

Differentially Quantized Gradient Methods

Chung-Yi Lin, Victoria Kostina, and Babak Hassibi

arXiv:2002.02508v3 [cs.LG] 22 Jun 2021

Abstract—Consider the following distributed optimization scenario. A worker has access to training data that it uses to compute the gradients while a server decides when to stop iterative computation based on its target accuracy or delay constraints. The only information that the server knows about the problem instance is what it receives from the worker via a rate-limited noiseless communication channel. We introduce the principle we call *differential quantization* (DQ) that prescribes that the past quantization errors should be compensated in such a way as to direct the descent trajectory of a quantized algorithm towards that of its unquantized counterpart. Assuming that the objective function is smooth and strongly convex, we prove that *differentially quantized gradient descent* (DQ-GD) attains a linear contraction factor of $\max\{\sigma_{\text{GD}}, \rho_n 2^{-R}\}$, where σ_{GD} is the contraction factor of unquantized gradient descent (GD), $\rho_n \geq 1$ is the covering efficiency of the quantizer, and R is the bitrate per problem dimension n . Thus at any $R \geq \log_2 \rho_n / \sigma_{\text{GD}}$ bits, the contraction factor of DQ-GD is the same as that of unquantized GD, i.e., there is no loss due to quantization. We show a converse demonstrating that no GD-like quantized algorithm can converge faster than $\max\{\sigma_{\text{GD}}, 2^{-R}\}$. Since quantizers exist with $\rho_n \rightarrow 1$ as $n \rightarrow \infty$ (Rogers, 1963), this means that DQ-GD is asymptotically optimal. In contrast, naively quantized GD where the worker directly quantizes the gradient barely attains $\sigma_{\text{GD}} + \rho_n 2^{-R}$. The principle of differential quantization continues to apply to gradient methods with momentum such as Nesterov’s accelerated gradient descent, and Polyak’s heavy ball method. For these algorithms as well, if the rate is above a certain threshold, there is no loss in contraction factor obtained by the differentially quantized algorithm compared to its unquantized counterpart, and furthermore, the differentially quantized heavy ball method attains the optimal contraction achievable among all (even unquantized) gradient methods. Experimental results on both simulated and real-world least-squares problems validate our theoretical analysis.

Index Terms—gradient descent, quantized gradient descent, accelerated gradient descent, heavy ball method, error compensation, sigma-delta modulation, federated learning, linear convergence.

I. INTRODUCTION

A. Motivation and related work

Distributed optimization plays a central role in large-scale machine learning where gradient descent (GD) and its stochastic variant SGD are employed to minimize an objective function [2]–[9]. Despite the scalability of parallel gradient training, the frequent exchange of high-dimensional gradients has become a communication overhead that slows down the overall learning process [3], [6], [10]–[13].

The authors are with California Institute of Technology (chungyi@caltech.edu, vkostina@caltech.edu, hassibi@caltech.edu). This work was supported in part by the National Science Foundation (NSF) under grants CCF-1751356, CCF-1956386, CNS-0932428, CCF-1018927, CCF-1423663 and CCF-1409204, by a grant from Qualcomm Inc., by NASA’s Jet Propulsion Laboratory through the President and Director’s Fund, and by King Abdullah University of Science and Technology. This paper is presented in part at ISIT 2021 [1].

To reduce the communication cost, one line of research focuses on gradient quantization with a fixed number of bits per problem dimension. Scalar quantizers, which quantize each coordinate of the input vector separately, are often used to address the communication bottleneck in distributed SGD algorithms. For each coordinate, the quantization levels are distributed either uniformly [10], [14]–[17] or non-uniformly [18] within the dynamic-range interval. Convergence of such quantized gradient methods is typically established by relating the variance of the quantized (thus noisy) gradient to the bit rate, and the communication cost is further reduced using an efficient integer coding scheme such as the Elias encoding [14], [18]. Due to scalar quantization, the obtained contraction factor would pay an extra dimension-dependent factor. On the other hand, vector quantizers are hard to implement in practice and hence are less considered in this context. In this direction, [19], [20] construct vector quantizers from the convex hull of specifically structured point sets.

Another common way to reduce the communication bandwidth in parallel SGD training is to sparsify the gradient vectors. For example, the top- k sparsifier (or *compressor*) preserves the k coordinates of the largest magnitude and sends them with full precision [12], [22]–[26]. The compressors used are either biased [12], [23], [25], [27]–[31] or unbiased [32]–[34], and the compression error is controlled by a user-specific accuracy parameter (e.g. k for the top- k compressor). Similar to the failure of 1-bit SGD (without mini-batching) due to the biased quantization error [10], [21], it is observed in [31] that compressed GD with the top-1 compressor does not always converge. For an empirical risk minimization problem where the global objective function is the average of local objective functions, recent works [35]–[37] perform analog gradient compression and communication by taking the physical superposition nature of the underlying multiple-access channel into the account.

Most distributed SGD algorithms with biased compressors in the literature apply the idea of error compensation that can be traced back to $\Sigma\Delta$ modulation [38]. To form the compressor input at each iteration, there are various ways to add past compression errors back to the computed gradients. While [39] weights all the past errors in a time-decaying fashion, the *error-feedback scheme* (EF) uses only the very last error and provides convergence for the single-worker setting [21], [25], [40] as well as for the multi-worker setting [31], [41].

Although convergence rates of quantized gradient methods depend on the bit rate R [14], [17]–[19], few existing works provide convergence lower bounds in terms of R that apply to any algorithm within a specified class. For quantized projected SGD, [17], [19] give lower bounds to a minimax expected estimation error (i.e. difference between the output function

value and the optimal one), which is in the same order of convergence as that of the unquantized SGD over convex functions. However, the allowable quantizer input in [17], [19] is fixed to be the gradient of the current iterate. Besides, standard assumptions such as unbiasedness and boundedness for the stochastic gradients [42]–[44] are crucial for SGD and in fact simplify the analysis of the quantized gradient algorithms. In this paper, we show that the performance of quantized (non-stochastic) gradient descent can be improved if this restriction is relaxed.

B. Contributions

We consider the single-worker scenario of the parameter server framework [11], [14]–[16], [18], [45], [46] consisting of a worker that computes the gradients and a server that successively refines the model parameter (i.e. the iterate) and decides when to stop the distributed iterative algorithm based on its target accuracy or delay constraints. See Fig. 1.

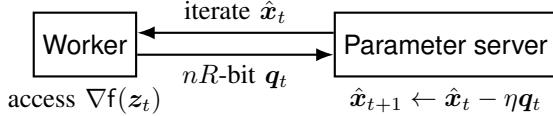


Fig. 1: Quantized gradient descent in a single-worker remote training setting. At each iteration t , the server first sends the current iterate \hat{x}_t to the worker noiselessly, who computes the gradient at some point z_t that is a function of (but not necessarily equal to) \hat{x}_t . Then, the worker forms a descent direction q_t and pushes it back to the server under the nR bits per iteration constraint.

We study the fundamental tradeoff between the convergence rate and the communication rate of quantized gradient descent. We focus on the class \mathcal{F}_n of smooth and strongly convex objective functions $f: \mathbb{R}^n \mapsto \mathbb{R}$ whose minimizers are bounded in the Euclidean norm. For a quantized iterative algorithm A, its worst-case linear *contraction factor* over \mathcal{F}_n at rate R bits per problem dimension is defined as

$$\sigma_A(n, R) \triangleq \inf_{R' \leq R} \sup_{f \in \mathcal{F}_n} \limsup_{T \rightarrow \infty} \|\hat{x}_T(R') - x_f^*\|^{1/T} \quad (1)$$

where x_f^* is the optimizer, and $\hat{x}_0(R'), \hat{x}_1(R'), \hat{x}_2(R'), \dots$ is the sequence of iterates generated by A in response to $f \in \mathcal{F}_n$ when it operates at R' bits per problem dimension.

We consider three popular algorithms that converge linearly:¹ the classical gradient descent (GD) with fixed step size, the accelerated gradient descent (AGD) [47], and the heavy ball method (HB) [48]. We propose a principle for error feedback we call *differential quantization* (DQ) that says that the quantizer input should be formed in such a way as to guide the descent trajectory of the quantized algorithm towards the descent trajectory of its unquantized counterpart. By applying the DQ principle to the GD, AGD, and HB algorithms, we construct three new quantized iterative optimization algorithms:

¹The term “linear convergence” is used in the literature as a synonym for convergence with the rate of geometric progression. Note that SGD converges only sub-linearly over smooth and strongly convex functions [42]–[44].

DQ-GD, DQ-AGD, and DQ-HB. By analyzing them, we show achievability bounds of the form²

$$\sigma_A(n, R) \leq \max \{ \sigma_A, \rho_n 2^{-R} \phi_A(n, R) \}, \quad (2)$$

where $A \in \{\text{DQ-GD, DQ-AGD, DQ-HB}\}$, σ_A is the contraction factor of the unquantized counterpart of A, $\rho_n \geq 1$ is the covering efficiency of the quantizer, and $\phi_A(n, R) \geq 1$ is a function that we specify; for example,

$$\sigma_{\text{DQ-GD}}(n, R) \leq \max \{ \sigma_{\text{GD}}, \rho_n 2^{-R} \}. \quad (3)$$

As (2) indicates, each of the novel DQ algorithms achieves the corresponding σ_A once the rate passes a hard threshold. In other words, there is no loss at all due to quantization once the rate is high enough.

We show an information-theoretic converse of the form

$$\sigma_A(n, R) \geq \max \{ \sigma_{\text{GD}}, 2^{-R} \}, \quad (4)$$

which applies to any “quantized gradient descent” algorithm A (in the class of “quantized gradient descent” algorithms, summarized in Fig. 1, the server can utilize only the last quantized input to form the next iterate). Recalling the classical result of Rogers [49, Th. 3] that shows the existence of quantizers with covering efficiency $\rho_n \rightarrow 1$ as $n \rightarrow \infty$ and comparing (2) and (4), one can deduce the asymptotic optimality of DQ-GD within the class of “GD-like” algorithms. In contrast, the widely adopted method that quantizes the gradient of its current iterate directly [14]–[16], [50] referred to as naively quantized (NQ) GD in this paper, has contraction factor

$$\sigma_{\text{NQ-GD}}(n, R) \leq \sigma_{\text{GD}} + \frac{2\kappa}{\kappa + 1} \rho_n 2^{-R} \quad (5)$$

where $\kappa \geq 1$ is the condition number of f . The guarantee (5) is significantly worse than (3).

Our numerical results indicate that the upper bounds (2) and (5) accurately represent the actual achieved contraction factors.

Within a wider class of quantized gradient methods (the server can utilize full memory of the past), the converse (4) can be surpassed. Once the rate passes the threshold mentioned earlier, DQ-HB attains the minimum possible contraction factor among all algorithms in that wider class, even unquantized ones.

The rest of the paper is organized as follows. Differentially quantized algorithms are presented in Section II. Their convergence analyses and an experimental validation on least-squares problems are shown in Section III. The converses are presented in Section IV. The multi-worker setting is discussed in Section V.

II. DIFFERENTIALLY QUANTIZED ALGORITHMS

A. Quantizers employed in DQ algorithms

A quantizer of dimension n and rate R is a function $q: \mathcal{D} \rightarrow \mathbb{R}^n$, where $\mathcal{D} \subseteq \mathbb{R}^n$ is the domain, such that the image of q satisfies

$$|\text{Im}(q)| = 2^{nR}. \quad (6)$$

²The convergence result on DQ-HB in (2) requires that the function $f \in \mathcal{F}_n$ is twice continuously differentiable.

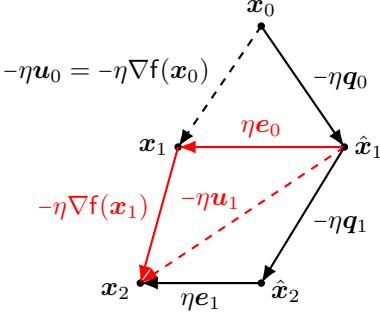


Fig. 2: Illustration of the DQ-GD algorithm (Algorithm 1).

We fix a dimension- n , rate- R quantizer q , and we set up quantizer q_t to be used at iteration t as

$$q_t(\cdot) = r_t q(\cdot/r_t) \quad (7)$$

for a properly chosen sequence of shrinkage factors $\{r_t\}$ (see (22), (34), and (45), below). Therefore, each quantizer q_t has the same geometric structure but different resolution.

B. Differentially Quantized Gradient Descent

The (unquantized) gradient descent algorithm searches along the direction of the negative gradient toward which the function value decreases:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t), \quad (8)$$

where $\eta > 0$ is the constant stepsize chosen to minimize the function value along the search direction.

In Fig. 2, we illustrate an application of differential quantization (DQ) to GD (8), which yields the DQ-GD algorithm (Algorithm 1). At each iteration t , DQ-GD first determines the iterate \mathbf{x}_t associated with the corresponding unquantized algorithm, i.e., GD, by compensating previous scaled quantization error ηe_{t-1} (Line 4). It then computes the gradient at $\mathbf{z}_t = \mathbf{x}_t$ (Lemma B.1) and sets the quantizer input as (Line 5)

$$\mathbf{u}_t = \nabla f(\hat{\mathbf{x}}_t + \eta e_{t-1}) - \mathbf{e}_{t-1}, \quad (9)$$

which in the absence of quantization error e_t would guide the iterate $\hat{\mathbf{x}}_t$ back to \mathbf{x}_{t+1} (see Fig. 2). The recorded scaled quantization error ηe_t captures exactly the difference between $\hat{\mathbf{x}}_{t+1}$ and \mathbf{x}_{t+1} for the next iteration.

See Appendix A for the DQ-GD algorithm with varying stepsize η_t .

C. Differentially Quantized Accelerated Gradient Descent

Nesterov's Accelerated Gradient Descent (AGD) [51] keeps track of two iterate sequences

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) \quad (10)$$

$$\mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \gamma (\mathbf{y}_{t+1} - \mathbf{y}_t). \quad (11)$$

It first performs the gradient descent step (10), and then adds the momentum term $\gamma (\mathbf{y}_{t+1} - \mathbf{y}_t)$ (11) to form a projection \mathbf{x}_{t+1} of the GD iterate \mathbf{y}_{t+1} to its near future. The momentum term incorporates second-order effects by leveraging the past

Algorithm 1: DQ-GD

```

1 Initialize  $\mathbf{e}_{-1} = \mathbf{0}$ 
2 for  $t = 0, 1, 2, \dots$  do
3   Worker:
4      $\mathbf{z}_t = \hat{\mathbf{x}}_t + \eta \mathbf{e}_{t-1}$ 
5      $\mathbf{u}_t = \nabla f(\mathbf{z}_t) - \mathbf{e}_{t-1}$ 
6      $\mathbf{q}_t = q_t(\mathbf{u}_t)$ 
7      $\mathbf{e}_t = \mathbf{q}_t - \mathbf{u}_t$ 
8   Server:  $\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t - \eta \mathbf{q}_t$ 
9 end

```

y_t . The AGD is the first algorithm that achieved the contraction factor that is order-wise optimal (in terms of the condition number of f) among all first-order (gradient) optimization methods [47] (Lemma B.5). There are various interpretations of Nesterov's acceleration phenomenon. We refer the reader to [52] for a connection to the mirror descent algorithm and to [53] for an interpretation in terms of differential equation.

Differentially Quantized AGD algorithm is presented as Algorithm 2. At each iteration t , DQ-AGD uses the past two quantization errors $\mathbf{e}_{t-1}, \mathbf{e}_{t-2}$ to determine the gradient-compute point \mathbf{z}_t (Line 4) and the quantizer input \mathbf{u}_t (Line 5). As dictated by the principle of differential quantization, DQ-AGD computes the gradient at the same point as unquantized AGD, i.e., $\mathbf{z}_t = \mathbf{x}_t$ (Lemma B.4).

Algorithm 2: DQ-AGD

```

1 Initialize  $\mathbf{e}_{-2} = \mathbf{e}_{-1} = \mathbf{0}$ ,  $\hat{\mathbf{y}}_0 = \hat{\mathbf{x}}_0$ 
2 for  $t = 0, 1, 2, \dots$  do
3   Worker:
4      $\mathbf{z}_t = \hat{\mathbf{x}}_t + \eta [\mathbf{e}_{t-1} + \gamma (\mathbf{e}_{t-1} - \mathbf{e}_{t-2})]$ 
5      $\mathbf{u}_t = \nabla f(\mathbf{z}_t) - [\mathbf{e}_{t-1} + \gamma (\mathbf{e}_{t-1} - \mathbf{e}_{t-2})]$ 
6      $\mathbf{q}_t = q_t(\mathbf{u}_t)$ 
7      $\mathbf{e}_t = \mathbf{q}_t - \mathbf{u}_t$ 
8   Server:
9      $\hat{\mathbf{y}}_{t+1} = \hat{\mathbf{x}}_t - \eta \mathbf{q}_t$ 
10     $\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{y}}_{t+1} + \gamma (\hat{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t)$ 
11 end

```

D. Differentially Quantized Heavy Ball Method

Polyak's Heavy Ball (HB) algorithm [48] iterates

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) + \gamma (\mathbf{x}_t - \mathbf{x}_{t-1}), \quad (12)$$

where $\gamma (\mathbf{x}_t - \mathbf{x}_{t-1})$ is the momentum term that nudges \mathbf{x}_{t+1} in the direction of the previous step, and accelerates convergence to the optimizer. In contrast to AGD, the HB method only uses the gradient at the current iterate. The HB method derives from the analogy with physics, since the continuous-time counterpart of (12) is a second-order ODE that models the motion of a body ("the heavy ball") in a field with potential f under the force of friction. At the expense of requiring function f in \mathcal{F}_n to be further twice continuously differentiable, the HB algorithm can be shown to converge

with the optimal contraction factor achievable among all first-order optimization methods [48, Th. 3.1] (Lemma B.8), [47, Th. 2.1.13] (Lemma C.2). In comparison, the AGD approaches it only order-wise, but it does not require the second derivative of f to exist, a significant restriction in practical applications.

Differentially Quantized HB algorithm is presented as Algorithm 3. In accordance with the principle of differential quantization, the worker computes the gradient at $\mathbf{z}_t = \hat{\mathbf{x}}_t$ (Lemma B.7). Note that DQ-HB has the same expression for its quantizer input \mathbf{u}_t (Line 4) as DQ-AGD (Line 5).

Algorithm 3: Differentially Quantized Heavy Ball Method (DQ-HB)

```

1 Initialize  $\mathbf{e}_{-2} = \mathbf{e}_{-1} = \mathbf{0}$ ,  $\hat{\mathbf{x}}_{-1} = \hat{\mathbf{x}}_0$ 
2 for  $t = 0, 1, 2, \dots$  do
3   Worker:
4      $\mathbf{z}_t = \hat{\mathbf{x}}_t + \eta \mathbf{e}_{t-1}$ 
5      $\mathbf{u}_t = \nabla f(\mathbf{z}_t) - [\mathbf{e}_{t-1} + \gamma (\mathbf{e}_{t-1} - \mathbf{e}_{t-2})]$ 
6      $\mathbf{q}_t = \mathbf{q}_t(\mathbf{u}_t)$ 
7      $\mathbf{e}_t = \mathbf{q}_t - \mathbf{u}_t$ 
8   Server:  $\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t - \eta \mathbf{q}_t + \gamma (\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t-1})$ 
9 end
```

III. CONVERGENCE RATES OF DQ ALGORITHMS

A. Definitions

We denote by $\|\cdot\|$ the Euclidean norm, and by $\mathcal{B}(r) \triangleq \{\mathbf{u} \in \mathbb{R}^n : \|\mathbf{u}\| \leq r\}$ the Euclidean ball in \mathbb{R}^n with radius r and center at $\mathbf{0}$.

We fix positive scalars L , and μ , and D , and we say that a continuously differentiable function $f: \mathbb{R}^n \mapsto \mathbb{R}$ is in class \mathcal{F}_n if

i) f is L -smooth, i.e.,

$$\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq L \|\mathbf{v} - \mathbf{w}\|; \quad (13)$$

ii) f is μ -strongly convex, i.e.,

$$\text{function } \mathbf{v} \mapsto f(\mathbf{v}) - \frac{\mu}{2} \|\mathbf{v}\|^2 \text{ is convex}; \quad (14)$$

iii) the minimizer $\mathbf{x}_f^* \triangleq \arg \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$ satisfies

$$\|\mathbf{x}_f^* - \hat{\mathbf{x}}_0\| \leq D, \quad (15)$$

where $\hat{\mathbf{x}}_0$ is the starting location of iterative algorithms.

We say that $f: \mathbb{R}^n \mapsto \mathbb{R}$ is in class \mathcal{F}_n^2 if it is in \mathcal{F}_n and is in addition twice continuously differentiable.

We denote the *condition number* of an $f \in \mathcal{F}_n$ by

$$\kappa \triangleq \frac{L}{\mu}. \quad (16)$$

Note that $\kappa \geq 1$ due to (13) and (14).

For a bounded-domain quantizer $\mathbf{q}: \mathcal{D} \rightarrow \mathbb{R}^n$, we refer to

$$r(\mathbf{q}) \triangleq \max \{\delta : \mathcal{B}(\delta) \subseteq \mathcal{D}\} \quad (17)$$

as the *dynamic range* of \mathbf{q} , to

$$d(\mathbf{q}) \triangleq \min \{d : \forall \mathbf{x} \in \mathcal{D}, \|\mathbf{x} - \mathbf{q}(\mathbf{x})\| \leq d\} \quad (18)$$

as its *covering radius*, and to

$$\rho(\mathbf{q}) \triangleq |\text{Im}(\mathbf{q})|^{1/n} \frac{d(\mathbf{q})}{r(\mathbf{q})} \quad (19)$$

as its *covering efficiency*.³ A scalar uniform quantizer \mathbf{q}_u has domain $[-r(\mathbf{q}_u), r(\mathbf{q}_u)]^n$ and covering efficiency \sqrt{n} . This is wasteful: the classical result of Rogers [49, Th. 3] implies that there exists a sequence of n -dimensional quantizers \mathbf{q}_n with $\rho(\mathbf{q}_n) \rightarrow 1$ as $n \rightarrow \infty$, while definition (19) implies that $\rho(\mathbf{q}) \geq 1$ for any quantizer \mathbf{q} .

B. DQ-GD: convergence analysis and simulation results

Unquantized gradient descent with the optimal stepsize given by

$$\eta = \eta_{\text{GD}} \triangleq \frac{2}{L + \mu} \quad (20)$$

achieves contraction factor

$$\sigma_{\text{GD}} \triangleq \frac{\kappa - 1}{\kappa + 1} \quad (21)$$

over \mathcal{F}_n [48, Th. 1.4], [47, Th. 2.1.15] (Lemma B.2). The following result provides a convergence guarantee for DQ-GD.

Theorem III.1 (Convergence of DQ-GD). *Fix a dimension- n , rate- R quantizer \mathbf{q} with dynamic range 1 and covering efficiency ρ_n . Then, Algorithm 1 with stepsize (20) and dynamic ranges $r_0 = LD$,*

$$r_{t+1} = \sigma_{\text{GD}}^{t+1} LD + r_t \rho_n 2^{-R}, \quad t = 1, 2, \dots \quad (22)$$

in the definition of \mathbf{q}_t (7) achieves the following contraction factor over \mathcal{F}_n (1):

$$\sigma_{\text{DQ-GD}}(n, R) \leq \max \{\sigma_{\text{GD}}, \rho_n 2^{-R}\}. \quad (23)$$

Proof sketch. The path of DQ-GD and that of GD are related as (see Fig. 2, Lemma B.1)

$$\hat{\mathbf{x}}_t = \mathbf{x}_t - \eta \mathbf{e}_{t-1} \quad (24)$$

Comparing (24) and Line 4 in Algorithm 1, we see that $\mathbf{z}_t = \hat{\mathbf{x}}_t$, i.e., DQ-GD computes the gradient at the unquantized trajectory $\{\mathbf{x}_t\}$. The convergence guarantee of GD [48, Th. 1.4], [47, Theorem 2.1.15] (Lemma B.2) controls the difference between the first term in the recursion (24) and the optimizer \mathbf{x}_f^* . To bound the second term in (24), we observe using (19) that for any $r_t > 0$ in (7),

$$\max_{\mathbf{u} \in \mathcal{B}(r_t)} \|\mathbf{q}_t(\mathbf{u}) - \mathbf{u}\| = r_t \max_{\mathbf{u} \in \mathcal{B}(1)} \|\mathbf{q}(\mathbf{u}) - \mathbf{u}\| \quad (25)$$

$$= r_t \rho_n 2^{-R}, \quad (26)$$

i.e. quantizer \mathbf{q}_t used at iteration t has dynamic range r_t and covering radius (26). To complete the proof, we show by induction that with r_t in (22), the input \mathbf{u}_t to the quantizer \mathbf{q}_t generated by Algorithm 1 always lies within $\mathcal{B}(r_t)$. Since recurrence relation (22) represents a geometric sequence, (26) implies that the quantization error decays exponentially fast.

³Covering efficiency introduced in (19) extends the notion of covering efficiency of an infinite lattice [54], which measures how well that lattice covers the whole space, to bounded-domain quantizers.

The stepsize (20) is optimal both for GD [47, Theorem 2.1.15] and for DQ-GD. See Appendix B-A for details. \square

The bound in (23) exhibits a phase-transition behavior: at any $R \geq \log_2 \frac{\rho_n}{\sigma_{\text{GD}}}$, achieving the contraction factor of unquantized GD is possible, while at any $R < \log_2 \frac{\rho_n}{\sigma_{\text{GD}}}$, the achievable contraction factor is only $\rho_n 2^{-R} = \frac{d(q)}{r(q)}$. The algorithm converges linearly as long as $\rho_n 2^{-R} < 1$.

A common approach to quantizing descent algorithms [14]–[18], [50] we refer to as *naive quantization* has the worker directly quantize the gradient of its current iterate. Applied to GD, it leads to the Naively Quantized Gradient Descent (NQ-GD) with the quantizer input (cf. (9))

$$\mathbf{u}_t = \nabla f(\hat{\mathbf{x}}_t). \quad (27)$$

In Theorem V.1 in Section V below, we show that

$$\sigma_{\text{NQ-GD}}(n, R) \leq \sigma_{\text{GD}} + \frac{2\kappa}{\kappa+1} \rho_n 2^{-R}. \quad (28)$$

which is strictly greater than (23). In Fig. 3, we numerically compare the contraction factor of DQ-GD (Algorithm 1), the NQ-GD, and the unquantized GD (8) on least-squares problems

$$f(\mathbf{x}) = \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 \quad (29)$$

where $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, with $m \geq n$. We use the uniform scalar quantizer for the ease of implementation and take as a consequence a space-filling loss of \sqrt{n} .

We observe that DQ-GD has a significantly faster contraction factor than NQ-GD, and that the empirical results closely track our analytical convergence bounds (23) and (28). The contraction factor of unquantized GD serves as a lower bound to both quantized algorithms.

C. DQ-AGD: convergence analysis

Unquantized accelerated gradient descent with stepsize

$$\eta = \eta_{\text{AGD}} \triangleq \frac{1}{L} \quad (30)$$

and momentum coefficient

$$\gamma = \gamma_{\text{AGD}} \triangleq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \quad (31)$$

achieves contraction factor

$$\sigma_{\text{AGD}} \triangleq \sqrt{1 - \frac{1}{\sqrt{\kappa}}} \quad (32)$$

over \mathcal{F}_n (1) [43, Th. 3.18] (Lemma B.5), which improves the contraction factor of gradient descent $\sigma_{\text{GD}} = 1 - \frac{1}{\kappa} + O\left(\frac{1}{\kappa^2}\right)$ (21) to $\sigma_{\text{AGD}} = 1 - \frac{1}{2\sqrt{\kappa}} + O\left(\frac{1}{\kappa}\right)$, a significant improvement if κ is large and optimal order-wise (the converse to the optimal contraction factor expands as $1 - \frac{4}{\sqrt{\kappa}} + O\left(\frac{1}{\kappa}\right)$ [47] (Lemma C.2) and is attained in \mathcal{F}_n^2 by the heavy ball method [48] (Lemma B.8)).

Denote for brevity the constant

$$\lambda \triangleq (1 + \gamma_{\text{AGD}} + \gamma_{\text{AGD}} \sigma_{\text{AGD}}^{-1}) \sqrt{\kappa + 1}. \quad (33)$$

The following result extends (32) to DQ-AGD.

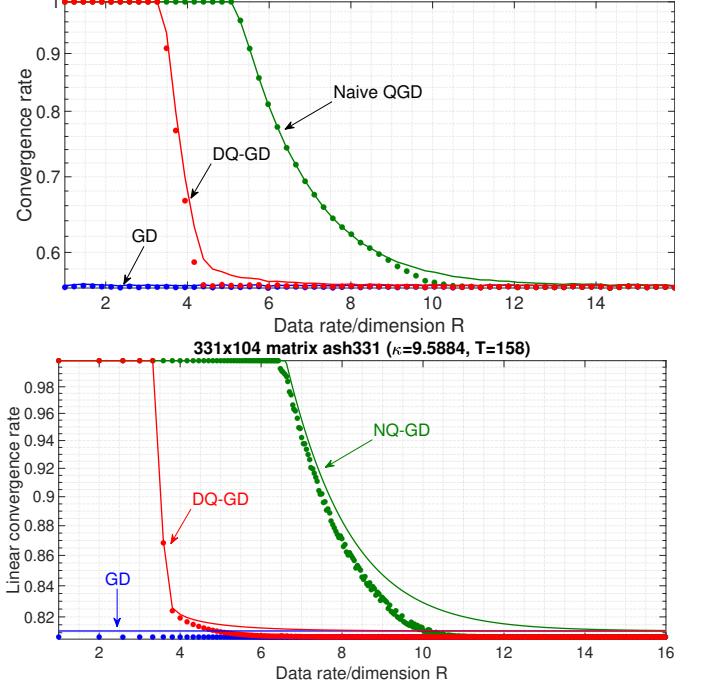


Fig. 3: Empirical contraction factors (as circles) and corresponding upper bounds (21), (23), and (28) (as lines). The real-world least-squares matrix ash331 is extracted from the online repository SuiteSparse [55]. We employ a uniform scalar quantizer. For smaller values of the data rate R , quantized GD may not even converge as $\sqrt{n}2^{-R} \geq 1$. In that case, we clip off the contraction factor at 1. For each per-dimension quantization rate $R \geq 1$, we generate 500 instances of the vector \mathbf{y} and $\hat{\mathbf{x}}_0$ with i.i.d. standard normal entries. In the case of Gaussian ensemble, we also generate 500 matrices \mathbf{A} 's with i.i.d. standard normal entries, one for each \mathbf{y} , and rescale the spectrum of \mathbf{A} so that it has a prescribed condition number κ . We run the iterative algorithms for as many iterations T as possible until reaching the machine's floating point precision.

Theorem III.2 (Convergence of DQ-AGD). *Fix a dimension- n , rate- R quantizer q with dynamic range 1 and covering efficiency ρ_n . Then, Algorithm 2 with stepsize (30), momentum coefficient (31), and dynamic ranges $r_{-2} = r_{-1} = 0$,*

$$r_t = \sigma_{\text{AGD}}^t L D \lambda + (r_{t-1} + \gamma_{\text{AGD}}(r_{t-1} + r_{t-2})) \rho_n 2^{-R}, \quad (34)$$

t = 1, 2, … in the definition of q_t (7) achieves the following contraction factor over \mathcal{F}_n (1):

$$\sigma_{\text{DQ-AGD}}(n, R) \leq \max \{ \sigma_{\text{AGD}}, \rho_n 2^{-R} \phi(n, R, \gamma_{\text{AGD}}) \} \quad (35)$$

where

$$\phi(n, R, \gamma) \triangleq \frac{1}{2}(1 + \gamma) + \frac{1}{2} \sqrt{(1 + \gamma)^2 + \frac{4\gamma}{\rho_n 2^{-R}}}. \quad (36)$$

Proof sketch. The proof follows the roadmap of the proof of Theorem III.1 with the following complication. Where in Algorithm 1 the quantizer input depends on the previous quantization error e_{t-1} , the quantizer input in Algorithm 2 depends on the past two quantization errors e_{t-1} and e_{t-2} (Line 5). The resulting recursion (34) is a second-order linear

non-homogeneous recurrence relation, which unlike (22) does not simply represent a geometric sequence. The characteristic polynomial of (34) is

$$p(r) \triangleq r^2 - r\rho_n 2^{-R}(1 + \gamma_{\text{AGD}}) - \rho_n 2^{-R}\gamma_{\text{AGD}}, \quad (37)$$

and $\rho_n 2^{-R}\phi_{\text{DQ-AGD}}(n, R)$ in (35) is its positive, larger-magnitude root. This implies that the quantization error decays with the contraction factor in the right side of (35). See Appendix B-B for details. \square

Define the functions

$$R_1(n, \gamma) \triangleq \log_2(1 + 2\gamma) + \log_2 \rho_n \quad (38)$$

$$R_2(n, \sigma, \gamma) \triangleq \log_2 \frac{(1 + \gamma)\sigma + \gamma}{\sigma^2} + \log_2 \rho_n \quad (39)$$

The achievability bound (35) exhibits two phase transitions. The first one is at $\rho_n 2^{-R}\phi(n, R, \gamma_{\text{AGD}}) < 1$, which is equivalent to $p(1) > 0$: if

$$R > R_1(n, \gamma_{\text{AGD}}) \text{ bits / dimension,} \quad (40)$$

then DQ-AGD enjoys linear convergence. The second one is at $\rho_n 2^{-R}\phi(n, R, \gamma_{\text{AGD}}) \leq \sigma_{\text{AGD}}$, which is equivalent to $p(\sigma_{\text{AGD}}) \geq 0$: if

$$R \geq R_2(n, \sigma_{\text{AGD}}, \gamma_{\text{AGD}}) \text{ bits / dimension,} \quad (41)$$

then there is no loss in the long-term convergence behavior of the DQ-AGD compared to AGD.

Curiously, $R_1(n, 0)$ and $R_2(n, \sigma_{\text{GD}}, 0)$ express the two phase transitions of DQ-DG that were determined in Section III-B.

D. DQ-HB: convergence analysis and numerical comparison

Unquantized heavy ball method with stepsize

$$\eta = \eta_{\text{HB}} \triangleq \left(\frac{2}{\sqrt{L} + \sqrt{\mu}} \right)^2 \quad (42)$$

and momentum coefficient

$$\gamma = \gamma_{\text{HB}} \triangleq \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^2 \quad (43)$$

achieves contraction factor

$$\sigma_{\text{HB}} \triangleq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \quad (44)$$

over \mathcal{F}_n^2 (1) [48] (Lemma B.8), which is optimal among all gradient methods [47, Th. 2.1.13] (Lemma C.2).

The following convergence analysis of DQ-HB applies to smooth and strongly convex functions that are in addition twice continuously differentiable.

Theorem III.3 (Convergence of DQ-HB). *Fix a dimension n , rate- R quantizer q with dynamic range 1 and covering efficiency ρ_n . Then, there exists a constant $\alpha > 0$ such that Algorithm 3 with stepsize (42), momentum coefficient (43) and dynamic ranges $r_{-1} = r_{-2} = 0$,*

$$r_t = \sigma_{\text{HB}}^t t^\alpha e^\alpha \sqrt{2} LD + (r_{t-1} + \gamma_{\text{HB}}(r_{t-1} + r_{t-2})) \rho_n 2^{-R}, \quad (45)$$

$t = 1, 2, \dots$ achieves the following contraction factor over \mathcal{F}_n^2 :

$$\sigma_{\text{DQ-HB}}(n, R) \leq \max \{ \sigma_{\text{HB}}, \rho_n 2^{-R} \phi(n, R, \gamma_{\text{HB}}) \}, \quad (46)$$

where $\phi(n, R, \gamma)$ is defined in (36).

Proof sketch. The proof is similar to the proof of Theorem III.2. The recurrence relation (45) differs from (34) in only the presence of the subexponential factor t^α , which does not matter when we take $t \rightarrow \infty$ to obtain (46). This factor arises from our nonasymptotic sharpening of Polyak's convergence result for the unquantized HB (Lemma B.8). See Appendix B-C for details. \square

DQ-HB exhibits two phase transitions, a behavior similar to DQ-HB and DQ-AGD. The two thresholds are given by R_1 (38) and R_2 (39) evaluated with $\sigma = \sigma_{\text{HB}}$ and $\gamma = \gamma_{\text{HB}}$.

Plugging the parameters γ and σ into (39), we can infer that DQ-HB always has the largest R_2 (39) for any condition number $\kappa \geq 1$ among the three DQ schemes. On the other hand, whether R_2 of DQ-AGD is smaller than that of DQ-GD depends on whether κ is smaller than a threshold that is roughly 2.18. For the unquantized algorithms, contraction factor σ_{HB} of HB is always the smallest among the three for any condition number $\kappa \geq 1$. On the other hand, whether σ_{AGD} of AGD is smaller than σ_{GD} of GD depends on whether κ is greater than a threshold that is roughly 11.83. For the differentially quantized algorithms, DQ-GD actually has the best convergence behavior in the transient regime where $R > \log_2 \rho_n$ so that GD converges linearly and R is small enough so that DQ-HB does not yet outperform GD, i.e., $\rho_n 2^{-R} \phi(n, R, \gamma_{\text{HB}}) > \sigma_{\text{GD}}$. This is because DQ-GD is the first among the three DQ algorithms to pass R_1 (38) above which it has linear convergence.

In Fig. 4, we compare the performance of the differentially quantized algorithms DQ-HB (Algorithm 3), DQ-AGD (Algorithm 2) and DQ-GD (Algorithm 1) on least-squares problems (29). We set $\alpha = 0$ for Algorithm 3, and DQ-HB still converges empirically for this parameter. We also record the performance of the corresponding unquantized gradient methods HB, AGD and GD. The curves exhibit the two phase transitions and comparative performance as discussed above. The level lines that the contraction factors of these DQ schemes rest on for $R \geq R_2$ are almost the same as the corresponding linear convergence rates σ of their unquantized counterparts. We observe that there is a gap between the worst-case linear convergence rate σ_{AGD} that we design DQ-AGD to follow and the empirical convergence rate of AGD. This is because AGD applies for functions that are not necessarily twice continuously differentiable, and the least-squares problems (29) happen not to be a worst-case problem class for AGD.

IV. CONVERSES

A. Quantized gradient descent algorithms

In this section, we characterize the optimal contraction factor achievable within class \mathcal{A}_{GD} of quantized gradient descent algorithms, formally defined next.

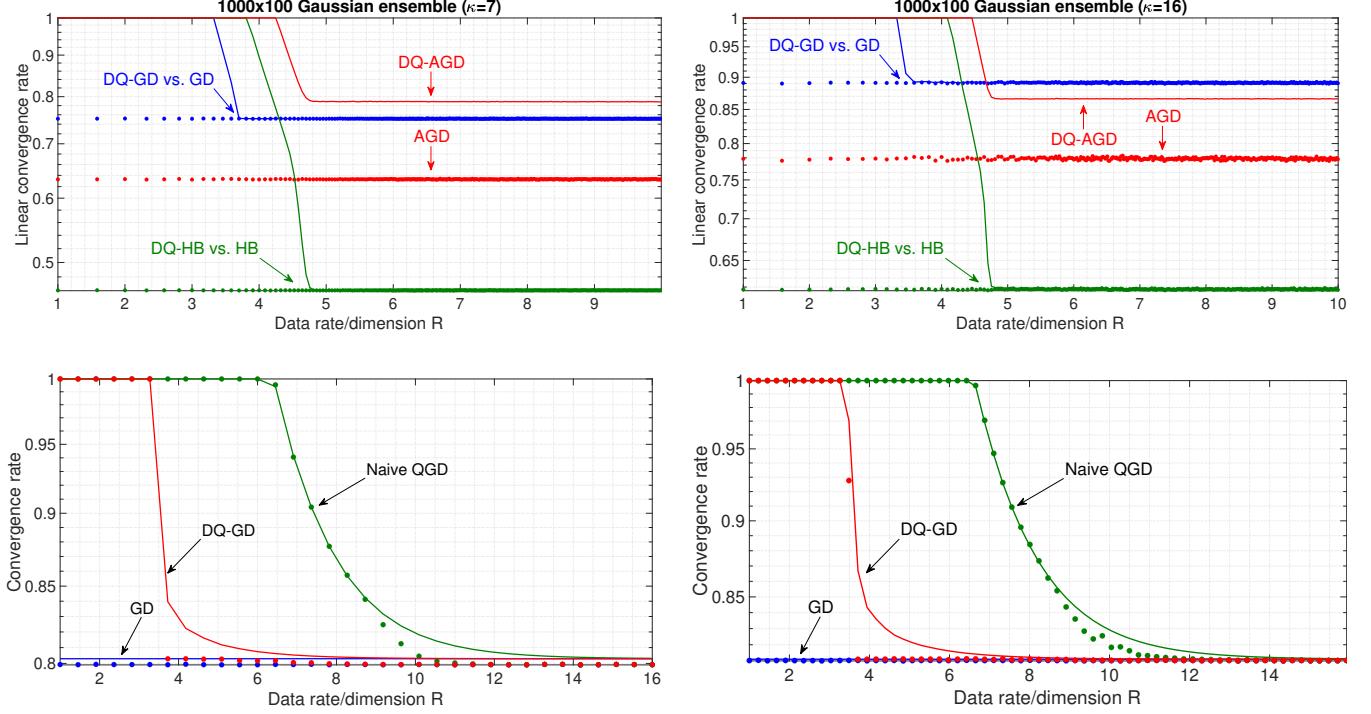


Fig. 4: Empirical contraction factors of various DQ algorithms (plotted as lines) and their corresponding unquantized counterparts (plotted as circles). Same setup as in Fig. 3.

Definition IV.1 (Class \mathcal{A}_{GD} of quantized gradient descent algorithms). A *quantized gradient descent* algorithm $A \in \mathcal{A}_{\text{GD}}$ consists of a central server and an end worker. The algorithm is initialized with a collection of quantizers q indexed by rate R such that $d(q) \rightarrow 0$ (17) as $R \rightarrow \infty$ and a sequence of dynamic ranges $\{r_t\}_{t=1}^\infty$. The worker has access to the function f . At each iteration t , the server first sends \hat{x}_t to the worker noiselessly, starting from some $\hat{x}_0 \in \mathbb{R}^n$. The worker then determines its gradient-access point z_t and its quantizer input u_t under the structural constraints

$$z_t \in \hat{x}_t + \text{span}\{e_0, \dots, e_{t-1}\} \quad (47)$$

$$u_t \in \nabla f(z_t) + \text{span}\{e_0, \dots, e_{t-1}\}, \quad (48)$$

where $e_i \triangleq q_i - u_i, i = 0, \dots, t-1$ are the past quantization errors before iteration t , and $+$ denotes Minkowski's sum. Upon receiving $q_t = q_t(u_t)$ (7) from the worker, the server performs the update

$$\hat{x}_{t+1} = \hat{x}_t - \eta q_t \quad (49)$$

with a fixed stepsize $\eta > 0$.

Due to conditions (47) and (48), if there is no quantization error at each iteration (i.e., if $R = \infty$), then any quantized algorithm in \mathcal{A}_{GD} reduces to the unquantized gradient descent. Both DQ-GD and NQ-GD fall in the class \mathcal{A}_{GD} .

Theorem IV.1 (Converse within class \mathcal{A}_{GD}). *The contraction factor achievable over functions $f \in \mathcal{F}_n$ within class \mathcal{A}_{GD} of algorithms satisfies*

$$\inf_{A \in \mathcal{A}_{\text{GD}}} \sigma_A(n, R) \geq \max\{\sigma_{\text{GD}}, 2^{-R}\} \quad (50)$$

Proof sketch. We fix an $A \in \mathcal{A}_{\text{GD}}$, and we lower-bound the contraction factor it achieves at rate R in two different ways. On one hand, we show that A cannot converge faster than the unquantized GD. Then, we use an argument similar to [56] to craft a worst-case problem instance $g \in \mathcal{F}_n$ for which the iterates of the unquantized GD satisfy $\|\hat{x}_{t+1} - \hat{x}_t^*\| = \sigma_{\text{GD}} \|\hat{x}_t - \hat{x}_t^*\|$, which ensures that $\inf_{A \in \mathcal{A}_{\text{GD}}} \sigma_A(n, R) \geq \sigma_{\text{GD}}$. On the other hand, we notice that if A is applied at dimension n and rate R , then the set $\mathcal{S}_A \subseteq \mathbb{R}^n$ of all possible locations of the iterate \hat{x}_T after T iterations of A has cardinality at most 2^{nRT} , and we apply a volume-division argument to claim that $\inf_{A \in \mathcal{A}_{\text{GD}}} \sigma_A(n, R) \geq 2^R$. See Appendix C-B for details. \square

Applying Theorem III.1 with Rogers-optimal quantizers with $\rho_n \rightarrow 1$ [49, Th. 3] and juxtaposing with Theorem IV.1, we characterize the optimal contraction factor achievable by quantized gradient descent in the limit of large problem dimension as

$$\lim_{n \rightarrow \infty} \inf_{A \in \mathcal{A}_{\text{GD}}} \sigma_A(n, R) = \max\{\sigma_{\text{GD}}, 2^{-R}\}. \quad (51)$$

In other words, DQ-DG achieves the best possible contraction factor within \mathcal{A}_{GD} , in the limit of large problem dimension. This is rather remarkable: it means not only that DQ-DG compensates previous quantization errors optimally so that no rate is wasted, but that our convergence analysis in Theorem III.1 is tight enough to capture this optimality. Furthermore, notice that the right side of (51) is < 1 at any $R > 0$. This means that at any $R > 0$ however small, DQ-DG with Rogers-optimal quantizers converges linearly at a large

enough problem dimension n , i.e. the first phase transition (38) disappears.

B. Quantized gradient methods

All quantized algorithms considered in this paper fall in the following class.

Definition IV.2 (Class \mathcal{A}_{GM} of quantized gradient methods). A *quantized gradient method* $A \in \mathcal{A}_{\text{GM}}$ follows Definition IV.1 with (49) relaxed to

$$\hat{\mathbf{x}}_{t+1} \in \hat{\mathbf{x}}_0 + \text{span}\{\mathbf{q}_0, \dots, \mathbf{q}_t\}. \quad (52)$$

In the absence of rate constraints, there are no quantization errors, i.e. $e_t = \mathbf{0}$ for all t , and the class of quantized gradient methods reduces to the class of unquantized gradient methods satisfying

$$\mathbf{x}_{t+1} \in \mathbf{x}_0 + \text{span}\{\nabla f(\mathbf{x}_0), \dots, \nabla f(\mathbf{x}_t)\}. \quad (53)$$

To present our converse result for \mathcal{A}_{GM} , we consider functions f defined on the square-summable Hilbert space⁴

$$\mathbb{L}_2 \triangleq \{\mathbf{x} = [\mathbf{x}(1), \mathbf{x}(2), \dots] : \sum_{i=1}^{\infty} \mathbf{x}(i)^2 < \infty\}. \quad (54)$$

We say that continuously differentiable function $f: \mathbb{L}_2 \mapsto \mathbb{R}$ is in class \mathcal{F}_{∞} if it is L -smooth, μ -strongly convex and its minimizer is bounded, i.e., f satisfies i)-iii) in Section III-A.

To quantize an infinitely long vector $\mathbf{u} \in \mathbb{L}_2$ to $\mathbf{q} \in \mathbb{L}_2$, we fix a free parameter $n \in \mathbb{N}$, apply a rate- R quantizer q in \mathbb{R}^n (6) to the first n coordinates of \mathbf{u} , and set the remaining coordinates to 0, i.e.,

$$\begin{cases} [\mathbf{q}(1), \dots, \mathbf{q}(n)] = q([\mathbf{u}(1), \dots, \mathbf{u}(n)]) \\ \mathbf{q}(i) = 0 \quad \forall i > n, \end{cases} \quad (55)$$

where $\mathbf{u}(i)$ denotes i -th coordinate of vector $\mathbf{u} \in \mathbb{L}_2$.

Although only n coordinates $\mathbf{u} \in \ell_2$ are quantized, we can still control the overall quantization error since in \mathbb{L}_2 ,

$$\sum_{i>n} \mathbf{u}(i)^2 = o_n(1) \quad (56)$$

due to the Cauchy convergence criterion. Here $o_n(1)$ denotes a function that vanishes as $n \rightarrow \infty$. Thus, (26) continues to hold for quantization in \mathbb{L}_2 with ρ_n replaced by $\rho_n + o_n(1)$. It follows that the achievability bounds in Theorems III.2 and III.3 with with ρ_n replaced by $\rho_n + o_n(1)$ apply to functions $f \in \mathcal{F}_{\infty}$.

Contraction factor $\sigma_A(n, R)$ over \mathcal{F}_{∞} is defined in the same way as that over \mathcal{F}_n (1) except that n is now a parameter of the employed quantizer (like ρ_n) rather than the dimension of the problem, and the total number of bits sent per iteration is nR , where R is the quantizer's rate (6).

⁴We do so to take advantage of the sharpest converse in the literature on the convergence of unquantized gradient methods (53) [47] (Lemma C.2), which applies to functions on \mathbb{L}_2 . Convergence lower bounds for smooth and strongly convex functions on \mathbb{R}^n (rather than \mathbb{L}_2) are also known [57]. However, [57] considers only quadratic functions as objectives, and the considered class of iterative algorithms is more restrictive than (53) in that the next iterate \mathbf{x}_{t+1} depends on the past p terms $\mathbf{x}_t, \dots, \mathbf{x}_{t-p+1}$ for some fixed $p \in \mathbb{N}$.

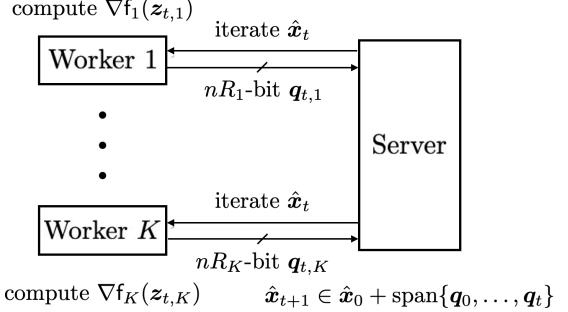


Fig. 5: K -worker quantized gradient method. At each iteration t , the server broadcasts the current iterate $\hat{\mathbf{x}}_t$. Worker k computes the gradient at some point $\mathbf{z}_{t,k}$ that is a function of (but not necessarily equal to) $\hat{\mathbf{x}}_t$. Then, worker k forms a descent direction $\mathbf{q}_{t,k}$ and pushes it back to the server under an nR_k -bit rate constraint.

Theorem IV.2 (Converse within class \mathcal{A}_{GM}). *The contraction factor achievable over functions $f \in \mathcal{F}_{\infty}$ within class \mathcal{A}_{GM} of algorithms satisfies*

$$\inf_{A \in \mathcal{A}_{\text{GM}}} \sigma_A(n, R) \geq \max\{\sigma_{\text{HB}}, 2^{-R}\} \quad (57)$$

where σ_{HB} is given in (44).

Proof sketch. The proof is similar to that of Theorem IV.1: we apply a volume-division argument to recover the 2^{-R} in the right side of (57), and we apply a known result on unquantized gradient methods that states that the best contraction factor achievable over \mathcal{F}_{∞} is that of the heavy ball method, σ_{HB} [47] (Lemma C.2). See Appendix C-C. \square

Together, Theorems IV.2 and III.3 imply that for any $R \geq R_2(\infty, \sigma_{\text{HB}}, \gamma_{\text{HB}})$, DQ-HB attains the optimal contraction factor within \mathcal{A}_{GM} (under the additional assumption that $f \in \mathcal{F}_{\infty}$ is twice continuously differentiable).

V. MULTI-WORKER GRADIENT METHODS

A. Problem setup

In empirical risk minimization [58], [59], the sample average of the loss function on the data points

$$f(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K f_k(\mathbf{x}) \quad (58)$$

arises as a substitute for the expected loss on the true data distribution that is often unknown. In multi-worker distributed empirical risk minimization, each worker has access to only one of the summands in (58), and they communicate to the parameter server under rate constraints. See Figure 5.

B. Converges

Definitions IV.1 and IV.2 extend naturally to the K -worker setting. Converges in Theorems IV.1 and IV.2 extend verbatim to the multi-worker setting where the workers' rates satisfy the sum-rate constraint

$$\sum_{k=1}^K R_k \leq R. \quad (59)$$

C. Differential quantization

Differential quantization does not apply to K -worker quantized gradient methods since each worker does not know the local quantization errors stored by the others, and thus cannot guide the descent trajectory back to the unquantized path. Thus, whether (50) and (57) are attainable in the multiworker setting, and how each worker should optimally compensate its own past quantization errors, remain open problems.

D. Naively Quantized Gradient Descent

The Naively Quantized Gradient Descent applies a common method of quantizing distributed gradient algorithms [14]–[18], [50] in which each worker quantizes the gradient of the current iterate, to GD. It is summarized as Algorithm 4.

Algorithm 4: K -worker NQ-GD

```

1 for  $t = 0, 1, 2, \dots$  do
2   for  $k = 1$  to  $K$  do
3     Worker  $k$ :
4        $\underline{q}_{t,k} = q_{t,k}(\nabla f_k(\hat{x}_t))$ 
5     Server:  $\hat{x}_{t+1} \leftarrow \hat{x}_t - \frac{\eta}{K} \sum_{k=1}^K \underline{q}_{t,k}$ 
6   end
7 end
```

Our convergence result for NQ-GD holds under the following assumptions. We assume that continuously differentiable summands f_k in (58) are (i) L_k -smooth and (ii) μ_k -strongly convex, and we continue to assume that (iii) the optimizer of f is bounded as in (15). Note that f is L -smooth and μ -strongly convex with

$$L \triangleq \frac{1}{K} \sum_{k=1}^K L_k \quad \text{and} \quad \mu \triangleq \frac{1}{K} \sum_{k=1}^K \mu_k. \quad (60)$$

Further, we focus on the *interpolation setting* [60]–[62] that assumes (iv)

$$\mathbf{x}_f^* = \mathbf{x}_{f_k}^* \quad \forall k = 1, \dots, K, \quad (61)$$

where

$$\mathbf{x}_{f_k}^* \triangleq \arg \min_{\mathbf{x} \in \mathbb{R}^n} f_k(\mathbf{x}). \quad (62)$$

The interpolation setting is motivated by the observation that almost all local minima are also global in an over-parametrized neural network with a very large data dimension n [62]. We denote by \mathcal{G}_n the class of functions f (58) that satisfy the assumptions (i)-(iv).

The minimum contraction factor achievable by K -worker NQ-GD under the sum rate constraint (59) is given by

$$\begin{aligned} & \sigma_{\text{NQ-GD}}(n, R) \\ & \triangleq \inf_{\sum_{k=1}^K R_k \leq R} \sup_{f \in \mathcal{G}_n} \limsup_{T \rightarrow \infty} \|\hat{x}_T(\{R_k\}_{k=1}^K) - \mathbf{x}_f^*\|^{\frac{1}{T}}, \end{aligned} \quad (63)$$

where $\hat{x}_0(\{R_k\}_{k=1}^K), \hat{x}_1(\{R_k\}_{k=1}^K), \hat{x}_2(\{R_k\}_{k=1}^K), \dots$ is the sequence of iterates generated by NQ-GD (Algorithm 4) in response to $f \in \mathcal{G}_n$ when the k -th worker operates at R_k bits per problem dimension, $k = 1, \dots, K$.

Theorem V.1 (Convergence of K -worker NQ-GD). *Fix a dimension- n , rate- R_k quantizer q_k with dynamic range 1 and covering efficiency ρ_n . Set up the quantizer to be used by worker k at iteration t as*

$$q_{t,k}(\cdot) = r_{t,k} q_k(\cdot / r_{t,k}), \quad (64)$$

where the dynamic ranges are given by

$$r_{t,k} = \left(\sigma_{\text{GD}} + \frac{\eta_{\text{GD}} \rho_n}{K} \sum_{k=1}^K \min\{\nu, L_k\} \right)^t L_k D, \quad (65)$$

and the optimum rate allocation is given by the waterfilling solution

$$R_k = |\log_2(L_k/\nu)|_+ \text{ bits}, \quad (66)$$

where ν is the water level found from the sum rate constraint

$$\sum_{k=1}^K |\log_2(L_k/\nu)|_+ = R, \quad (67)$$

and $|\cdot|_+ \triangleq \max\{0, \cdot\}$. Then, Algorithm 4 with stepsize (20) achieves the following contraction factor over \mathcal{G}_n (63):

$$\sigma_{\text{NQ-GD}}(n, R) \leq \sigma_{\text{GD}} + \frac{\eta_{\text{GD}} \rho_n}{K} \sum_{k=1}^K \min\{\nu, L_k\}. \quad (68)$$

Proof. Appendix D. □

According to (66), higher rates are allocated to users whose function gradients have higher Lipschitz constants, and if the Lipschitz constant is low enough in comparison to others no rate would be allocated at all. In the special case $L_k \equiv L$, (68) reduces to

$$\sigma_{\text{NQ-GD}}(n, R) \leq \sigma_{\text{GD}} + \frac{2\kappa}{\kappa+1} \frac{\rho_n}{2^{R/K}}. \quad (69)$$

The bound in Theorem V.1 approaches the converse only in the limit of large R .

VI. CONCLUSION

This paper formalizes the problem of finding the optimal contraction factor achievable within a class of rate-constrained iterative optimization algorithms ((1), Definitions IV.1 and IV.2). We show information-theoretic converses to that fundamental limit (Theorem IV.1, Theorem IV.2).

We introduce the principle of *differential quantization* that posits that the quantizer's input shall be constructed in such a way as to guide the quantized algorithm's trajectory towards the unquantized trajectory. Applied to gradient descent (Algorithm 1), differential quantization leads to the contraction factor that is optimal within the class of quantized gradient descent algorithms (51)). Thus, differential quantization leverages the memory of past quantized inputs in an optimal way, removing the impact of past quantization errors.

Beyond gradient descent, we apply differential quantization to gradient methods with momentum - the accelerated gradient descent (Algorithm 2) and the heavy ball method (Algorithm 3). In all three cases, differentially quantized algorithms attain the contraction factor of their unquantized counterparts

as long as the data rate exceeds the corresponding threshold R_2 (39).

Incidentally, in the course of the analysis, we provide a sharper bound on the convergence of the unquantized HB algorithm than available in the literature (Lemma B.8). We also provide a slightly more general worst-case problem instance for the unquantized GD than available (Lemma C.1).

Quantizers employed at each step have the same geometry (covering efficiency, (19)) but different resolution (covering radius, (18)). The resolution is controlled by scaling the quantizer's dynamic range (7). To attain the contraction factors in Theorems III.1, III.2, and III.3, the dynamic range is set to follow a recursion ((22), (34), (45)). That recursion shrinks the dynamic range at the fastest possible rate that still guarantees that the quantizer's input at each iteration falls within its dynamic range. This maximizes the usefulness of the bits exchanged at each iteration. While that recursion for DQ-GD (22) is simply a geometric sequence, those for DQ-AGD (34) and DQ-HB (45) are second-order linear non-homogeneous recurrence relations.

While DQ-GD attains the optimal contraction factor among quantized gradient descent algorithms (Definition IV.1) and DQ-HB attains the optimal contraction factor among all gradient methods, even unquantized, if $R \geq R_2(n, \sigma_{\text{HB}}, \gamma_{\text{HB}})$ (39), it remains an open problem whether the contraction factor of 2^{-R} dictated by the converse (Theorem IV.2) is achievable in the regime $R_2(n, \sigma_{\text{GD}}, 0) < R < R_2(n, \sigma_{\text{HB}}, \gamma_{\text{HB}})$ in the class of quantized gradient methods (Definition IV.2).

For multi-worker gradient descent, we provide a convergence result on naive quantization, in which the workers directly quantize their gradients, and show a waterfilling solution to optimize the allocation of data rates among the workers under the sum rate constraint (Theorem V.1). That result approaches the converse (Theorem IV.1) only in the limit $R \rightarrow \infty$, leaving open a tighter characterization of the optimum convergence factor in that scenario. Differential quantization does not directly apply to multi-worker optimization since the workers cannot compute the unquantized path without the knowledge of the local quantization errors stored by the others. We leave as future work the question of how they should optimally compensate their own quantization errors.

ACKNOWLEDGMENT

Helpful discussions with Dr. Himanshu Tyagi, Dr. Vincent Tan, and Dr. Victor Kozyakin are gratefully acknowledged.

APPENDIX A DQ-GD WITH VARYING STEPSIZE

APPENDIX B CONVERGENCE ANALYSES OF DQ ALGORITHMS

A. Proof of Theorem III.1

As mentioned in the proof sketch, relation (24) is key to showing Theorem III.1. The next lemma, which applies to the more general version of the DQ-DG algorithm shown in Appendix A, establishes (24).

Algorithm 5: DQ-GD with varying stepsize

```

1 Initialize  $\mathbf{e}_{-1} = \hat{\mathbf{x}}_0 = \mathbf{0}$ 
2 for  $t = 0, 1, 2, \dots$  do
3   Worker:
4      $\mathbf{z}_t = \hat{\mathbf{x}}_t + \eta_{t-1} \mathbf{e}_{t-1}$ 
5      $\mathbf{u}_t = \nabla f(\mathbf{z}_t) - (\eta_{t-1}/\eta_t) \mathbf{e}_{t-1}$ 
6      $\mathbf{q}_t = \mathbf{q}_t(\mathbf{u}_t)$ 
7      $\mathbf{e}_t = \mathbf{q}_t - \mathbf{u}_t$ 
8   Server:  $\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t - \eta_t \mathbf{q}_t$ 
9 end

```

Lemma B.1 (DQ-GD trajectory). *Consider descent trajectories $\{\hat{\mathbf{x}}_t\}$ of Algorithm 5 and $\{\mathbf{x}_t\}$ of unquantized GD (8) with the same sequence of stepsizes $\{\eta_t\}$ starting at the same location $\hat{\mathbf{x}}_0 = \mathbf{x}_0$. Then, at each iteration t ,*

$$\hat{\mathbf{x}}_t = \mathbf{x}_t - \eta_{t-1} \mathbf{e}_{t-1}. \quad (70)$$

Proof. We prove (70) via mathematical induction.

- Base case: (70) holds for $t = 0$ since the starting location is the same.
- Inductive step: Suppose (70) holds for iteration t . First, the induction hypothesis, the quantizer input at Line 5 and the quantizer output at Line 6 of Algorithm 5 together imply

$$\mathbf{u}_t = \nabla f(\mathbf{x}_t) - \frac{\eta_{t-1}}{\eta_t} \mathbf{e}_{t-1}. \quad (71)$$

(We define $0/0 \triangleq 0$ for the very first iteration when $\eta_{-1} = 0$.) We then have

$$\hat{\mathbf{x}}_{t+1} = \hat{\mathbf{x}}_t - \eta_t \mathbf{q}_t \quad (72)$$

$$= \hat{\mathbf{x}}_t - \eta_t (\mathbf{u}_t + \mathbf{e}_t) \quad (73)$$

$$= \hat{\mathbf{x}}_t - \eta_t \left(\nabla f(\mathbf{x}_t) - \frac{\eta_{t-1}}{\eta_t} \mathbf{e}_{t-1} \right) - \eta_t \mathbf{e}_t \quad (74)$$

$$= [\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)] - \eta_t \mathbf{e}_t \quad (75)$$

$$= \mathbf{x}_{t+1} - \eta_t \mathbf{e}_t, \quad (76)$$

where (75) is due to the induction hypothesis. \square

Relation (70) implies for the constant stepsize $\eta_t \equiv \eta$

$$\|\hat{\mathbf{x}}_t - \mathbf{x}_t^*\| \leq \|\mathbf{x}_t - \mathbf{x}_t^*\| + \eta \|\mathbf{e}_{t-1}\| \quad (77)$$

We use the contraction factor of unquantized GD to bound the first term of (77):

Lemma B.2 (Convergence of GD [47, Theorem 2.1.15]). *For any L -smooth and μ -strongly convex function f on \mathbb{R}^n , GD (8) with stepsize (20) satisfies*

$$\|\mathbf{x}_t - \mathbf{x}_t^*\| \leq \sigma_{\text{GD}}^t \|\mathbf{x}_0 - \mathbf{x}_t^*\|, \quad (78)$$

where σ_{GD} is defined in (21).

We use the following bound on quantization error to bound the second term in (77):

Lemma B.3 (DQ-GD quantization error). *Let $f \in \mathcal{F}_n$. Quantization errors $\{e_t\}$ in Algorithm 1 with stepsize (20) and dynamic ranges (22) satisfy*

$$\|e_t\| \leq r_t \rho_n 2^{-R} \quad (79)$$

$$\leq \max \{\sigma_{\text{GD}}, \rho_n 2^{-R}\}^t LD \sigma_{\text{GD}} b_t, \quad (80)$$

where

$$b_t \triangleq \begin{cases} \frac{\rho_n 2^{-R}}{|\sigma_{\text{GD}} - \rho_n 2^{-R}|} & \sigma_{\text{GD}} \neq \rho_n 2^{-R} \\ t+1 & \sigma_{\text{GD}} = \rho_n 2^{-R} \end{cases} \quad (81)$$

Proof. Once we show that quantizer inputs $\{u_t\}$ satisfy

$$u_t \in \mathcal{B}(r_t), \quad (82)$$

(79) will follow from (26).

We prove (82) via induction.

- Base case: (82) holds for $t = 0$ since

$$\|\nabla f(\hat{x}_0) - e_{-1}\| = \|\nabla f(\hat{x}_0)\| \quad (83)$$

$$\leq L \|\hat{x}_0 - x_f^*\| \quad (84)$$

$$\leq LD, \quad (85)$$

where (84) is due to $\nabla f(x_f^*) = \mathbf{0}$ and L -smoothness (13), and (85) is due to assumption (15).

- Inductive step: Suppose (82) holds for iteration t . Applying triangle inequality and (70) to the expression in Line 5 yields

$$\|u_{t+1}\| \leq \|\nabla f(x_{t+1})\| + \|e_t\|. \quad (86)$$

The first term is bounded as

$$\|\nabla f(x_{t+1})\| \leq L \|x_{t+1} - x_f^*\| \quad (87)$$

$$\leq \sigma_{\text{GD}}^{t+1} L \|x_0 - x_f^*\| \quad (88)$$

$$\leq \sigma_{\text{GD}}^{t+1} LD, \quad (89)$$

where (87) is due to $\nabla f(x_f^*) = \mathbf{0}$ and L -smoothness (13); (88) is due to (78); and (89) is due to assumption (15). Quantization error term $\|e_t\|$ in (86) is bounded by (26) due to the induction hypothesis. Plugging (89) and (26) into (86) gives

$$\|u_{t+1}\| \leq \sigma_{\text{GD}}^{t+1} LD + r_t \rho_n 2^{-R} \quad (90)$$

$$= r_{t+1}, \quad (91)$$

where (91) is due to the choice of the dynamic ranges (22).

This concludes the proof of (79). To establish (80), we unwrap recursion (22) as the geometric sum

$$r_t = LD \sum_{\tau=0}^t \sigma_{\text{GD}}^\tau (\rho_n 2^{-R})^{t-\tau} \quad (92)$$

$$= LD \cdot \begin{cases} \sigma_{\text{GD}}^t \frac{1 - [\rho_n 2^{-R} / \sigma_{\text{GD}}]^{t+1}}{1 - \rho_n 2^{-R} / \sigma_{\text{GD}}} & \sigma_{\text{GD}} \neq \rho_n 2^{-R} \\ t+1 & \sigma_{\text{GD}} = \rho_n 2^{-R} \end{cases} \quad (93)$$

$$\leq LD \cdot \begin{cases} \sigma_{\text{GD}}^t \left(1 - \frac{\rho_n 2^{-R}}{\sigma_{\text{GD}}}\right)^{-1} & \sigma_{\text{GD}} > \rho_n 2^{-R} \\ (\rho_n 2^{-R})^t \left(\frac{\rho_n 2^{-R}}{\sigma_{\text{GD}}} - 1\right)^{-1} & \sigma_{\text{GD}} < \rho_n 2^{-R} \\ t+1 & \sigma_{\text{GD}} = \rho_n 2^{-R} \end{cases} \quad (94)$$

where the bound for the case $\sigma_{\text{GD}} > \rho_n 2^{-R}$ is obtained by lower-bounding $[\rho_n 2^{-R} / \sigma_{\text{GD}}]^{t+1}$ by 0, and the bound for $\sigma_{\text{GD}} < \rho_n 2^{-R}$ is obtained by lower-bounding $[\sigma_{\text{GD}} / \rho_n 2^{-R}]^{t+1}$ by 0. \square

Putting together the results in Lemmas B.1, B.2, and B.3, we show the following nonasymptotic (in the iteration number) convergence result for the DQ-GD:

Theorem B.1 (Convergence of DQ-GD). *In the setting of Theorem III.1, the difference between the iterate and the optimizer at step t satisfies*

$$\|\hat{x}_t - x_f^*\| \leq \max \{\sigma_{\text{GD}}, \rho_n 2^{-R}\}^t [1 + \eta_{\text{GD}} L b_{t-1}] D. \quad (95)$$

Proof. Plugging (78) and (80) into (77), we obtain

$$\begin{aligned} \|\hat{x}_t - x_f^*\| &\leq \sigma_{\text{GD}}^t D \\ &\quad + \max \{\sigma_{\text{GD}}, \rho_n 2^{-R}\}^{t-1} \eta_{\text{GD}} L D \sigma_{\text{GD}} b_{t-1}, \end{aligned} \quad (96)$$

which leads to (95) by an elementary weakening. \square

Bound (23) in Theorem III.1 follows immediately by applying (95) to definition (1) of the contraction factor.

B. Proof of Theorem III.2

The proof follows steps similar to those in the proof of Theorem III.1. First, we prove that, as prescribed by the principle of differential quantization, DQ-AGD compensates quantization errors by directing the quantized trajectory back to the trajectory of AGD:

Lemma B.4 (DQ-AGD trajectory). *Iterate sequences $\{\hat{y}_t, \hat{x}_t\}$ of Algorithm 2 and $\{y_t, x_t\}$ of AGD (10) starting at the same location $(\hat{y}_0, \hat{x}_0) = (y_0, x_0)$ are related as, $\forall t = 0, 1, 2, \dots$,*

$$\hat{y}_t = y_t - \eta e_{t-1} \quad (97)$$

$$\hat{x}_t = x_t - \eta e_{t-1} - \eta \gamma (e_{t-1} - e_{t-2}). \quad (98)$$

Proof. We prove (97) and (98) by induction.

- Base case: (97) and (98) hold for $t = 0$ since both algorithms start at the same location.
- Inductive step: Line 9 yields

$$\hat{y}_{t+1} = \hat{x}_t - \eta q_t \quad (99)$$

$$= x_t - \eta [e_{t-1} + \gamma (e_{t-1} - e_{t-2})] \quad (100)$$

$$- \eta [\nabla f(x_t) - [e_{t-1} + \gamma (e_{t-1} - e_{t-2})]] + e_t]$$

$$= y_{t+1} - \eta e_t, \quad (101)$$

where (100) is due to induction hypothesis (98) and Lines 5 and 7, and (101) is due to (10). On the other hand, plugging (97) and (101) into Line 10 yields

$$\hat{x}_{t+1} = \hat{y}_{t+1} + \gamma (\hat{y}_{t+1} - \hat{y}_t) \quad (102)$$

$$= y_{t+1} + \gamma (y_{t+1} - y_t) - \eta [e_t + \gamma (e_t - e_{t-1})]$$

$$= x_{t+1} - \eta [e_t + \gamma (e_t - e_{t-1})], \quad (103)$$

where (103) is due to (11). \square

Via triangle inequality, relation (97) implies

$$\|\hat{y}_t - x_f^*\| \leq \|y_t - x_f^*\| + \eta \|e_{t-1}\| \quad (104)$$

The following simple corollary to a known bound on the contraction factor of unquantized AGD controls the first term of (104):

Lemma B.5 (Convergence of AGD). *For any L -smooth and μ -strongly convex function f on \mathbb{R}^n , AGD ((10), (11) starting at $x_0 = y_0$) with stepsize (30) and momentum coefficient (31) satisfies*

$$\|y_t - x_f^*\| \leq \sigma_{\text{AGD}}^t \sqrt{\kappa + 1} \|x_0 - x_f^*\| \quad (105)$$

$$\|x_t - x_f^*\| \leq \sigma_{\text{AGD}}^t \lambda \|x_0 - x_f^*\|. \quad (106)$$

Proof. According to [43, Theorem 3.18],

$$f(y_t) - f(x_f^*) \leq \frac{L + \mu}{2} \sigma_{\text{AGD}}^{2t} \|x_0 - x_f^*\|^2. \quad (107)$$

Convergence bound (105) w.r.t. $\{y_t\}$ is due to (107) and

$$f(x) - f(x_f^*) \geq \frac{\mu}{2} \|x - x_f^*\|^2 \quad (108)$$

implied by μ -strong convexity [47, Theorem 2.1.8]. On the other hand, applying triangle inequality to (11) yields

$$\|x_{t+1} - x_f^*\| \leq (1 + \gamma) \|y_{t+1} - x_f^*\| + \gamma \|y_t - x_f^*\|, \quad (109)$$

and (106) then follows from (105) and (109). \square

The following bound on the quantization error controls the second term in (77):

Lemma B.6 (DQ-AGD quantization error). *Let $f \in \mathcal{F}_n$. Quantization errors $\{e_t\}$ in Algorithm 2 with stepsize (30), momentum coefficient (31) and dynamic ranges (34) satisfy*

$$\|e_t\| \leq r_t \rho_n 2^{-R} \quad (110)$$

$$\leq \sigma_{\text{AGD}}^t c_0 + \phi_+^t c_+ + \phi_-^t c_-, \quad (111)$$

where $\phi_{\pm} \triangleq \phi_{\pm}(\gamma_{\text{AGD}})$ with

$$\phi_{\pm}(\gamma) \triangleq \rho_n 2^{-R} \left(\frac{1}{2}(1 + \gamma) \pm \frac{1}{2} \sqrt{(1 + \gamma)^2 + \frac{4\gamma}{\rho_n 2^{-R}}} \right) \quad (112)$$

and c_0, c_+, c_- are specified below in (117), (121) and (122) respectively.

Proof. Like in the proof of Lemma B.3, to establish (110), in view of (26) it is enough to show (82). Applying triangle inequality and (98) to the expression in Line 5 yields

$$\|u_t\| \leq \|\nabla f(x_t)\| + \|e_{t-1}\| + \gamma (\|e_{t-1}\| + \|e_{t-2}\|). \quad (113)$$

The first term in (113) is bounded as

$$\|\nabla f(x_t)\| \leq L \|x_t - x_f^*\| \quad (114)$$

$$\leq \sigma_{\text{AGD}}^t L D \lambda \quad (115)$$

where (114) is due to L -smoothness (13), and (115) is due to (106). Plugging (115) into (113) and applying (26) to bound quantization error terms in (113), we conclude via induction (similar to the proof of Lemma B.3) that setting the sequence of dynamic ranges recursively as (34) ensures (82).

To show (111), we proceed to solve recursion (34). This step is significantly different from the corresponding step in the proof of Lemma B.3 since now r_t depends not just on r_{t-1}

but also on r_{t-2} . More precisely, recursion (34) is a second-order linear non-homogeneous recurrence relation.

- Particular solution: Plugging the candidate

$$p_t = \sigma_{\text{AGD}}^t c_0 \quad (116)$$

into (34), we solve for the constant

$$c_0 = \frac{\sigma_{\text{AGD}}^2}{p(\sigma_{\text{AGD}})} L D \lambda, \quad (117)$$

where $p(r)$ is the characteristic polynomial in (37) associated with recursion (34).

- Homogeneous solution: Since ϕ_+ and ϕ_- are roots of the quadratic polynomial in (37), the homogeneous solution is given by

$$\phi_t = \phi_+^t c_+ + \phi_-^t c_-, \quad (118)$$

- General solution: constants c_+, c_- in (118) are determined by plugging initial conditions $r_{-2} = r_{-1} = 0$ into the general solution

$$r_t = p_t + \phi_t \quad (119)$$

$$= \sigma_{\text{AGD}}^t c_0 + \phi_+^t c_+ + \phi_-^t c_-, \quad (120)$$

which are

$$c_+ = -\frac{c_0 \phi_+^2}{\sigma_{\text{AGD}}^2} \frac{\sigma_{\text{AGD}} - \phi_-}{\phi_+ - \phi_-} \quad (121)$$

$$c_- = \frac{c_0 \phi_-^2}{\sigma_{\text{AGD}}^2} \frac{\sigma_{\text{AGD}} - \phi_+}{\phi_+ - \phi_-}. \quad (122)$$

\square

Lemmas B.4, B.5, and B.6 lead to the following finite- t convergence bound for the DQ-AGD:

Theorem B.2 (Convergence of DQ-AGD). *In the setting of Theorem III.2, the difference between the iterate and the optimizer at step t satisfies*

$$\|\hat{y}_t - x_f^*\| \leq \sigma_{\text{AGD}}^t c + \phi_+^{t-1} c_+ \eta + \phi_-^{t-1} c_- \eta \quad (123)$$

where $c = \sqrt{\kappa + 1} D + \eta c_0 \sigma_{\text{AGD}}^{-1}$, and c_0, c_+, c_- are specified in (117), (121) and (122) respectively.

Proof. Plugging (105) and (111) into (104) immediately leads to (123). \square

Note that if $\sigma_{\text{AGD}} \geq \phi_+$, which is equivalent to $R \geq R_2$ (39), then $c_0 \geq 0$, $c_+ \leq 0$, $c_- \geq 0$; and $c_0 \leq 0$, $c_+ \geq 0$, $c_- \leq 0$ otherwise. Asymptotic convergence guarantee (35) in Theorem III.2 follows by plugging (123) into definition (1) of the contraction factor.

C. Proof of Theorem III.3

The proof of the DQ-HB convergence result in Theorem III.3 follows the same recipe as the proofs of Theorems III.1 and III.2. Lemma B.7 below states that the path of DQ-HB tracks that of the unquantized HB.

Lemma B.7 (DQ-HB trajectory). *Path $\{\hat{x}_t\}$ of DQ-HB (Line 7) and path $\{x_t\}$ of HB (12) starting at the same location $\hat{x}_{-1} = x_{-1} = \hat{x}_0 = x_0$ are related as, $\forall t = 0, 1, 2, \dots$,*

$$\hat{x}_t = x_t - \eta e_{t-1}. \quad (124)$$

Proof. We prove (124) via induction.

- Base case: (124) holds for $t = 0$ by the initialization $\mathbf{e}_{-2} = \mathbf{e}_{-1} = \mathbf{0}$ in Line 1 of Algorithm 3 and since the starting location is the same.
- Inductive step: Plugging expressions on Line 5 and 7 into Line 8 yields

$$\hat{\mathbf{x}}_{t+1} \quad (125)$$

$$= \hat{\mathbf{x}}_t - \eta \mathbf{q}_t + \gamma (\hat{\mathbf{x}}_t - \hat{\mathbf{x}}_{t-1}) \quad (126)$$

$$= \mathbf{x}_t - \eta \mathbf{e}_{t-1} \quad (127)$$

$$- \eta [\nabla f(\mathbf{x}_t) - [\mathbf{e}_{t-1} + \gamma (\mathbf{e}_{t-1} - \mathbf{e}_{t-2})] + \mathbf{e}_t] \\ + \gamma (\mathbf{x}_t - \mathbf{x}_{t-1} - \eta (\mathbf{e}_{t-1} - \mathbf{e}_{t-2})) \quad (127)$$

$$= \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) + \gamma (\mathbf{x}_t - \mathbf{x}_{t-1}) - \eta \mathbf{e}_t \quad (128)$$

$$= \mathbf{x}_{t+1} - \eta \mathbf{e}_t, \quad (129)$$

where (127) is due to the induction hypothesis and (129) is due to (12).

It follows that

$$\left\| \begin{bmatrix} \mathbf{x}_{t+1} - \mathbf{x}_f^* \\ \mathbf{x}_t - \mathbf{x}_f^* \end{bmatrix} \right\|_2 \leq \|A_t \cdot \dots \cdot A_0\|_2 \left\| \begin{bmatrix} \mathbf{x}_0 - \mathbf{x}_f^* \\ \mathbf{x}_{-1} - \mathbf{x}_f^* \end{bmatrix} \right\|_2. \quad (136)$$

It is shown in [63, Lemma 2.3] that if matrices A_1, \dots, A_t all belong to a bounded set \mathcal{A} , then there exists a constant $\alpha > 0$ such that $\forall t = 0, 1, \dots,$

$$\|A_t \cdot \dots \cdot A_0\|_2 \leq \rho(\mathcal{A})^{t+1} (t+1)^\alpha e^\alpha \quad (137)$$

where

$$\rho(\mathcal{A}) \triangleq \limsup_{t \rightarrow \infty} \sup_t \rho(A_t), \quad (138)$$

where $\rho(A_t)$ is the spectral radius of A_t . It is shown in [48, Proof of Th. 3.1] that if

$$\gamma = \max \left\{ \left(1 - \sqrt{\eta L}\right)^2, \left(1 - \sqrt{\eta \mu}\right)^2 \right\}, \quad (139)$$

then

$$\rho(A_t) \leq \sqrt{\gamma}. \quad (140)$$

With the optimal choice of η (42) the right side of (140) is equal to σ_{HB} (44).

In our setting \mathcal{A} is bounded since for twice continuously differentiable functions, L -smoothness and μ -strong convexity are equivalent to

$$\mu I \preceq \nabla^2 f(v) \preceq L I, \quad (141)$$

in the positive semidefinite order, thus (137) applies. Inequality (131) follows after applying (140) to (137) and the latter to (136). \square

We ensure that the quantization error term $\|\mathbf{e}_{t-1}\|$ in (130) decays exponentially fast by adjusting the sequence of dynamic ranges (7) carefully:

Lemma B.9 (DQ-HB quantization error). *Let $f \in \mathcal{F}_n^2$. Quantization errors $\{\mathbf{e}_t\}$ in Algorithm 3 with stepsize (42), momentum coefficient (43) and dynamic ranges (45) satisfy*

$$\|\mathbf{e}_t\| \leq r_t \rho_n 2^{-R} \quad (142)$$

$$\leq (\sigma_{HB}^t c_0 + \phi_+^t c_+ + \phi_-^t c_-) t^\alpha, \quad (143)$$

where $\phi_\pm \triangleq \phi_\pm(\gamma_{HB})$ (112), and c_0, c_+, c_- are as in (117), (121) and (122) respectively, with $LD\lambda$ replaced by $e^\alpha \sqrt{2} LD$, and σ_{AGD} by σ_{HB} .

Proof. Applying triangle inequality and (124) to the expression in Line 5 yields the same expression as in the analysis of DQ-AGD (113):

$$\|\mathbf{u}_t\| \leq \|\nabla f(\mathbf{x}_t)\| + \|\mathbf{e}_{t-1}\| + \gamma (\|\mathbf{e}_{t-1}\| + \|\mathbf{e}_{t-2}\|). \quad (144)$$

The first term in (144) is bounded as

$$\|\nabla f(\mathbf{x}_t)\| \leq L \|\mathbf{x}_t - \mathbf{x}_f^*\| \quad (145)$$

$$\leq \sigma_{HB}^t t^\alpha e^\alpha \sqrt{2} LD \quad (146)$$

where (114) is due to L -smoothness (13), and (146) is due to (131). Applying the argument used to show (110) in the proof of Lemma B.6 leads to (142).

$$\begin{aligned} & \left[\begin{array}{c} \mathbf{x}_{t+1} - \mathbf{x}_f^* \\ \mathbf{x}_t - \mathbf{x}_f^* \end{array} \right] \\ &= \left[\begin{array}{c} \mathbf{x}_t + \gamma(\mathbf{x}_t - \mathbf{x}_{t-1}) - \mathbf{x}_f^* \\ \mathbf{x}_t - \mathbf{x}_f^* \end{array} \right] - \eta \begin{bmatrix} \nabla f(\mathbf{x}_t) \\ 0 \end{bmatrix} \quad (132) \\ &= \begin{bmatrix} (1+\gamma)I & -\gamma I \\ I & 0 \end{bmatrix} \left[\begin{array}{c} \mathbf{x}_t - \mathbf{x}_f^* \\ \mathbf{x}_{t-1} - \mathbf{x}_f^* \end{array} \right] - \eta \begin{bmatrix} \nabla^2 f(v_t)(\mathbf{x}_t - \mathbf{x}_f^*) \\ 0 \end{bmatrix} \quad (133) \\ &= \begin{bmatrix} (1+\gamma)I & -\gamma I - \eta \nabla^2 f(v_t) \\ I & 0 \end{bmatrix} \left[\begin{array}{c} \mathbf{x}_t - \mathbf{x}_f^* \\ \mathbf{x}_{t-1} - \mathbf{x}_f^* \end{array} \right], \quad (134) \end{aligned}$$

where (133) holds for some v_t on the line segment between \mathbf{x}_t and \mathbf{x}_f^* by the mean value theorem, since f is twice continuously differentiable by the assumption. Denoting the matrix in (134) by A_t , we unroll the recursion as

$$\left[\begin{array}{c} \mathbf{x}_{t+1} - \mathbf{x}_f^* \\ \mathbf{x}_t - \mathbf{x}_f^* \end{array} \right] = A_t \cdot A_{t-1} \cdot \dots \cdot A_0 \left[\begin{array}{c} \mathbf{x}_0 - \mathbf{x}_f^* \\ \mathbf{x}_{-1} - \mathbf{x}_f^* \end{array} \right]. \quad (135)$$

To show (143), consider the recursion $r'_{-1} = r'_{-2} = 0$,

$$r'_t = \sigma_{\text{HB}}^t e^\alpha \sqrt{2} LD + (r'_{t-1} + \gamma_{\text{HB}}(r'_{t-1} + r_{t-2})) \rho_n 2^{-R}, \quad (147)$$

We show by strong induction that

$$r_t \leq t^\alpha r'_t, \quad (148)$$

where r_t solves (45). Base case $r_0 = r'_0$ holds by the initial conditions. Assuming that (148) holds for $1, \dots, t-1$, we establish (148) for t using (45) and the fact that t^α is increasing in t :

$$r_t \leq \sigma_{\text{HB}}^t t^\alpha e^\alpha \sqrt{2} LD \quad (149)$$

$$+ ((t-1)^\alpha r'_{t-1} + \gamma_{\text{HB}}((t-1)^\alpha r'_{t-1} + (t-2)^\alpha r'_{t-2})) \rho_n 2^{-R} \leq t^\alpha r'_t. \quad (150)$$

Using (142) and (148), we can show (143) by solving the recursion (147). But this is the same recursion as in (34), up to the constants, thus the solution in the proof of Lemma B.6 applies. \square

We now apply Lemmas B.7, B.8 and B.9 to state a finite- t refinement of Theorem III.3.

Theorem B.3 (Convergence of DQ-HB). *In the setting of Theorem III.3, the difference between the iterate and the optimizer at step t satisfies*

$$\|\hat{x}_t - x_f^*\| \leq (\sigma_{\text{HB}}^t c + \phi_+^{t-1} c_+ + \phi_-^{t-1} c_-) t^\alpha \quad (151)$$

where $c = e^\alpha \sqrt{2} D + \eta c_0 \sigma_{\text{HB}}^{-1}$, and $\phi_+, \phi_-, c_0, c_+, c_-$ are as in Lemma B.9.

Proof. Plugging (131) and (143) into (130) and using $(t-1)^\alpha < t^\alpha$ leads to (151). \square

APPENDIX C CONVERSES

A. Converse for unquantized GD

As explained in the proof sketch, the lower bound in (50) is a combination of an unquantized GD converse and a volume-division converse. The former relies on the following converse result, obtained by constructing a least-square problem instance that satisfies Nesterov's upper bound in Lemma B.2 with equality.

Lemma C.1 (Optimality of σ_{GD}). *Consider GD (8) with starting point x_0 and any constant stepsize η . Then, there exists a problem instance $f \in \mathcal{F}_n$ such that the distance to the optimizer at each iteration t of GD satisfies*

$$\|x_{t+1} - x_f^*\| \geq \sigma_{\text{GD}} \|x_t - x_f^*\|, \quad (152)$$

with equality if η is the optimal stepsize given in (20).

Proof of Lemma C.1. We first find an $x_f^* \in \mathcal{B}(D)$ such that

$$\|x_0 - x_f^*\| \geq D \quad (153)$$

and then construct a least-squares problem instance $f \in \mathcal{F}_n$ (29) that admits x_f^* as a unique minimizer and satisfies (152). Note that f is

$$\sigma_1^2(\mathbf{A})\text{-smooth and } \sigma_n^2(\mathbf{A})\text{-strongly convex} \quad (154)$$

where we denote by $\sigma_i(\mathbf{A})$ the i -th largest singular value of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$. The gradient of f at iteration t is

$$\nabla f(x_t) = \mathbf{A}^\top (\mathbf{A}x_t - \mathbf{y}). \quad (155)$$

The first-order optimality condition $\nabla f(x_f^*) = \mathbf{0}$ implies

$$\mathbf{A}^\top \mathbf{y} = \mathbf{A}^\top \mathbf{A} x_f^*. \quad (156)$$

To each $x_f^* \in \mathcal{B}(D)$ that satisfies (153), there corresponds a $\mathbf{y} \in \mathbb{R}^m$ such that (156) holds. This is because $m \geq n$, i.e. we have more degrees of freedom than the problem dimension when selecting the vector \mathbf{y} . Since we can always select an \mathbf{A} with $\sigma_1(\mathbf{A}) = \sqrt{L}$ and $\sigma_n(\mathbf{A}) = \sqrt{\mu}$ and a \mathbf{y} to ensure (156), we have $f \in \mathcal{F}_n$. It remains to show how to set the right singular vectors of \mathbf{A} to ensure (152).

Plugging (155) into (8) yields

$$x_{t+1} = x_t - \eta \mathbf{A}^\top \mathbf{A} (x_t - x_f^*). \quad (157)$$

Subtracting x_f^* from both sides of (157), we conclude that the distance to the optimizer x_f^* satisfies

$$\|x_{t+1} - x_f^*\| \leq \sigma_1(\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A}) \|x_t - x_f^*\|, \quad (158)$$

where equality is achieved if $x_t - x_f^*$ points in the direction corresponding to the largest singular vector of the matrix $\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A}$. Since

$$\sigma_1(\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A}) = \max \{ |1 - \eta \sigma_n^2(\mathbf{A})|, |1 - \eta \sigma_1^2(\mathbf{A})| \}, \quad (159)$$

we designate the unit vector

$$\mathbf{v}_1 \triangleq \frac{x_0 - x_f^*}{\|x_0 - x_f^*\|} \quad (160)$$

as the right singular vector of \mathbf{A} corresponding to either $\sigma_1(\mathbf{A})$ if $|1 - \eta \sigma_1^2(\mathbf{A})|$ achieves the maximum in (159) or $\sigma_n(\mathbf{A})$ otherwise. This is determined solely by the stepsize η ; the optimal stepsize (20) ensures that $|1 - \eta \sigma_1^2(\mathbf{A})| = |1 - \eta \sigma_n^2(\mathbf{A})| = \sigma_{\text{GD}}$. To complete the construction of \mathbf{A} , we complement \mathbf{v}_1 with $n-1$ orthonormal vectors to form an orthonormal basis $\{\mathbf{v}_i\}_{i=1}^n$ of \mathbb{R}^n . Then,

$$x_1 - x_f^* = \sigma_1(\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A})(x_0 - x_f^*), \quad (161)$$

and using (157) it is easy to show by induction that

$$x_{t+1} - x_f^* = \sigma_1(\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A})(x_t - x_f^*), \quad (162)$$

which implies that (158) holds with equality $\forall t = 0, 1, \dots$. \square

Remark C.1. While variants of Lemma C.1 are known in the literature (e.g. [56, Example 1.3]), they are not directly applicable because of our need to satisfy (15). Furthermore, Lemma C.1 constructs a worst-case problem instance for any initial point and constant stepsize chosen by the GD, whereas the worst-case problem instance constructed in [56, Example 1.3] is tailored to a particular starting point $x_0 \neq \mathbf{0}$ and the optimal η (20).

B. Proof of Theorem IV.1

On one hand, we have

$$\inf_{A \in \mathcal{A}_{GD}} \sigma_A(n, R) \geq \inf_{A \in \mathcal{A}_{GD}} \sigma_A(n, \infty) \quad (163)$$

$$\geq \sigma_{GD}, \quad (164)$$

where (163) holds because the left side of (163) is non-increasing in the data rate R by definition (1), and (164) is by Lemma C.1 since the infinite-rate algorithm in \mathcal{A}_{GD} is the one incurring no quantization error at each iteration, i.e., the GD itself.

On the other hand, to show

$$\inf_{A \in \mathcal{A}_{GD}} \sigma_A(n, R) \geq 2^{-R}, \quad (165)$$

we fix an algorithm $A \in \mathcal{A}_{GD}$ operating at $R' \leq R$ bits per dimension. The set of possible outputs of A after T iterations

$$\mathcal{S}_A \triangleq \left\{ \hat{x}_t \in \mathbb{R}^n \mid \hat{x}_t \text{ is the output of } A \text{ after } T \text{ iterations} \right\} \quad (166)$$

has cardinality

$$|\mathcal{S}_A| = 2^{nR'T}. \quad (167)$$

Given \mathcal{S}_A , consider the minimum-distance quantizer q_A

$$q_A(x) = \arg \min_{\hat{x} \in \mathcal{S}_A} \|\hat{x} - x\| \quad (168)$$

with dynamic range D and covering radius $d(q_A)$ (18). In other words, $2^{nR'T}$ Euclidean balls of radius $d(q_A)$ with centers in \mathcal{S}_A cover $\mathcal{B}(D)$; therefore

$$\frac{D}{d(q_A)} \leq 2^{R'T}. \quad (169)$$

(This classical volume-division argument also shows that $\rho(q_A) \geq 1$ (19)).

Since one can construct an $f \in \mathcal{F}_n$ such that

$$\|q_A(x_f^*) - x_f^*\| = d(q_A), \quad (170)$$

(165) follows by rearranging (169) and taking the limit $T \rightarrow \infty$.

C. Proof of Theorem IV.2

The following lemma shows that gradient iterative methods cannot achieve an arbitrarily low contraction factor.

Lemma C.2 ([47, Th. 2.1.13]). *For any gradient method (53), there exists an L -smooth and μ -strongly convex function $f: \mathbb{L}_2 \rightarrow \mathbb{R}$ such that, $\forall t = 0, 1, \dots$,*

$$\|x_t - x_f^*\| \geq \sigma_{HB}^t \|x_0 - x_f^*\|. \quad (171)$$

The proof of Theorem IV.2 follows the same steps as the proof of Theorem IV.1, with the replacement of the converse for unquantized GD (Lemma C.1) by Lemma C.2.

APPENDIX D

CONVERGENCE ANALYSIS OF K-WORKER NQ-GD

The path of NQ-GD satisfies the following recursive relation.

Lemma D.1 (NQ-GD trajectory). *At each iteration $t = 0, 1, 2, \dots$, the path of NQ-GD with stepsize (20) satisfies*

$$\|\hat{x}_{t+1} - x_f^*\| \leq \sigma_{GD} \|\hat{x}_t - x_f^*\| + \frac{\eta_{GD}}{K} \sum_{k=1}^K \|e_{t,k}\|, \quad (172)$$

where $e_{t,k} \triangleq q_{t,k} - \nabla f_k(\hat{x}_t)$, and σ_{GD} is defined in (21).

Proof of Lemma D.1. The update rule (Line 5) and the quantizer inputs (Line 4) together imply

$$\hat{x}_{t+1} = \hat{x}_t - \frac{\eta_{GD}}{K} \sum_{k=1}^K (\nabla f_k(\hat{x}_t) + e_{t,k}) \quad (173)$$

$$= \hat{x}_t - \eta_{GD} \nabla f(\hat{x}_t) - \frac{\eta_{GD}}{K} \sum_{k=1}^K e_{t,k} \quad (174)$$

where (174) is due to (58). Triangle inequality now implies

$$\|\hat{x}_{t+1} - x_f^*\| \quad (175)$$

$$\leq \|\hat{x}_t - \eta_{GD} \nabla f(\hat{x}_t) - x_f^*\| + \frac{\eta_{GD}}{K} \sum_{k=1}^K \|e_{t,k}\|. \quad (176)$$

Applying the coercive property of smooth and strongly convex functions [47, Th. 2.1.12] in the same manner as in [47, Proof of Th. 2.1.5] to further upper-bound the first term in (176) gives (172). \square

Denote for brevity

$$\sigma \triangleq \sigma_{GD} + \frac{\eta_{GD} \rho_n}{K} \sum_{k=1}^K \frac{L_k}{2^{R_k}} \quad (177)$$

Minimizing (177) under the sum rate constraint (59) is a convex optimization problem whose solution is given by (66) and whose optimal value is given by the right side of (68). The following nonasymptotic result, which with the optimal rate allocation immediately yields Theorem V.1, uses an inductive argument to simultaneously bound both terms in the right-hand side of (172).

Theorem D.1 (Convergence of NQ-GD). *In the setting of Theorem V.1,*

$$\|\hat{x}_t - x_f^*\| \leq \sigma^t D. \quad (178)$$

Proof. We prove (178) by induction.

- Base case: for $t = 0$, (178) holds by (15).
- Inductive step: Suppose (178) holds for iteration t . Then, by L_k -smoothness of f_k , (61) and the inductive hypothesis,

$$\|\nabla f_k(\hat{x}_t)\| \leq L_k \|\hat{x}_t - x_f^*\| \quad (179)$$

$$\leq \sigma^t L_k D, \quad (180)$$

which due to the choice of dynamic ranges (65) and (26) leads to

$$\|e_{t,k}\| \leq \frac{\rho_n}{2^{R_k}} \sigma^t L_k D. \quad (181)$$

Applying (181) and the inductive hypothesis to (172) yields

$$\|\hat{\mathbf{x}}_{t+1} - \mathbf{x}_f^*\| \leq \sigma^t \left[\sigma_{\text{GD}} + \frac{\eta_{\text{GD}} \rho_n}{K} \sum_{k=1}^K \frac{L_k}{2^{R_k}} \right] D, \quad (182)$$

which is exactly (178) for $t + 1$. \square

REFERENCES

- [1] C.-Y. Lin, V. Kostina, and B. Hassibi, “Differentially quantized gradient descent,” in *Proceedings 2021 IEEE International Symposium on Information Theory*, July 2021.
- [2] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, “Parallelized stochastic gradient descent,” in *Advances in Neural Information Processing Systems 23*, Vancouver, British Columbia, Canada, Dec. 2010, pp. 2595–2603.
- [3] B. Recht, C. Re, S. Wright, and F. Niu, “Hogwild: A lock-free approach to parallelizing stochastic gradient descent,” in *Advances in Neural Information Processing Systems 24*, Granada, Spain, Dec. 2011, pp. 693–701.
- [4] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling Up Machine Learning: Parallel and Distributed Approaches*. New York, NY, USA: Cambridge University Press, Dec. 2011.
- [5] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q. V. Le, and A. Y. Ng, “Large scale distributed deep networks,” in *Advances in Neural Information Processing Systems 25*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1223–1231.
- [6] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, “Project Adam: Building an efficient and scalable deep learning training system,” in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, Broomfield, CO, Oct. 2014, pp. 571–582.
- [7] C. M. De Sa, C. Zhang, K. Olukotun, C. Ré, and C. Ré, “Taming the wild: A unified analysis of hogwild-style algorithms,” in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Dec. 2015, pp. 2674–2682.
- [8] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” in *NIPS Workshop on Private Multi-Party Machine Learning*, Dec. 2016.
- [9] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, “Optimal algorithms for smooth and strongly convex distributed optimization in networks,” in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 3027–3036.
- [10] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, “1-bit stochastic gradient descent and application to data-parallel distributed training of speech DNNs,” in *Interspeech 2014*, Sep. 2014.
- [11] M. Li, D. G. Andersen, A. J. Smola, and K. Yu, “Communication efficient distributed machine learning with the parameter server,” in *Advances in Neural Information Processing Systems 27*, Montreal, QB, Canada, Dec. 2014, pp. 19–27.
- [12] N. Strom, “Scalable distributed DNN training using commodity GPU cloud computing,” in *INTERSPEECH*, Sep. 2015.
- [13] S. Zhang, A. E. Choromanska, and Y. LeCun, “Deep learning with elastic averaging SGD,” in *Advances in Neural Information Processing Systems 28*, Montreal, QB, Canada, Dec. 2015, pp. 685–693.
- [14] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, “QSGD: Communication-efficient SGD via gradient quantization and encoding,” in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, Dec. 2017, pp. 1709–1720.
- [15] 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,” in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, Dec. 2017, pp. 1509–1519.
- [16] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, “signSGD: Compressed optimisation for non-convex problems,” in *Proceedings of the 35th International Conference on Machine Learning*, Long Beach, CA, USA, Jul. 2018, pp. 560–569.
- [17] P. Mayekar and H. Tyagi, “RATQ: A universal fixed-length quantizer for stochastic optimization,” *arXiv*, vol. 1908.08200, Dec. 2019.
- [18] A. Ramezani-Kebrya, F. Faghri, and D. M. Roy, “NUQSGD: Improved communication efficiency for data-parallel SGD via nonuniform quantization,” Aug. 2019.
- [19] P. Mayekar and H. Tyagi, “Limits on gradient compression for stochastic optimization,” *arXiv*, vol. 2001.09032, Jan. 2020.
- [20] V. Gandikota, D. Kane, R. K. Maity, and A. Mazumdar, “vqSGD: Vector quantized stochastic gradient descent,” Nov. 2019.
- [21] S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi, “Error feedback fixes SignSGD and other gradient compression schemes,” in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, Long Beach, CA, USA, June 2019, pp. 3252–3261.
- [22] N. Dryden, S. A. Jacobs, T. Moon, and B. Van Essen, “Communication quantization for data-parallel training of deep neural networks,” in *Proceedings of the Workshop on Machine Learning in High Performance Computing Environments*, 2016, pp. 1–8.
- [23] A. F. Aji and K. Heafield, “Sparse communication for distributed gradient descent,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sep. 2017, pp. 440–445.
- [24] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, “The convergence of sparsified gradient methods,” in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Dec. 2018, pp. 5973–5983.
- [25] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, “Sparsified SGD with memory,” in *Advances in Neural Information Processing Systems 31*, Montréal, Canada, Dec. 2018, pp. 4447–4458.
- [26] M. Yu, Z. Lin, K. Narra, S. Li, Y. Li, N. S. Kim, A. Schwing, M. Annavaram, and S. Avestimehr, “GradiVeQ: Vector quantization for bandwidth-efficient gradient aggregation in distributed CNN training,” *Advances in Neural Information Processing Systems*, vol. 31, pp. 5123–5133, Jan. 2018.
- [27] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, “Deep gradient compression: Reducing the communication bandwidth for distributed training,” in *International Conference on Learning Representations*, Vancouver, BC, Canada, Apr. 2018.
- [28] J. Wangni, J. Wang, J. Liu, and T. Zhang, “Gradient sparsification for communication-efficient distributed optimization,” in *Advances in Neural Information Processing Systems 31*, Montréal, Canada, Dec. 2018, pp. 1299–1309.
- [29] H. Wang, S. Sievert, S. Liu, Z. Charles, D. Papailiopoulos, and S. Wright, “ATOMO: Communication-efficient learning via atomic sparsification,” in *Advances in Neural Information Processing Systems 31*, Montréal, Canada, Dec. 2018, pp. 9850–9861.
- [30] D. Basu, D. Data, C. Karakus, and S. Diggavi, “Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Dec. 2019, pp. 14 695–14 706.
- [31] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan, “On biased compression for distributed learning,” *arXiv:2002.12410*, Feb. 2020.
- [32] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik, “Distributed learning with compressed gradient differences,” Jan. 2019.
- [33] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik, “Stochastic distributed learning with gradient quantization and variance reduction,” Apr. 2019.
- [34] S. Horváth, C.-Y. Ho, L. Horvath, A. N. Sahu, M. Canini, and P. Richtarik, “Natural compression for distributed deep learning,” May 2019.
- [35] M. M. Amiri and D. Gündüz, “Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air,” in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1432–1436.
- [36] G. Zhu, Y. Wang, and K. Huang, “Broadband analog aggregation for low-latency federated edge learning,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 491–506, 2020.
- [37] K. Yang, T. Jiang, Y. Shi, and Z. Ding, “Federated learning via over-the-air computation,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.
- [38] R. M. Gray, *Source Coding Theory*. Kluwer Academic Publishers, Oct. 1989.
- [39] J. Wu, W. Huang, J. Huang, and T. Zhang, “Error compensated quantized SGD and its applications to large-scale distributed optimization,” in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, Stockholm, Sweden, July 2018, pp. 5325–5333.

- [40] S. U. Stich and S. P. Karimireddy, “The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication,” *arXiv:1909.05350*, Sep. 2019.
- [41] S. Zheng, Z. Huang, and J. Kwok, “Communication-efficient distributed blockwise momentum SGD with error-feedback,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Dec. 2019, pp. 11450–11460.
- [42] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, Jan. 2009.
- [43] S. Bubeck, “Convex optimization: Algorithms and complexity,” *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, pp. 231–357, Nov. 2015.
- [44] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” *SIAM Review*, vol. 60, no. 2, pp. 223–311, May 2018.
- [45] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, “Scaling distributed machine learning with the parameter server,” in *11th USENIX Symposium on Operating Systems Design and Implementation*, Broomfield, CO, Oct. 2014, pp. 583–598.
- [46] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson, “Distributed learning with compressed gradients,” *arXiv: 1806.06573*, June 2018.
- [47] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, Dec. 2014.
- [48] B. T. Polyak, *Introduction to optimization*, ser. Translations Series in Mathematics and Engineering. Optimization Software, May 1987.
- [49] C. A. Rogers, “Covering a sphere with spheres,” *Mathematika*, vol. 10, no. 2, pp. 157–164, Dec. 1963.
- [50] M. Friedlander and M. Schmidt, “Hybrid deterministic-stochastic methods for data fitting,” *SIAM Journal on Scientific Computing*, vol. 34, no. 3, pp. A1380–A1405, May 2012.
- [51] Y. Nesterov, “A method of solving a convex programming problem with convergence rate $O(1/k^2)$,” in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1982, pp. 372–376.
- [52] Z. Allen-Zhu and L. Orecchia, “Linear coupling: An ultimate unification of gradient and mirror descent,” *arXiv: 1407.1537*, July 2014.
- [53] W. Su, S. Boyd, and E. J. Candès, “A differential equation for modeling Nesterov’s accelerated gradient method: Theory and insights,” *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.
- [54] R. Zamir, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. Cambridge University Press, Sep. 2014.
- [55] S. Kolodziej, M. Aznaveh, M. Bullock, J. David, T. Davis, M. Henderson, Y. Hu, and R. Sandstrom, “The SuiteSparse matrix collection website interface,” *Journal of Open Source Software*, vol. 4, no. 35, p. 1244, Mar. 2019.
- [56] E. de Klerk, F. Glineur, and A. B. Taylor, “On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions,” *Optim. Lett.*, vol. 11, no. 7, pp. 1185–1199, Oct. 2017.
- [57] Y. Arjevani, S. Shalev-Shwartz, and O. Shamir, “On lower and upper bounds in smooth and strongly convex optimization,” *Journal of Machine Learning Research*, vol. 17, no. 126, pp. 1–51, 2016.
- [58] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [59] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.
- [60] M. Schmidt and N. L. Roux, “Fast convergence of stochastic gradient descent under a strong growth condition,” Aug. 2013.
- [61] D. Needell, R. Ward, and N. Srebro, “Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm,” in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Dec. 2014, pp. 1017–1025.
- [62] S. Ma, R. Bassily, and M. Belkin, “The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning,” in *International Conference on Machine Learning*, vol. 80, Stockholm, Sweden, July 2018, pp. 3325–3334.
- [63] F. Wirth, “On the calculation of time-varying stability radii,” *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, vol. 8, no. 12, pp. 1043–1058, 1998.