a rate-distortion analysis has not been done in [25]. Remarkably, an early instance of the "rotated dithering" scheme for distributing energy equally appears in the image compression literature in [40], albeit without formal error or performance analysis. Another interesting scheme was proposed in [8] where nonuniform quantization (using *companding*) was combined with dithering. Our adaptive choice of dynamic range for uniform quantizers is similar, in essence, to companding. But our scheme differs from the one in [8] in several ways: First, [8] uses the knowledge of input distribution to design their companding function, whereas we only need knowledge of the tail behaviour of the input distribution in our setting; second, we apply a random rotation to our input leading to a universal quantizer, which is not needed in [8]; and finally, the specific structure of our quantizer with adaptive dynamic ranges makes it amenable to mean square error analysis for a large variety of sources.

Nevertheless, our proposed scheme has elements of all these approaches. We build on ideas similar to these works and develop a new approach for relating the mean square error to the dynamic range of our adaptive uniform quantizer. Interestingly, it leads to new results even for the well-studied Gaussian rate-distortion problem. Namely, we show that our scheme with constant computational complexity per dimension almost achieves the Gaussian rate distortion function, and that too universally among subgaussian sources. Moreover, the scheme and its analysis can be easily extended to sources with other tail behaviour. We believe that this approach will yield very efficient rate-distortion codes for various sources, answering a question of fundamental interest and having many applications.

Returning to the literature on quantizers for first order stochastic optimization, prior works including [9] remain vague about the analysis for mean square bounded oracles. Most of the works use gain-shape quantizers that separately quantize the Euclidean norm (*gain*) and the normalized vector (*shape*). But they operate under an engineering assumption: "the standard 32 bit precision suffices for describing the gain." One of our goals in this work is to evaluate how to efficiently use these 32 bits. For instance, can we use a simple uniform quantizer for gain?

To study such questions, we need a lower bound for gap to optimality for any optimization protocol using a uniform gain quantizer. However, such a lower bound is not available. Indeed, all prior lower bounds use almost surely bounded oracles and cannot establish an additional limitation for mean square bounded oracles with heavy tails. In fact, while information theoretic lower bounds for SGD are well-known (*cf.* [4]), even for the almost surely bounded oracle setting, bounds for quantized oracles (similar to [18]) have not been reported anywhere. We note that lower bounds for first-order optimization using quantized gradients are related to that of statistical learning and estimation when each sample must be quantized to a few bits (*cf.* [2, 22, 30, 43, 44, 54]).

Also, it is interesting to compare our qualitative results with those in [3]. The focus of [3] was to design algorithms that attain the optimal convergence rate using communication that is sublinear in d. As our lower bounds show, this is impossible in our setting. Rather, we provide a new scheme to reduce the dependence of the number of bits per dimension on T.

Finally, the related problem of memory constrained optimization was stated as an open problem in [50]. In this setting, we are only allowed to use a 2^M state machine to implement the optimization algorithm. While there is a high level connection between this problem and our problem, the memory constrained setting is more restrictive since the state of the algorithm must be restricted to M bits at every instance, as opposed to our setting where only the oracle output is restricted to a finite precision of r bits.

1.2 Our contributions

We start with almost surely bounded oracles and consider first order optimization protocols for d dimensional problems with T iterations. We begin by deriving a simple information theoretic, precision-dependent lower bound which shows that no optimization protocol using a first order oracle and gradient updates of precision $r < d$ bits can have gap to optimality smaller than roughly $\sqrt{d}/\sqrt{rT}$. In particular, we need precision exceeding $\Omega(d)$ bits to get the classic convergence rate of $1/\sqrt{T}$ for convex functions.

As our main contribution, we propose a new fixed-length quantizer we term *Rotated Adaptive Tetra-iterated Quantizer* (RATQ) that along with projected subgradient descent (PSGD) is merely a factor of $O(\log\log\log\log^* d)$ far from this minimum precision required to attain the $O(1/\sqrt{T})$ convergence rate. In a different setting, when the precision is fixed upfront to r, we modify RATQ by roughly quantizing and sending only a subset of coordinates of the rotated vector. We show that this modified version of RATQ is only a factor $O(\log\log^* d)$ far from the optimal convergence rate.

For the case of mean square bounded oracles, we establish an information theoretic lower bound in Section 4.1 which shows (using a heavy-tailed oracle) that the precision used for gain quantizer must exceed $\log T$ when the gain is quantized uniformly for T iterations and we seek $O(1/\sqrt{T})$ optimization accuracy. Thus, 32 bits are good for roughly a billion iterations with uniform gain quantizers, but not beyond that. Interestingly, we present a new, adaptive gain quantizer which can attain the same performance using only $\log\log T$ bits for quantizing gain. If one has 32 bits to spare for gain, then by using our quantizer we can handle algorithms with $2^{2^{32}}$ iterations, sufficient for any practical application.

As an application of our general construction, we revisit the distributed mean estimation problem considered in [47]. We show that using RATQ at each client requires fixed-length communication that is roughly the same as the optimal variable-length communication from [47]. Furthermore, we show that RATQ yields a fixed-length quantizer for the unit Euclidean ball that is only a factor $O(\log\log\log\log^* d)$ from the optimal communication, must better in comparison to the prior known quantizer with $O(\log\log d)$ factor gap to optimality.

Also, we consider the Gaussian rate-distortion problem and evaluate the performance of a subroutine of RATQ (without rotation). We show that this efficient quantizer requires a minuscule excess rate over the classic $(1/2)\log(\sigma^2/D)$ to get a normalized mean square error less than D. Further, our proposed quantizer is universal and applies to any random vector with centered subgaussian entries.

1.3 Remarks on techniques

In this work we use adaptive quantizers with multiple dynamic-ranges $\{[-M_i, M_i] : i \in [h]\}$, with possibly a different dynamic range chosen for each coordinate. Once a dynamic-range $[-M_i, M_i]$ is chosen for a coordinate, the coordinate is represented using a quantized uniformly within this dynamic-range using k levels. Using a different dynamic-range for each coordinate allows us to reduce error per coordinate, but costs us in communication since we need to communicate which M_i is used for each coordinate. In devising our scheme, we need to carefully balance this tradeoff. We do this by taking recourse to the following observation: when the same dynamic range is chosen for all coordinates, the mean square error per coordinate roughly grows as

$$O\left(\frac{\sum_{i\in[h]} M_i^2 \cdot p(M_{i-1})}{(k-1)^2}\right),$$

where $p(M)$ is the probability of the ℓ_∞ norm of the input vector exceeding M and k denotes the number of levels of the uniform quantizer. This observation allows us to relate the mean square error to the tail-probabilities of the ℓ_∞ norm of the input vector. In particular, we exploit it to decide on the subvectors which we quantize using the same dynamic range.

We use another classic trick (see [23]): we transform the input vector before we apply our adaptive quantizer. In particular, we use a randomized transform that expresses the input vector over a random basis. The specific choice of our random transform is determined by our assumption for the gradients, namely that their ℓ_2 norms are almost surely bounded by B.

Drawing from these ideas, we propose the quantizer RATQ for quantizing random vectors with ℓ_2 norm almost surely bounded by B. The main steps in RATQ are as follows:

1. *Rotate.* RATQ transforms the input vector by rotating it by multiplying with the randomized Hadamard transform which preservers the Euclidean norm. This specific random transform was also used in [7] for the Fast J-L transform. More recently, [47] used it to build a fixed length quantizer for distributed mean estimation. Incidentally, RATQ improves upon the performance of this fixed length quantizer for the problem of distributed mean estimation, as can be seen in Section 6.1. In both these works, the randomized Hadamard transform is used to control the infinity norm of the output vector.

2. *Adaptively quantized subvectors.* RATQ groups coordinates of the input vector to form smaller dimensional subvectors, after preprocessing the input vector using random rotation. Then, for each subvector the smallest dynamic range from the set $\{[-M_i, M_i] : i \in [h]\}$ is selected so that all the coordinates of that subvector lie within that range. Within this selected dynamic-range, each coordinate of the subvector is quantized uniformly. A key distinguishing feature of RATQ is choosing the set of M_is to grow as a tetration, roughly as $M_{i+1} = e^{M_i}$. The large growth rate of a tetration allows us to cover the complete range of each coordinate using only a small number of dynamic ranges, which leads to an unbiased quantizer and reduces the communication. Also, after random rotation, each coordinate of the vector is a centered subgaussian random variable with a variance-parameter of $O(B^2/d)$, which, despite the large growth rate of a tetration, ensures that the per coordinate mean square error between the quantized output and the input is almost a constant.

We remark that using an adaptively chosen dynamic-range can alternatively be implemented by transforming the input using a monotone function. This, too, is a classic technique in quantization known as *companding* (*cf.* [23]). Companding is known as a popular alternative to entropic coding for fixed-length codes. However, to the best of our knowledge, our paper is the first to combine it with other techniques and rigorously analyze it for the ℓ_2 norm bounded vector quantization problem. Perhaps it is a bit surprising that this combination of classic technique was not analysed for constructing an efficient covering of the unit Euclidean ball, the problem underlying our quantization problem. We separately highlight the performance of RATQ as a covering for the unit Euclidean ball in Section 6.1, in the context of distributed mean estimation.

Moving to oracles with mean square bounded ℓ_2 norms, we take recourse to gain-shape quantizers and quantize the (normalized) shape vector using RATQ. However, unlike prior work, we rigorously treat gain quantization. Our proposed quantizer for gain is once again an adaptive uniform quantizer, but this time we cannot use a tetration for selecting possible dynamic-ranges M_is since gain need not be subgaussian. We now only have tail-probability bounds determined by the Markov inequality (heavy-tails) and can only increase M_is geometrically.

In fact, the choice of M_is for both the gain-quantizer and RATQ above is based on our general procedure for selecting M_is based on the tail-probability bounds for the coordinates. As another instantiation of this principle, we study the Gaussian rate-distortion problem where we have a handle over these tail-probabilities, even without any additional transforms applied to the input random vector.

To extend our quantizers (for both almost surely and mean square bounded oracles) to the fixed precision setting where only r less than d bits can be sent, we take recourse to the standard uniform subsampling technique. Specifically, we uniformly sample $O(r)$ coordinates from $[d]$ (without replacement) and communicate quantized values only for these coordinates.

Our lower bounds draw from an oracle complexity lower bound derived in [5] and use a strong data processing inequality from [18]. Similar ideas have appeared in lower bounds for communication constrained statistics; see, for instance, [1, 12, 52, 54]. However, this only allows us to obtain lower bounds for the almost surely bounded setting. For the mean square bounded setting, we need a new construction with "heavy tails". In particular, our proposed heavy-tailed construction shows a bottleneck for uniform gain quantizers which can be circumvented by our proposed quantizer, thereby establishing a strict improvement over uniform gain quantizers.

We remark that independent of our work, an adaptive quantizer similar to the one with we use for gain-quantization with geometrically increasing M_is appears in [41]. Note that we use this quantizer for gain-quantization, while [41] uses it to quantize the shape. However, the setting considered is that of [9] where quantization is followed by entropic compression. In particular, the fixed-length performance is suboptimal and mean square bounded oracles are not handled in the worst-case. Another recent independent work [21] presents a different scheme where a different random transform is used instead of random rotation. However, the goal of this work is different from ours, and in particular, it has much worse communication requirement compared to our scheme.

1.4 Organization

We formalize our problem in the next section and describe our results for almost surely and mean square bounded oracles in Sections 3 and 4, respectively, along with some of the shorter proofs. The more elaborate proofs are provided in Section 5, with additional details relegated to the appendix. We present the application of our quantizers to the problem distributed mean estimation in Section 6.1 and Gaussian rate-distortion in Section 6.2.

2 The setup and preliminaries

2.1 Problem setup

We fix the number of iterations T of the optimization algorithm (the number of times the first order oracle is accessed) and the precision r allowed to describe each subgradient. Our fundamental metric of performance is the minimum error (as a function of T and r) with which such an algorithm can find the optimum value.

Formally, we want to find the minimum value of an unknown convex function $f : \mathcal{X} \to \mathbb{R}$ using *oracle access* to noisy subgradients of the function (*cf.* [13, 38]). We assume that the function f is convex over the compact, convex domain $\mathcal{X}$ such that $\sup_{x,y\in\mathcal{X}} \|x - y\|_2 \le D$; we denote the set of all such $\mathcal{X}$ 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

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

$$\mathbb{E}\left[\|\hat{g}(x)\|_2^2|x\right] \le B^2, \tag{2}$$

where $\partial f(x)$ denotes the set of subgradients of f at x.

Definition 2.1 (Mean square bounded oracle)**.** A first order oracle which upon a query x outputs the subgradient estimate $\hat{g}(x)$ satisfying the assumptions (1) and (2) is termed a mean square bounded oracle. We denote by $\mathcal{O}$ the set of pairs (f, O) with a convex function f and a mean square bounded oracle O.

The variant with almost surely bounded oracles has also been considered (*cf.* [4, 38]), where we assume for all $x \in \mathcal{X}$

$$P(\|\hat{g}(x)\|_2^2 \le B^2|x) = 1. \tag{3}$$

Definition 2.2 (Almost surely bounded oracle)**.** A first order oracle which upon a query x outputs only the subgradient estimate $\hat{g}(x)$ satisfying the assumptions (1) and (3) is termed an almost surely bounded oracle. We denote the class of convex functions and oracle's satisfying assumptions (1) and (3) by $\mathcal{O}_0$.

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 $Q = Q^{\mathsf{d}} \circ Q^{\mathsf{e}}$. Denote by $\mathcal{Q}_r$ the set of all such r-bit quantizers.

For an oracle $(f, O) \in \mathcal{O}$ 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.

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 $\mathcal{E}(f, \pi^{QO})$ defined as $\mathcal{E}(f, \pi^{QO}) := \mathbb{E}\left[f(x^*(\pi^{QO})) - \min_{x\in\mathcal{X}} f(x)\right]$. The fundamental quantity of interest in this work are minmax errors

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

$$\mathcal{E}^*(T, r) := \sup_{\mathcal{X}\in\mathbb{X}} \inf_{\pi\in\Pi_T} \inf_{Q\in\mathcal{Q}_r} \sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{QO}).$$

Clearly, $\mathcal{E}^*(T,r) \geq \mathcal{E}_0^*(T,r)$.

Remark 1. We restrict to *memoryless* quantization schemes where the same quantizer will be applied to each new gradient vector, without using any information from the previous updates. Specifically, at each instant t and for any precision r, the quantizers in $\mathcal{Q}_r$ do not use any information from the previous time instants to quantize the subgradient outputted by O at t.

2.2 A benchmark from prior results

We recall results for the classic setting with $r = \infty$. Prior work gives a complete characterization of the minmax errors $\mathcal{E}_0^*(T,\infty)$ and $\mathcal{E}^*(T,\infty)$ for this setting; see, for instance, [5, 37, 38]. We summarize these well-known results below (*cf.* [38], [4, Theorem 1a]).

Theorem 2.3. *For an absolute constant* c_0*, we have*

$$\frac{DB}{\sqrt{T}} \geq \mathcal{E}^*(T,\infty) \geq \mathcal{E}_0^*(T,\infty) \geq \frac{c_0 DB}{\sqrt{T}}.$$

2.3 Quantizer performance for finite precision optimization

Our overall optimization protocol throughout is the *projected SGD* (PSGD) (see [13]). In fact, we establish lower bound showing roughly the optimality of PSGD with our quantizers.

In PSGD the standard SGD updates are projected back to the domain using the projection map $\Gamma_{\mathcal{X}}$ given by $\Gamma_{\mathcal{X}}(y) := \min_{x \in \mathcal{X}} \|x - y\|_2$. We use the *quantized PSGD* algorithm described in Algorithm 1.

Require: $x_0 \in \mathcal{X}, \eta \in \mathbb{R}^+$, T and access to composed oracle QO
1: **for** $t = 0$ to $T-1$ **do**
   $x_{t+1} = \Gamma_{\mathcal{X}}\left(x_t - \eta Q(\hat{g}(x_t))\right)$
2: **Output:** $\frac{1}{T} \cdot \sum_{t=1}^{T} x_t$

Algorithm 1: Quantized PSGD with quantizer Q

The quantized output $Q(\hat{g}(x_t))$, too, constitutes a noisy oracle, but it can be biased for mean square bounded oracles. Though biased first-order oracles were considered in [32], the effect of quantizer-bias has not been studied in the past. The performance of a quantizer Q, when it is used with PSGD for mean square bounded oracles, is controlled by the worst-case L_2 norm $\alpha(Q)$ of its output and the worst-case bias $\beta(Q)$ defined as[1]

$$\begin{aligned}
\alpha(Q) &:= \sup_{Y \in \mathbb{R}^d : \mathbb{E}\left[\|Y\|_2^2\right] \leq B^2} \sqrt{\mathbb{E}\left[\|Q(Y)\|_2^2\right]}, \\
\beta(Q) &:= \sup_{Y \in \mathbb{R}^d : \mathbb{E}\left[\|Y\|_2^2\right] \leq B^2} \|\mathbb{E}\left[Y - Q(Y)\right]\|_2.
\end{aligned} \tag{4}$$

[1]We omit the dependence on B and d from our notation.

The corresponding quantities for almost surely bounded oracles are

$$\alpha_0(Q) := \sup_{Y \in \mathbb{R}^d : \|Y\|_2 \le B \text{ a.s.}} \sqrt{\mathbb{E}\left[\|Q(Y)\|_2^2\right]},$$
$$\beta_0(Q) := \sup_{Y \in \mathbb{R}^d : \|Y\|_2 \le B \text{ a.s.}} \|\mathbb{E}\left[Y - Q(Y)\right]\|_2. \tag{5}$$

Using a slight modification of the standard proof of convergence for PSGD, we get the following result.

Theorem 2.4. *For any quantizer Q, the output x_T of optimization protocol π given in Algorithm 1 satisfies*

$$\sup_{(f,O)\in\mathcal{O}_0} \mathcal{E}(f, \pi^{QO}) \le D\left(\frac{\alpha_0(Q)}{\sqrt{T}} + \beta_0(Q)\right),$$
$$\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{QO}) \le D\left(\frac{\alpha(Q)}{\sqrt{T}} + \beta(Q)\right),$$

when the parameter η is set to $D/(\alpha_0(Q)\sqrt{T})$ and $D/(\alpha(Q)\sqrt{T})$, respectively.

See Appendix A for the proof.

Remark 2 (Choice of learning rate). We fix the parameter η of Algorithm 1 to $D/(\alpha_0(Q)\sqrt{T})$ and $D/(\alpha(Q)\sqrt{T})$ for all the results in Section 3 and Section 4, respectively.

3 Main results for almost surely bounded oracles

Our main results will be organized along two regimes: the high-precision and the low-precision regime. For the high-precision regime, we seek to attain the optimal convergence rate of $1/\sqrt{T}$ using the minimum precision possible. For the low-precision regime, we seek to attain the fastest convergence rate possible for a given, fixed precision r.

3.1 A precision-dependent lower bound

We begin with a simple refinement of the lower bound implied by Theorem 2.3 The proof of this result is obtained by appropriately modifying the proof in [4], along with the strong data processing inequality in [18].

Theorem 3.1. *There exists an absolute constant c, independent of d, T, and r such that*

$$\mathcal{E}^*(T, r) \ge \mathcal{E}_0^*(T, r) \ge \frac{cDB}{\sqrt{T}} \cdot \sqrt{\frac{d}{\min\{d, r\}}}.$$

Proof. The proof of lower bound in Theorem 3.1 is a slight extension of the standard proof of Theorem 2.3. We provide a sketch for completeness. For simplicity, we assume $\mathcal{X} = \{x : \|x\|_\infty \le D/(2\sqrt{d})\}$. 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 $\alpha, \alpha' \in V$, $d_H(\alpha, \alpha') \ge d/4$. 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 (3) given below:

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

Note that the gradient of $f_\alpha(x)$ is given by $B\alpha/\sqrt{d}$ 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 $B/\sqrt{d}$ and $-B/\sqrt{d}$ 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.

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

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

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 $-(D/2\sqrt{d})\alpha$. 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 $\|(2\sqrt{d}/D)x_T + \alpha\|_1 < d/4$. The probability of error for this test is bounded above by $P\left(\|(2\sqrt{d}/D)x_T + \alpha\|_1 \geq \frac{d}{4}\right)$, whereby the previous bound and Fano's inequality give

$$\mathbb{E}\left[f_\alpha(x_T) - f_\alpha(x^*)\right] \geq \frac{DB\delta}{8}\left(1 - \frac{TI(V \wedge Q(Y)) + 1}{\log|\mathcal{V}|}\right).$$

For a quantizer Q with precision r, using the strong data processing inequality bound from [18, Proposition 2], we have $I(V \wedge Q(Y)) \leq 360\delta^2 \min\{r, d\}$. Therefore,

$$\max_\alpha \mathcal{E}_0(f, \pi^{QO}) \geq \frac{DB\delta}{8}\left(1 - \frac{1}{c_2 d} - \frac{360T\delta^2 \min\{r, d\}}{c_2 d}\right).$$

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

As a corollary, we get that there is no hope of getting the desired convergence rate of $1/\sqrt{T}$ by using a precision of less than d.

Corollary 3.2. *For $\mathcal{E}_0^*(T, r)$ or $\mathcal{E}^*(T, r)$ to be less than $DB/\sqrt{T}$, the precision r must be at least $\Omega(d)$.*

3.2 RATQ: Our quantizer for the ℓ_2 ball

We propose *Rotated Adaptive Tetra-iterated Quantizer* (RATQ) to quantize any random vector Y with $\|Y\|_2^2 \le B^2$, which is what we need for almost surely bounded oracles. RATQ first rotates the input vector, then divides the coordinates of the rotated vectors into smaller groups, and finally quantizes each subgroup-vector using a *Coordinate-wise Uniform Quantizer* (CUQ). However, the dynamic-range used for each subvector is chosen adaptively from a set of tetra-iterated levels. We call this adaptive quantizer *Adaptive Tetra-iterated Uniform Quantizer* (ATUQ), and it is the main workhorse of our construction. The encoder and decoder for RATQ are given in Algorithm 2 and Algorithm 3, respectively. The details of all the components involved are described below.

Require: Input $Y \in \mathbb{R}^d$, rotation matrix R
1: Compute $\tilde{Y} = RY$
2: **for** $i \in [d/s]$ **do**
  $\tilde{Y}_i^T = [\tilde{Y}((i-1)s+1), \cdots \tilde{Y}(\min\{is, d\})]^T$
3: **Output:** $Q^{\mathtt{e}}_{\mathtt{at},R}(Y) = \{Q^{\mathtt{e}}_{\mathtt{at}}(\tilde{Y}_1) \cdots Q^{\mathtt{e}}_{\mathtt{at}}(\tilde{Y}_{\lceil d/s \rceil})\}$

Algorithm 2: Encoder $Q^{\mathtt{e}}_{\mathtt{at},R}(Y)$ for RATQ

Require: Input $\{Z_i, j_i\}$ for $i \in [\lceil d/s \rceil]$, rotation matrix R
1: $Y^T = [Q^{\mathtt{d}}_{\mathtt{at}}(Z_1, j_1), \cdots, Q^{\mathtt{d}}_{\mathtt{at}}(Z_{\lceil d/s \rceil}, j_{\lceil d/s \rceil})]^T$
2: **Output:** $Q^{\mathtt{d}}_{\mathtt{at},R}(\{Z_i, j_i\}_{i=1}^{\lceil d/s \rceil}) = R^{-1}Y$

Algorithm 3: Decoder $Q^{\mathtt{d}}_{\mathtt{at},R}(Z, j)$ for RATQ

Rotation and division into subvectors. RATQ first rotates the input vector by multiplying it with a random Hadamard matrix. Specifically, denoting by H the $d \times d$ Walsh-Hadamard Matrix (see [31])[2], define

$$R := \frac{1}{\sqrt{d}} \cdot HD, \tag{6}$$

where D is a diagonal matrix with each diagonal entry generated uniformly from $\{-1, +1\}$. The input vector y is multiplied by R in the rotation step. The matrix D can be generated using shared randomness between the encoder and decoder.

Next, the rotated vector of dimension d is partitioned into $\lceil d/s \rceil$ smaller subvectors. The i^{th} subvector comprises the coordinates $\{(i-1)s+1, \cdots, \min\{is, d\}\}$, for all $i \in [d/s]$. Note that the dimension of all the sub vectors except the last one is s, with the last one having a dimension of $d - s\lfloor d/s \rfloor$.

Coordinate-wise Uniform Quantizer (CUQ). RATQ uses CUQ as a subroutine; we describe the latter for d dimensional inputs, but it will only be applied to subvectors of lower dimension in RATQ. CUQ has a dynamic range $[-M, M]$ associated with it, and it uniformly quantizes

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

each coordinate of the input to k-levels as long as the component is within the dynamic-range $[-M, M]$. Specifically, it partitions the interval $[-M, M]$ into parts $I_\ell := (B_{M,k}(\ell), B_{M,k}(\ell + 1)]$, $\ell \in \{0, \ldots, k-1\}$, where $B_{M,k}(\ell)$ are given by

$$B_{M,k}(\ell) := -M + \ell \cdot \frac{2M}{k-1}, \quad \forall \ell \in \{0, \ldots, k-1\}.$$

Note that we need to communicate $k+1$ symbols per coordinate – k of these symbols correspond to the k uniform levels and the additional symbol corresponds to the overflow symbol $\emptyset$. Thus we need a total precision of $d \lceil \log(k+1) \rceil$ bits to represent the output of the CUQ encoder. The encoder and decoders used in CUQ are given in Algorithms 4 and 5, respectively. In the decoder, we have set $B_{M,k}(\emptyset)$ to 0.

Require: Parameters $M \in \mathbb{R}^+$ and input $Y \in \mathbb{R}^d$

1: **for** $i \in [d]$ **do**
2: **if** $|Y(i)| > M$ **then**
$Z(i) = \emptyset$
3: **else**
4: **for** $\ell \in \{0, \ldots, k-1\}$ **do**
5: **if** $Y(i) \in (B_{M,k}(\ell), B_{M,k+1}(\ell + 1)]$ **then**

$$Z(i) = \begin{cases} \ell + 1, & w.p. \ \frac{Y(i) - B_{M,k}(\ell)}{B_{M,k}(\ell+1) - B_{M,k}(\ell)} \\ \ell, & w.p. \ \frac{B_{M,k}(\ell+1) - Y(i)}{B_{M,k}(\ell+1) - B_{M,k}(\ell)} \end{cases}$$

6: **Output:** $Q_{\mathtt{u}}^{\mathtt{e}}(Y; M) = Z$

Algorithm 4: Encoder $Q_{\mathtt{u}}^{\mathtt{e}}(Y; M)$ of CUQ

Require: Parameters $M \in \mathbb{R}^+$and input $Z \in \{0, \ldots, k-1, \emptyset\}^d$

1: Set $\hat{Y}(i) = B_{M,k}(Z(i))$, for all $i \in [d]$
2: **Output:** $Q_{\mathtt{u}}^{\mathtt{d}}(Z; M) = \hat{Y}$

Algorithm 5: Decoder $Q_{\mathtt{u}}^{\mathtt{d}}(Z; M)$ of CUQ

Adaptive Tetra-iterated Uniform Quantizer (ATUQ). The quantizer ATUQ is CUQ with its dynamic-range chosen in an adaptive manner. In order to a quantize a particular input vector, it first chooses a dynamic range from $[-M_i, M_i]$, $1 \le i \le h$. To describe these M_is, we first define the i^{th} tetra-iteration for e, denoted by e^{*i}, recursively as follows:

$$e^{*1} := e, \quad e^{*i} := e^{e^{*(i-1)}}, \quad i \in \mathbb{N}.$$

Also, for any non negative number b, we define $\ln^* b := \inf\{i \in \mathbb{N} : e^{*i} \ge b\}$. With this notation, the values M_is are defined in terms of m and m_0 as follows:

$$M_0^2 = m + m_0, \quad M_i^2 = m \cdot e^{*i} + m_0, \quad \forall i \in \{1, \ldots, h-1\}.$$

ATUQ finds the smallest level M_i which bounds the infinity norm of the input vector; if no such M_i exists, it simply uses M_{h-1}. It then uses CUQ with dynamic range $[-M_i, M_i]$ to quantize the input vector. In RATQ, we apply ATUQ to each subvector. The decoder of ATUQ is simply the decoder of CUQ using the dynamic range outputted by the ATUQ encoder.

Note that in order to represent the output of ATUQ for d dimensional inputs, we need a precision of at the most $\lceil \log h \rceil + d \lceil \log(k+1) \rceil$ bits: $\lceil \log h \rceil$ bits to represent the dynamic range and at the most $d \lceil \log(k+1) \rceil$ bits to represent the output of CUQ. The encoder and decoder for ATUQ are given in Algorithms 6 and 7, respectively.

Require: Input $Y \in \mathbb{R}^d$
1: **if** $\|Y\|_\infty > M_{h-1}$ **then**
Set $M^* = M_{h-1}$
2: **else**
Set $j^* = \min\{j : \|Y\|_\infty \leq M_j\}$, $M^* = M_{j^*}$
3: Set $Z = Q_{\mathtt{u}}^{\mathtt{e}}(Y; M^*, k)$
4: **Output:** $Q_{\mathtt{at}}^{\mathtt{e}}(Y) = \{Z, j^*\}$

Algorithm 6: Encoder $Q_{\mathtt{at}}^{\mathtt{e}}(Y)$ for ATUQ

Require: Input $\{Z, j\}$ with $Z \in \{0, \dots, k-1, \emptyset\}^d$ and $j \in \{0, \dots h-1\}$
1: **Output:** $Q_{\mathtt{at}}^{\mathtt{d}}(Z, j) = Q_{\mathtt{u}}^{\mathtt{d}}(Z; M_j)$

Algorithm 7: Decoder $Q_{\mathtt{at}}^{\mathtt{d}}(Z, j)$ for ATUQ

When ATUQ is applied to each subvector in RATQ, each of the $\lceil d/s \rceil$ subvectors are represented using less than $\lceil \log h \rceil + s \lceil \log(k+1) \rceil$ bits. Thus, the overall precision for RATQ is less than[3]

$$\lceil d/s \rceil \cdot \lceil \log h \rceil + d \lceil \log(k+1) \rceil$$

bits. The decoder of RATQ is simply formed by collecting the output of the ATUQ decoders for all the subvectors to form a d-dimensional vector, and rotating it back using the matrix R^{-1} (the inverse of the rotation matrix used at the encoder).

Choice of parameters. Throughout the remainder of this section, we set our parameters m, m_0, and h as follows

$$m = \frac{3B^2}{d}, \quad m_0 = \frac{2B^2}{d} \cdot \ln s, \quad \log h = \lceil \log(1 + \ln^*(d/3)) \rceil . \tag{7}$$

In particular, this results in $M_{h-1} \geq B$ whereby, for an input Y with $\|Y\|_2^2 \leq B^2$, RATQ outputs an unbiased estimate of Y.

[3] log denotes the logarithm to the base 2, ln denotes logarithm to the base e.

3.3 RATQ in the high-precision regime

The following result shows that RATQ is unbiased for almost surely bounded inputs and provides a bound for its worst-case second order moment; this constitutes a key technical tool for characterizing the performance of RATQ.

Theorem 3.3 (Performance of RATQ)**.** *Let $Q_{\mathtt{at},R}$ be the quantizer RATQ with $M_j s$ set by* (7)*. Then, for all $s, k \in \mathbb{N}$,*

$$\alpha_0(Q_{\mathtt{at},R}) \leq B\sqrt{\frac{9+3\ln s}{(k-1)^2}+1}, \quad \beta_0(Q_{\mathtt{at},R}) = 0. \tag{8}$$

The proof is deferred to Section 5.1.

Thus, α_0 is lower when s is small, but the overall precision needed grows since the number of subvectors increases. The following choice of parameters yields almost optimal performance:

$$s = \log h, \quad \log(k+1) = \left\lceil \log(2+\sqrt{9+3\ln s}) \right\rceil. \tag{9}$$

For these choices, we obtain the following.

Corollary 3.4. *The overall precision r used by the quantizer $Q = Q_{\mathtt{at},R}$ with parameters set as in* (7)*,* (9) *satisfies*

$$r \leq d(1+\Delta_1) + \Delta_2,$$

where $\Delta_1 = \left\lceil \log\left(2+\sqrt{9+3\ln \Delta_2}\right)\right\rceil$ and $\Delta_2 = \lceil \log(1+\ln^*(d/3))\rceil$.
Furthermore, the optimization protocol π given in Algorithm 1 satisfies

$$\sup_{(f,O)\in\mathcal{O}_0} \mathcal{E}(f, \pi^{QO}) \leq \frac{\sqrt{2}DB}{\sqrt{T}}.$$

Proof. By the description RATQ, it encodes the subgradients using a fixed-length code of at the most $\lceil d/s\rceil \cdot \lceil \log h\rceil + d\lceil \log(k+1)\rceil$ bits. Upon substituting s, $\log h$, and $\log(k+1)$ as in (9) and (7), we obtain that the total precision is bounded above by $d(1+\Delta_1)+\Delta_2$.

For the second statement of the corollary, we have

$$\begin{aligned}\sup_{(f,O)\in\mathcal{O}_0} \mathcal{E}(f, \pi^{QO}) &\leq D\left(\frac{\alpha_0(Q_{\mathtt{at},R})}{\sqrt{T}} + \beta_0(Q_{\mathtt{at},R})\right)\\ &\leq \frac{DB}{\sqrt{T}} \cdot \sqrt{\frac{9+3\ln s}{(k-1)^2}+1}\\ &\leq \frac{\sqrt{2}DB}{\sqrt{T}},\end{aligned}$$

where the first inequality follows by Theorem 2.4, the second inequality follows by upper bounding $\alpha_0(Q_{\mathtt{at},R})$ and $\beta_0(Q_{\mathtt{at},R})$ using Theorem 3.3, and the third follows by substituting the parameters in the corollary statement. □

Remark 3. The precision requirement in Corollary 3.4 matches the d bit lower bound of Corollary 3.2 upto a multiplicative factor of $O\left(\log\log\log \ln^*(d/3)\right)$.

3.4 Comparison with Quantized Stochastic Gradient Descent (QSGD)

At this point, it will be instructive to compare our results with a state of the art scheme from [9] – Quantized Stochastic Gradient Descent(QSGD). We have the following result as a consequence of [9, Lemma 3.1, Corollary 3.3].

Theorem 3.5. *Under the assumption that the $\ell_2$norm of the subgradient estimate can be communicated using F bits of communication, QSGD achieves for any $(f, O) \in \mathcal{O}_0$*

$$\mathcal{E}(f, \pi^{QO}) \leq \frac{\sqrt{2}DB}{\sqrt{T}},$$

using a variable-length code of expected precision at every iteration less than $F + 2.8d$.

To compare this performance with that of our proposed RATQ, we will show that unless the dimension d exceeds a astronomically large we will only need around $4d$ bits of a fixed-length code to communicate the subgradient at each round. That is, our fixed-length requires only 4 bits dimension. Formally, we have the following corollary.

Corollary 3.6. *Suppose we have $d \geq 3$ and $\ln^*(\frac{d}{3}) \leq 2^{8\cdot 10^3}$. Then, the overall precision r used by the quantizer $Q = Q_{\mathtt{at},R}$ with parameters set as in (7), (9) satisfies*

$$r \leq d\left(4 + \frac{\lceil \log(1 + \ln^*(d/3)) \rceil}{d}\right).$$

Furthermore, the optimization protocol π given in Algorithm 1 satisfies

$$\sup_{(f,O)\in\mathcal{O}_0} \mathcal{E}(f, \pi^{QO}) \leq \frac{\sqrt{2}DB}{\sqrt{T}}.$$

Proof. To see the first statement, we have from Corollary 3.4 that $r \leq d(1 + \Delta_1) + \Delta_2$. Note that Δ_1 is a monotonic function of $\ln^*(d/3)$. Therefore, by upper bounding $\ln^*(d/3)$ by $2^{8\cdot 10^3}$ in the formula for Δ_1 completes the proof. The second statement follows from Corollary 3.4. □

Thus, our proposed RATQ offers roughly the same compression as QSGD from [9] and, in addition, has the advantage of using only a fixed-length code, whereas QSGD uses a variable-length code whose worst-case length becomes significantly higher than the worst-case length of RATQ as d increases.

3.5 RATQ in the low-precision regime

We present a general method for reducing precision to much below r. This scheme is applicable when the output of the quantizer's encoder is a d length vector, where each coordinate is a separate fixed-length code. We simply reduce the length of the output message vector from the quantizer's encoder by sub-sampling a subset of coordinates using shared randomness. The decoder obtains the values of these coordinates using the decoder for the original quantizer and sets the rest of the coordinate-values to zero. This subsampling layer, which we call the *Random Coordinate Sampler* (RCS), can be added to RATQ after applying random rotation. In particular, *RATQ* we need the parameter s of these quantizers to be set to 1. This requirement of setting $s = 1$ ensures that the subsampled coordinates of the rotated vector can be decoded separately. This is a randomized scheme and requires the encoder and the decoder to share a random set $S \subset [d]$ distributed uniformly over all subsets of $[d]$ of cardinality μd.

The encoder $Q_S^{\mathtt{e}}$ of RCS simply outputs the vector

$$Q_S^{\mathtt{e}}(Y) := \{Y(i), i \in S\},$$

and the decoder $Q_S^{\mathtt{d}}(\tilde{Y})$, when applied to a vector $\tilde{Y} \in \mathbb{R}^{\mu d}$, outputs

$$Q_S^{\mathtt{d}}(\tilde{Y}) := \mu^{-1} \sum_{i \in S} \tilde{Y}(i) e_i,$$

where e_i denotes the ith element of standard basis for $\mathbb{R}^d$.

We can compose RCS with RATQ with parameter $s = 1$ by setting the encoder to $Q_S^{\mathtt{e}} \circ Q^{\mathtt{e}}$, and setting the decoder to $Q^{\mathtt{d}} \circ Q_S^{\mathtt{d}}$. Here we follow the convention that all 0-coordinates outputted by $Q_S^{\mathtt{d}}$ are decoded as 0 by $Q^{\mathtt{d}}$. Note that since we need to retain RATQ encoder output for only μd coordinates, the overall precision of the quantizer is reduced by a factor of μ. We analyze the performance of this combined quantizer in the following theorem.

Theorem 3.7. *Let $Q_{\mathtt{at},R}$ be RATQ with $s = 1$ and $\tilde{Q}$ be the combination of RCS and $Q_{\mathtt{at},R}$ as described above. Then,*

$$\mathbb{E}\left[\tilde{Q}(Y)|Y\right] = \mathbb{E}\left[Q_{\mathtt{at},R}(RY)|Y\right] \quad \textit{and} \quad \mathbb{E}\left[\|\tilde{Q}(Y)\|_2^2|Y\right] = \frac{1}{\mu}\mathbb{E}\left[\|Q_{\mathtt{at},R}(RY)\|_2^2|Y\right],$$

which further leads to

$$\alpha_0(\tilde{Q}) \le \frac{\alpha_0(Q_{\mathtt{at},R})}{\sqrt{\mu}} \quad \textit{and} \quad \beta_0(\tilde{Q}) = \beta_0(Q_{\mathtt{at},R}).$$

Proof. By the description of $Q_{\mathtt{at},R}$, we have

$$\tilde{Q}(Y) = \frac{1}{\mu} R^{-1} \sum_{i \in S} Q_{\mathtt{at},I}(RY)(i) e_i,$$

where $Q_{\mathtt{at},I}$ is the output vector formed by combining the d quantized values outputted by ATUQ ($Q_{\mathtt{at}}$) when input is the rotated vector. Namely,

$$Q_{\mathtt{at},I}(RY) = [Q_{\mathtt{at}}(RY(1)), \cdots, Q_{\mathtt{at}}(RY(d))]^T.$$

For the mean of $\tilde{Q}(Y)$, it holds that

$$\begin{aligned}
\mathbb{E}\left[\tilde{Q}(Y)|Y\right] &= \mathbb{E}\left[R^{-1} \sum_{i \in d} Q_{\mathtt{at},I}(RY)(i) e_i \frac{1}{\mu} \mathbb{1}_{i \in S}|Y\right] \\
&= \sum_{i \in d} \mathbb{E}\left[R^{-1} Q_{\mathtt{at},I}(RY)(i) e_i|Y\right] \cdot \frac{1}{\mu}\mathbb{E}\left[\mathbb{1}_{i \in S}|Y\right] \\
&= \sum_{i \in d} \mathbb{E}\left[R^{-1} Q_{\mathtt{at},I}(RY)(i) e_i|Y\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[R^{-1} \sum_{i \in d} Q_{\mathtt{at},I}(RY)(i) e_i | Y \right] \\
&= \mathbb{E}\left[R^{-1} Q_{\mathtt{at},I}(RY) | Y \right] \\
&= \mathbb{E}\left[Q_{\mathtt{at},R}(RY) | Y \right], \qquad (10)
\end{aligned}
$$

where the second identity follows from the fact that randomness used to generate a set S is independent of the randomness used in the quantizer and the randomness of Y; the third identity holds since $P(i \in S) = \mu$.

Next, moving to the computation of the second moment of the output of $\tilde{Q}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\|\tilde{Q}(Y)\|_2^2 | Y \right] &= \mathbb{E}\left[\| \frac{1}{\mu} R^{-1} \sum_{i \in S} Q_{\mathtt{at},I}(RY)(i) e_i \|_2^2 | Y \right] \\
&= \frac{1}{\mu^2} \mathbb{E}\left[\| \sum_{i \in S} Q_{\mathtt{at},I}(RY)(i) e_i \|_2^2 | Y \right] \\
&= \frac{1}{\mu^2} \sum_{i \in [d]} \mathbb{E}\left[Q_{\mathtt{at},I}(RY)(i)^2 | Y \right] \mathbb{E}\left[\mathbb{1}_{i \in S} | Y \right] \\
&= \frac{1}{\mu} \mathbb{E}\left[\| Q_{\mathtt{at}}(RY) \|_2^2 | Y \right] \\
&= \frac{1}{\mu} \mathbb{E}\left[\| Q_{\mathtt{at},R}(RY) \|_2^2 | Y \right], \qquad (11)
\end{aligned}
$$

where the second identity follows from the fact that R is a unitary matrix and the remaining steps follow simply by the description of the quantizers used. It follows that

$$
\alpha(\tilde{Q}) = \frac{1}{\sqrt{\mu}} \alpha(Q_{\mathtt{at},R}), \quad \beta(\tilde{Q}) = \beta(Q_{\mathtt{at},R}).
$$

□

We now set the parameter k to be a constant and sample roughly r coordinates. Specifically, we set

$$
\begin{aligned}
s &= 1, \quad \log(k+1) = 3, \\
\mu d &= \min\{d, \lfloor r/(3 + \lceil \log(1 + \ln^*(d/3)) \rceil) \rfloor\}. \qquad (12)
\end{aligned}
$$

For these choices, we have the following corollary.

Corollary 3.8. *For $r \geq 3 + \lceil \log(1 + \ln^*(d/3)) \rceil$, let Q be the composition of RCS and RATQ with parameters set as in* (7), (12). *Then, the optimization protocol π in Algorithm 1 satisfies*

$$
\sup_{(f,O) \in \mathcal{O}_0} \mathcal{E}(f, \pi^{QO}) \leq \frac{\sqrt{2} DB}{\sqrt{\mu T}}
$$

Proof. When Q is a composition of RCS and RATQ, from Theorem 3.7 $\alpha(Q) \leq \frac{1}{\sqrt{\mu}}\alpha(Q_{\mathtt{at},R})$, $\quad \beta(Q) \leq \beta(Q_{\mathtt{at},R})$, which by Theorem 2.4 yields

$$
\begin{aligned}
\sup_{(f,O)\in\mathcal{O}_0} \mathcal{E}(f,\pi^{QO}) &\leq D\left(\frac{\alpha_0(Q_{\mathtt{at},R})}{\sqrt{\mu T}} + \beta_0(Q_{\mathtt{at},R})\right) \\
&\leq \frac{DB}{\sqrt{\mu T}} \cdot \sqrt{\frac{9}{(k-1)^2} + 1} \\
&\leq \frac{\sqrt{2}DB}{\sqrt{T}} \cdot \frac{\sqrt{d}}{\sqrt{\min\{d, \lfloor r/(3+\log\ln^*(d/3))\rfloor\}}},
\end{aligned}
$$

where the second inequality follows from Theorem 3.3 with $s = 1$, and the final inequality is obtained upon substituting the parameters as in the statement of the result.

□

Remark 4. Note that the convergence rate slows down by a μ specified in (12), which matches the lower bound in Theorem 3.1 upto a multiplicative factor of $O(\log \ln^*(d/3))$

4 Main results for mean square bounded oracles

Moving to oracles satisfying the mean square bounded assumption, we now need to quantize random vectors Y such that $\mathbb{E}\left[\|Y\|_2^2\right] \leq B^2$. We take recourse to the standard *gain-shape* quantization paradigm in vector quantization ($cf.$[23]).

Definition 4.1 (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.

Specifically, we separately quantize the *gain* $\|Y\|_2$ and the *shape*[4] $Y/\|Y\|_2$ of Y, and form the estimate of Y by simply multiplying the estimates for the gain and the shape. Note that we already have a good shape quantizer: RATQ. We only need to modify the parameters in (7) to make it work for the unit sphere; we set

$$m = \frac{3}{d}, \quad m_0 = \frac{2}{d} \cdot \ln s, \quad \log h = \lceil \log(1 + \ln^*(d/3)) \rceil . \tag{13}$$

We now proceed to derive the worst-case α and β for a general gain-shape. In order to make clear the dependence on B and d, we refine our notations for $\{\alpha(Q), \beta(Q)\}$ and $\{\alpha_0(Q), \beta_0(Q)\}$, defined in (4) and (5), respectively, to $\{\alpha(Q; B, d), \beta(Q; B, d)\}$ and $\{\alpha_0(Q; B, d), \beta_0(Q; B, d)\}$.

Theorem 4.2. *Let $Q(Y) = Q_1(\|Y\|_2) \cdot Q_2(Y/\|Y\|_2)$, where Q_1 is any gain quantizer and Q_2 is any shape quantizer. Also, suppose $Q_1(\|Y\|_2)$ and $Q_2(Y/\|Y\|_2)$ are conditionally independent given Y. Then,*

$$\alpha(Q; B, d) \leq \alpha(Q_1; B, 1) \cdot \alpha_0(Q_2; 1, d).$$

[4]For the event $\|Y\|_2 = 0$, we follow the convention that $Y/\|Y\|_2 = e_1$.

Furthermore, suppose that Q_2 satisfies

$$\mathbb{E}\left[Q_2(y_s)\right] = y_s, \quad \forall y_s \quad s.t. \quad \|y_s\|_2^2 \le 1.$$

Then, we have[5]

$$\beta(Q; B, d) \le \sup_{Y \in \mathbb{R}^d : \mathbb{E}\left[\|Y\|_2^2\right] \le B^2} \mathbb{E}\left[\left|\mathbb{E}\left[Q_1(\|Y\|_2) - \|Y\|_2 \mid Y\right]\right|\right].$$

Proof. Denote by Y_s the shape of the vector Y given by

$$Y_s := \frac{Y}{\|Y\|_2}.$$

The worst-case second moment: Towards evaluating $\alpha(Q; B, d)$, we have

$$\begin{aligned}
\mathbb{E}\left[\|Q(Y)\|_2^2\right] &= \mathbb{E}\left[Q_1(\|Y\|_2)^2 \|Q_2(Y_s)\|_2^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)^2 \|Q_2(Y_s)\|_2^2 | Y\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)^2 | Y\right] \mathbb{E}\left[\|Q_2(Y_s)\|_2^2 | Y\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)^2 | Y\right] \mathbb{E}\left[\|Q_2(Y_s)\|_2^2 | Y_s\right]\right],
\end{aligned}$$

where the third identity follows by conditional independence of $Q_1(\|Y\|_2)^2$ and $\|Q_2(Y_s)\|_2^2$ given Y and the fourth follows from the law of iterated expectations.

Consider the random variable $\mathbb{E}\left[\|Q_2(Y_s)\|_2^2 | Y_s\right]$. We claim that this is less than $\alpha_0(Q_2; 1, d)$ almost surely. Towards this end, note that

$$\mathbb{E}\left[\|Q_2(Y_s)\|_2^2 | Y_s = y\right] = \mathbb{E}\left[\|Q_2(y)\|_2^2\right],$$

since the randomness used in implementation of Q_2 is independent of the input random variable Y. Moreover, for any y with $\|y\|_2^2 \le 1$, we have from the definition of $\alpha_0(Q_2; 1, d)$ that $\mathbb{E}\left[\|Q_2(y)\|_2^2\right] \le \alpha_0(Q_2; 1, d)^2$. Therefore, for any Y with $\mathbb{E}\left[\|Y\|_2^2\right] \le B^2$, we have

$$\begin{aligned}
\mathbb{E}\left[\|Q(Y)\|_2^2\right] &= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)^2 | Y\right] \mathbb{E}\left[\|Q_2(Y_s)\|_2^2 | Y_s\right]\right] \\
&\le \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)^2 | Y\right]\right] \cdot \alpha_0(Q_2; 1, d)^2 \\
&= \mathbb{E}\left[Q_1(\|Y\|_2)^2\right] \cdot \alpha_0(Q_2; 1, d)^2 \\
&\le \alpha(Q_1; B, 1)^2 \cdot \alpha_0(Q_2; 1, d)^2. \qquad (14)
\end{aligned}$$

Taking the supremum of the left-side over all random vectors Y with $\mathbb{E}\left[\|Y\|_2^2\right] \le B^2$ gives us the desired bound for $\alpha(Q)$.

[5] The quantity on the right-side of this bound exceeds the bias $\beta(Q_1; B, 1)$. Nonetheless, in all our bounds for bias, this is the quantity we have been handling.

The worst-case bias: Towards evaluating $\beta(Q)$, we note from our hypothesis that $\mathbb{E}\left[Q_2(Y_s)|Y\right] = \mathbb{E}\left[Q_2(Y_s)|Y_s\right] = Y_s$, which further yields

$$\begin{aligned}
\mathbb{E}\left[Q(Y) - Y\right] &= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)Q_2(Y_s) - Y|Y\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)|Y\right]\mathbb{E}\left[Q_2(Y_s)|Y\right] - Y\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2)|Y\right]Y_s - \|Y\|_2 Y_s\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2) - \|Y\|_2|Y\right]Y_s\right], \qquad (15)
\end{aligned}$$

where the second identity uses conditional independence of $Q_1(\|Y\|_2)$ and $Q_2(Y_s)$. By using the conditional Jensen's inequality, we get

$$\begin{aligned}
\|\mathbb{E}\left[Q(Y) - Y\right]\|_2 &= \|\mathbb{E}\left[\mathbb{E}\left[Q_1(\|Y\|_2) - \|Y\|_2|Y\right]Y_s\right]\|_2 \\
&\leq \mathbb{E}\left[\|\mathbb{E}\left[Q_1(\|Y\|_2) - \|Y\|_2|Y\right]Y_s\|_2\right] \\
&= \mathbb{E}\left[\left|\mathbb{E}\left[Q_1(\|Y\|_2) - \|Y\|_2|Y\right]\right|\right].
\end{aligned}$$

□

We remark that quantizers proposed in most of the prior work can be cast in this gain-shape framework. Most works simply state that gain is a single parameter which can be quantized using a fixed number of bits; for instance, a single double precision number is prescribed for storing the gain. However, the quantizer is not specified. We carefully analyze this problem and establish lower bounds when a uniform quantizer with a fixed dynamic range is used for quantizing the gain. Further, we present our own quantizer which significantly outperforms uniform gain quantization.

4.1 Limitation of uniform gain quantization

We establish lower bounds for a general class of gain-shape quantizers $Q(y) = Q_g(\|y\|_2)Q_s(y/\|y\|_2)$ of precision r that satisfy the following *structural assumptions*:

1. **(Independent gain-shape quantization)** For any given $y \in \mathbb{R}^d$, the output of the gain and the shape quantizers are independent.
2. **(Bounded dynamic-range)** There exists $M > 0$ such that $y \in \mathbb{R}^d$ such that whenever $\|y\|_2 > M$, $Q(y)$ has a fixed distribution $P_\emptyset$.
3. **(Uniformity)** There exists $m \in [M/2^r, M]$ such that for every t in $[0, m]$,
   (a) $\mathtt{supp}(Q_g(t)) \subseteq \{0, m\}$;
   (b) If $P(Q_g(t) = m) > 0$, then

$$\frac{P(Q_g(t_2) = m)}{P(Q_g(t_1) = m)} \leq \frac{t_2}{t_1}, \quad \forall\, 0 \leq t_1 \leq t_2 \leq m.$$

The first two assumptions are perhaps clear and hold for a large class of quantizers. The third one is the true limitation and is satisfied by different forms of uniform gain quantizers. For instance, for the one-dimensional version of CUQ with dynamic range $[0, M]$, which is an unbiased, uniform gain quantizer with k_g levels, it holds with $m = M/(k_g - 1)$ (corresponding to the innermost level

$[0, M/(k_g - 1)]$). It can also be shown to include a deterministic uniform quantizer that rounds-off at the mid-point. The third condition, in essence, captures the unbiasedness requirement that the probability of declaring higher level is proportional to the value. Note that (t_2/t_1) on the right-side can be replaced with any constant multiple of (t_2/t_1).

Below we present lower bounds for performance of any optimization protocol using a gain-shape quantizer that satisfies the assumptions above. We present separate results for high-precision and low-precision regimes, but both are obtained using a general construction that exploits the admissibility of heavy-tail distributions for mean square bounded oracles. This construction is new and may be of independent interest.

Theorem 4.3. *Consider a gain-shape quantizer Q satisfying the assumptions above. Suppose that for $\mathcal{X} = \{x : \|x\|_2 \le D/2\}$ we can find an optimization protocol π which, using at most T iterations, achieves $\sup_{f,O\in\mathcal{O}} \mathcal{E}(f, \pi^{QO}) \le \frac{3DB}{\sqrt{T}}$. Then, we can find a universal constant c such that the overall precision r of the quantizer must satisfy*

$$r \ge c(d + \log T).$$

Theorem 4.4. *Consider a gain-shape quantizer Q satisfying the assumptions above . Suppose that the number of bits r_g used by the gain quantizer are fixed independently of T. Then, for $\mathcal{X} = \{x : \|x\|_2 \le D/2\}$, there exists $(f, O) \in \mathcal{O}$ such that for any optimization protocol π using at most T iterations, we must have*

$$\mathcal{E}(f, \pi^{QO}) \ge \frac{c(r_g)DB}{T^{1/3}},$$

where $c(r_g)$ is a constant depending only on the number of bits used by the gain quantizer (but not on T).

The proofs of Theorems 4.3 and 4.4 are technical and long; we defer them to Section 5.3.

4.2 A-RATQ in the high precision regime

Instead of quantizing the gain uniformly, we propose to use an adaptive quantizer termed *Adaptive Geometric Uniform Quantizer* (AGUQ) for gain. AGUQ operates similar to the one-dimensional ATUQ, except the possible dynamic-ranges $M_{g,0}, \ldots, M_{g,h}$ grow geometrically (and not using tetra-iterations of ATUQ) as follows:

$$M_{g,j}^2 = B^2 \cdot a_g^j, \quad 0 \le j \le h_g - 1. \tag{16}$$

Specifically, for a given gain $G \ge 0$, AGUQ first identifies the smallest j such that $G \le M_{g,j}$ and then represents G using the one-dimensional version of CUQ with a dynamic range $[0, M_{g,j}]$ and k_g uniform levels

$$B_{M_{g,j},k}(\ell) := \ell \cdot \frac{M_{g,j}}{k_g - 1}, \quad \forall\, \ell \in \{0, \ldots, k - 1\}.$$

As in ATUQ, if $G > M_{h_g - 1}$, the overflow $\emptyset$ symbol is used and the decoder simply outputs 0. The overall procedure is the similar to Algorithms (6) and (7) for $s = 1, h = h_g$, and $M_j = M_{g,j}$, $0 \le j \le h_g - 1$; the only changes is that now we restrict to nonnegative interval $[0, M_{g,j}]$ for the one-dimensional version of CUQ with uniform levels k_g. The following result characterizes the performance of one-dimensional quantizer AGUQ; it is the only component missing in the analysis of A-RATQ.

Lemma 4.5. *Let $Q_{\mathtt{a}}$ be the quantizer AGUQ described above, with $h_g \geq 2$. Then,*

$$\alpha(Q_{\mathtt{a}}; B, 1) \leq B\sqrt{\frac{1}{4(k_g-1)^2} + \frac{a_g(h_g-1)}{4(k_g-1)^2} + 1},$$

$$\beta(Q_{\mathtt{a}}; B, 1) \leq \sup_{Y \geq 0 \ a.s. \ :\mathbb{E}[Y^2] \leq B^2} \mathbb{E}\left[\left|\mathbb{E}\left[Q_1(Y) - Y | Y\right]\right|\right] \leq \frac{B^2}{M_{g,h_g-1}}.$$

Proof of this result, too, is deferred to Section 5.2. Note that we have derived a bound for a quantity that is slightly larger than the bias of $Q_{\mathtt{a}}$, since we want to use this result along with Theorem 4.2.

Thus, our overall quantizer termed the *adaptive-RATQ* (A-RATQ) is given by

$$Q(Y) := Q_a(\|Y\|_2) \cdot Q_{\mathtt{at},R}(Y/\|Y\|_2),$$

where Q_a denotes the one dimensional AGUQ and $Q_{\mathtt{at},R}$ denotes the d-dimensional RATQ. Note that we use independent randomness for $Q_a(\|Y\|_2)$ and $Q_{\mathtt{at},R}(Y/\|Y\|_2)$, rendering them conditionally independent given Y.

The parameters s, k for RATQ and a_g, k_g for AGUQ are yet to be set. We first present a result which holds for all choices of these parameters.

Theorem 4.6 (Performance of A-RATQ)**.** *For Q set to A-RATQ with parameters set as in* (13), (16), *we have*

$$\alpha(Q; B, d) \leq B\sqrt{\frac{1}{4(k_g-1)^2} + \frac{a_g(h_g-1)}{4(k_g-1)^2} + 1} \cdot \sqrt{\frac{9 + 3\ln s}{(k-1)^2} + 1},$$

$$\beta(Q; B, d) \leq \frac{B^2}{M_{g,h_g-1}}.$$

Proof.

The worst-case second moment of A-RATQ: By Theorem 4.2 we have

$$\begin{aligned}\alpha(Q; B, d) &\leq \alpha(Q_{\mathtt{a}}; B, 1) \cdot \alpha_0(Q_{\mathtt{at},R}; 1, d) \\ &\leq \alpha(Q_{\mathtt{a}}; B, 1) \cdot \sqrt{\frac{9 + 3\ln s}{(k-1)^2} + 1} \\ &\leq B\sqrt{\frac{1}{4(k_g-1)^2} + \frac{a_g(h_g-1)}{4(k_g-1)^2} + 1} \cdot \sqrt{\frac{9 + 3\ln s}{(k-1)^2} + 1},\end{aligned}$$

where the second inequality used Theorem 3.3 with $B = 1$, and the third follows by Lemma 4.5.

The worst-case bias of A-RATQ: With parameters of RATQ set as in (13), we have that

$$\mathbb{E}\left[Q_{\mathtt{at},R}(y)\right] = y, \quad \forall y \quad s.t \quad \|y\|_2^2 \leq 1.$$

Therefore, by Theorem 4.2 it follows that

$$\beta(Q; B, d) \leq \sup_{Y:\mathbb{E}\left[\|Y\|_2^2\right] \leq B^2} \mathbb{E}\left[\left|\mathbb{E}\left[Q_{\mathtt{a}}(\|Y\|_2) - \|Y\|_2 | Y\right]\right|\right] \leq \frac{B^2}{M_{g,h_g-1}},$$

where the second inequality follows from Lemma 4.5. ∎

Note that RATQ yields an unbiased estimator; the bias in A-RATQ arises from AGUQ since the gain is not bounded. Further, AGUQ uses a precision of $\lceil \log h_g \rceil + \lceil \log(k_g+1) \rceil$ bits, and therefore, the overall precision of A-RATQ is $\lceil \log h_g \rceil + \lceil \log(k_g+1) \rceil + \lceil d/s \rceil \lceil \log h \rceil + d \lceil \log(k+1) \rceil$ bits.

In the high-precision regime, we set

$$a_g = 2, \quad \log h_g = \left\lceil \log(1 + \frac{1}{2}\log T) \right\rceil,$$
$$\log(k_g+1) = \left\lceil \log\left(2 + \frac{1}{2}\sqrt{\log T + 1}\right) \right\rceil. \tag{17}$$

Corollary 4.7. *Denote by Q the quantizer A-RATQ with parameters set as in* (13)*,* (9)*, and* (17)*. Then, the overall precision r used by Q is less than*

$$d(1+\Delta_1) + \Delta_2 + \left\lceil \log\left(2 + \sqrt{\log T + 1}\right) \right\rceil,$$

where $\Delta_1 = \left\lceil \log\left(2 + \sqrt{9 + 3\ln \Delta_2}\right) \right\rceil$ and $\Delta_2 = \lceil \log(1 + \ln^*(d/3)) \rceil$, the same as Corollary 3.4. Furthermore, the optimization protocol π given in algorithm 1 satisfies $\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{QO}) \leq 3DB/\sqrt{T}$.

Proof. The proof is similar to the proof of Corollary 3.4. The first statement follows by simply upper bounding the precision of the fixed-length code for A-RATQ with parameters as in the statement. The second statement follows by bounding $\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{QO})$ using Theorem 2.4, using the upper bounds for α and β given in Theorem 4.6, and finally substituting the parameters. □

Remark 5. The precision used in Corollary 4.7 matches the lower bound in Corollary 3.2 upto an additive factor of $\log\log T$ (ignoring the mild factor of $\log\log\ln^*(d/3)$), which is much lower than the $\log T$ lower bound we established for uniform gain quantizers.

4.3 A-RATQ in the low precision regime

In order to operate with a fixed precision r, we combine A-RATQ with RCS. We simply combine RCS with RATQ as in Section 3.5 to limit the precision and use AGUQ as the gain quantizer. Note that we use independent randomness in our gain quantizer $Q_a(\|Y\|_2)$ and our shape quantizer $\tilde{Q}(Y/\|Y\|_2)$, rendering them conditionally independent given Y. We have the following theorem characterizing α and β for this quantizer.

Theorem 4.8. *Let $Q(Y) = Q_a(\|Y\|) \cdot \tilde{Q}(Y/\|Y\|_2)$, where $\tilde{Q}$ is the composition of RCS and RATQ described in Theorem 3.7 with parameters m, m_0, and h of RATQ as in* (13) *and Q_a is AGUQ. Then,*

$$\alpha(Q; B, d) \leq B\sqrt{\frac{1}{4(k_g-1)^2} + \frac{a_g(h_g-1)}{4(k_g-1)^2} + 1} \cdot \frac{1}{\sqrt{\mu}}\sqrt{\frac{9 + 3\ln s}{(k-1)^2} + 1},$$
$$\beta(Q; B, d) \leq \frac{B^2}{M_{g,h_g-1}}.$$

Proof.

The worst-case second moment: Starting by applying Theorem 4.2, we have

$$
\begin{aligned}
\alpha(Q;B,d) &\leq \alpha(Q_{\mathtt{a}};B,1)\cdot \alpha_0(\tilde{Q};1,d)\\
&\leq \alpha(Q_{\mathtt{a}};B,1)\cdot \frac{1}{\sqrt{\mu}}\alpha_0(Q_{\mathtt{at},R};1,d)\\
&\leq B\sqrt{\frac{1}{4(k_g-1)^2}+\frac{a_g(h_g-1)}{4(k_g-1)^2}+1}\cdot \frac{1}{\sqrt{\mu}}\sqrt{\frac{9+3\ln s}{(k-1)^2}+1},
\end{aligned}
$$

where the second inequality follows by Theorem 3.7 and the third follows by Theorem 3.3 and Lemma 4.5.

The worst-case bias: With parameters of RATQ set as in (13), we have that

$$
\mathbb{E}\left[\tilde{Q}(y)\right] = y, \quad \forall y \quad s.t \quad \|y\|_2^2 \leq 1.
$$

Therefore, by Theorem 4.2 we get

$$
\begin{aligned}
\beta(Q;B,d) &\leq \sup_{Y:\mathbb{E}\left[\|Y\|_2^2\right]\leq B^2} \mathbb{E}\left[\left|\mathbb{E}\left[Q_{\mathtt{a}}(\|Y\|_2)-\|Y\|_2|Y\right]\right|\right]\\
&\leq \frac{B^2}{M_{g,h_g-1}},
\end{aligned}
$$

where the second inequality follows from Lemma 4.5.

□

We divide the total precision r into r_g and r_s bits: r_g to quantize the gain, r_s to quantize the subsampled shape vector. We set

$$
\begin{aligned}
&s, k, \text{ and } \mu d \text{ as in (12), with } r_s \text{ replacing } r,\\
&\log h_g = \log(k_g+1) = \frac{r_g}{2},\ a_g = (\mu T)^{\frac{1}{h_g+1}}
\end{aligned}
\tag{18}
$$

That is, our shape quantizer simply quantizes μd randomly chosen coordinates of the rotated vector using ATUQ with r_s bits, and the remaining bits are used by the gain quantizer AGUQ. The result below shows the performance of this quantizer.

Corollary 4.9. *For any r with gain quantizer being assigned $r_g \geq 4$ bits and shape quantizer being assigned $r_s \geq 3+\lceil\log(1+\ln^*(d/3))\rceil$, let Q be the combination of RCS and A-RATQ with parameters set as in* (13), (16), (18). *Then for $\mu T \geq 1$, the optimization protocol π in Algorithm 1 can obtain*

$$
\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f,\pi^{QO}) \leq O\left(DB\left(\frac{d}{T\min\{d,\frac{r_s}{\log\ln^*(d/3)}\}}\right)^{\frac{1}{2}\cdot\frac{2^{r_g/2}-1}{2^{r_g/2}+1}}\right).
$$

Proof. By using Theorem 2.4 to upper bound $\sup_{(f,O)\in\mathcal{O}}\mathcal{E}(f,\pi^{QO})$ and then Theorem 4.8 to upper-bound α and β, we get

$$
\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f,\pi^{QO}) \leq D\left(\frac{1}{\sqrt{\mu T}}\sqrt{\frac{B^2}{4(k_g-1)^2}+\frac{a_g(h_g-1)B^2}{4(k_g-1)^2}+B^2}\sqrt{\frac{9+3\ln s}{(k-1)^2}+1}+\frac{B^2}{M_{g,h-1}}\right).
$$

By substituting the parameters as in the statement and using the fact that $\mu T \geq 1$ completes the proof. $\square$

Remark 6. Our fixed precision quantizer in Corollary 4.9 establishes that using only a constant number of bits for gain-quantization, we get very close to the lower bound in Theorem 3.1. For instance, given access to a large enough precision r, if we set r_g to be 16 bits, we get

$$\sup_{(f,O)\in\mathcal{O}} \mathcal{E}(f, \pi^{QO}) \leq O\left(DB\left(\frac{d}{T\min\{d, \frac{r-16}{\log\ln^*(d/3)}\}}\right)^{\frac{1}{2}\cdot\frac{255}{257}}\right).$$

Here, the ratio of $d/(\min\{d, \frac{r-16}{\log\ln^*(d/3)}\})$ is very close to the optimal ratio of $d/(\min\{d, r\})$, and the exponent $255/(2\cdot 257)$ is close to the optimal exponent $1/2$.

Remark 7. We remark that A-RATQ satisfies Assumptions (1) and (2) in Section 4.1 but not (3), and breaches the lower bound for uniform gain quantizers established in Section 4.1.

5 Main proofs

5.1 Proof of Theorem 3.3

Step 1: Analysis of CUQ. We first prove a result for CUQ (with a dynamic range of $[-M, M]$) which will bound the expected value of

$$\sum_{i\in[d]} \left(Q_{\mathtt{u}}(Y)(i) - Y(i)\right)^2 \mathbb{1}_{\{Y(i)\leq M\}},$$

namely the mean square error when there is no overflow. This will be useful in the analysis of RATQ, too.

Lemma 5.1. *For an $\mathbb{R}^d$-valued random variable Y and $Q_{\mathtt{u}}$ denoting the quantizer CUQ with parameters M (with dynamic range $[-M, M]$) and k, let $Q_{\mathtt{u}}(Y)$ be the quantized value of Y. Then,*

$$\mathbb{E}\left[\sum_{i\in[d]} \left(Q_{\mathtt{u}}(Y)(i) - Y(i)\right)^2 \mathbb{1}_{\{|Y(i)|\leq M\}} \mid Y\right] \leq \frac{dM^2}{(k-1)^2}\left(\frac{1}{d}\sum_{j\in[d]} \mathbb{1}_{\{|Y(j)|\leq M\}}\right).$$

The proof is relatively straightforward with the calculations similar to [47, Theorem 2]; it is deferred to Appendix B.1.

Also, the quantizer AGUQ in Section 4.2 uses the one-dimensional CUQ with dynamic range $[0, M]$ as a subroutine. The uniform levels for this variant of CUQ are given by

$$B_{M,k}(\ell) = \ell \cdot \frac{M}{k-1}, \forall \ell \in [k-1].$$

We have the following lemma for this variant of CUQ.

Lemma 5.2. *For an $\mathbb{R}$-valued random variable Y which is almost surely nonnegative and the quantizer $Q_{\mathtt{u}}$ with dynamic range $[0, M]$ and parameter k, let $Q_{\mathtt{u}}(Y)$ denote the quantized value of Y. Then,*

$$\mathbb{E}\left[\left(Q_{\mathtt{u}}(Y) - Y\right)^2 \mathbb{1}_{\{|Y|\leq M\}} \mid Y\right] \leq \frac{M^2}{4(k-1)^2}\left(\mathbb{1}_{\{|Y|\leq M\}}\right).$$

The proof is very similar to the proof of Lemma 5.1 and is deferred to Appendix B.1.

Step 2: Mean square error for adaptive quantizers. The quantizers RATQ and A-RATQ use ATUQ as subroutine; in addition, A-RATQ uses AGUQ for gain quantization. Thus, in order to analyze RATQ and A-RATQ, we need to analyze ATUQ and AGUQ first.

In this step we provide a general bound on the mean square error of adaptive quantizers. We capture the performances of ATUQ and AGUQ in two separate results below.

Lemma 5.3. *For an $\mathbb{R}^d$-valued random variable Y and Q denoting the quantizer ATUQ with dynamic-range parameters M_js, we have*

$$\mathbb{E}\left[\sum_{i\in[d]}\big(Q(Y)(i)-Y(i)\big)^2\mathbb{1}_{\{|Y(i)|\le M_{h-1}\}}\right]\le\frac{d}{(k-1)^2}\left(m+m_0+\sum_{j=1}^{h-1}M_j^2P\left(\|Y\|_\infty>M_{j-1}\right)\right).$$

Proof. Consider the events A_js corresponding to different levels used by the adaptive quantizer of the norm, defined as follows:

$$\begin{aligned}A_0&:=\{\|Y\|_\infty\le m\},\\ A_j&:=\{M_{j-1}<\|Y\|_\infty\le M_j\},\quad\forall j\in[h-2],\\ A_{h-1}&:=\{M_{h-2}<\|Y\|_\infty\}.\end{aligned}$$

By construction, $\sum_{j=0}^{h-1}\mathbb{1}_{A_j}=1$ a.s.. Therefore, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in[d]}\big(Q(Y)(i)-Y(i)\big)^2\mathbb{1}_{\{|Y(i)|\le M_{h-1}\}}\right]&=\mathbb{E}\left[\|Q(Y)-Y\|_2^2\mathbb{1}_{A_0}\right]+\sum_{j=1}^{h-2}\mathbb{E}\left[\|Q(Y)-Y\|_2^2\mathbb{1}_{A_j}\right]\\&\quad+\mathbb{E}\left[\sum_{i\in[d]}\big(Q_{\mathtt{u}}(Y)(i)-Y(i)\big)^2\mathbb{1}_{\{|Y(i)|\le M_{h-1}\}}\mathbb{1}_{A_{h-1}}\right].\end{aligned}$$

Note that $\mathbb{1}_{A_0}$ implies that we are using a k-level uniform quantization with a dynamic range of $[-m,m]$. Therefore, this term can be bounded by Lemma 5.1 as follows:

$$\mathbb{E}\left[\|Q(Y)-Y\|_2^2\mathbb{1}_{A_0}\right]\le\frac{dm}{(k-1)^2}.$$

Under the event A_j with $j\in[h-1]$, we use a k-level uniform quantization with a dynamic range of $[-M_j,M_j]$. Therefore, by Lemma 5.1, we have

$$\begin{aligned}\mathbb{E}\left[\|Q(Y)-Y\|_2^2\mathbb{1}_{A_j}\right]&\le\frac{dM_j^2}{(k-1)^2}\mathbb{E}\left[\mathbb{1}_{A_j}\right]\\&\le\frac{dM_j^2}{(k-1)^2}P\left(\|Y\|_\infty>M_{j-1}\right).\end{aligned}$$

□

Note that the proof above does note use specific form of M_j's and therefore applies as it is for the one-dimensional AGUQ gain quantizer used in A-RATQ; the only change is the fact that instead of using Lemma 5.1 for uniform quantization we use Lemma 5.2. This leads to the following lemma, which will be useful later in the analysis of A-RATQ.

Lemma 5.4. *For an $\mathbb{R}$-valued random variable Y which is almost surely nonnegative and Q denoting the quantizer AGUQ with dynamic-range parameters $M_{g,j}s$, we have*

$$\mathbb{E}\left[(Q(Y)-Y)^2 \mathbb{1}_{\{|Y| \le M_{g,h-1}\}}\right] \le \frac{1}{4(k-1)^2}\left(B^2 + \sum_{j=1}^{h-1} M_{g,j}^2 P\left(|Y| > M_{g,j-1}\right)\right).$$

The proof is similar to that of Lemma 5.3 and is omitted.

Step 3: Mean square error of ATUQ for a subgaussian input vector. In our analysis, we need to evaluate the performance of ATUQ for *subgaussian* input vectors.

Definition 5.5 (*cf.* [11])**.** A centered random variable X is said to be subgaussian with variance factor v if for all λ in $\mathbb{R}$, we have

$$\ln \mathbb{E}\left[e^{\lambda X}\right] \le \frac{\lambda^2 v}{2}.$$

The following well-known fact (*cf.* [11, Chapter 2]) will be used throughout.

Lemma 5.6. *For a centered subgaussian random variable X with variance factor v the*

$$P(|X| > x) \le 2e^{-x^2/2v}, \quad \forall\, x > 0,$$
$$\mathbb{E}\left[X^2\right] \le 4v, \quad \mathbb{E}\left[X^4\right] \le 32v^2.$$

Next, consider the quantizer $Q_{\mathtt{at},I}$ which is similar to RATQ but skips the rotation step. Specifically, $Q_{\mathtt{at},I}$ is obtained by replacing the random matrix R in the encoder and decoder of RATQ (given in Algorithms 2 and 3, respectively) by the identity matrix I. Symbolically, the quantizer $Q_{\mathtt{at},I}$ can be described as follows for the d-dimensional input vector Y

$$Q_{\mathtt{at},I}(Y) = [Q_{\mathtt{at}}(Y_1)^T, \cdots, Q_{\mathtt{at}}(Y_{\lceil d/s \rceil})^T], \tag{19}$$

where $Q_{\mathtt{at}}$ is the quantizer ATUQ and Y_i is the i^{th} subvector of Y. Recall that the i^{th} subvector Y_i comprises the coordinates $\{(i-1)s+1, \cdots, \min\{is, d\}\}$, for all $i \in [d/s]$. Also, recall that the dimension of all the sub vectors except the last one is s, with the last one having dimension $d - s\lfloor d/s \rfloor$.

Notice that like RATQ, $Q_{at,I}$ has parameters k, h, s, m, and m_0 which need to be set. We set the parameters m and m_0 to be $3v$ and $2v \ln s$, respectively, and prove a general lemma in terms of the other parameters of $Q_{at,I}$ for a subgaussian input vector.

Lemma 5.7. *Consider $Y = [Y(1), \ldots, Y(d))]^T$, where for all i in $[d]$, $Y(i)$ is a centered subgaussian random variable with variance factor v. Let Q denote the quantizer $Q_{\mathtt{at},I}$ with parameters m and m_0 set to $3v$ and $2v \ln s$, respectively. Then, for every $s, k, h \in \mathbb{N}$, we have*

$$\frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} (Y(i) - Q(Y)(i))^2 \mathbb{1}_{\{|Y(i)| \le M_{h-1}\}}\right] \le v \cdot \frac{9 + 3\ln s}{(k-1)^2}.$$

Proof. Since

$$\mathbb{E}\left[\sum_{i\in[d]}(Y(i)-Q(Y)(i))^2\mathbb{1}_{\{|Y(i)|\le M_{h-1}\}}\right]=\sum_{i=1}^{\lceil \frac{d}{s}\rceil}\sum_{j=(i-1)s+1}^{\min\{is,d\}}\mathbb{E}\left[(Q_{\mathtt{at}}(Y)(j)-Y(j))^2\,\mathbb{1}_{\{|Y(j)|\le M_{h-1}\}}\right],$$

by using Lemma 5.3 for each of the $\lceil d/s\rceil$ subvectors, we get

$$\begin{aligned}&\mathbb{E}\left[\sum_{i\in[d]}(Y(i)-Q(Y)(i))^2\mathbb{1}_{\{|Y(i)|\le M_{h-1}\}}\right]\\&\le\frac{s}{(k-1)^2}\sum_{i\in1}^{\lfloor\frac{d}{s}\rfloor}\left(m+m_0+\sum_{j\in[h-1]}M_j^2P\left(\|Y_{i,s}\|_\infty>M_{j-1}\right)\right)\\&\quad+\frac{(d-s\lfloor\frac{d}{s}\rfloor)}{(k-1)^2}\left(m+m_0+\sum_{j\in[h-1]}M_j^2P\left(\|Y_{\lceil d/s\rceil,s}\|_\infty>M_{j-1}\right)\right).\end{aligned}$$

For all $i\in\lfloor d/s\rfloor$, it follows from the union bound that

$$P\left(\|Y_{1,s}\|_\infty>M_{j-1}\right)\le 2se^{\frac{-M_{j-1}^2}{2v}}.$$

Also, since $d-s\lfloor d/s\rfloor\le s$, we have

$$P\left(\|Y_{\lceil d/s\rceil,s}\|_\infty>M_{j-1}\right)\le 2se^{\frac{-M_{j-1}^2}{2v}}.$$

Using these tail bounds in the previous inequality, we get

$$\mathbb{E}\left[\sum_{i\in[d]}(Y(i)-Q(Y)(i))^2\mathbb{1}_{\{|Y(i)|\le M_{h-1}\}}\right]\le\frac{d}{(k-1)^2}\left(m+m_0+2s\sum_{j\in[h-1]}M_j^2e^{\frac{-M_{j-1}^2}{2v}}\right).$$

Setting $m=3v$ and $m_0=2v$, the summation on the right-side is bounded further as

$$\begin{aligned}&2s\left(\frac{3v}{s}\sum_{j=1}^{h-1}(e^{*j})\cdot e^{-1.5e^{*(j-1)}}\right)+2s\left(\frac{2v}{s}\sum_{j=1}^{h-1}e^{-1.5e^{*(j-1)}}\right)\\&=6v\sum_{j=1}^{h-1}e^{-0.5e^{*(j-1)}}+4v\ln s\sum_{j=1}^{h-1}e^{-1.5e^{*(j-1)}}\\&\le 6v\sum_{j=1}^{\infty}e^{-0.5e^{*(j-1)}}+4v\ln s\sum_{j=1}^{h-1}e^{-1.5e^{*(j-1)}}\\&\le 6v+v\ln s,\end{aligned}$$

where we use a bound of 1 for $\sum_{j=1}^{\infty} e^{-0.5e^{*(j-1)}}$, whose validity can be seen as follows[6]

$$\begin{aligned}\sum_{j=1}^{\infty} e^{-0.5e^{*(j-1)}} &= e^{-0.5} + e^{-0.5e} + e^{-0.5e^e} + \sum_{j=3}^{\infty} e^{-0.5e^{*(j)}} \\ &\leq e^{-0.5} + e^{-0.5e} + e^{-0.5e^e} + \sum_{j=3}^{\infty} e^{-0.5je^e} \\ &\leq e^{-0.5} + e^{-0.5e} + e^{-0.5e^e} + \frac{1}{e^{e^e} - 1} \\ &\leq 1,\end{aligned}$$

and 1/4 for $\sum_{j=1}^{h-1} e^{-1.5e^{*(j-1)}}$, whose validity can be seen as follows

$$\begin{aligned}\sum_{j=1}^{\infty} e^{-1.5e^{*(j-1)}} &= e^{-1.5} + e^{-1.5e} + e^{-1.5e^e} + \sum_{j=3}^{\infty} e^{-1.5e^{*(j)}} \\ &\leq e^{-1.5} + e^{-1.5e} + e^{-1.5e^e} + \sum_{j=3}^{\infty} e^{-1.5je^e} \\ &\leq e^{-1.5} + e^{-1.5e} + e^{-1.5e^e} + \frac{1}{e^{3e^e} - 1/e^{1.5e^e}} \\ &\leq 0.2401.\end{aligned}$$

Therefore, we obtain

$$\frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} (Y(i) - Q(Y)(i))^2 \mathbb{1}_{|Y(i)| \leq M_{h-1}}\right] \leq v \cdot \frac{9 + 3\ln s}{(k-1)^2}.$$

□

We remark that calculations present in this lemma are at the heart of the analysis of RATQ. Also, this lemma will be useful for other applications discussed in Section 6.

Step 4: Completing the proof. Recall that the random matrix R defined in (6) is used at the encoder of RATQ to randomly rotate the input vector. We observe that the rotated vector has subgaussian entries.

Lemma 5.8. *For an $\mathbb{R}^d$-valued random variable Y such that $\|Y\|_2^2 \leq B^2$ a.s., all coordinates of the rotated vector RY are centered subgaussian random variables with a variance factor of B^2/d, whereby*

$$P(|RY(j)| \geq M) \leq 2e^{-dM^2/2B^2}, \quad \forall\, j \in [d],$$

where $RY(j)$ is the j^{th} coordinate of the rotated vector.

The proof uses similar calculations as [7] and [47]; it is deferred to the Appendix B.2.

Intuitively, the Lemma 5.8 highlights the fact that overall energy $\|Y\|_2^2$ in the input vector Y is divided equally among all the coordinates after random rotation.

[6] In fact, these bounds motivate the use of tetration as our choice for M_js.

The worst-case second moment of RATQ. Note that by the description of RATQ which will be denoted by $Q_{\texttt{at},R}(RY)$, we have that

$$Q_{\texttt{at},R}(Y) = R^{-1}Q_{\texttt{at},I}(RY),$$

where $Q_{\texttt{at},I}$ is as defined in (19). Thus,

$$Q_{\texttt{at},I}(RY) = [Q_{\texttt{at}}(RY_{1,s})^T, \cdots, Q_{\texttt{at}}(RY_{\lceil d/s\rceil,s})^T]^T, \tag{20}$$

where the subvector $RY_{i,s}$ is given by

$$RY_{i,s} = [RY((i-1)s+1), \cdots, RY(\min\{is, d\})]^T.$$

To compute $\alpha(Q_{\texttt{at},R}(Y))$, we will first compute the second moment for the output of RATQ. Specifically, using the fact R is a unitary transform, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|Q_{\texttt{at},R}(Y)\|_2^2\right] &= \mathbb{E}\left[\|R^{-1}Q_{\texttt{at},I}(RY)\|_2^2\right] \\
&= \mathbb{E}\left[\|Q_{\texttt{at},I}(RY)\|_2^2\right] \\
&= \sum_{j\in[d]} \mathbb{E}\left[(Q_{\texttt{at},I}(RY)(j))^2\right] \\
&= \sum_{i=1}^{\lceil \frac{d}{s}\rceil} \sum_{j=(i-1)s+1}^{\min\{is,d\}} \mathbb{E}\left[(Q_{\texttt{at},I}(RY)(j))^2\right].
\end{aligned}$$

We now observe that for our choice of m and h for RATQ given by (7), we have

$$M_{h-1}^2 \geq m(e^{*\log_e^*(d/3)}) = (3B^2/d).(d/3) = B^2.$$

Using this observation and noting that R is a unitary matrix, we have that

$$\mathbb{1}_{\{\|RY\|_2 \leq M_{h-1}\}} = 1 \text{ a.s.}.$$

Also, noting that $|RY(j)| \leq \|RY\|_2 = \|Y\|_2 = B$ a.s., for all $j \in [d]$, we get

$$\mathbb{1}_{\{|RY(j)| \leq M_{h-1}\}} = 1 \text{ a.s.}, \forall j \in [d]. \tag{21}$$

Proceeding with these observations, we get

$$\begin{aligned}
\mathbb{E}\left[\|Q_{\texttt{at},R}(Y)\|_2^2\right] &\leq \sum_{i=1}^{\lceil \frac{d}{s}\rceil} \sum_{j=(i-1)s+1}^{\min\{is,d\}} \mathbb{E}\left[(Q_{\texttt{at},I}(RY)(j))^2 \mathbb{1}_{\{|RY(j)|\leq M_{h-1}\}}\right] \\
&= \sum_{i=1}^{\lceil \frac{d}{s}\rceil} \sum_{j=(i-1)s+1}^{\min\{is,d\}} \mathbb{E}\left[(Q_{\texttt{at},I}(RY)(j) - RY(j) + RY(j))^2 \mathbb{1}_{\{|RY(j)|\leq M_{h-1}\}}\right] \\
&\leq \sum_{i=1}^{\lceil \frac{d}{s}\rceil} \sum_{j=(i-1)s+1}^{\min\{is,d\}} \mathbb{E}\left[\left((Q_{\texttt{at},I}(RY)(j) - RY(j))^2 + RY(j)^2\right) \mathbb{1}_{\{|RY(j)|\leq M_{h-1}\}}\right],
\end{aligned}$$

where the previous inequality uses the fact that, under the event $\{|RY(j)| \leq M_{h-1}\}$, $Q_{\texttt{at},I}(RY)(j)$ is an unbiased estimate of $RY(j)$. Namely,

$$\mathbb{E}\left[Q_{\texttt{at},I}(RY)(j)\mathbb{1}_{\{|RY(j)|\leq M_{h-1}\}} \mid R, Y\right] = \mathbb{E}\left[RY(j)\mathbb{1}_{\{|RY(j)|\leq M_{h-1}\}} \mid R, Y\right].$$

Therefore, noting that R is a unitary matrix, we have

$$\mathbb{E}\left[\|Q_{\texttt{at},R}(Y)\|_2^2\right] \leq \mathbb{E}\left[\sum_{j\in[d]} (RY(i) - Q_{\texttt{at},I}(RY)(j))^2 \mathbb{1}_{\{|RY(j)|\leq M_{h-1}\}}\right] + \mathbb{E}\left[\|Y\|_2^2\right].$$

To bound the first term on the right-side we have the following lemma, which will also be useful later in Section 6.1.

Lemma 5.9. *For an $\mathbb{R}^d$-valued random variable Y such that $\|Y\|_2^2 \leq B^2$ a.s.. Then, for m and m_0 set to be $3B^2/d$ and $(2B^2/d)\ln s$, respectively, we have that*

$$\mathbb{E}\left[\sum_{j\in[d]} (RY(i) - Q_{\texttt{at},I}(RY)(j))^2 \mathbb{1}_{\{|RY(j)|\leq M_{h-1}\}}\right] \leq B^2 \cdot \frac{9 + 3\ln s}{(k-1)^2}.$$

Proof. By Lemma 5.8 we have that all coordinates $RY(j)$ are centered subgaussian random variable with variance factor B^2/d. Thus, the parameters m and m_0 of RATQ set as in (7), and equal $3v$ and $2v\ln s$, respectively, where v is the variance factor of each subgaussian coordinate. The result follows by invoking Lemma 5.7. □

Therefore, for any Y such that $\|Y\|_2^2 \leq B^2$, we have

$$\mathbb{E}\left[\|Q_{\texttt{at},R}(Y)\|_2^2\right] \leq B^2 \cdot \frac{9 + 3\ln s}{(k-1)^2} + B^2,$$

whereby

$$\alpha_0(Q_{\texttt{at},R}) \leq B\sqrt{\frac{9 + 3\ln s}{(k-1)^2} + 1}.$$

The worst-case bias of RATQ. By (21) we have that the input always remains in the dynamic-range of the quantizer, resulting in unbiased quantized values. In other words, $\beta_0(Q_{\texttt{at,R}}) = 0$.

5.2 Proof of Lemma 4.5

We first note AGUQ is used to quantize a scalar Y. It follows from the description of the quantizer that

$$\mathbb{1}_{\{|Y|\leq M_{g,h_g-1}\}}\mathbb{E}\left[Q_{\texttt{a}}(Y)|Y\right] = \mathbb{1}_{\{|Y|\leq M_{g,h_g-1}\}}Y, \tag{22}$$

and that[7]

$$\mathbb{1}_{\{|Y|> M_{g,h_g-1}\}}Q_{\texttt{a}}(Y) = 0. \tag{23}$$

[7]Once again, this follows from our convention that the outflow symbol is evaluated to 0.

The worst-case second moment of AGUQ. Towards evaluating $\alpha(Q_{\mathtt{a}})$ for AGUQ, for any $Y \in \mathbb{R}$ we have

$$\begin{aligned}\mathbb{E}\left[Q_{\mathtt{a}}(Y)^2\right] &= \mathbb{E}\left[Q_{\mathtt{a}}(Y)^2 \mathbb{1}_{\{|Y| \le M_{g,h_g-1}\}}\right] + \mathbb{E}\left[Q_{\mathtt{a}}(Y)^2 \mathbb{1}_{\{|Y| > M_{g,h-1}\}}\right] \\ &= \mathbb{E}\left[(Q_{\mathtt{a}}(Y) - Y + Y)^2 \mathbb{1}_{\{|Y| \le M_{g,h_g-1}\}}\right] + \mathbb{E}\left[Q_{\mathtt{a}}(Y)^2 \mathbb{1}_{\{|Y| > M_{g,h_g-1}\}}\right] \\ &= \mathbb{E}\left[(Q_{\mathtt{a}}(Y) - Y)^2 \mathbb{1}_{\{|Y| \le M_{g,h_g-1}\}}\right] + \mathbb{E}\left[Y^2 \mathbb{1}_{\{|Y| \le M_{g,h_g-1}\}}\right],\end{aligned}$$

where the last identity uses (23), and the fact that $\mathbb{E}\left[(Q_{\mathtt{a}}(Y) - Y) Y \mathbb{1}_{\{|Y| \le M_{g,h-1}\}} | Y\right] = 0$, which follows from (22). From Lemma 5.4 it follows that

$$\mathbb{E}\left[(Q_{\mathtt{a}}(Y) - Y)^2 \mathbb{1}_{\{|Y| \le M_{g,h-1}\}}\right] \le \frac{1}{4(k_g - 1)^2} \left(B^2 + \sum_{j=1}^{h-1} M_j^2 P\left(|Y| > M_{g,j-1}\right)\right).$$

By Markov's inequality we get that for any random variable Y with $\mathbb{E}\left[Y^2\right] \le B^2$, we have $P(|Y| > M_{g,j-1}) \le B^2 / M_{g,j-1}^2$, which further leads to

$$\begin{aligned}\mathbb{E}\left[(Q_{\mathtt{a}}(Y) - Y)_2^2 \mathbb{1}_{\{|Y| \le M_{g,h-1}\}}\right] &\le \frac{B^2}{4(k_g - 1)^2} + \sum_{j=1}^{h_g - 1} \frac{M_{g,j}^2}{4(k_g - 1)^2} \frac{B^2}{M_{g,j-1}^2} \\ &= \frac{B^2}{4(k_g - 1)^2} + \frac{a_g (h_g - 1) B^2}{4(k_g - 1)^2}.\end{aligned}$$

Therefore, we have

$$\mathbb{E}\left[Q_{\mathtt{a}}(Y)^2\right] \le \frac{B^2}{4(k_g - 1)^2} + \frac{a_g (h_g - 1) B^2}{4(k_g - 1)^2} + \mathbb{E}\left[Y^2 \mathbb{1}_{\{|Y| \le M_{g,h-1}\}}\right].$$

The result follows upon taking the supremum of the left-side over all random variables Y with $\mathbb{E}\left[Y^2\right] \le B^2$.

The worst-case bias of AGUQ. Towards evaluating $\beta(Q_{\mathtt{a}})$, we note first using Jensen's inequality that

$$\left|\mathbb{E}\left[Q_{\mathtt{a}}(Y) - Y\right]\right| \le \mathbb{E}\left[\left|\mathbb{E}\left[Q_{\mathtt{a}}(Y) - Y | Y\right]\right|\right].$$

Then, for Y with $\mathbb{E}\left[Y^2\right] \le B^2$, using (22) and Markov's inequality, we get

$$\begin{aligned}\mathbb{E}\left[\left|\mathbb{E}\left[Q_{\mathtt{a}}(Y) - Y | Y\right]\right|\right] &= \mathbb{E}\left[|Y| \mathbb{1}_{\{|Y| \ge M_{g,h-1}\}}\right] \\ &\le \sqrt{\mathbb{E}\left[Y^2\right] P(|Y| \ge M_{g,h-1})} \\ &\le \frac{B^2}{M_{g,h-1}}.\end{aligned} \tag{24}$$

Therefore, for any Y with $\mathbb{E}\left[Y^2\right] \le B^2$, we have

$$\left|\mathbb{E}\left[Q_{\mathtt{a}}(Y) - Y\right]\right| \le \sup_{Y \ge 0 \text{a.s.}: \mathbb{E}[Y^2] \le B^2} \mathbb{E}\left[\left|\mathbb{E}\left[Q_{\mathtt{a}}(Y) - Y | Y\right]\right|\right] \le \frac{B^2}{M_{g,h-1}}.$$

The result follows upon taking the supremum of left-side over all random variables Y with $\mathbb{E}\left[Y^2\right] \le B^2$. □

5.3 Proof of Theorems 4.3 and 4.4

Before we proceed with our lower bounds, we will set up some notation. We consider quantizers of the form

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

Let $W(\cdot|y)$, $W_g(\cdot|y)$, and $W_s(\cdot|y)$, respectively, denote the distribution of the output of quantizers $Q(y)$, $Q_g(y)$, and $Q_s(y)$. We prove a general lower bound for a quantizer satisfying Assumptions 1-3 in Section 4.1 in terms of the precision r; Theorems 4.3 and 4.4 are obtained as corollaries of this general lower bound.

Theorem 5.10. *Suppose that $\mathcal{X}$ contains the set $\{x \in \mathbb{R}^d : \|x\|_2 \le D/2\}$. Consider a gain-shape quantizer Q of precision r satisfying the Assumptions 1-3 in Section 4.1. Then, there exists an oracle $(f, O) \in \mathcal{O}$ such that for any optimization protocol π using T iterations, we have*

$$\mathcal{E}(f, \pi^{QO}) \ge \frac{DB}{2\sqrt{2}} \min\left\{\frac{1}{2^r}, \frac{1}{4\cdot 2^{r/3}T^{1/3}}, \frac{1}{2(2T)^{1/3}}\right\}.$$

Proof. Consider the function $f_\alpha : \mathbb{R}^d \to \mathbb{R}$, $\alpha \in \{-1, 1\}$ given by

$$f_\alpha(x) := \delta\frac{B}{\sqrt{2}}|x(1) - \alpha D/2|, \quad \alpha \in \{-1, 1\}.$$

Note that the functions f_1 and f_{-1} are convex and depend only on the first coordinate of x. Further, for $x \in \mathcal{X}$, the gradient of f_α is $\delta\alpha Be_1/\sqrt{2}$, where e_1 is the vector $[1, 0, 0, \ldots, 0]^T$. We consider oracles O_α, $\alpha \in \{-1, 1\}$, that produce noisy gradient updates with distribution

$$P_\alpha\left(\frac{B}{\sqrt{2}}e_1\right) = \frac{1-\delta^2}{2}, \quad P_\alpha\left(\frac{-B}{\sqrt{2}}e_1\right) = \frac{1-\delta^2}{2}, \quad P_\alpha\left(\frac{\alpha B}{\sqrt{2}\delta}e_1\right) = \delta^2.$$

It is easy to check that the oracle outputs satisfy (1) and (2) described in Section 2.1. That is, the output of O_α is an unbiased estimate of the subgradient of f_α, and the expected Euclidean norm square of the oracle output is bounded by B^2.

We now take recourse to the standard reduction of optimization to hypothesis testing: To estimate the optimal value of f_1 and f_{-1} to an accuracy δ, the optimization protocol must determine if the oracle outputs are generated by P_1 or P_{-1}. However in order to distinguish between P_1 or P_{-1}, the optimization protocol only has access to the quantized oracle outputs. Specifically, the protocol sees the samples from $Q(Y)$ at every time step, where Y has distribution either P_1 or P_{-1}.

Denoting by $P_\alpha W$ the distribution of the output $Q(Y)$ when the input Y is generated from P_α, we have from the standard reduction (see, for instance, [17, Theorem 5.2.4]) that

$$\max_{\alpha\in\{-1,1\}} \mathcal{E}(f, \pi^{QO}) \ge \frac{DB}{2\sqrt{2}}\delta\left(1 - \sqrt{\frac{T}{2}\chi^2(P_1W, P_{-1}W}\right),$$

where $\chi^2(P, Q) = \sum_x (P(x) - Q(x))^2/Q(x)$ denotes the chi-squared divergence between P and Q.

Note that Assumption 2 on the structure of the quantizer implies that when $M < B/\delta\sqrt{2}$, the distributions P_1W and $P_{-1}W$ are the same. It follows that for every $\delta < \min\{\sqrt{B^2/2M^2}, 1\}$, the left-side of the previous inequality exceeds $(DB/2\sqrt{2})\delta$, whereby

$$\max_{\alpha\in\{-1,1\}} \mathcal{E}(f, \pi^{QO}) \ge \frac{DB}{2\sqrt{2}} \min\left\{\frac{B}{\sqrt{2}M}, 1\right\}. \tag{25}$$

Next, we consider the following modification of the previous construction in the case when $B/\sqrt{2} < m$:

$$P_\alpha\left(\frac{B}{\sqrt{2}}e_1\right) = \frac{1-\delta^{1+y}}{2}, \quad P_\alpha\left(\frac{-B}{\sqrt{2}}e_1\right) = \frac{1-\delta^{1+y}}{2}, \quad P_\alpha\left(\frac{\alpha B}{\sqrt{2}\delta^y}e_1\right) = \delta^{1+y}.$$

for $y \in [0,1]$. Once again, the oracle outputs satisfy (1) and (2) described in Section 2.1. In this case, the vector $Y \sim P_\alpha$ has entries with ℓ_2 norm at the most $B/(\sqrt{2}\delta^y)$. We set y such that this value is less than m and $\chi^2(P_1W, P_{-1}W)$ is minimized. Note that if $B/(\delta^y\sqrt{2d}) < m$, then $\mathtt{supp}(Q_g(\|a\|)) \subseteq \{0, m\}$ for all the a's in the support of P_1 or P_{-1}.

For all $z \neq 0$, $z \in \mathtt{supp}(Q(a))$, when a is in the support of P_1 or P_{-1}, we have

$$W\left(z \middle| a\right) = W_g\left(m|\|a\|_2\right)W_s\left(\frac{z}{m}|a/\|a\|_2\right).$$

Therefore,

$$P_1W(z) - P_{-1}W(z) = \delta^{1+y}W_g\left(m|\frac{B}{\sqrt{2}\delta^y}\right)\left(W_s\left(\frac{z}{m}|e_1\right) - W_s\left(\frac{z}{m}|-e_1\right)\right)$$

$$P_{-1}W(z) \geq \frac{1-\delta^{1+y}}{2}W_g\left(m|\frac{B}{\sqrt{2}}\right)W_s\left(\frac{z}{m}|e_1\right) + \frac{1-\delta^{1+y}}{2}W_g\left(m|\frac{B}{\sqrt{2}}\right)W_s\left(\frac{z}{m}|-e_1\right).$$

Using the preceding two inequalities

$$\begin{aligned}
\frac{(P_1W(z) - P_{-1}W(z))^2}{P_{-1}W(z)} &\leq \frac{\delta^{2+2y}W_g\left(m|\frac{B}{\sqrt{2}\delta^y}\right)^2\left(W_s\left(\frac{z}{m}|e_1\right) - W_s\left(\frac{z}{m}|-e_1\right)\right)^2}{\frac{1-\delta^{1+y}}{2}W_g\left(m|\frac{B}{\sqrt{2}}\right)W_s\left(\frac{z}{m}|e_1\right) + \frac{1-\delta^{1+y}}{2}W_g\left(m|\frac{B}{\sqrt{2}}\right)W_s\left(\frac{z}{m}|-e_1\right)} \\
&\leq \frac{2\delta^{2+2y}}{1-\delta^{1+y}} \cdot \frac{W_g\left(m|\frac{B}{\sqrt{2}\delta^y}\right)^2\left(W_s\left(\frac{z}{m}|e_1\right) + W_s\left(\frac{z}{m}|-e_1\right)\right)}{W_g\left(m|\frac{B}{\sqrt{2}}\right)} \\
&\leq \frac{2\delta^{2+y}}{1-\delta^{1+y}} \cdot W_g\left(m|\frac{B}{\sqrt{2}\delta^y}\right)\left(W_s\left(\frac{z}{m}|e_1\right) + W_s\left(\frac{z}{m}|-e_1\right)\right) \\
&\leq \frac{2\delta^{2+y}}{1-\delta^{1+y}} \cdot \left(W_s\left(\frac{z}{m}|e_1\right) + W_s\left(\frac{z}{m}|-e_1\right)\right),
\end{aligned}$$

where the third inequality uses Assumption 3b for the quantizer in Section 4.1, i.e., it uses

$$\frac{W_g\left(m|\frac{B}{\sqrt{2}\delta^y}\right)}{W_g\left(m|\frac{B}{\sqrt{2}}\right)} \leq \delta^{-y}.$$

For all $z = 0$, $z \in \mathtt{supp}(Q(a))$, when a is in the support of P_1 or P_{-1}, we have

$$W\left(0 \middle| a\right) = W_g\left(0|\|a\|_2\right) + W_g\left(m|\|a\|_2\right)W_s\left(0|a/\|a\|_2\right).$$

Therefore, by similar calculations for $z \neq 0$, we have

$$\frac{(P_1 W(0) - P_{-1}W(0))^2}{P_{-1}W(0)} \leq \frac{\delta^{2+2y} W_g\left(m|\frac{B}{\sqrt{2}\delta^y}\right)^2 \left(W_s\left(0|e_1\right) + W_s\left(0|-e_1\right)\right)^2}{(\frac{1-\delta^{1+y}}{2})W_g\left(m|\frac{B}{\sqrt{2}}\right)W_s\left(0|e_1\right) + (\frac{1-\delta^{1+y}}{2})W_g\left(m|\frac{B}{\sqrt{2}}\right)W_s\left(0|-e_1\right)}$$
$$\leq \frac{2\delta^{2+y}}{1-\delta^{1+y}}\left(W_s\left(0|e_1\right) + W_s\left(0|-e_1\right)\right).$$

In conclusion,

$$\chi^2(P_1W, P_{-1}W) \leq \frac{4\delta^{2+y}}{1-\delta^{1+y}}.$$

Now, if $\delta < 1/2$, we have

$$\chi^2(P_1W, P_{-1}W) \leq 8\delta^{2+y}.$$

Upon setting $\delta = (16T)^{-1/(2+y)}$, which satisfies $\delta < 1/2$ for all T, we get

$$\max_{\alpha \in \{-1,1\}} \mathcal{E}(f, \pi^{QO}) \geq \frac{DB}{2\sqrt{2}}\delta(1 - \sqrt{4T\delta^{2+y}}) = \frac{DB}{4\sqrt{2}}\left(\frac{1}{16T}\right)^{\frac{1}{2+y}}. \quad (26)$$

But we can only set δ to this value if

$$\frac{B}{\sqrt{2}} \cdot (16T)^{\frac{y}{2+y}} < m. \quad (27)$$

Thus, for each y such that (27) holds, we get (26). Taking the the supremum of RHS in (26) over all $y \in [0,1]$ such that (27) holds, we obtain whenever $B/\sqrt{2} \leq m$,

$$\max_{\alpha \in \{-1,1\}} \mathcal{E}(f, \pi^{QO}) \geq \frac{DB}{2\sqrt{2}} \cdot \min\left\{\frac{1}{8}\sqrt{\frac{m\sqrt{2}}{BT}}, \frac{1}{2(2T)^{1/3}}\right\},$$

where we use the following lemma proved in Appendix B.3.

Lemma 5.11. *For $a, c > 0$, and $b > 1$.*

$$\sup_{y \in [0,1]: a(b)^{y/(2+y)} < c.} a\left(\frac{1}{b}\right)^{\frac{1}{2+y}} = \min\left\{\sqrt{\frac{ca}{b}}, \frac{a}{b^{\frac{1}{3}}}\right\}$$

Upon combining this bound with (25), we obtain

$$\sup_{(f,O)\in\mathcal{O}} \varepsilon(f_\alpha, \pi^{QO}) \geq \frac{DB}{2\sqrt{2}} \max\left\{\min\left\{\frac{cm}{M}, 1\right\}, \min\left\{\frac{1}{8}\sqrt{\frac{1}{cT}}, \frac{1}{2(2T)^{1/3}}\right\}\mathbb{1}_{\{c<1\}}\right\},$$

where $c = B/(m\sqrt{2})$. By making cases $1 \leq c$, $\frac{1}{8(2T)^{1/3}} \leq c < 1$, and $c < \frac{1}{8(2T)^{1/3}}$, and using the fact that for $a, b \geq 0$, $\max\{a, b\} \geq a^{1/3}b^{2/3}$ in the second case, we get

$$\sup_{(f,O)\in\mathcal{O}} \varepsilon(f_\alpha, \pi^{QO}) \geq \frac{DB}{2\sqrt{2}} \min\left\{1, \frac{1}{(M/m)}, \frac{1}{4(M/m)^{1/3}T^{1/3}}, \frac{1}{2(2T)^{1/3}}\right\}.$$

By Assumption 3 in Section 4.1, we know that $\frac{M}{m} \leq 2^r$. Therefore,

$$\sup_{(f,O)\in\mathcal{O}} \varepsilon(f_\alpha, \pi^{QO}) \geq \frac{DB}{2\sqrt{2}} \min\left\{\frac{1}{2^r}, \frac{1}{4(2)^{r/3}T^{1/3}}, \frac{1}{2(2T)^{1/3}}\right\}.$$

□

Theorem 4.4 follows as an immediate corollary; Theorem 4.3, too, is obtained by noting that

$$\sup_{(f,O)\in\mathcal{O}} \varepsilon(f_\alpha, \pi^{QO}) < \frac{3DB}{\sqrt{T}}$$

holds only if $\sqrt{T} < 2^r$.

6 Other applications of our quantizers

In this section we will discuss applications of our quantizers for other related problems.

6.1 RATQ and distributed mean estimation

We now discuss the performance of RATQ for distributed mean estimation with limited communication, considered in [47]. To state our results formally, we describe the setting. Consider n vectors $\{x_i\}_{i=1}^n$ with each x_i in $\mathbb{R}^d$ and vector x_i available to client i. Each client communicates to a fusion center using r bits to enable the center to compute the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

We seek to design quantizers which can be used to express x_is using r bits each and yield a high accuracy estimate of $\bar{x}$ at the center.

Specifically, consider two cases for the quantization schemes: The case of *fixed-length* codes where client i uses a randomized encoding mapping $e_i : \mathbb{R}^d \to \{0,1\}^r$, $i \in [n]$, and the center uses a decoding mapping $d : \{0,1\}^{nr} \to \mathbb{R}^d$; and the case of *variable-length* codes where the encoder mappings are $e_i : \mathbb{R}^d \to \{0,1\}^*$ and must satisfy $\mathbb{E}\left[|e_i(x_i)|\right] \leq r$ for each $x_i \in \mathbb{R}^d$ and each $i \in [n]$, where $|b|$ denotes the length of a binary vector b. A distributed mean estimation code π is specified by the encoder mappings e_i and the decoder mapping d. We emphasize that the encoder mappings and the decoder mapping are allowed to be randomized using shared randomness; namely, we allow public-coin simultaneous message passing protocols. We denote the set of all fixed-length codes by $\Pi(r)$ and the set of all variable-length codes by $\Pi_*(r)$.

We measure the performance of a code π by the mean square error (MSE) between $\bar{x}$ and $\hat{\bar{x}} = d(e_1(x_1), ..., e_n(x_n)))$, for a fixed input vector $x = (x_1, ..., x_n)$, given by

$$\mathcal{E}(\pi, x) = \mathbb{E}\left[\|\hat{\bar{x}} - \bar{x}\|_2^2\right].$$

We consider a minmax setting where we allow the input vectors $x = (x_1, ..., x_n)$ to be chosen arbitrarily from the unit Euclidean ball B^d. That is, consider the worst-case MSE over all vectors in B^d given by

$$\mathcal{E}(\pi, B^d) = \max_{x_i \in B^d, \forall i \in [n]} \mathcal{E}(\pi, x)$$

The minimum error attained by fixed-length codes is given by

$$\mathcal{E}(\Pi(r), B^d) = \min_{\pi \in \Pi(r)} \mathcal{E}(\pi, B^d),$$

and that by variable-length codes is given by

$$\mathcal{E}(\Pi_*(r), B^d) = \min_{\pi \in \Pi_*(r)} \mathcal{E}(\pi, B^d),$$

The following lower bound is from [47, Theorem 5] (where it was shown using a construction from [54]).

Theorem 6.1 ([47,54]). *There exists a constant t such that for every $r \leq ndt/4$ and $n \geq 4/t$, we have*

$$\mathcal{E}(\Pi_*(r), B^d) \geq \mathcal{E}(\Pi_*(r), B^d) \geq \frac{t}{4} \min\{1, \frac{d}{r}\}.$$

As a corollary, we have the following alternative form of the same lower bound.

Corollary 6.2. *For $\mathcal{E}(\Pi_*(r), B^d) = O(1/n)$, we must have r to be $\Omega(nd)$.*

The protocol π_{srk} proposed in [47] for this problem achieves $\mathcal{E}(\pi_{srk}, B^d) = O(1/n)$ with $r = \Omega(nd \log\log(d))$. This scheme uses a quantizer that randomly rotates a input vector, similar to RATQ, before quantizing it uniformly. A simpler quantizer similar to CUQ with a variable-length entropic compression code, denoted by π_{svk}, achieves $\mathcal{E}(\pi_{svk}, B^d) = O(1/n)$ with $r = \Omega(nd)$. This establishes the orderwise optimality of π_{svk}. Thus, prior to our work, the best known fixed-length scheme for distributed mean estimation was π_{srk} which was off from the optimal performance attained by a variable-length code by a factor of $\log\log d$.

We now consider performance of a protocol π_{RATQ} in which RATQ is employed by all the clients, and the center declares the average of the quantized values as its mean estimate. First, we have the following lemma which describes the mean square performance of RATQ.

Lemma 6.3. *Let $Q_{\mathtt{at},R}$ be the quantizer RATQ with parameters m, m_0, and h as in* (13)*. Then, for every $\mathbb{R}^d$-valued random variable Y such that $\|Y\|_2^2 \leq 1$ a.s., we have*

$$\mathbb{E}\left[\|Q_{\mathtt{at},R}(Y) - Y\|_2^2\right] \leq \frac{9 + 3\ln s}{(k-1)^2}.$$

See Appendix C.1 for the proof.

Theorem 6.4. *Let π_{RATQ} be a protocol for distributed mean estimation in which RATQ with parameters m, h as in* (13) *and k, s as in* (9) *is employed by all the clients. Then, $\mathcal{E}(\pi_{RATQ}, B^d) = O(1/n)$ with a total precision of*

$$r' = n\left(d(1 + \Delta_1) + \Delta_2\right),$$

where Δ_1 and Δ_2 are as in Corollary 3.4. This further yields $\mathcal{E}(\Pi(r), B^d) = O(1/n)$ for every $r \geq r'$.

See Appendix C.2 for the proof.

Thus, RATQ enjoys the fixed length structure of π_{srk}, while being only $O(\log\log\log \ln^*(d/3))$ away from the expected length of π_{svk}.

6.2 ATUQ and Gaussian rate distortion

In this final section, we consider the classic Gaussian rate-distortion problem. We first describe this problem.

Consider a random vector $X = [X(1), \cdots, X(d)]^T$ with *iid* components $X(1), \cdots, X(d)$ generated from a zero-mean Gaussian distribution with variance σ^2. For pair (R, D) of nonnegative numbers is an *achiveable* rate-distortion pair if we can find a quantizer Q_d of precision dR and with mean square error $\mathbb{E}\left[\|X - Q_d(X)\|_2^2\right] \leq dD$. For $D > 0$, denote by $R(D)$ the infimum over all R such that (R, D) constitute an achievable rate-distortion pair for all d sufficiently large. A well-known result in information theory characterizes $R(D)$ as follows ($cf.$ [14]):

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & \text{if } D \leq \sigma^2, \\ 0 & \text{if } D > \sigma^2. \end{cases}$$

The function $R(D)$ is called the *Gaussian rate-distortion function.*

Over the years, several constructions using error correcting codes and lattices have evolved that attain the rate-distortion function, asymptotically for large d. In this section, we show that a slight variant of ATUQ, too, attains a rate very close to the Gaussian rate-distortion function, when applied to Gaussian random vectors.

Specifically, consider the quantizer $Q_{\mathtt{at},I}$ described earlier in (19). Recall that $Q_{\mathtt{at},I}$ can be described by algorithm 2 and 3 with random matrix R replaced with I. That is, we divide the input vector in $\lceil d/s \rceil$ and employ ATUQ to quantize them. In fact, we will apply this quantizer not only to a Gaussian random vector, but any random vector with subgaussian components; the components need not even be independent. Thus, we show that our quantizer is almost optimal *universally* for all subgaussian random vectors.

We set the parameters m, m_0, h, s, and $\log(k+1)$ of $Q_{at,I}$ as follows:

$$\begin{aligned} m = 3v, \quad m_0 = 2v\ln s, \quad \log h = \left\lceil \log\left(1 + \ln^*\left(\frac{4\ln(8\sqrt{2}v/D)}{3}\right)\right)\right\rceil, \\ s = \min\{\log h, d\}, \text{ and } \log(k+1) = \left\lceil \log\left(2 + \sqrt{\frac{18v + 6v\ln s}{D}}\right)\right\rceil. \end{aligned} \tag{28}$$

Theorem 6.5. *Consider a random vector X taking values in $\mathbb{R}^d$ and with components X_i, $1 \leq i \leq d$ such that each X_i is a centered subgaussian random variables with a variance factor v. Let Q_d be the d-dimensional $Q_{at,I}$ with parameters as in* (28)*. Then, for $d \geq \log h$ and $D < v/4$, Q_d gets the mean square error less than dD using rate R satisfying*

$$R \leq \frac{1}{2}\log\frac{v}{D} + O\left(\log\log\log\log^*\log\left(\frac{v}{D}\right)\right).$$

We provide the proof in Appendix D. We remark that the additional term is a small constant for reasonable values of the parameters v and D. Note that our proposed quantizer just uses uniform quantizers with different dynamic ranges, and yet is almost universally rate optimal.

Acknowledgement

Authors would like to thank Jayadev Acharya and Ananda Theertha Suresh for useful discussions, and Alexander Barg for sharing his knowledge on covering codes and pointing to related work.

Appendices

A Analysis of PSGD with quantized subgradients: Proof of Theorem 2.4

We proceed as in the standard proof of convergence (see, for instance, [13]): Denoting by $\Gamma_{\mathcal{X}}(x)$ the projection of x on the set $\mathcal{X}$ (in the Euclidean norm), the error at time t can be bounded as

$$\begin{aligned}
\|x_t - x^*\|_2^2 &= \|\Gamma_{\mathcal{X}}\big(x_{t-1} - \eta Q(\hat{g}(x_{t-1}))\big) - x^*\|_2^2 \\
&\le \|\big(x_{t-1} - \eta Q(\hat{g}(x_{t-1}))\big) - x^*\|_2^2 \\
&= \|x_{t-1} - x^*\|_2^2 + \|\eta Q(\hat{g}(x_{t-1}))\|_2^2 - 2\eta (x_{t-1} - x^*)^T Q(\hat{g}(x_{t-1})) \\
&= \|x_{t-1} - x^*\|_2^2 + \|\eta Q(\hat{g}(x_{t-1}))\|_2^2 - 2\eta (x_{t-1} - x^*)^T \big(Q(\hat{g}(x_{t-1})) - \hat{g}(x_{t-1})\big) \\
&\quad - 2\eta (x_{t-1} - x^*)^T \hat{g}(x_{t-1}),
\end{aligned}$$

where the first inequality is a well known property of the projection operator Γ (see, for instance, Lemma 3.1, [13]). By rearranging the terms, we have

$$\begin{aligned}
2\eta (x_{t-1} - x^*)^T \hat{g}(x_{t-1}) \le \|x_{t-1} - x^*\|_2^2 - \|x_t - x^*\|_2^2 + \|\eta Q(\hat{g}(x_{t-1}))\|_2^2 \\
- 2\eta (x_{t-1} - x^*)^T \left(Q(\hat{g}(x_{t-1})) - \hat{g}(x_{t-1})\right).
\end{aligned}$$

Also, since $\mathbb{E}\left[\hat{g}(x_{t-1}) | x_{t-1}\right]$ is a subgradient at x_{t-1} for the convex function f, upon taking expectation we get

$$\mathbb{E}\left[f(x_{t-1}) - f(x^*)\right] \le \mathbb{E}\left[(x_{t-1} - x^*)^T \mathbb{E}\left[\hat{g}(x_{t-1}) | x_{t-1}\right]\right],$$

which with the previous bound yields

$$\begin{aligned}
2\eta \mathbb{E}\left[f(x_{t-1}) - f(x^*)\right] \le \mathbb{E}\left[\|x_{t-1} - x^*\|_2^2\right] - \mathbb{E}\left[\|x_t - x^*\|_2^2\right] + \eta^2 \mathbb{E}\left[\|Q(\hat{g}(x_{t-1}))\|_2^2\right] \\
- 2\eta \mathbb{E}\left[(x_{t-1} - x^*)^T \big(Q(\hat{g}(x_{t-1})) - \hat{g}(x_{t-1})\big)\right].
\end{aligned}$$

Next, by the Cauchy-Schwarz inequality and the assumption in (1), the third term on the right-side above can be bounded further to obtain

$$\begin{aligned}
2\eta \mathbb{E}\left[f(x_{t-1}) - f(x^*)\right] \le \mathbb{E}\left[\|x_{t-1} - x^*\|_2^2\right] - \mathbb{E}\left[\|x_t - x^*\|_2^2\right] + \eta^2 \mathbb{E}\left[\|Q(\hat{g}(x_{t-1}))\|_2^2\right] \\
+ 2\eta \cdot D \cdot \mathbb{E}\left[\|\mathbb{E}\left[Q(\hat{g}(x_{t-1})) - \hat{g}(x_{t-1}) | x_{t-1}\right]\|_2\right].
\end{aligned}$$

Finally, we note that, by the definition of α and β, for L_2-bounded oracles we have

$$\begin{aligned}
\mathbb{E}\left[\|Q(\hat{g}(x_{t-1}))\|_2^2\right] \le \alpha(Q)^2, \\
\|\mathbb{E}\left[Q(\hat{g}(x_{t-1})) - \hat{g}(x_{t-1}) | x_{t-1}\right]\|_2 \le \beta(Q),
\end{aligned}$$

which gives

$$2\eta \mathbb{E}\left[f(x_{t-1}) - f(x^*)\right] \le \mathbb{E}\left[\|x_{t-1} - x^*\|_2^2\right] - \mathbb{E}\left[\|x_t - x^*\|_2^2\right] + \eta^2 \alpha(Q)^2 + 2\eta D \beta(Q).$$

Therefore, by summing from $t = 2$ to $T + 1$, dividing by T, and using assumption that the domain $\mathcal{X}$ has diameter at the most D, we have

$$2\eta\mathbb{E}\left[f(\bar{x}_T) - f(x^*)\right] \le \frac{D^2}{T} + \eta^2\alpha(Q)^2 + 2\eta D\beta(Q).$$

The first statement of Theorem 2.4 follows upon dividing by η and setting the value of η as in the statement. The second statement holds in a similar manner by replacing α and β with α_0 and β_0, respectively. □

B Remaining proofs for the main results

B.1 Analysis of CUQ: Proof of Lemmas 5.1 and 5.2

Proof of Lemma 5.1: Denoting by $\mathcal{B}_{j,\ell}$ the event $\left\{Y(j) \in [B_{M,k}(\ell), B_{M,k}(\ell+1))\right\}$, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{j\in[d]} \left(Q_{\mathtt{u}}(Y)(j) - Y(j)\right)^2 \mathbb{1}_{\{|Y(j)|\le M\}} \mid Y\right] \\
&= \sum_{j\in[d]} \sum_{\ell=0}^{k-1} \mathbb{E}\left[\left(Q_{\mathtt{u}}(Y)(j) - Y(j)\right)^2 \mathbb{1}_{\mathcal{B}_{j,\ell}} \mid Y\right] \mathbb{1}_{\{|Y(j)|\le M\}}.
\end{aligned}$$

For the summand on the right-side, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(Q_{\mathtt{u}}(Y)(j) - Y(j)\right)^2 \mathbb{1}_{\mathcal{B}_{j,\ell}} \mid Y\right] \\
&= \left((B_{M,k}(\ell+1) - Y(j))^2 \frac{Y(j) - B_{M,k}(\ell)}{B_{M,k}(\ell+1) - B_{M,k}(\ell)}\right) \mathbb{1}_{\mathcal{B}_{j,\ell}} \\
&\qquad + \left((B_{M,k}(\ell) - Y(j))^2 \frac{B_{M,k}(\ell+1) - Y(j)}{B_{M,k}(\ell+1) - B_{M,k}(\ell)}\right) \mathbb{1}_{\mathcal{B}_{j,\ell}} \\
&= (B_{M,k}(\ell+1) - Y(j))(Y(j) - B_{M,k}(\ell))\, \mathbb{1}_{\mathcal{B}_{j,\ell}} \\
&\le \frac{1}{4}\left(B_{M,k}(\ell+1) - B_{M,k}(\ell)\right)^2 \\
&= \frac{M^2}{(k-1)^2},
\end{aligned} \tag{29}$$

where the inequality uses the GM-AM inequality and the final identity is simply by the definition of $B_{M,k}(\ell)$. Upon combining the bounds above, we obtain

$$\mathbb{E}\left[\sum_{j\in[d]} \left(Q_{\mathtt{u}}(Y)(j) - Y(j)\right)^2 \mathbb{1}_{\{|Y(j)|\le M\}} \mid Y\right] \le \frac{dM^2}{(k-1)^2} \cdot \frac{1}{d}\sum_{j\in[d]} \mathbb{1}_{\{|Y(j)|\le M\}}.$$

□

Proof of Lemma 5.2 The proof is the same as the proof of Lemma 5.1, except that we need to set $d = 1$ and replace the identity used in(29) with

$$B_{M,k}(\ell+1) - B_{M,k}(\ell) = \frac{M}{k-1}.$$

B.2 Proof of Lemma 5.8

For the rotation matrix $R = (1/\sqrt{d})HD$, each entry of $RY(j)$ of the rotated matrix has the same distribution as $(1/\sqrt{d})V^TY$, where $V = [V(1), ..., V(d)]^T$ has independent Rademacher entries. We will use this observation to bound the moment generating function of $RY(i)$ conditioned on Y. Towards that end, we have

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda RY(i)} \mid Y\right] &= \prod_{i=1}^{d} \mathbb{E}\left[e^{\lambda V(i)Y(i)/\sqrt{d}} \mid Y\right] \\
&= \prod_{i=1}^{d} \frac{e^{\lambda Y(i)/\sqrt{d}} + e^{-\lambda Y(i)/\sqrt{d}}}{2} \\
&\leq \prod_{i=1}^{d} e^{\lambda^2 Y(i)^2/2d} \\
&= e^{\lambda^2 \|Y\|_2^2/2d},
\end{aligned}$$

where the first identity follows from independence of $V(i)$s and the first inequality follows by the fact that $(e^x + e^{-x})/2$ is less than $e^{x^2/2}$, which in turn can be seen from the Taylor series expansion of these terms. Thus, we have proved the following:

$$\mathbb{E}\left[e^{\lambda RY(i)} \mid Y\right] \leq e^{\lambda^2 \|Y\|_2^2/2d}, \quad \forall \lambda \in \mathbb{R}, \forall i \in [d]. \tag{30}$$

Note that $\|Y\|_2^2$ can be further bounded by B^2, which along with (30) leads to

$$\mathbb{E}\left[e^{\lambda RY(i)}\right] \leq e^{\lambda^2 B^2/2d} \quad \forall \lambda \in \mathbb{R}, \forall i \in [d].$$

Using this inequality and the observation that $\mathbb{E}\left[RY(i)\right] = 0$, we note that $RY(i)$ is a centered subgaussian with a variance parameter B^2/d. The second statement of the lemma trivially follows from Lemma 5.6. □.

B.3 Proof of Lemma 5.11

For any $y \in [0,1]$ such that $ab^{y/(2+y)} < c$, we have $ab^{y/(2+y)} < \min\left\{c, ab^{1/3}\right\}$. By multiplying by a/b on both sides and taking square root, we get

$$\frac{a}{b^{\frac{1}{2+y}}} < \min\left\{\sqrt{\frac{ca}{b}}, \frac{a}{b^{1/3}}\right\},$$

which gives

$$\sup_{y\in[0,1]:a(b)^{y/(2+y)}<c.} \frac{a}{b^{\frac{1}{2+y}}} \leq \min\left\{\sqrt{\frac{ca}{b}}, \frac{a}{b^{1/3}}\right\}.$$

Making cases $ab^{1/3} \geq c$ and $ab^{1/3} < c$, we note that the supremum on the left-side equals the right-side in both the cases. □

C Analysis of distributed mean estimation

C.1 Proof of Lemma 6.3

By the description of RATQ we have that $Q_{\texttt{at},R} = R^{-1} Q_{\texttt{at},I}(RY)$, where $Q_{\texttt{at},I}$ is as defined in (19). Thus, using the fact that R is a unitary matrix

$$\mathbb{E}\left[\|Q_{\texttt{at},R}(Y) - Y\|_2^2\right] = \mathbb{E}\left[\|Q_{\texttt{at},I}(RY) - RY\|_2^2\right].$$

When the parameters are set as in (13), we get

$$RY(j) \leq M_{h-1} \text{ a.s.}, \forall j \in [d],$$

whereby

$$\mathbb{E}\left[\|Q_{\texttt{at},R}(Y) - Y\|_2^2\right] = \mathbb{E}\left[\sum_{j\in[d]} (Q_{\texttt{at},I}(RY)(j) - RY(j))^2 \mathbb{1}_{RY(j)\leq M_{h-1}}\right].$$

The proof is completed by noting that Y satisfies $\|Y\|_2 \leq 1$ a.s., setting $m = 3/d$ and $m_0 = (2/d)\ln s$, and applying Lemma 5.9. □

C.2 Proof of Theorem 6.4

Proof. When RATQ is employed by all the clients, the MSE between the sample mean of the quantized vectors and the sample mean of the input is bounded as

$$\begin{aligned}
\mathcal{E}(\pi_{RATQ}, B^d) &= \max_{x: x_i \in B^d, \forall i \in [d]} \mathbb{E}\left[\|\frac{\sum_{i=1}^n Q_{\texttt{at},R_i}(x_i)}{n} - \frac{\sum_{i=1}^n x_i}{n}\|_2^2\right] \\
&= \max_{x: x_i \in B^d, \forall i \in [d]} \sum_{i=1}^n \frac{1}{n^2} \mathbb{E}\left[\|Q_{\texttt{at},R_i}(x_i) - x_i\|_2^2\right] \\
&\leq \frac{9 + 3\ln s}{n(k-1)^2} \\
&= \frac{1}{n},
\end{aligned}$$

where the second identity uses the unbiasedness of quantizers used by all the clients[8] and the independence of randomness used by the quantizers of different clients, the third inequality uses Lemma (6.3), and the final identity follows by substituting for k. □

[8]Note that $\{x_i\}_{i=1}^n$ are deterministic vectors, each with $\|x_i\|_2 \leq 1$. Thus, setting $B = 1$ and choosing h so that $M_{h-1} \geq 1$, we get the desired unbiased estimators.

D Proof of Theorem 6.5

We split the overall mean square error into two terms and derive upper bounds for each of them. Specifically, we have

$$\frac{1}{d} \cdot \mathbb{E}\left[\|X_d - Q_d(X_d)\|_2^2\right] = \frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} (X_d(i) - Q_d(X_d)(i))^2 \mathbb{1}_{\{|X_d(i)| \leq M_{h-1}\}}\right]$$
$$+ \frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} (X_d(i) - Q_d(X_d)(i))^2 \mathbb{1}_{\{|X_d(i)| > M_{h-1}\}}\right].$$

The second term on the right-side above can be bounded as follows:

$$\begin{aligned}\frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} (X_d(i) - Q_d(X_d)(i))^2 \mathbb{1}_{\{|X_d(i)| > M_{h-1}\}}\right] &= \frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} X_d(i)^2 \mathbb{1}_{\{|X_d(i)| > M_{h-1}\}}\right] \\ &\leq \mathbb{E}\left[X_d(1)^4\right]^{1/2} P\left(|X_d(1)| > M_{h-1}\right)^{1/2} \\ &\leq 4\sqrt{2} v e^{-\frac{M_{h-1}^2}{4v}},\end{aligned}$$

where the first inequality follows by the Cauchy-Schwarz inequality and the second follows by Lemma 5.6. Note that $M_{h-1}^2 \geq m e^{*(h-1)} \geq 3v e^{*\ln^*(4\ln(8\sqrt{2}v/D)/3)} = 4v \ln(8\sqrt{2}v/D)$, which with the previous bound leads to

$$\frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} (X_d(i) - Q_d(X_d)(i))^2 \mathbb{1}_{\{|X_d(i)| > M_{h-1}\}}\right] \leq \frac{D}{2}.$$

Furthermore, by Lemma 5.7 we have

$$\frac{1}{d} \cdot \mathbb{E}\left[\sum_{i \in [d]} (X_d(i) - Q_d(X_d)(i))^2 \mathbb{1}_{\{|X_d(i)| \leq M_{h-1}\}}\right] \leq \frac{9v + 3v \ln s}{(k-1)^2} \leq \frac{D}{2},$$

where the last equality holds since $k \geq 1 + \sqrt{\frac{18v + 6v \ln s}{D}}$.

It remains to bound the rate. Note that the overall resolution used for the entire vector is

$$d \log(k+1) + \left\lceil \frac{d}{s} \right\rceil \log h \leq d \left\lceil \log\left(2 + \sqrt{\frac{18v + 6v \ln s}{D}}\right)\right\rceil + d + \log h$$

Therefore, for $d \geq \log h$ and $D < v/4$, the proof is completed by bounding the rate R as

$$\begin{aligned}R &\leq \log\left(2 + \sqrt{\frac{v}{D}}\sqrt{18 + 6\ln\log h}\right) + 3 \\ &\leq \frac{1}{2}\log\frac{v}{D} + \log\left(1 + \sqrt{18 + 6\ln\left\lceil \log\left(1 + \ln^*\left(\frac{4\ln(8\sqrt{2}v/D)}{3}\right)\right)\right\rceil}\right) + 3 \\ &\leq \frac{1}{2}\log\frac{v}{D} + O\left(\log\log\log\log^*\log\frac{v}{D}\right).\end{aligned}$$

References

[1] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under Information Constraints I: Lower Bounds from Chi-Square Contraction," *arXiv:1812.11476*, 2018.

[2] ——, "Inference under Information Constraints II: Communication Constraints and Shared Randomness," *arXiv:1905.08302*, 2019.

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

[4] 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.

[5] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar, "Information-theoretic lower bounds on the oracle complexity of convex optimization," *Advances in Neural Information Processing Systems*, pp. 1–9, 2009.

[6] 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.

[7] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast johnson-lindenstrauss transform," *Proceedings of the ACM symposium on Theory of computing (STOC'06)*, pp. 557–563, 2006.

[8] E. Akyol and K. Rose, "On constrained randomized quantization," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3291–3302, July 2013.

[9] 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.

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

[11] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[12] M. Braverman, A. Garg, T. Ma, H. L. Nguyen, and D. P. Woodruff, "Communication lower bounds for statistical estimation problems via a distributed data processing inequality," *Proceedings of ACM Symposium on the Theory of Computing (STOC' 16)*, pp. 1011–1020, 2016.

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

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

[15] I. Csiszár and P. Narayan, "Capacity of the gaussian arbitrarily varying channel," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 18–26, 1991.

[16] 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.

[17] J. C. Duchi, "Introductory lectures on stochastic optimization," 2017, Available Online http://stanford.edu/~jduchi/PCMIConvex/Duchi16.pdf.

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

[19] I. Dumer, "Covering spheres with spheres," *Discrete & Computational Geometry*, vol. 38, no. 4, pp. 665–679, Dec 2007.

[20] R. G. Gallager, *Information theory and reliable communication.* Springer, 1968, vol. 2.

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

[22] A. Garg, T. Ma, and H. L. Nguyen, "On Communication Cost of Distributed Statistical Estimation and Dimensionality," *Advances in Neural Information Processing Systems*, pp. 2726–2734, 2014.

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

[24] 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.

[25] R. Hadad and U. Erez, "Dithered quantization via orthogonal transformations," *IEEE Transactions on Signal Processing*, vol. 64, pp. 5887–5900, 11 2016.

[26] J. Hamkins and K. Zeger, "Asymptotically dense spherical codes. I. Wrapped spherical codes," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1774–1785, Nov 1997.

[27] ——, "Asymptotically dense spherical codes. II. Laminated spherical codes," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1786–1798, Nov 1997.

[28] J. Hamkins, "Design and analysis of spherical codes," *Ph. D. Dissertation, UCSD*, 1996.

[29] J. Hamkins and K. Zeger, "Gaussian source coding with spherical codes," *IEEE Transactions on Information Theory*, vol. 48, no. 11, pp. 2980–2989, 2002.

[30] Y. Han, A. Özgür, and T. Weissman, "Geometric Lower Bounds for Distributed Parameter Estimation under Communication Constraints," *Proceedings of the International Conference on Computational Learning Theory (COLT' 18)*, vol. 75, pp. 3163–3188, 2018.

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

[32] X. Hu, L. Prashanth, A. György, and C. Szepesvári, "(Bandit) Convex Optimization with Biased Noisy Gradient Oracles," *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS' 16)*, pp. 819–828, 2016.

[33] B. Hughes and P. Narayan, "Gaussian arbitrarily varying channels," *IEEE Transactions on Information Theory*, vol. 33, no. 2, pp. 267–284, 1987.

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

[35] A. Lapidoth, "On the role of mismatch in rate distortion theory," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 38–47, 1997.

[36] E. Martinian and M. Wainwright, "Low density codes achieve the rate-distortion bound," *Proc. IEEE Data Compression Conference*, pp. 153–162, 2006.

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

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

[39] Y. Nesterov, "Introductory lectures on convex optimization: A basic course," *Springer Science and Business Media*, vol. 87, 2013.

[40] V. Ostromoukhov, R. D. Hersch, and I. Amidror, "Rotated dispersed dither: A new technique for digital halftoning," in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '94. New York, NY, USA: ACM, 1994, pp. 123–130.

[41] 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.

[42] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, pp. 400–407, 1951.

[43] K. Sahasranand and H. Tyagi, "Extra samples can reduce the communication for independence testing," *Proc. International Symposium of Information Theory*, pp. 2316–2320, 2018.

[44] O. Shamir, "Fundamental limits of online and distributed algorithms for statistical learning and estimation," *Advances in Neural Information Processing Systems*, pp. 163–171, 2014.

[45] N. Sommer, M. Feder, and O. Shalvi, "Low-density lattice codes," *IEEE Transactions on Information Theory*, vol. 54, no. 4, pp. 1561–1585, April 2008.

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

[47] 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.

[48] 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.

[49] 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.

[50] B. Woodworth and N. Srebro, "Open Problem: The Oracle Complexity of Convex Optimization with Limited Memory," *arXiv:1907.00762*, 2019.

[51] A. D. Wyner, "Random packings and coverings of the unit n-sphere," *The Bell System Technical Journal*, vol. 46, no. 9, pp. 2111–2118, 1967.

[52] A. Xu and M. Raginsky, "Information-theoretic lower bounds on Bayes risk in decentralized estimation," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1580–1600, 2017.

[53] Y. Yan, C. Ling, and X. Wu, "Polar lattices: where arıkan meets forney," *Proc. IEEE International Symposium of Information Theory*, pp. 1292–1296, 2013.

[54] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," *Advances in Neural Information Processing Systems*, pp. 2328–2336, 2013.

[55] J. Ziv, "On universal quantization," *IEEE Transactions on Information Theory*, vol. 31, no. 3, pp. 344–347, 1985.