Differentially Quantized Gradient Methods

Chung-Yi Lin, Victoria Kostina, and Babak Hassibi

***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 server receives all its information about the problem instance from the worker via a rate-limited noiseless communication channel.**

We introduce the principle we call *differential quantization* (DQ) that prescribes compensating the past quantization errors 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 $\max\{\sigma_{\mathrm{GD}}, \rho_n 2^{-R}\}$, 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 $R \geq \log_2 \rho_n/\sigma_{\mathrm{GD}}$ 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 algorithm within a certain class can converge faster than $\max\{\sigma_{\mathrm{GD}}, 2^{-R}\}$. Since quantizers exist with $\rho_n \to 1$ as $n \to \infty$ (Rogers, 1963), this means that DQ-GD is asymptotically optimal. In contrast, naively quantized GD where the worker directly quantizes the gradient barely attains $\sigma_{\mathrm{GD}} + \rho_n 2^{-R}$.

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 least-squares problems validate our theoretical analysis.

***Index Terms*—gradient descent, quantized gradient descent, accelerated gradient descent, heavy ball method, error compensation, error feedback, 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 between distributed agents in the federated learning setting has become a communication bottleneck that slows down the overall learning process [3], [6], [10]–[13] .

A natural approach to alleviating that communication bottleneck is to quantize the gradients with a limited number of bits per problem dimension. Its power was first demonstrated by Seide et al. [10], where the gradient computed by stochastic gradient descent (SGD) [14] is quantized down to just one bit per dimension and the quantization error is carried forward across mini-batches, resulting in almost no loss in empirical convergence performance compared to the unquantized algorithm. Wen at al. [15] propose a ternary quantizer for SGD and prove that it converges almost surely under the assumption of bounded gradients. Bernstein et al. [16] propose a sign-based quantizer for mini-batch SGD, give its convergence analysis on nonconvex problems, and extend it to accelerated gradient descent and to a multi-worker setting. Alistarh et al. [17] propose a quantized SGD algorithm that compresses the gradient using a stochastic scalar quantizer with an adjustable number of quantization levels, and provide convergence guarantees that depend on this variable compression rate on smooth convex and non-convex functions. The quantizer in [10], [15]–[17] is a uniform scalar quantizer, which simply rounds the binary representation of each coordinate to a fixed number of bits, while [18] considers a non-uniform scalar quantizer, and [19], [20] construct vector quantizers from the convex hull of specifically structured point sets.

A different approach to addressing the communication bottleneck in parallel SGD training is to sparsify the gradient vectors [12], [21]–[25]. For example, the top-k sparsifier (or *compressor*) preserves the k coordinates of the largest magnitude and sends them with full precision [12], [21], [22], [26]–[28]. A user-specified parameter (e.g. k for the top-k compressor) serves as a proxy for the communication rate in this line of work.

For an empirical risk minimization problem where the global objective function is the average of local objective functions, recent works [29]–[31] perform analog gradient compression and communication by taking the physical superposition nature of the underlying multiple-access channel into the account.

The assumption of unbiased compression error [32]–[43] is commonly imposed to enable convergence analyses of compressed SGD. Employing biased compressors in compressed SGD can lead to divergence: for example, both the 1-bit SGD without mini-batching [10], [44] and the top-1 compressed SGD [45] diverge on some problem instances. A set of sufficient conditions on the compression operators to ensure convergence of SGD is put forth in [46].

The same paper - [10] - that initiates the study of quantized SGD is also the first to introduce the idea of adding back previous quantization errors before quantizing the gradient

Chung-Yi Lin (hsnu1220@gmail.com) is with Kronos Research, Taiwan. V. Kostina (vkostina@caltech.edu) and B. Hassibi (hassibi@caltech.edu) are with California Institute of Technology. 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 was presented in part at ISIT 2021 [1].

at the next step of iterative optimization, which fixes the divergence issue mentioned above. The idea, referred to as *error compensation*, or *error feedback*, in the federated learning literature, has been long known as Σ-Δ modulation [47] in the information theory literature. Stich et al. [22] apply the mechanism of error feedback in [10] to a more general setting of SGD and show that it converges with the same order as unquantized SGD on strongly convex and smooth functions, providing the first theoretical performance guarantee of that error feedback strategy. Karimireddy et al. [44] extend the analysis of [22] to the non-convex and weakly convex objective functions, while Zheng et al. [48] and Gorbunov et al. [40] prove its convergence in the multi-worker setting. Wu et al. [49] propose an error feedback mechanism different from [10] and prove its convergence on quadratic functions using the same quantizer as in [17]. Past quantization errors in the algorithm of [49] accumulate from one iteration to another and are weighted by time-decaying factors. The momentum correction used in [25], as well as the distributed SGD with skipped communication rounds in [50], are also variants of error compensation. Qian et al. [51] propose an error-compensated accelerated SGD, while Richrárik et al. [52] propose an error-compensated SGD that achieves the same order of convergence as SGD with unbiased compressors [42]. The analyses in [22], [40], [44], [52] assume that the compressor is a contraction operator, while [49] also assumes its unbiasedness. Horváth and Richtárik [53] construct an unbiased compressor from a contractive compressor and employ the resulting unbiased compressor within SGD as an alternative to error feedback to overcome the divergence issue with biased compressors.

Although a number of works provide convergence analyses of their proposed methods, showing that convergence rates of quantized gradient methods depend on the bit rate R [17]–[19], [54], [55], there are few existing convergence lower bounds in terms of R that apply to any algorithm within a specified class. For quantized projected SGD, [19], [55] 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 [19], [55] is fixed to be the gradient of the current iterate, precluding the use of error compensation.

The parameter server framework that we consider in this paper is somewhat different from the *distributed estimation* or optimization setting [56], where there has also been great interest in communication-efficient algorithms to account for the distributed nature of these problems. In such applications, all parties in a connected network communicate back and forth in order to estimate the mean of a distribution [57] (or a population [58]) or to solve a convex optimization cooperatively with quantization effects [59], [60]. Information-theoretic lower bounds have also been established either in the minimax sense for distributed statistical estimation problem [61] or in terms of the communication complexity for the distributed convex learning problem [58], [62], [63].

B. Contributions

In this paper, we provide a lower bound on quantized non-stochastic gradient descent, and we show a single-worker algorithm that achieves the lower bound with equality, thereby establishing an information-theoretic fundamental limit of quantized gradient descent. In other words, we quantify exactly the minimum amount of information required to achieve a desired convergence speed (within a class of algorithms), and we show an algorithm that achieves it. Because the algorithm achieves the information-theoretic converse with equality, no other algorithm can surpass its performance. It is remarkable that only a finite bit rate is required to achieve the optimal convergence speed achievable with an infinite rate. Our analysis is sharper than existing analyses because we identify constants and not just the order of convergence. We focus on (nonstochastic) GD and not on SGD as most prior work. We do not assume that the quantizer is unbiased or is a contraction operator - our information-theoretic lower bound applies to any quantizer, and any quantizer can be inserted into our algorithm, although our achievability result suggests that picking a (scalar or vector) quantizer with good covering efficiency would perform best. Our mechanism for error compensation (that we call "differential quantization") differs from prior works in the gradient-compute point, which is crucial for achieving our sharp information-theoretic lower bound. Although our information-theoretic lower bound applies to the multi-worker setting as well, our best achievability bound comes short of it at a finite R. Thus, it remains an open problem whether the lower bound is achievable in the multi-worker setting. While our main results are presented in terms of a quantity that is asymptotic in the number of iterations T, the analyses that lead to these results are nonasymptotic. Incidentally, we discover two new results on the classical (unquantized) gradient methods: a slightly more general converse for the gradient descent, and a nonasymptotic global convergence bound on Polyak's heavy ball method.

We consider the single-worker scenario of the parameter server framework [11], [15]–[18], [64], [65] 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.

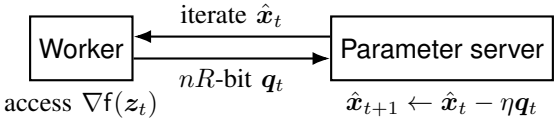

Fig. 1: Quantized gradient descent in a single-worker remote training setting. At each iteration t, the server first sends the current iterate $\hat{\boldsymbol{x}}_t$ to the worker noiselessly, who computes the gradient at some point $\boldsymbol{z}_t$ that is a function of (but not necessarily equal to) $\hat{\boldsymbol{x}}_t$. Then, the worker forms a descent direction $\boldsymbol{q}_t$ 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 $\mathsf{f}\colon \mathbb{R}^n \mapsto \mathbb{R}$ 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

$$\sigma_{\mathrm{A}}(n,R) \triangleq \inf_{R'\le R} \sup_{\mathsf{f}\in\mathcal{F}_n} \limsup_{T\to\infty} \|\hat{\boldsymbol{x}}_T(R') - \boldsymbol{x}^*_{\mathsf{f}}\|^{\frac{1}{T}} \tag{1}$$

where $\boldsymbol{x}^*_{\mathsf{f}}$ is the optimizer, and $\hat{\boldsymbol{x}}_0(R'), \hat{\boldsymbol{x}}_1(R'), \hat{\boldsymbol{x}}_2(R'), \ldots$ is the sequence of iterates generated by A in response to $\mathsf{f}\in\mathcal{F}_n$ 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) [66], and the heavy ball method (HB) [67]. 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: DQ-GD, DQ-AGD, and DQ-HB. By analyzing them, we show achievability bounds of the form[2]

$$\sigma_{\mathrm{A}}(n,R) \le \max\left\{\sigma_{\mathrm{A}}, \rho_n 2^{-R}\phi_{\mathrm{A}}(n,R)\right\}, \tag{2}$$

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

$$\sigma_{\text{DQ-GD}}(n,R) \le \max\left\{\sigma_{\mathrm{GD}}, \rho_n 2^{-R}\right\}. \tag{3}$$

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

$$\sigma_{\mathrm{A}}(n,R) \ge \max\left\{\sigma_{\mathrm{GD}}, 2^{-R}\right\}, \tag{4}$$

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 [68, Th. 3] that shows the existence of quantizers with covering efficiency $\rho_n \to 1$ as $n\to\infty$ and comparing (2) and (4), one can deduce the asymptotic optimality of DQ-GD within the class of "GD-like" algorithms. In contrast, the natural method that quantizes the gradient of its current iterate directly [15]–[17], [69] referred to as naively quantized (NQ) GD in this paper, has contraction factor (in the single-worker scenario; see Section V for the multi-worker result)

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

[1]The 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 [32]–[34].

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

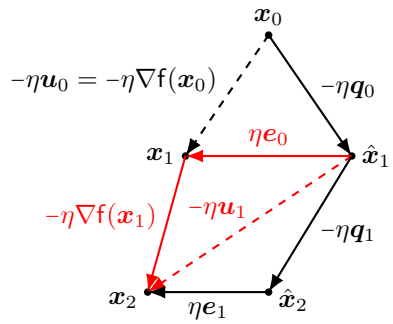

Fig. 2: Illustration of the DQ-GD algorithm (Algorithm 1).

where $\kappa\ge 1$ 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 $\mathsf{q}\colon \mathcal{D}\to\mathbb{R}^n$, where $\mathcal{D}\subseteq\mathbb{R}^n$ is the domain, such that the image of q satisfies

$$|\mathrm{Im}(\mathsf{q})| = 2^{nR}. \tag{6}$$

This is the classical general fixed-rate quantizer in the information theory literature. We fix a dimension-n, rate-R quantizer q, and we set up quantizer q_t to be used at iteration t as

$$\mathsf{q}_t(\cdot) = r_t\mathsf{q}(\cdot/r_t) \tag{7}$$

for a properly chosen sequence of shrinkage factors $\{r_t\}$ (see (22), (35), and (46), 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:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\nabla\mathsf{f}(\boldsymbol{x}_t), \tag{8}$$

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 $\boldsymbol{x}_t$ associated with the corresponding unquantized algorithm, i.e., GD, by compensating previous scaled quantization error $\eta \boldsymbol{e}_{t-1}$ (Line 4). It then computes the gradient at $\boldsymbol{z}_t = \boldsymbol{x}_t$ (Lemma B.1) and sets the quantizer input as (Line 5)

$$\boldsymbol{u}_t = \nabla \mathsf{f}\left(\hat{\boldsymbol{x}}_t + \eta \boldsymbol{e}_{t-1}\right) - \boldsymbol{e}_{t-1}, \tag{9}$$

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

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

Algorithm 1: DQ-GD

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

C. Differentially Quantized Accelerated Gradient Descent

Nesterov's Accelerated Gradient Descent (AGD) [70] keeps track of two iterate sequences

$$\boldsymbol{y}_{t+1} = \boldsymbol{x}_t - \eta \nabla \mathsf{f}(\boldsymbol{x}_t) \tag{10}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{y}_{t+1} + \gamma\left(\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\right). \tag{11}$$

It first performs the gradient descent step (10), and then adds the momentum term $\gamma\left(\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\right)$ (11) to form a projection $\boldsymbol{x}_{t+1}$ of the GD iterate $\boldsymbol{y}_{t+1}$ to its near future. The momentum term incorporates second-order effects by leveraging the past $\boldsymbol{y}_t$. 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 [66] (Lemma B.5). There are various interpretations of Nesterov's acceleration phenomenon. We refer the reader to [71] for a connection to the mirror descent algorithm and to [72] 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 $\boldsymbol{e}_{t-1}, \boldsymbol{e}_{t-2}$ to determine the gradient-compute point $\boldsymbol{z}_t$ (Line 4) and the quantizer input $\boldsymbol{u}_t$ (Line 5). As dictated by the principle of differential quantization, DQ-AGD computes the gradient at the same point as unquantized AGD, i.e., $\boldsymbol{z}_t = \boldsymbol{x}_t$ (Lemma B.4).

D. Differentially Quantized Heavy Ball Method

Polyak's Heavy Ball (HB) algorithm [67] iterates

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla \mathsf{f}(\boldsymbol{x}_t) + \gamma\left(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\right), \tag{12}$$

Algorithm 2: DQ-AGD

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

where $\gamma\left(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\right)$ is the momentum term that nudges $\boldsymbol{x}_{t+1}$ 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 [67, Th. 3.1] (Lemma B.8), [66, 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 $\boldsymbol{z}_t = \boldsymbol{x}_t$ (Lemma B.7). Note that DQ-HB has the same expression for its quantizer input $\boldsymbol{u}_t$ (Line 4) as DQ-AGD (Line 5).

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

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

III. Convergence Rates of DQ Algorithms

A. Definitions

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

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

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

$$\|\nabla \mathsf{f}(\boldsymbol{v}) - \nabla \mathsf{f}(\boldsymbol{w})\| \leq L \|\boldsymbol{v} - \boldsymbol{w}\|; \tag{13}$$

ii) f is μ-strongly convex, i.e.,

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

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

$$\|\boldsymbol{x}_{\mathsf{f}}^* - \hat{\boldsymbol{x}}_0\| \leq D, \tag{15}$$

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

We say that $\mathsf{f}\colon \mathbb{R}^n \mapsto \mathbb{R}$ 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 $\mathsf{f} \in \mathcal{F}_n$ by

$$\kappa \triangleq \frac{L}{\mu}. \tag{16}$$

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

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

$$r(\mathsf{q}) \triangleq \max\{\delta\colon \mathcal{B}(\delta) \subseteq \mathcal{D}\} \tag{17}$$

as the *dynamic range* of q, to

$$\mathsf{d}(\mathsf{q}) \triangleq \min\{d\colon \forall \boldsymbol{x} \in \mathcal{D},\ \|\boldsymbol{x} - \mathsf{q}(\boldsymbol{x})\| \leq d\} \tag{18}$$

as its *covering radius*, and to

$$\rho(\mathsf{q}) \triangleq |\mathrm{Im}(\mathsf{q})|^{1/n} \frac{\mathsf{d}(\mathsf{q})}{r(\mathsf{q})} \tag{19}$$

as its *covering efficiency*.[3] A scalar uniform quantizer q_{u} has domain $[-r(\mathsf{q}_{\mathsf{u}}), r(\mathsf{q}_{\mathsf{u}})]^n$ and covering efficiency $\sqrt{n}$. This is wasteful: the classical result of Rogers [68, Th. 3] implies that there exists a sequence of n-dimensional quantizers q_n with $\rho(\mathsf{q}_n) \to 1$ as $n \to \infty$, while definition (19) implies that $\rho(\mathsf{q}) \geq 1$ for any quantizer q.

B. DQ-GD: convergence analysis and simulation results

Unquantized gradient descent with the optimal stepsize given by

$$\eta = \eta_{\mathrm{GD}} \triangleq \frac{2}{L+\mu} \tag{20}$$

achieves contraction factor

$$\sigma_{\mathrm{GD}} \triangleq \frac{\kappa - 1}{\kappa + 1} \tag{21}$$

over $\mathcal{F}_n$ [67, Th. 1.4], [66, 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 q with dynamic range 1 and covering efficiency ρ_n. Then, Algorithm 1 with stepsize (20) and dynamic ranges $r_0 = LD$,*

$$r_{t+1} = \sigma_{\mathrm{GD}}^{t+1} LD + r_t \rho_n 2^{-R}, \quad t = 1, 2, \ldots \tag{22}$$

[3]Covering efficiency introduced in (19) extends the notion of covering efficiency of an infinite lattice [73], which measures how well that lattice covers the whole space, to bounded-domain quantizers.

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

$$\sigma_{\mathrm{DQ-GD}}(n, R) \leq \max\left\{\sigma_{\mathrm{GD}}, \rho_n 2^{-R}\right\}. \tag{23}$$

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

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \eta \boldsymbol{e}_{t-1} \tag{24}$$

Comparing (24) and Line 4 in Algorithm 1, we see that $\boldsymbol{z}_t = \boldsymbol{x}_t$, i.e., DQ-GD computes the gradient at the unquantized trajectory $\{\boldsymbol{x}_t\}$. The convergence guarantee of GD [67, Th. 1.4], [66, Theorem 2.1.15] (Lemma B.2) controls the difference between the first term in the recursion (24) and the optimizer $\boldsymbol{x}_{\mathsf{f}}^*$. To bound the second term in (24), we observe using (19) that for any $r_t > 0$ in (7),

$$\max_{\boldsymbol{u}\in\mathcal{B}(r_t)} \|\mathsf{q}_t(\boldsymbol{u}) - \boldsymbol{u}\| = r_t \max_{\boldsymbol{u}\in\mathcal{B}(1)} \|\mathsf{q}(\boldsymbol{u}) - \boldsymbol{u}\| \tag{25}$$

$$= r_t \rho_n 2^{-R}, \tag{26}$$

i.e. quantizer 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 $\boldsymbol{u}_t$ to the quantizer 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. The stepsize (20) is optimal both for GD [66, Theorem 2.1.15] and for DQ-GD. See Appendix B-A for details. □

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

A common approach to quantizing descent algorithms [15]–[18], [55], [69] 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))

$$\boldsymbol{u}_t = \nabla \mathsf{f}(\hat{\boldsymbol{x}}_t). \tag{27}$$

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

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

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

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

where $\boldsymbol{y} \in \mathbb{R}^m, \mathbf{A} \in \mathbb{R}^{m\times n}$, with $m \geq n$. We generate 500 matrices $\mathbf{A}$'s with i.i.d. standard normal entries, one for each $\boldsymbol{y}$, and rescale the spectrum of $\mathbf{A}$ so that it has a prescribed condition number κ. We also run the algorithm on the real-world least-squares matrix `ash331` extracted from the online repository *SuiteSpare* [74]. For each per-dimension quantization rate $R \geq 1$, we generate 500 instances of the vector $\boldsymbol{y}$ and $\hat{\boldsymbol{x}}_0$ with i.i.d. standard normal entries. We run the

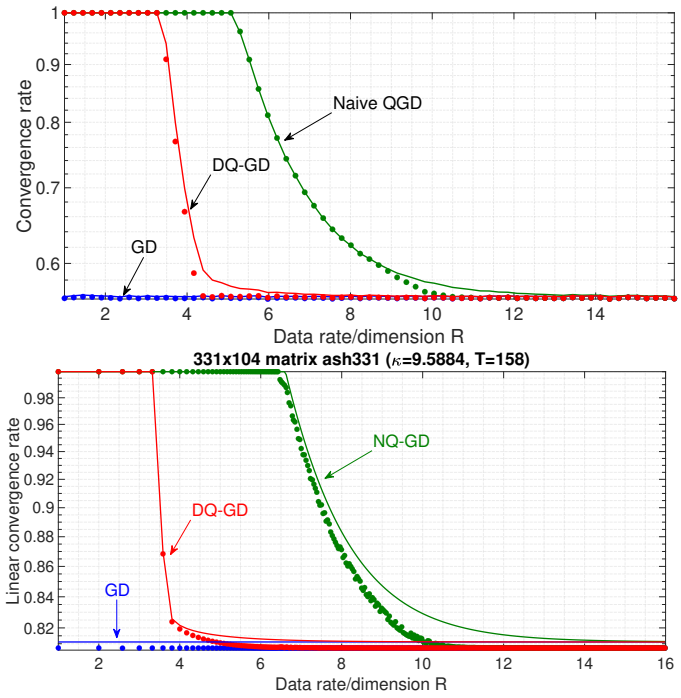

Fig. 3: Empirical contraction factors (as circles) and corresponding upper bounds (21), (23), and (28) (as lines).

iterative algorithms for as many iterations T as possible until reaching the machine's floating point precision, and report the average contraction factor. We use the uniform scalar quantizer for the ease of implementation and take as a consequence a space-filling loss of $\sqrt{n}$. 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 in the plots. We set the stepsize and the quantizer's dynamic range in the DQ-GD algorithm as prescribed by Theorem III.1, and in the NQ-GD algorithm as prescribed by Theorem V.1 in Section V below.

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.

Applying the error feedback mechanism of [10], [22], [44], developed for SGD, to GD results in an algorithm that forms the quantizer input as

$$\boldsymbol{u}_t = \nabla \mathsf{f}(\hat{\boldsymbol{x}}_t) - \boldsymbol{e}_{t-1}. \tag{30}$$

Unlike DQ-GD (9), error feedback in (30) results in computing the gradient along the quantized trajectory $\{\hat{\boldsymbol{x}}_t\}$, and it is unclear whether it can even improve upon NQ-GD (28) in the setting of our paper - nonstochastic GD with a worst-case performance criterion and without further assumptions on the quantizer (Appendix E).

C. DQ-AGD: convergence analysis

Unquantized accelerated gradient descent with stepsize

$$\eta = \eta_{\mathrm{AGD}} \triangleq \frac{1}{L} \tag{31}$$

and momentum coefficient

$$\gamma = \gamma_{\mathrm{AGD}} \triangleq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \tag{32}$$

achieves contraction factor

$$\sigma_{\mathrm{AGD}} \triangleq \sqrt{1-\frac{1}{\sqrt{\kappa}}} \tag{33}$$

over $\mathcal{F}_n$ (1) [33, Th. 3.18] (Lemma B.5), which improves the contraction factor of gradient descent $\sigma_{\mathrm{GD}} = 1 - \frac{1}{\kappa} + O\left(\frac{1}{\kappa^2}\right)$ (21) to $\sigma_{\mathrm{AGD}} = 1 - \frac{1}{2\sqrt{\kappa}} + O\left(\frac{1}{\kappa}\right)$, a significant improvement if κ is large and optimal order-wise (the converse to the optimal contraction factor expands as $1 - \frac{4}{\sqrt{\kappa}} + O\left(\frac{1}{\kappa}\right)$ [66] (Lemma C.2) and is attained in $\mathcal{F}_n^2$ by the heavy ball method [67] (Lemma B.8).

Denote for brevity the constant

$$\lambda \triangleq \left(1 + \gamma_{\mathrm{AGD}} + \gamma_{\mathrm{AGD}}\sigma_{\mathrm{AGD}}^{-1}\right)\sqrt{\kappa+1}. \tag{34}$$

The following result extends (33) to DQ-AGD.

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* (31)*, momentum coefficient* (32)*, and dynamic ranges* $r_{-2} = r_{-1} = 0$*,*

$$r_t = \sigma_{\mathrm{AGD}}^t LD\lambda + \left(r_{t-1} + \gamma_{\mathrm{AGD}}(r_{t-1} - r_{t-2})\right)\rho_n 2^{-R}, \tag{35}$$

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

$$\sigma_{\mathrm{DQ-AGD}}(n,R) \leq \max\left\{\sigma_{\mathrm{AGD}}, \rho_n 2^{-R}\phi(n,R,\gamma_{\mathrm{AGD}})\right\} \tag{36}$$

where

$$\phi(n,R,\gamma) \triangleq \frac{1}{2}(1+\gamma) + \frac{1}{2}\sqrt{(1+\gamma)^2 + \frac{4\gamma}{\rho_n 2^{-R}}}. \tag{37}$$

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 $\boldsymbol{e}_{t-1}$, the quantizer input in Algorithm 2 depends on the past two quantization errors $\boldsymbol{e}_{t-1}$ and $\boldsymbol{e}_{t-2}$ (Line 5). The resulting recursion (35) is a second-order linear non-homogeneous recurrence relation, which unlike (22) does not simply represent a geometric sequence. The characteristic polynomial of (35) is

$$\mathsf{p}(r) \triangleq r^2 - r\rho_n 2^{-R}(1+\gamma_{\mathrm{AGD}}) - \rho_n 2^{-R}\gamma_{\mathrm{AGD}}, \tag{38}$$

and $\rho_n 2^{-R}\phi_{\mathrm{DQ-AGD}}(n,R)$ in (36) is its positive, larger-magnitude root. This implies that the quantization error decays with the contraction factor in the right side of (36). See Appendix B-B for details. □

Define the functions

$$R_1(n,\gamma) \triangleq \log_2(1+2\gamma) + \log_2 \rho_n \tag{39}$$

$$R_2(n,\sigma,\gamma) \triangleq \log_2 \frac{(1+\gamma)\sigma + \gamma}{\sigma^2} + \log_2 \rho_n \tag{40}$$

The achievability bound (36) exhibits two phase transitions. The first one is at $\rho_n 2^{-R}\phi(n,R,\gamma_{\mathrm{AGD}}) < 1$, which is equivalent to $\mathsf{p}(1) > 0$: if

$$R > R_1(n,\gamma_{\mathrm{AGD}}) \quad \text{bits / dimension}, \tag{41}$$

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

$$R \ge R_2(n,\sigma_{\mathrm{AGD}},\gamma_{\mathrm{AGD}}) \quad \text{bits / dimension}, \tag{42}$$

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

Curiously, $R_1(n,0)$ and $R_2(n,\sigma_{\mathrm{GD}},0)$ 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

$$\eta = \eta_{\mathrm{HB}} \triangleq \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2 \tag{43}$$

and momentum coefficient

$$\gamma = \gamma_{\mathrm{HB}} \triangleq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2 \tag{44}$$

achieves contraction factor

$$\sigma_{\mathrm{HB}} \triangleq \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \tag{45}$$

over $\mathcal{F}_n^2$ (1) [67] (Lemma B.8), which is optimal among all gradient methods [66, 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* (43)*, momentum coefficient* (44) *and dynamic ranges* $r_{-1} = r_{-2} = 0$*,*

$$r_t = \sigma_{\mathrm{HB}}^t\, t^\alpha e^\alpha \sqrt{2}\, LD + \left(r_{t-1} + \gamma_{\mathrm{HB}}(r_{t-1}+r_{t-2})\right)\rho_n 2^{-R}, \tag{46}$$

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

$$\sigma_{\mathrm{DQ-HB}}(n,R) \le \max\left\{\sigma_{\mathrm{HB}}, \rho_n 2^{-R}\phi(n,R,\gamma_{\mathrm{HB}})\right\}, \tag{47}$$

where $\phi(n,R,\gamma)$ *is defined in* (37)*.*

Proof sketch. The proof is similar to the proof of Theorem III.2. The recurrence relation (46) differs from (35) in only the presence of the subexponential factor t^α, which does not matter when we take $t \to \infty$ to obtain (47). 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. □

DQ-HB exhibits two phase transitions, a behavior similar to DQ-HB and DQ-AGD. The two thresholds are given by R_1 (39) and R_2 (40) evaluated with $\sigma = \sigma_{\mathrm{HB}}$ and $\gamma = \gamma_{\mathrm{HB}}$.

Plugging the parameters γ and σ into (40), we can infer that DQ-HB always has the largest R_2 (40) for any condition number $\kappa \ge 1$ 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 $\kappa \ge 1$. 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., $\rho_n 2^{-R}\phi(n,R,\gamma_{\mathrm{HB}}) > \sigma_{\mathrm{GD}}$. This is because DQ-GD is the first among the three DQ algorithms to pass R_1 (39) 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 use the same experimental setup as in Fig. 3, with the uniform scalar quantizer. The stepsize, the interpolation coefficient and the dynamic ranges are set to the values prescribed by Theorems III.1 (DQ-GD), III.2 (ADQ-GD), and III.3 (ADQ-HB). 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 $R \ge R_2$ 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 $\mathcal{A}_{\mathrm{GD}}$ of quantized gradient descent algorithms, formally defined next.

Definition IV.1 (Class $\mathcal{A}_{\mathrm{GD}}$ of quantized gradient descent algorithms)**.** A *quantized gradient descent* algorithm $\mathrm{A} \in \mathcal{A}_{\mathrm{GD}}$ 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 $d(\mathsf{q}) \to 0$ (17) as $R \to \infty$ 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 $\hat{\boldsymbol{x}}_t$ to the worker noiselessly, starting from some $\hat{\boldsymbol{x}}_0 \in \mathbb{R}^n$. The worker then determines its gradient-access point $\boldsymbol{z}_t$ and its quantizer input $\boldsymbol{u}_t$ under the structural constraints

$$\boldsymbol{z}_t \in \hat{\boldsymbol{x}}_t + \mathrm{span}\left\{\boldsymbol{e}_0,\ldots,\boldsymbol{e}_{t-1}\right\} \tag{48}$$

$$\boldsymbol{u}_t \in \nabla \mathsf{f}(\boldsymbol{z}_t) + \mathrm{span}\left\{\boldsymbol{e}_0,\ldots,\boldsymbol{e}_{t-1}\right\}, \tag{49}$$

where $\boldsymbol{e}_i \triangleq \boldsymbol{q}_i - \boldsymbol{u}_i, i = 0,\ldots,t-1$ are the past quantization errors before iteration t, and $+$ denotes Minkowski's sum.

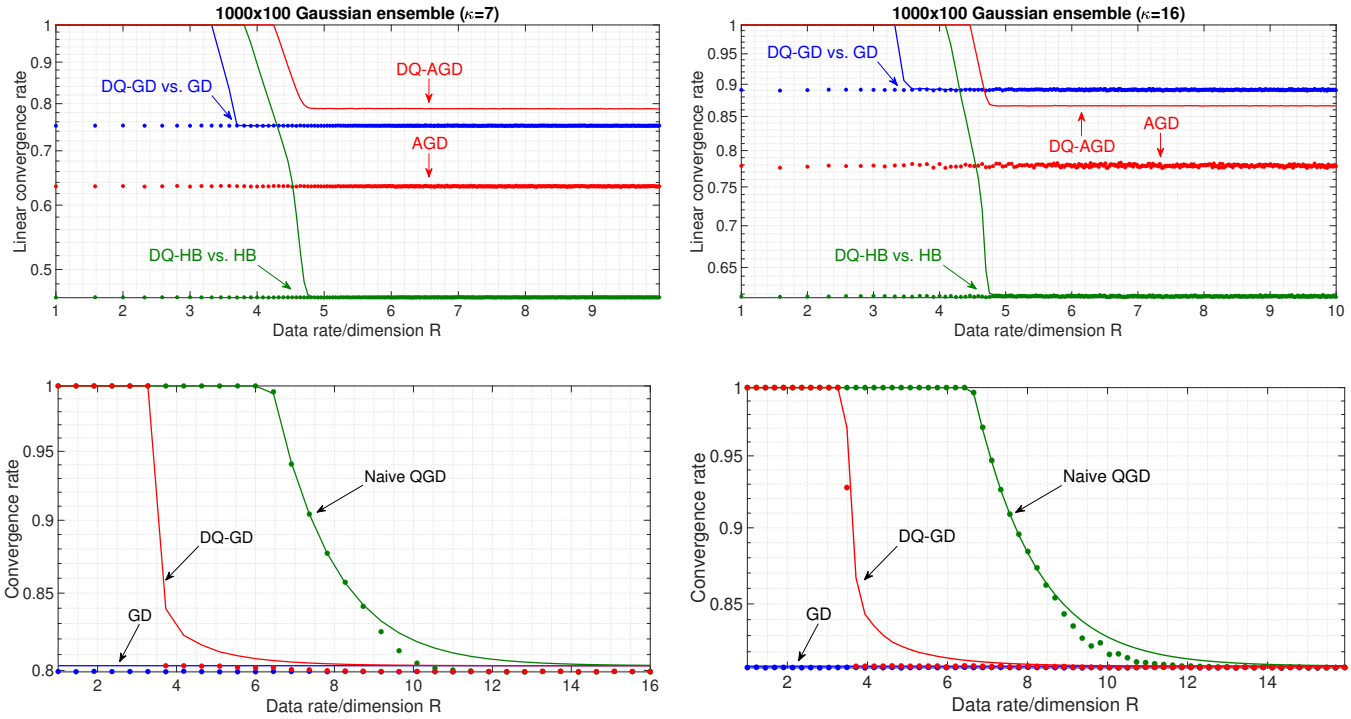

Fig. 4: Empirical contraction factors of various DQ algorithms (plotted as lines) and their corresponding unquantized counterparts (plotted as circles).

Upon receiving $\boldsymbol{q}_t = \mathsf{q}_t(\boldsymbol{u}_t)$ (7) from the worker, the server performs the update

$$\hat{\boldsymbol{x}}_{t+1} = \hat{\boldsymbol{x}}_t - \eta \boldsymbol{q}_t \tag{50}$$

with a fixed stepsize $\eta > 0$.

Due to conditions (48) and (49), if there is no quantization error at each iteration (i.e., if $R = \infty$), then any quantized algorithm in $\mathcal{A}_{\mathrm{GD}}$ reduces to the unquantized gradient descent. Both DQ-GD and NQ-GD fall in the class $\mathcal{A}_{\mathrm{GD}}$.

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

$$\inf_{A \in \mathcal{A}_{\mathrm{GD}}} \sigma_A(n, R) \geq \max\left\{\sigma_{\mathrm{GD}}, 2^{-R}\right\} \tag{51}$$

Proof sketch. We fix an $\mathrm{A} \in \mathcal{A}_{\mathrm{GD}}$, 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 [75] to craft a worst-case problem instance $\mathsf{g} \in \mathcal{F}_n$ for which the iterates of the unquantized GD satisfy $\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*_{\mathsf{g}}\right\| = \sigma_{\mathrm{GD}} \left\|\boldsymbol{x}_t - \boldsymbol{x}^*_{\mathsf{g}}\right\|$, which ensures that $\inf_{\mathrm{A} \in \mathcal{A}_{\mathrm{GD}}} \sigma_{\mathrm{A}}(n, R) \geq \sigma_{\mathrm{GD}}$ (Lemma C.1). On the other hand, we notice that if A is applied at dimension n and rate R, then the set $\mathcal{S}_{\mathrm{A}} \subseteq \mathbb{R}^n$ of all possible locations of the iterate $\hat{\boldsymbol{x}}_T$ after T iterations of A has cardinality at most 2^{nRT}, and we apply a volume-division argument to claim that $\inf_{\mathrm{A} \in \mathcal{A}_{\mathrm{GD}}} \sigma_{\mathrm{A}}(n, R) \geq 2^R$. See Appendix C-B for details. □

Applying Theorem III.1 with Rogers-optimal quantizers with $\rho_n \to 1$ [68, 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

$$\lim_{n \to \infty} \inf_{\mathrm{A} \in \mathcal{A}_{\mathrm{GD}}} \sigma_{\mathrm{A}}(n, R) = \max\left\{\sigma_{\mathrm{GD}}, 2^{-R}\right\}. \tag{52}$$

In other words, DQ-DG achieves the best possible contraction factor within $\mathcal{A}_{\mathrm{GD}}$, 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 (52) 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 (39) dissappears.

Although the notion of a contraction factor (1) and thus the result in (52) are asymptotic in the number of iterations T, the achievability results in Appendix B-A and converse results in Appendix B-A used to derive (52) are nonasymptotic. They show that gap between the achievability and converse bounds on the finite-T counterpart of (52) is $O\left(\frac{1}{T}\right)$. Whether DQ-GD remains optimal at finite T remains an open problem.

B. Quantized gradient methods

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

Definition IV.2 (Class $\mathcal{A}_{\rm GM}$ of quantized gradient methods)**.** A *quantized gradient method* $\mathrm{A} \in \mathcal{A}_{\rm GM}$ follows Definition IV.1 with (50) relaxed to

$$\hat{\boldsymbol{x}}_{t+1} \in \hat{\boldsymbol{x}}_0 + \operatorname{span}\{\boldsymbol{q}_0, \dots, \boldsymbol{q}_t\}. \tag{53}$$

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

$$\boldsymbol{x}_{t+1} \in \boldsymbol{x}_0 + \operatorname{span}\{\nabla \mathsf{f}(\boldsymbol{x}_0), \dots, \nabla \mathsf{f}(\boldsymbol{x}_t)\}. \tag{54}$$

To present our converse result for $\mathcal{A}_{\rm GM}$, we consider functions f defined on the square-summable Hilbert space[4]

$$\mathbb{L}_2 \triangleq \left\{\boldsymbol{x} = [\boldsymbol{x}(1), \boldsymbol{x}(2), \dots] \colon \textstyle\sum_{i=1}^{\infty} \boldsymbol{x}(i)^2 < \infty\right\}. \tag{55}$$

We say that continuously differentiable function $\mathsf{f} \colon \mathbb{L}_2 \mapsto \mathbb{R}$ is in class $\mathcal{F}_\infty$ 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 $\boldsymbol{u} \in \mathbb{L}_2$ to $\boldsymbol{q} \in \mathbb{L}_2$, 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 $\boldsymbol{u}$, and set the remaining coordinates to 0, i.e.,

$$\begin{cases} [\boldsymbol{q}(1), \dots, \boldsymbol{q}(n)] & = \mathsf{q}\left([\boldsymbol{u}(1), \dots, \boldsymbol{u}(n)]\right) \\ \boldsymbol{q}(i) & = 0 \quad \forall i > n, \end{cases} \tag{56}$$

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

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

$$\sum_{i>n} \boldsymbol{u}(i)^2 = o_n(1) \tag{57}$$

due to the Cauchy convergence criterion. Here $o_n(1)$ denotes a function that vanishes as $n \to \infty$. 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 $\mathsf{f} \in \mathcal{F}_\infty$.

Contraction factor $\sigma_{\rm A}(n, R)$ over $\mathcal{F}_\infty$ 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).

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

$$\inf_{A \in \mathcal{A}_{\rm GM}} \sigma_A(n, R) \geq \max\left\{\sigma_{\rm HB}, 2^{-R}\right\} \tag{58}$$

where $\sigma_{\rm HB}$ is given in (45).

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 (58), and we apply a known result on unquantized gradient methods that states that the best contraction factor achievable over $\mathcal{F}_\infty$ is that of the heavy ball method, $\sigma_{\rm HB}$ [66] (Lemma C.2). See Appendix C-C. □

[4]We do so to take advantage of the sharpest converse in the literature on the convergence of unquantized gradient methods (54) [66] (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 [76]. However, [76] considers only quadratic functions as objectives, and the considered class of iterative algorithms is more restrictive than (54) in that the next iterate $\boldsymbol{x}_{t+1}$ depends on the past p terms $\boldsymbol{x}_t, \dots, \boldsymbol{x}_{t-p+1}$ for some fixed $p \in \mathbb{N}$.

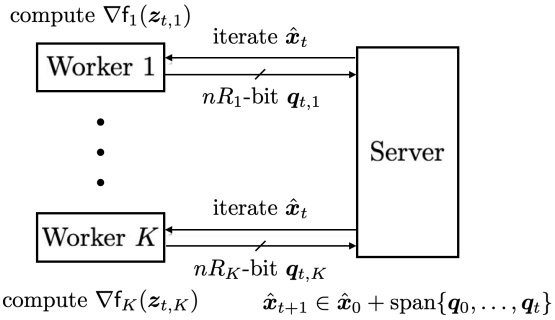

Fig. 5: K-worker quantized gradient method. At each iteration t, the server broadcasts the current iterate $\hat{\boldsymbol{x}}_t$. Worker k computes the gradient at some point $\boldsymbol{z}_{t,k}$ that is a function of (but not necessarily equal to) $\hat{\boldsymbol{x}}_t$. Then, worker k forms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

Together, Theorems IV.2 and III.3 imply that for any $R \geq R_2(\infty, \sigma_{\rm HB}, \gamma_{\rm HB})$, DQ-HB attains the optimal contraction factor within $\mathcal{A}_{\rm GM}$ (under the additional assumption that $\mathsf{f} \in \mathcal{F}_\infty$ is twice continuously differentiable). The nonasymptotic achievability and converse results in Appendices B-C and C.2 used to show this result leave open the question of whether DQ-HB is optimal at finite number of iterations T, as they determine the finite-length analog of the optimal contraction factor only with accuracy $O\left(\frac{\log T}{T}\right)$.

V. Multi-worker gradient methods

A. Problem setup

In empirical risk minimization [77], [78], the sample average of the loss function on the data points

$$\mathsf{f}(\boldsymbol{x}) = \frac{1}{K} \sum_{k=1}^{K} \mathsf{f}_k(\boldsymbol{x}) \tag{59}$$

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 (59), and they communicate to the parameter server under rate constraints. See Figure 5.

B. Converses

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

$$\sum_{k=1}^{K} R_k \leq R. \tag{60}$$

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 (51) and (58) 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 [15]–[18], [55], [69] 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

for $t = 0, 1, 2, \dots$ **do**
  for $k = 1$ **to** K **do**
    Worker k**:**
      $\boldsymbol{q}_{t,k} = \mathsf{q}_{t,k}(\nabla \mathsf{f}_k(\hat{\boldsymbol{x}}_t))$
    Server: $\hat{\boldsymbol{x}}_{t+1} \leftarrow \hat{\boldsymbol{x}}_t - \frac{\eta}{K} \sum_{k=1}^K \boldsymbol{q}_{t,k}$
  end
end

Our convergence result for NQ-GD holds under the following assumptions. We assume that continuously differentiable summands f_k in (59) 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

$$L \triangleq \frac{1}{K} \sum_{k=1}^K L_k \quad \text{and} \quad \mu \triangleq \frac{1}{K} \sum_{k=1}^K \mu_k. \tag{61}$$

Further, we focus on the *interpolation setting* [79]–[81] that assumes (iv)

$$\boldsymbol{x}_{\mathsf{f}}^* = \boldsymbol{x}_{\mathsf{f}_k}^* \quad \forall k = 1, \dots, K, \tag{62}$$

where

$$\boldsymbol{x}_{\mathsf{f}_k}^* \triangleq \arg\min_{\boldsymbol{x} \in \mathbb{R}^n} \mathsf{f}_k(\boldsymbol{x}). \tag{63}$$

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 [81], and is implied We denote by $\mathcal{G}_n$ the class of functions f (59) that satisfy the assumptions (i)-(iv).

The minimum contraction factor achievable by K-worker NQ-GD under the sum rate constraint (60) is given by

$$\sigma_{\text{NQ-GD}}(n, R) \triangleq \inf_{\sum_{k=1}^K R_k \le R} \sup_{\mathsf{f} \in \mathcal{G}_n} \limsup_{T \to \infty} \left\| \hat{\boldsymbol{x}}_T(\{R_k\}_{k=1}^K) - \boldsymbol{x}_{\mathsf{f}}^* \right\|^{\frac{1}{T}}, \tag{64}$$

where $\hat{\boldsymbol{x}}_0(\{R_k\}_{k=1}^K), \hat{\boldsymbol{x}}_1(\{R_k\}_{k=1}^K), \hat{\boldsymbol{x}}_2(\{R_k\}_{k=1}^K), \dots$ is the sequence of iterates generated by NQ-GD (Algorithm 4) in response to $\mathsf{f} \in \mathcal{G}_n$ 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*

$$\mathsf{q}_{t,k}(\cdot) = r_{t,k} \mathsf{q}_k(\cdot / r_{t,k}), \tag{65}$$

where the dynamic ranges are given by

$$r_{t,k} = \left(\sigma_{\text{GD}} + \frac{\eta_{\text{GD}}\, \rho_n}{K} \sum_{k=1}^K \min\{\nu, L_k\} \right)^t L_k D, \tag{66}$$

and the optimum rate allocation is given by the waterfilling solution

$$R_k = |\log_2(L_k/\nu)|_+ \quad \textit{bits}, \tag{67}$$

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

$$\sum_{k=1}^K |\log_2(L_k/\nu)|_+ = R, \tag{68}$$

and $|\cdot|_+ \triangleq \max\{0, \cdot\}$. Then, Algorithm 4 with stepsize (20) achieves the following contraction factor over $\mathcal{G}_n$ (64):

$$\sigma_{\text{NQ-GD}}(n, R) \le \sigma_{\text{GD}} + \frac{\eta_{\text{GD}}\, \rho_n}{K} \sum_{k=1}^K \min\{\nu, L_k\}. \tag{69}$$

Proof. Appendix D. □

According to (67), 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 $L_k \equiv L$, (69) reduces to

$$\sigma_{\text{NQ-GD}}(n, R) \le \sigma_{\text{GD}} + \frac{2\kappa}{\kappa + 1} \frac{\rho_n}{2^{R/K}}. \tag{70}$$

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

Without assumption (62) that all summands share the minimizer, NQ-GD converges only to a neighborhood of $\boldsymbol{x}_{\mathsf{f}}^*$; the size of the neighborhood is controlled by the quantization error and vanishes as $R \to \infty$ (Theorem D.2 in Appendix D).

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 ((52)). 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 (40).

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), (35), (46)). 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 (35) and DQ-HB (46) 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 $R \geq R_2(n, \sigma_{\mathrm{HB}}, \gamma_{\mathrm{HB}})$ (40), it remains an open problem whether the contraction factor of 2^{-R} dictated by the converse (Theorem IV.2) is achievable in the regime $R_2(n, \sigma_{\mathrm{GD}}, 0) < R < R_2(n, \sigma_{\mathrm{HB}}, \gamma_{\mathrm{HB}})$ 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 $R \to \infty$, 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

The authors would like to thank Dr. Himanshu Tyagi for pointing out related works [19], [55]; Dr. Vincent Tan for bringing a known result on the worst-case contraction factor of unquantized GD [75] to our attention; Dr. Victor Kozyakin for a helpful discussion about joint spectral radius; and two anonymous reviewers for detailed comments.

Appendix A
DQ-GD with varying stepsize

See Algorithm 5, below.

Algorithm 5: DQ-GD with varying stepsize

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

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).

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

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \eta_{t-1}\boldsymbol{e}_{t-1}. \tag{71}$$

Proof. We prove (71) via mathematical induction.

- Base case: (71) holds for $t = 0$ since the starting location is the same.
- Inductive step: Suppose (71) 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

$$\boldsymbol{u}_t = \nabla f(\boldsymbol{x}_t) - \frac{\eta_{t-1}}{\eta_t}\boldsymbol{e}_{t-1}\,. \tag{72}$$

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

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t+1} &= \hat{\boldsymbol{x}}_t - \eta_t \boldsymbol{q}_t && (73)\\
&= \hat{\boldsymbol{x}}_t - \eta_t(\boldsymbol{u}_t + \boldsymbol{e}_t) && (74)\\
&= \hat{\boldsymbol{x}}_t - \eta_t\left(\nabla f(\boldsymbol{x}_t) - \frac{\eta_{t-1}}{\eta_t}\boldsymbol{e}_{t-1}\right) - \eta_t\boldsymbol{e}_t && (75)\\
&= \left[\boldsymbol{x}_t - \eta_t \nabla f(\boldsymbol{x}_t)\right] - \eta_t\boldsymbol{e}_t && (76)\\
&= \boldsymbol{x}_{t+1} - \eta_t\boldsymbol{e}_t, && (77)
\end{aligned}$$

where (76) is due to the induction hypothesis.

□

Relation (71) implies for the constant stepsize $\eta_t \equiv \eta$

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\| + \eta\,\|\boldsymbol{e}_{t-1}\| \tag{78}$$

We use the contraction factor of unquantized GD to bound the first term of (78):

Lemma B.2 (Convergence of GD [66, Theorem 2.1.15])**.** *For any L-smooth and μ-strongly convex function f on $\mathbb{R}^n$, GD (8) with stepsize (20) satisfies*

$$\|\boldsymbol{x}_t - \boldsymbol{x}_\mathsf{f}^*\| \le \sigma_{\mathrm{GD}}^t \|\boldsymbol{x}_0 - \boldsymbol{x}_\mathsf{f}^*\|, \tag{79}$$

where σ_{GD} is defined in (21).

We use the following bound on quantization error to bound the second term in (78):

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

$$\|\boldsymbol{e}_t\| \le r_t\, \rho_n 2^{-R} \tag{80}$$

$$\le \max\left\{\sigma_{\mathrm{GD}},\, \rho_n 2^{-R}\right\}^t LD\, \sigma_{\mathrm{GD}} b_t, \tag{81}$$

where

$$b_t \triangleq \begin{cases} \frac{\rho_n 2^{-R}}{|\sigma_{\mathrm{GD}} - \rho_n 2^{-R}|} & \sigma_{\mathrm{GD}} \neq \rho_n 2^{-R} \\ t+1 & \sigma_{\mathrm{GD}} = \rho_n 2^{-R} \end{cases} \tag{82}$$

Proof. Once we show that quantizer inputs $\{\boldsymbol{u}_t\}$ satisfy

$$\boldsymbol{u}_t \in \mathcal{B}(r_t), \tag{83}$$

(80) will follow from (26).

We prove (83) via induction.

- Base case: (83) holds for $t = 0$ since

$$\|\nabla \mathsf{f}(\hat{\boldsymbol{x}}_0) - \boldsymbol{e}_{-1}\| = \|\nabla \mathsf{f}(\hat{\boldsymbol{x}}_0)\| \tag{84}$$

$$\le L \|\hat{\boldsymbol{x}}_0 - \boldsymbol{x}_\mathsf{f}^*\| \tag{85}$$

$$\le LD, \tag{86}$$

  where (85) is due to $\nabla \mathsf{f}(\boldsymbol{x}_\mathsf{f}^*) = \boldsymbol{0}$ and L-smoothness (13), and (86) is due to assumption (15).
- Inductive step: Suppose (83) holds for iteration t. Applying triangle inequality and (71) to the expression in Line 5 yields

$$\|\boldsymbol{u}_{t+1}\| \le \|\nabla \mathsf{f}(\boldsymbol{x}_{t+1})\| + \|\boldsymbol{e}_t\|. \tag{87}$$

  The first term is bounded as

$$\|\nabla \mathsf{f}(\boldsymbol{x}_{t+1})\| \le L \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_\mathsf{f}^*\| \tag{88}$$

$$\le \sigma_{\mathrm{GD}}^{t+1} L \|\boldsymbol{x}_0 - \boldsymbol{x}_\mathsf{f}^*\| \tag{89}$$

$$\le \sigma_{\mathrm{GD}}^{t+1} LD, \tag{90}$$

  where (88) is due to $\nabla \mathsf{f}(\boldsymbol{x}_\mathsf{f}^*) = \boldsymbol{0}$ and L-smoothness (13); (89) is due to (79); and (90) is due to assumption (15). Quantization error term $\|\boldsymbol{e}_t\|$ in (87) is bounded by (26) due to the induction hypothesis. Plugging (90) and (26) into (87) gives

$$\|\boldsymbol{u}_{t+1}\| \le \sigma_{\mathrm{GD}}^{t+1} LD + r_t \rho_n 2^{-R} \tag{91}$$

$$= r_{t+1}, \tag{92}$$

  where (92) is due to the choice of the dynamic ranges (22).

This concludes the proof of (80). To establish (81), we unwrap recursion (22) as the geometric sum

$$r_t = LD \sum_{\tau=0}^{t} \sigma_{\mathrm{GD}}^\tau \left(\rho_n 2^{-R}\right)^{t-\tau} \tag{93}$$

$$= LD \cdot \begin{cases} \sigma_{\mathrm{GD}}^t \frac{1 - \left[\rho_n 2^{-R}/\sigma_{\mathrm{GD}}\right]^{t+1}}{1 - \rho_n 2^{-R}/\sigma_{\mathrm{GD}}} & \sigma_{\mathrm{GD}} \neq \rho_n 2^{-R} \\ t+1 & \sigma_{\mathrm{GD}} = \rho_n 2^{-R} \end{cases} \tag{94}$$

$$\le LD \cdot \begin{cases} \sigma_{\mathrm{GD}}^t \left(1 - \frac{\rho_n 2^{-R}}{\sigma_{\mathrm{GD}}}\right)^{-1} & \sigma_{\mathrm{GD}} > \rho_n 2^{-R} \\ \left(\rho_n 2^{-R}\right)^t \left(\frac{\rho_n 2^{-R}}{\sigma_{\mathrm{GD}}} - 1\right)^{-1} & \sigma_{\mathrm{GD}} < \rho_n 2^{-R}, \\ t+1 & \sigma_{\mathrm{GD}} = \rho_n 2^{-R} \end{cases} \tag{95}$$

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

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*

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_\mathsf{f}^*\| \le \max\left\{\sigma_{\mathrm{GD}},\, \rho_n 2^{-R}\right\}^t \left[1 + \eta_{\mathrm{GD}} L b_{t-1}\right] D. \tag{96}$$

Proof. Plugging (79) and (81) into (78), we obtain

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_\mathsf{f}^*\| \le \sigma_{\mathrm{GD}}^t D \tag{97}$$

$$+ \max\left\{\sigma_{\mathrm{GD}}, \rho_n 2^{-R}\right\}^{t-1} \eta_{\mathrm{GD}} LD\, \sigma_{\mathrm{GD}} b_{t-1},$$

which leads to (96) by an elementary weakening. □

Bound (23) in Theorem III.1 follows immediately by applying (96) 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 $\{\hat{\boldsymbol{y}}_t, \hat{\boldsymbol{x}}_t\}$ of Algorithm 2 and $\{\boldsymbol{y}_t, \boldsymbol{x}_t\}$ of AGD (10) starting at the same location $(\hat{\boldsymbol{y}}_0, \hat{\boldsymbol{x}}_0) = (\boldsymbol{y}_0, \boldsymbol{x}_0)$ are related as, $\forall t = 0, 1, 2, \ldots$,*

$$\hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \eta \boldsymbol{e}_{t-1} \tag{98}$$

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \eta \boldsymbol{e}_{t-1} - \eta\gamma \left(\boldsymbol{e}_{t-1} - \boldsymbol{e}_{t-2}\right). \tag{99}$$

Proof. We prove (98) and (99) by induction.

- Base case: (98) and (99) hold for $t = 0$ since both algorithms start at the same location.

- Inductive step: Line 9 yields

$$\begin{aligned}\hat{\boldsymbol{y}}_{t+1} &= \hat{\boldsymbol{x}}_t - \eta \boldsymbol{q}_t && (100)\\ &= \boldsymbol{x}_t - \eta\left[\boldsymbol{e}_{t-1} + \gamma\left(\boldsymbol{e}_{t-1} - \boldsymbol{e}_{t-2}\right)\right] && (101)\\ &\quad - \eta\left[\nabla \mathsf{f}(\boldsymbol{x}_t) - \left[\boldsymbol{e}_{t-1} + \gamma\left(\boldsymbol{e}_{t-1} - \boldsymbol{e}_{t-2}\right)\right] + \boldsymbol{e}_t\right]\\ &= \boldsymbol{y}_{t+1} - \eta \boldsymbol{e}_t, && (102)\end{aligned}$$

where (101) is due to induction hypothesis (99) and Lines 5 and 7, and (102) is due to (10). On the other hand, plugging (98) and (102) into Line 10 yields

$$\begin{aligned}\hat{\boldsymbol{x}}_{t+1} &= \hat{\boldsymbol{y}}_{t+1} + \gamma\left(\hat{\boldsymbol{y}}_{t+1} - \hat{\boldsymbol{y}}_t\right) && (103)\\ &= \boldsymbol{y}_{t+1} + \gamma\left(\boldsymbol{y}_{t+1} - \boldsymbol{y}_t\right) - \eta\left[\boldsymbol{e}_t + \gamma\left(\boldsymbol{e}_t - \boldsymbol{e}_{t-1}\right)\right]\\ &= \boldsymbol{x}_{t+1} - \eta\left[\boldsymbol{e}_t + \gamma\left(\boldsymbol{e}_t - \boldsymbol{e}_{t-1}\right)\right], && (104)\end{aligned}$$

where (104) is due to (11).

□

Via triangle inequality, relation (98) implies

$$\left\|\hat{\boldsymbol{y}}_t - \boldsymbol{x}_{\mathsf{f}}^*\right\| \leq \left\|\boldsymbol{y}_t - \boldsymbol{x}_{\mathsf{f}}^*\right\| + \eta\left\|\boldsymbol{e}_{t-1}\right\| \tag{105}$$

The following simple corollary to a known bound on the contraction factor of unquantized AGD controls the first term of (105):

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 $\boldsymbol{x}_0 = \boldsymbol{y}_0$) with stepsize (31) and momentum coefficient (32) satisfies*

$$\begin{aligned}\left\|\boldsymbol{y}_t - \boldsymbol{x}_{\mathsf{f}}^*\right\| &\leq \sigma_{\mathrm{AGD}}^t \sqrt{\kappa + 1}\left\|\boldsymbol{x}_0 - \boldsymbol{x}_{\mathsf{f}}^*\right\| && (106)\\ \left\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\right\| &\leq \sigma_{\mathrm{AGD}}^t \lambda\left\|\boldsymbol{x}_0 - \boldsymbol{x}_{\mathsf{f}}^*\right\|. && (107)\end{aligned}$$

Proof. According to [33, Theorem 3.18],

$$\mathsf{f}(\boldsymbol{y}_t) - \mathsf{f}(\boldsymbol{x}_{\mathsf{f}}^*) \leq \frac{L+\mu}{2}\sigma_{\mathrm{AGD}}^{2t}\left\|\boldsymbol{x}_0 - \boldsymbol{x}_{\mathsf{f}}^*\right\|^2. \tag{108}$$

Convergence bound (106) w.r.t. $\{\boldsymbol{y}_t\}$ is due to (108) and

$$\mathsf{f}(\boldsymbol{x}) - \mathsf{f}(\boldsymbol{x}_{\mathsf{f}}^*) \geq \frac{\mu}{2}\left\|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{f}}^*\right\|^2 \tag{109}$$

implied by μ-strong convexity [66, Theorem 2.1.8]. On the other hand, applying triangle inequality to (11) yields

$$\left\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^*\right\| \leq (1+\gamma)\left\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^*\right\| + \gamma\left\|\boldsymbol{y}_t - \boldsymbol{x}_{\mathsf{f}}^*\right\|, \tag{110}$$

and (107) then follows from (106) and (110). □

The following bound on the quantization error controls the second term in (78):

Lemma B.6 (DQ-AGD quantization error)**.** *Let $\mathsf{f} \in \mathcal{F}_n$. Quantization errors $\{\boldsymbol{e}_t\}$ in Algorithm 2 with stepsize (31), momentum coefficient (32) and dynamic ranges (35) satisfy*

$$\begin{aligned}\left\|\boldsymbol{e}_t\right\| &\leq r_t \rho_n 2^{-R} && (111)\\ &\leq \sigma_{\mathrm{AGD}}^t c_0 + \phi_+^t c_+ + \phi_-^t c_-, && (112)\end{aligned}$$

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

$$\phi_\pm(\gamma) \triangleq \rho_n 2^{-R}\left(\frac{1}{2}(1+\gamma) \pm \frac{1}{2}\sqrt{(1+\gamma)^2 + \frac{4\gamma}{\rho_n 2^{-R}}}\right) \tag{113}$$

and c_0, c_+, c_- are specified below in (118), (122) and (123) respectively.

Proof. Like in the proof of Lemma B.3, to establish (111), in view of (26) it is enough to show (83). Applying triangle inequality and (99) to the expression in Line 5 yields

$$\left\|\boldsymbol{u}_t\right\| \leq \left\|\nabla \mathsf{f}(\boldsymbol{x}_t)\right\| + \left\|\boldsymbol{e}_{t-1}\right\| + \gamma\left(\left\|\boldsymbol{e}_{t-1}\right\| + \left\|\boldsymbol{e}_{t-2}\right\|\right). \tag{114}$$

The first term in (114) is bounded as

$$\begin{aligned}\left\|\nabla \mathsf{f}(\boldsymbol{x}_t)\right\| &\leq L\left\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\right\| && (115)\\ &\leq \sigma_{\mathrm{AGD}}^t L D \lambda && (116)\end{aligned}$$

where (115) is due to L-smoothness (13), and (116) is due to (107). Plugging (116) into (114) and applying (26) to bound quantization error terms in (114), we conclude via induction (similar to the proof of Lemma B.3) that setting the sequence of dynamic ranges recursively as (35) ensures (83).

To show (112), we proceed to solve recursion (35). 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 (35) is a second-order linear non-homogeneous recurrence relation.

- Particular solution: Plugging the candidate

$$p_t = \sigma_{\mathrm{AGD}}^t c_0 \tag{117}$$

into (35), we solve for the constant

$$c_0 = \frac{\sigma_{\mathrm{AGD}}^2}{\mathsf{p}(\sigma_{\mathrm{AGD}})} L D \lambda, \tag{118}$$

where $\mathsf{p}(r)$ is the characteristic polynomial in (38) associated with recursion (35).

- Homogeneous solution: Since ϕ_+ and ϕ_- are roots of the quadratic polynomial in (38), the homogeneous solution is given by

$$\phi_t = \phi_+^t c_+ + \phi_-^t c_-, \tag{119}$$

- General solution: constants c_+, c_- in (119) are determined by plugging initial conditions $r_{-2} = r_{-1} = 0$ into the general solution

$$\begin{aligned}r_t &= p_t + \phi_t && (120)\\ &= \sigma_{\mathrm{AGD}}^t c_0 + \phi_+^t c_+ + \phi_-^t c_-, && (121)\end{aligned}$$

which are

$$\begin{aligned}c_+ &= -\frac{c_0 \phi_+^2}{\sigma_{\mathrm{AGD}}^2}\frac{\sigma_{\mathrm{AGD}} - \phi_-}{\phi_+ - \phi_-} && (122)\\ c_- &= \frac{c_0 \phi_-^2}{\sigma_{\mathrm{AGD}}^2}\frac{\sigma_{\mathrm{AGD}} - \phi_+}{\phi_+ - \phi_-}. && (123)\end{aligned}$$

□

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*

$$\left\|\hat{\boldsymbol{y}}_t - \boldsymbol{x}_{\mathsf{f}}^*\right\| \leq \sigma_{\mathrm{AGD}}^t c + \phi_+^{t-1} c_+ \eta + \phi_-^{t-1} c_- \eta \tag{124}$$

where $c = \sqrt{\kappa+1}D + \eta c_0 \sigma_{\text{AGD}}^{-1}$, *and* c_0, c_+, c_- *are specified in* (118), (122) *and* (123) *respectively.*

Proof. Plugging (106) and (112) into (105) immediately leads to (124). □

Note that if $\sigma_{\text{AGD}} \geq \phi_+$, which is equivalent to $R \geq R_2$ (40), 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 (36) in Theorem III.2 follows by plugging (124) 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* $\{\hat{\boldsymbol{x}}_t\}$ *of DQ-HB (Line 7) and path* $\{\boldsymbol{x}_t\}$ *of HB* (12) *starting at the same location* $\hat{\boldsymbol{x}}_{-1} = \boldsymbol{x}_{-1} = \hat{\boldsymbol{x}}_0 = \boldsymbol{x}_0$ *are related as,* $\forall t = 0, 1, 2, \ldots$,

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \eta \boldsymbol{e}_{t-1}. \tag{125}$$

Proof. We prove (125) via induction.

- Base case: (125) holds for $t = 0$ by the initialization $\boldsymbol{e}_{-2} = \boldsymbol{e}_{-1} = \boldsymbol{0}$ 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

$$\begin{aligned}
&\hat{\boldsymbol{x}}_{t+1} && (126)\\
&= \hat{\boldsymbol{x}}_t - \eta \boldsymbol{q}_t + \gamma\left(\hat{\boldsymbol{x}}_t - \hat{\boldsymbol{x}}_{t-1}\right) && (127)\\
&= \boldsymbol{x}_t - \eta \boldsymbol{e}_{t-1} \\
&\quad - \eta\left[\nabla \mathsf{f}(\boldsymbol{x}_t) - \left[\boldsymbol{e}_{t-1} + \gamma\left(\boldsymbol{e}_{t-1} - \boldsymbol{e}_{t-2}\right)\right] + \boldsymbol{e}_t\right] \\
&\quad + \gamma\left(\boldsymbol{x}_t - \boldsymbol{x}_{t-1} - \eta\left(\boldsymbol{e}_{t-1} - \boldsymbol{e}_{t-2}\right)\right) && (128)\\
&= \boldsymbol{x}_t - \eta \nabla \mathsf{f}(\boldsymbol{x}_t) + \gamma\left(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\right) - \eta \boldsymbol{e}_t && (129)\\
&= \boldsymbol{x}_{t+1} - \eta \boldsymbol{e}_t, && (130)
\end{aligned}$$

  where (128) is due to the induction hypothesis and (130) is due to (12).

□

Applying triangle inequality to (125) gives

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\| + \eta \|\boldsymbol{e}_{t-1}\|. \tag{131}$$

The convergence result for unquantized HB in Lemma B.8, below, controls the first term in the right side of (131). Lemma B.8 is a nonasymptotic refinement of Polyak's original convergence result [67, Th. 1, Sec. 3.2]. Unlike the original, it does not require that the algorithm starts "sufficiently close" to the minimizer (i.e., it establishes global rather than local convergence), and it also clarifies that the subexponential factor in the bound is polynomial in t (see (132), below). This refinement is made possible by a result on joint spectral radius due to Wirth [82] that is more recent than [67, Th. 1, Sec. 3.2].

Lemma B.8 (Convergence of HB)**.** *For any* L*-smooth,* μ*-strongly convex, twice continuously differentiable function* f *on* $\mathbb{R}^n$*, there exists a constant* $\alpha > 0$ *such that the HB algorithm* (12) *starting at* $\boldsymbol{x}_{-1} = \boldsymbol{x}_0$ *with stepsize* (43) *and momentum coefficient* (44) *satisfies*

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \sigma_{\text{HB}}^t \, t^\alpha e^\alpha \sqrt{2} \|\boldsymbol{x}_0 - \boldsymbol{x}_{\mathsf{f}}^*\|. \tag{132}$$

Proof. Iterative process (12) can be written in the form

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{x}_t + \gamma(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix} - \eta \begin{bmatrix} \nabla \mathsf{f}(\boldsymbol{x}_t) \\ 0 \end{bmatrix} && (133)\\
&= \begin{bmatrix} (1+\gamma)\boldsymbol{I} & -\gamma \boldsymbol{I} \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_{t-1} - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix} - \eta \begin{bmatrix} \nabla^2 \mathsf{f}(\boldsymbol{v}_t)(\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*) \\ 0 \end{bmatrix} && (134)\\
&= \begin{bmatrix} (1+\gamma)\boldsymbol{I} & -\gamma \boldsymbol{I} - \eta \nabla^2 \mathsf{f}(\boldsymbol{v}_t) \\ \boldsymbol{I} & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_{t-1} - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix}, && (135)
\end{aligned}$$

where (134) holds for some $\boldsymbol{v}_t$ on the line segment between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{\mathsf{f}}^*$ by the mean value theorem, since f is twice continuously differentiable by the assumption. Denoting the matrix in (135) by $\boldsymbol{A}_t$, we unroll the recursion as

$$\begin{bmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix} = \boldsymbol{A}_t \cdot \boldsymbol{A}_{t-1} \cdot \ldots \cdot \boldsymbol{A}_0 \begin{bmatrix} \boldsymbol{x}_0 - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_{-1} - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix}. \tag{136}$$

It follows that

$$\left\| \begin{bmatrix} \boldsymbol{x}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix} \right\|_2 \leq \|\boldsymbol{A}_t \cdot \ldots \cdot \boldsymbol{A}_0\|_2 \left\| \begin{bmatrix} \boldsymbol{x}_0 - \boldsymbol{x}_{\mathsf{f}}^* \\ \boldsymbol{x}_{-1} - \boldsymbol{x}_{\mathsf{f}}^* \end{bmatrix} \right\|_2. \tag{137}$$

It is shown in [82, Lemma 2.3] that if matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_t$ all belong to a bounded set $\mathcal{A}$, then there exists a constant $\alpha > 0$ such that $\forall t = 0, 1, \ldots$,

$$\|\boldsymbol{A}_t \cdot \ldots \cdot \boldsymbol{A}_0\|_2 \leq \rho(\mathcal{A})^{t+1} (t+1)^\alpha e^\alpha \tag{138}$$

where

$$\rho(\mathcal{A}) \triangleq \limsup_{t \to \infty} \sup_t \rho(\boldsymbol{A}_t), \tag{139}$$

where $\rho(\boldsymbol{A}_t)$ is the spectral radius of $\boldsymbol{A}_t$. It is shown in [67, Proof of Th. 1, Sec. 3.2] that if

$$\gamma = \max\left\{ \left(1 - \sqrt{\eta L}\right)^2, \ \left(1 - \sqrt{\eta \mu}\right)^2 \right\}, \tag{140}$$

then

$$\rho(\boldsymbol{A}_t) \leq \sqrt{\gamma}. \tag{141}$$

With the optimal choice of η (43) the right side of (141) is equal to σ_{HB} (45).

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

$$\mu \boldsymbol{I} \preceq \nabla^2 \mathsf{f}(\boldsymbol{v}) \preceq L \boldsymbol{I}, \tag{142}$$

in the positive semidefinite order, thus (138) applies. Inequality (132) follows after applying (141) to (138) and the latter to (137). □

Remark B.1. Polyak's convergence result [67, Th. 1, Sec. 3.2] guarantees only the existence of $D > 0$ such that for all starting points $\boldsymbol{x}_{-1}, \boldsymbol{x}_0$ with $\|\boldsymbol{x}_{\mathsf{f}}^* - \boldsymbol{x}_{-1}\| \leq D$, $\|\boldsymbol{x}_{\mathsf{f}}^* - \boldsymbol{x}_0\| \leq D$

and all $0 < \epsilon < 1 - \sigma_{\mathrm{HB}}$, there exists a $c > 0$ such that (cf. (132))

$$\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \leq c(\sigma_{\mathrm{HB}} + \epsilon)^t \tag{143}$$

under the same assumptions on f and the same stepsize and momentum coefficient as in Lemma B.8. This is a local convergence result because convergence is not guaranteed for any starting location but only for locations in a small enough neighbourhood of $\boldsymbol{x}_{\mathsf{f}}^*$. Furthermore, (132) refines the subexponential factor in (143).

We ensure that the quantization error term $\|\boldsymbol{e}_{t-1}\|$ in (131) decays exponentially fast by adjusting the sequence of dynamic ranges (7) carefully:

Lemma B.9 (DQ-HB quantization error)**.** *Let $\mathsf{f} \in \mathcal{F}_n^2$. Quantization errors $\{\boldsymbol{e}_t\}$ in Algorithm 3 with with stepsize (43), momentum coefficient (44) and dynamic ranges (46) satisfy*

$$\|\boldsymbol{e}_t\| \leq r_t \rho_n 2^{-R} \tag{144}$$

$$\leq \left(\sigma_{\mathrm{HB}}^t c_0 + \phi_+^t c_+ + \phi_-^t c_-\right) t^\alpha, \tag{145}$$

where $\phi_\pm \triangleq \phi_\pm(\gamma_{\mathrm{HB}})$ (113), and c_0, c_+, c_- are as in (118), (122) and (123) respectively, with $LD\lambda$ replaced by $e^\alpha\sqrt{2}\,LD$, and σ_{AGD} by σ_{HB}.

Proof. Applying triangle inequality and (125) to the expression in Line 5 yields the same expression as in the analysis of DQ-AGD (114):

$$\|\boldsymbol{u}_t\| \leq \|\nabla \mathsf{f}(\boldsymbol{x}_t)\| + \|\boldsymbol{e}_{t-1}\| + \gamma\left(\|\boldsymbol{e}_{t-1}\| + \|\boldsymbol{e}_{t-2}\|\right). \tag{146}$$

The first term in (146) is bounded as

$$\|\nabla \mathsf{f}(\boldsymbol{x}_t)\| \leq L\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \tag{147}$$

$$\leq \sigma_{\mathrm{HB}}^t\, t^\alpha e^\alpha \sqrt{2}\, LD \tag{148}$$

where (115) is due to L-smoothness (13), and (148) is due to (132). Applying the argument used to show (111) in the proof of Lemma B.6 leads to (144).

To show (145), consider the recursion $r'_{-1} = r'_{-2} = 0$,

$$r'_t = \sigma_{\mathrm{HB}}^t\, e^\alpha \sqrt{2}\, LD + \left(r'_{t-1} + \gamma_{\mathrm{HB}}(r'_{t-1} + r_{t-2})\right)\rho_n 2^{-R}, \tag{149}$$

We show by strong induction that

$$r_t \leq t^\alpha r'_t, \tag{150}$$

where r_t solves (46). Base case $r_0 = r'_0$ holds by the initial conditions. Assuming that (150) holds for $1, \ldots, t-1$, we establish (150) for t using (46) and the fact that t^α is increasing in t:

$$r_t \leq \sigma_{\mathrm{HB}}^t\, t^\alpha e^\alpha \sqrt{2}\, LD \tag{151}$$

$$+\left((t-1)^\alpha r'_{t-1} + \gamma_{\mathrm{HB}}((t-1)^\alpha r'_{t-1} + (t-2)^\alpha r'_{t-2})\right)\rho_n 2^{-R}$$

$$\leq t^\alpha r'_t. \tag{152}$$

Using (144) and (150), we can show (145) by solving the recursion (149). But this is the same recursion as in (35), up to the constants, thus the solution in the proof of Lemma B.6 applies. □

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*

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \left(\sigma_{\mathrm{HB}}^t c + \phi_+^{t-1} c_+ + \phi_-^{t-1} c_-\right) t^\alpha \tag{153}$$

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

Proof. Plugging (132) and (145) into (131) and using $(t-1)^\alpha < t^\alpha$ leads to (153). □

Appendix C
Converses

A. Converse for unquantized GD

As explained in the proof sketch, the lower bound in (51) 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 $\boldsymbol{x}_0$ and any constant stepsize η. Then, there exists a problem instance $\mathsf{f} \in \mathcal{F}_n$ such that the distance to the optimizer at each iteration t of GD satisfies*

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^*\| \geq \sigma_{\mathrm{GD}}\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\|, \tag{154}$$

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

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

$$\|\boldsymbol{x}_0 - \boldsymbol{x}_{\mathsf{f}}^*\| \geq D \tag{155}$$

and then construct a least-squares problem instance $\mathsf{f} \in \mathcal{F}_n$ (29) that admits $\boldsymbol{x}_{\mathsf{f}}^*$ as a unique minimizer and satisfies (154). Note that f is

$$\sigma_1^2(\mathbf{A})\text{-smooth} \quad \text{and} \quad \sigma_n^2(\mathbf{A})\text{-strongly convex} \tag{156}$$

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

$$\nabla \mathsf{f}(\boldsymbol{x}_t) = \mathbf{A}^\top(\mathbf{A}\boldsymbol{x}_t - \boldsymbol{y}). \tag{157}$$

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

$$\mathbf{A}^\top \boldsymbol{y} = \mathbf{A}^\top \mathbf{A}\boldsymbol{x}_{\mathsf{f}}^*. \tag{158}$$

To each $\boldsymbol{x}_{\mathsf{f}}^* \in \mathcal{B}(D)$ that satisfies (155), there corresponds a $\boldsymbol{y} \in \mathbb{R}^m$ such that (158) holds. This is because $m \geq n$, i.e. we have more degrees of freedom than the problem dimension when selecting the vector $\boldsymbol{y}$. Since we can always select an $\mathbf{A}$ with $\sigma_1(\mathbf{A}) = \sqrt{L}$ and $\sigma_n(\mathbf{A}) = \sqrt{\mu}$ and a $\boldsymbol{y}$ to ensure (158), we have $\mathsf{f} \in \mathcal{F}_n$. It remains to show how to set the right singular vectors of $\mathbf{A}$ to ensure (154).

Plugging (157) into (8) yields

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \mathbf{A}^\top \mathbf{A}(\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*). \tag{159}$$

Subtracting $\boldsymbol{x}_{\mathsf{f}}^*$ from both sides of (159), we conclude that the distance to the optimizer $\boldsymbol{x}_{\mathsf{f}}^*$ satisfies

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \sigma_1\left(\mathbf{I} - \eta \mathbf{A}^\top \mathbf{A}\right)\|\boldsymbol{x}_t - \boldsymbol{x}_{\mathsf{f}}^*\|, \tag{160}$$

where equality is achieved if $\boldsymbol{x}_t - \boldsymbol{x}_\mathsf{f}^*$ points in the direction corresponding to the largest singular vector of the matrix $\mathbf{I} - \eta\mathbf{A}^\top\mathbf{A}$. Since

$$\sigma_1\left(\mathbf{I} - \eta\mathbf{A}^\top\mathbf{A}\right) = \max\left\{\left|1-\eta\sigma_n^2(\mathbf{A})\right|, \left|1-\eta\sigma_1^2(\mathbf{A})\right|\right\}, \tag{161}$$

we designate the unit vector

$$\boldsymbol{v}_1 \triangleq \frac{\boldsymbol{x}_0 - \boldsymbol{x}_\mathsf{f}^*}{\|\boldsymbol{x}_0 - \boldsymbol{x}_\mathsf{f}^*\|} \tag{162}$$

as the right singular vector of $\mathbf{A}$ corresponding to either $\sigma_1(\mathbf{A})$ if $\left|1-\eta\sigma_1^2(\mathbf{A})\right|$ achieves the maximum in (161) or $\sigma_n(\mathbf{A})$ otherwise. This is determined solely by the stepsize η; the optimal stepsize (20) ensures that $\left|1-\eta\sigma_1^2(\mathbf{A})\right| = \left|1-\eta\sigma_n^2(\mathbf{A})\right| = \sigma_{\mathrm{GD}}$. To complete the construction of $\mathbf{A}$, we complement $\boldsymbol{v}_1$ with $n-1$ orthonormal vectors to form an orthonormal basis $\{\boldsymbol{v}_i\}_{i=1}^n$ of $\mathbb{R}^n$. Then,

$$\boldsymbol{x}_1 - \boldsymbol{x}_\mathsf{f}^* = \sigma_1\left(\mathbf{I} - \eta\mathbf{A}^\top\mathbf{A}\right)\left(\boldsymbol{x}_0 - \boldsymbol{x}_\mathsf{f}^*\right), \tag{163}$$

and using (159) it is easy to show by induction that

$$\boldsymbol{x}_{t+1} - \boldsymbol{x}_\mathsf{f}^* = \sigma_1\left(\mathbf{I} - \eta\mathbf{A}^\top\mathbf{A}\right)\left(\boldsymbol{x}_t - \boldsymbol{x}_\mathsf{f}^*\right), \tag{164}$$

which implies that (160) holds with equality $\forall t = 0, 1, \ldots$. □

Remark C.1. While variants of Lemma C.1 are known in the literature (e.g. [75, 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 [75, Example 1.3] is tailored to a particular starting point $\boldsymbol{x}_0 \neq \mathbf{0}$ and the optimal η (20).

B. Proof of Theorem IV.1

On one hand, we have

$$\inf_{\mathsf{A}\in\mathcal{A}_{\mathrm{GD}}} \sigma_{\mathsf{A}}(n,R) \geq \inf_{\mathsf{A}\in\mathcal{A}_{\mathrm{GD}}} \sigma_{\mathsf{A}}(n,\infty) \tag{165}$$

$$\geq \sigma_{\mathrm{GD}}, \tag{166}$$

where (165) holds because the left side of (165) is non-increasing in the data rate R by definition (1), and (166) is by Lemma C.1 since the infinite-rate algorithm in $\mathcal{A}_{\mathrm{GD}}$ is the one incurring no quantization error at each iteration, i.e., the GD itself.

On the other hand, to show

$$\inf_{\mathsf{A}\in\mathcal{A}_{\mathrm{GD}}} \sigma_{\mathsf{A}}(n,R) \geq 2^{-R}, \tag{167}$$

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

$$\mathcal{S}_{\mathsf{A}} \triangleq \left\{\hat{\boldsymbol{x}}_t \in \mathbb{R}^n \mid \hat{\boldsymbol{x}}_t \text{ is the output of } \mathsf{A} \text{ after } T \text{ iterations}\right\} \tag{168}$$

has cardinality

$$|\mathcal{S}_{\mathsf{A}}| = 2^{nR'T}. \tag{169}$$

Given $\mathcal{S}_{\mathsf{A}}$, consider the minimum-distance quantizer q_{A}

$$\mathsf{q}_{\mathsf{A}}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\hat{\boldsymbol{x}}\in\mathcal{S}_{\mathsf{A}}} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\| \tag{170}$$

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

$$\frac{D}{\mathsf{d}\left(\mathsf{q}_{\mathsf{A}}\right)} \leq 2^{R'T}. \tag{171}$$

(This classical volume-division argument also shows that $\rho\left(\mathsf{q}_{\mathsf{A}}\right) \geq 1$ (19)).

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

$$\|\mathsf{q}_{\mathsf{A}}(\boldsymbol{x}_\mathsf{f}^*) - \boldsymbol{x}_\mathsf{f}^*\| = \mathsf{d}\left(\mathsf{q}_{\mathsf{A}}\right), \tag{172}$$

(167) follows by rearranging (171) and taking the limit $T \to \infty$.

C. Proof of Theorem IV.2

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

Lemma C.2 ([66, Th. 2.1.13])**.** *For any gradient method* (54)*, there exists an L-smooth and μ-strongly convex function $\mathsf{f}\colon \mathbb{L}_2 \to \mathbb{R}$ such that, $\forall t = 0, 1, \ldots$,*

$$\|\boldsymbol{x}_t - \boldsymbol{x}_\mathsf{f}^*\| \geq \sigma_{\mathrm{HB}}^t \|\boldsymbol{x}_0 - \boldsymbol{x}_\mathsf{f}^*\|. \tag{173}$$

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 $t = 0, 1, 2, \ldots$, the path of NQ-GD with stepsize* (20) *satisfies*

$$\|\hat{\boldsymbol{x}}_{t+1} - \boldsymbol{x}_\mathsf{f}^*\| \leq \sigma_{\mathrm{GD}} \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_\mathsf{f}^*\| + \frac{\eta_{\mathrm{GD}}}{K}\sum_{k=1}^K \|\boldsymbol{e}_{t,k}\|, \tag{174}$$

where $\boldsymbol{e}_{t,k} \triangleq \boldsymbol{q}_{t,k} - \nabla\mathsf{f}_k(\hat{\boldsymbol{x}}_t)$, 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

$$\hat{\boldsymbol{x}}_{t+1} = \hat{\boldsymbol{x}}_t - \frac{\eta_{\mathrm{GD}}}{K}\sum_{k=1}^K \left(\nabla\mathsf{f}_k(\hat{\boldsymbol{x}}_t) + \boldsymbol{e}_{t,k}\right) \tag{175}$$

$$= \hat{\boldsymbol{x}}_t - \eta_{\mathrm{GD}}\nabla\mathsf{f}(\hat{\boldsymbol{x}}_t) - \frac{\eta_{\mathrm{GD}}}{K}\sum_{k=1}^K \boldsymbol{e}_{t,k} \tag{176}$$

where (176) is due to (59). Triangle inequality now implies

$$\|\hat{\boldsymbol{x}}_{t+1} - \boldsymbol{x}_\mathsf{f}^*\| \tag{177}$$

$$\leq \|\hat{\boldsymbol{x}}_t - \eta_{\mathrm{GD}}\nabla\mathsf{f}(\hat{\boldsymbol{x}}_t) - \boldsymbol{x}_\mathsf{f}^*\| + \frac{\eta_{\mathrm{GD}}}{K}\sum_{k=1}^K \|\boldsymbol{e}_{t,k}\|. \tag{178}$$

Applying the coercive property of smooth and strongly convex functions [66, Th. 2.1.12] in the same manner as in [66, Proof of Th. 2.1.5] to further upper-bound the first term in (178) gives (174). □

Denote for brevity

$$\sigma \triangleq \sigma_{\mathrm{GD}} + C, \tag{179}$$

$$C \triangleq \frac{\eta_{\mathrm{GD}} \rho_n}{K} \sum_{k=1}^{K} \frac{L_k}{2^{R_k}}. \tag{180}$$

Minimizing (179) under the sum rate constrant (60) is a convex optimization problem whose solution is given by (67) and whose optimal value is given by the right side of (69). 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 (174).

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

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \sigma^t D, \tag{181}$$

where σ is defined in (179).

Proof. We prove (181) by induction.

- Base case: for $t = 0$, (181) holds by (15).
- Inductive step: Suppose (181) holds for iteration t. Then, by L_k-smoothness of f_k, (62) and the inductive hypothesis,

$$\|\nabla \mathsf{f}_k(\hat{\boldsymbol{x}}_t)\| \leq L_k \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \tag{182}$$

$$\leq \sigma^t L_k D, \tag{183}$$

which due to the choice of dynamic ranges (66) and (26) leads to

$$\|\boldsymbol{e}_{t,k}\| \leq \frac{\rho_n}{2^{R_k}} \sigma^t L_k D. \tag{184}$$

Applying (184) and the inductive hypothesis to (174) yields

$$\|\hat{\boldsymbol{x}}_{t+1} - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \sigma^t \left[\sigma_{\mathrm{GD}} + C\right] D, \tag{185}$$

which is exactly (181) for $t+1$ if the optimal rate allocation is employed.

□

Without assumption (iv) (62) that the summands f_k share the minimizer, NQ-GD converges only to a neighborhood of $\boldsymbol{x}_{\mathsf{f}}^*$, albeit exponentially fast. The radius of this neighborhood vanishes as the data rate $R \to \infty$:

Theorem D.2. *In the setting of Theorem V.1 but without assumption (iv) (62) on the objective function, setting the dynamic ranges to $r'_{t,k}$, where*

$$r'_{t,k} = r_{t,k} + 2L_k D \left(C \sum_{\tau=0}^{t-1} \sigma^\tau + 1 \right), \tag{186}$$

$r_{t.k}$ is defined in (66), and C is defined in (180), NQ-GD achieves

$$\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| \leq \sigma^t D + 2DC \sum_{\tau=0}^{t-1} \sigma^\tau, \tag{187}$$

where σ is defined in (179).

Proof. We follow the reasoning in the proof of Theorem D.1 until (182), which we replace by

$$\|\nabla \mathsf{f}_k(\hat{\boldsymbol{x}}_t)\| \leq L_k \left\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}_k}^*\right\| \tag{188}$$

$$\leq L_k \left(\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| + \left\|\boldsymbol{x}_{\mathsf{f}_k}^* - \boldsymbol{x}_{\mathsf{f}}^*\right\| \right) \tag{189}$$

$$\leq L_k \left(\|\hat{\boldsymbol{x}}_t - \boldsymbol{x}_{\mathsf{f}}^*\| + 2D \right) \tag{190}$$

$$\leq L_k \left(\sigma^t D + 2DC \sum_{\tau=0}^{t-1} \sigma^\tau + 2D \right), \tag{191}$$

where (191) applies inductive hypothesis (187). Due to (191), the choice of dynamic ranges (186) ensures that the quantizer input stays within its dynamic range, which limits the quantization error to $\rho_n 2^{-R_k}$ times the right-hand side of (191) (recall (26)). Plugging this bound on quantization error in (174) and applying inductive hypothesis (187) again establishes (187) for $t+1$.

□

Appendix E
Analysis of quantized GD with error feedback of [10], [22], [44]

We derive the following convergence guarantee on quantized GD with the error feedback mechanism of [10], [22], [44].

Lemma E.1 (Trajectory of GD with error feedback of [10], [22], [44])**.** *Let f be an L-smooth and μ-strongly convex function on $\mathbb{R}^n$. Then, the distance to the optimizer at each iteration $t \in \mathbb{N}$ of quantized gradient descent (50) with quantizer input (30) and stepsize (20) is bounded as*

$$\|\hat{\boldsymbol{x}}_{t+1} - \boldsymbol{x}^*\| \leq \sigma_{GD} \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}^*\| + \eta_{GD} \|\boldsymbol{e}_t\| + (\sigma_{GD} + \eta_{GD} L)\eta \|\boldsymbol{e}_{t-1}\|. \tag{192}$$

Proof. Using (30), the quantizer output can be computed as

$$\boldsymbol{q}_t = \left[\nabla \mathsf{f}(\hat{\boldsymbol{x}}_t) + \boldsymbol{e}_t\right] - \boldsymbol{e}_{t-1}, \tag{193}$$

and plugging (193) into (50) gives

$$\hat{\boldsymbol{x}}_{t+1} = \hat{\boldsymbol{x}}_t - \eta \nabla \mathsf{f}(\hat{\boldsymbol{x}}_t) - \eta \boldsymbol{e}_t + \eta \boldsymbol{e}_{t-1}. \tag{194}$$

Denoting a shifted trajectory by

$$\tilde{\boldsymbol{x}}_t \triangleq \hat{\boldsymbol{x}}_t + \eta \boldsymbol{e}_{t-1}, \tag{195}$$

we rewrite (194) as

$$\tilde{\boldsymbol{x}}_{t+1} = \tilde{\boldsymbol{x}}_t - \eta \nabla \mathsf{f}(\tilde{\boldsymbol{x}}_t - \eta \boldsymbol{e}_{t-1}), \tag{196}$$

which via triangle inequality implies

$$\|\tilde{\boldsymbol{x}}_{t+1} - \boldsymbol{x}^*\| \leq \|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^* - \eta \nabla \mathsf{f}(\tilde{\boldsymbol{x}}_t)\| \tag{197}$$

$$+ \eta \|\nabla \mathsf{f}(\tilde{\boldsymbol{x}}_t) - \nabla \mathsf{f}(\tilde{\boldsymbol{x}}_t - \eta \boldsymbol{e}_{t-1})\| \tag{198}$$

$$\leq \sigma_{GD} \|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^*\| + \eta^2 L \|\boldsymbol{e}_{t-1}\|, \tag{199}$$

where the first term in the right side of (199) is obtained by the convergence guarantee of GD (Lemma B.2), and the second term is due to L-smoothness (13). Since (195) implies

$$\|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^*\| - \eta \|\boldsymbol{e}_{t-1}\| \leq \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}^*\| \tag{200}$$

$$\leq \|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^*\| + \eta \|\boldsymbol{e}_{t-1}\|, \tag{201}$$

we leverage (199) to control $\|\hat{\boldsymbol{x}}_{t+1} - \boldsymbol{x}^*\|$ as follows:

$$\begin{aligned}
&\|\hat{\boldsymbol{x}}_{t+1} - \boldsymbol{x}^*\| \\
&\leq \sigma_{GD} \|\tilde{\boldsymbol{x}}_t - \boldsymbol{x}^*\| + \eta^2 L \|\boldsymbol{e}_{t-1}\| + \eta \|\boldsymbol{e}_t\| \qquad (202)\\
&\leq \sigma_{GD} \|\hat{\boldsymbol{x}}_t - \boldsymbol{x}^*\| + \eta \|\boldsymbol{e}_t\| + (\sigma_{GD} + \eta L)\eta \|\boldsymbol{e}_{t-1}\|, \qquad (203)
\end{aligned}$$

where (203) is due to (200). □

Compared to the guarantee on NQ-GD in Lemma D.1, that on GD with error feedback of [10], [22], [44] in Lemma E.1 has an extra error term. Thus, it is unclear whether the error feedback mechanism of [10], [22], [44] can even improve upon NQ-GD in our nonstochastic problem setting.

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. Richtárik, 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] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[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] 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.

[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] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] S. Bubeck, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3-4, pp. 231–357, Nov. 2015.

[34] 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.

[35] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik, "Distributed learning with compressed gradient differences," Jan. 2019.

[36] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik, "Stochastic distributed learning with gradient quantization and variance reduction," Apr. 2019.

[37] S. Horváth, C.-Y. Ho, L. Horváth, A. N. Sahu, M. Canini, and P. Richtárik, "Natural compression for distributed deep learning," May. 2019.

[38] C. Philippenko and A. Dieuleveut, "Bidirectional compression in heterogeneous settings for distributed or federated learning with partial partici-

pation: tight convergence guarantees," *arXiv preprint arXiv:2006.14591*, June 2020.
[39] E. Gorbunov, F. Hanzely, and P. Richtárik, "A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent," in *International Conference on Artificial Intelligence and Statistics*, June 2020, pp. 680–690.
[40] E. Gorbunov, D. Kovalev, D. Makarenko, and P. Richtárik, "Linearly converging error compensated SGD," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 889–20 900, Dec. 2020.
[41] A. Khaled, O. Sebbouh, N. Loizou, R. M. Gower, and P. Richtárik, "Unified analysis of stochastic gradient methods for composite convex and smooth optimization," *arXiv preprint arXiv:2006.11573*, June 2020.
[42] E. Gorbunov, K. P. Burlachenko, Z. Li, and P. Richtárik, "MARINA: Faster non-convex distributed learning with compression," in *International Conference on Machine Learning*. PMLR, July 2021, pp. 3788–3798.
[43] R. Islamov, X. Qian, and P. Richtárik, "Distributed second order methods with fast rates and compressed communication," in *International Conference on Machine Learning*. PMLR, July 2021, pp. 4617–4628.
[44] 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.
[45] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan, "On biased compression for distributed learning," *arXiv:2002.12410*, Feb. 2020.
[46] S. Magnússon, H. Shokri-Ghadikolaei, and N. Li, "On maintaining linear convergence of distributed learning and optimization under limited communication," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6101–6116, 2020.
[47] R. M. Gray, *Source Coding Theory*. Kluwer Academic Publishers, Oct. 1989.
[48] 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. 11 450–11 460.
[49] 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.
[50] J. Sun, T. Chen, G. B. Giannakis, Q. Yang, and Z. Yang, "Lazily aggregated quantized gradient innovation for communication-efficient federated learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[51] X. Qian, P. Richtárik, and T. Zhang, "Error compensated distributed sgd can be accelerated," *Advances in Neural Information Processing Systems*, vol. 34, Dec. 2021.
[52] P. Richtárik, I. Sokolov, and I. Fatkhullin, "EF21: A new, simpler, theoretically better, and practically faster error feedback," *Advances in Neural Information Processing Systems*, vol. 34, Dec. 2021.
[53] S. Horváth and P. Richtárik, "A better alternative to error feedback for communication-efficient distributed learning," in *Eighth International Conference on Learning Representations*, Apr. 2020.
[54] J. Acharya, C. De Sa, D. Foster, and K. Sridharan, "Distributed learning with sublinear communication," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, CA, USA: PMLR, 09–15 Jun. 2019, pp. 40–50.
[55] P. Mayekar and H. Tyagi, "RATQ: A universal fixed-length quantizer for stochastic optimization," *arXiv*, vol. 1908.08200, Dec. 2019.
[56] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan 2011.
[57] Y. Zhang, J. C. Duchi, and M. J. Wainwright, "Communication-efficient algorithms for statistical optimization," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 3321–3363, Jan 2013.
[58] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 06–11 Aug 2017, pp. 3329–3337.
[59] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2506–2517, Nov 2009.
[60] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, Oct 2019.
[61] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2328–2336.
[62] J. N. Tsitsiklis and Z.-Q. Luo, "Communication complexity of convex optimization," *Journal of Complexity*, vol. 3, no. 3, pp. 231 – 243, 1987.
[63] Y. Arjevani and O. Shamir, "Communication complexity of distributed convex learning and optimization," in *Advances in Neural Information Processing Systems 28*, Dec. 2015, pp. 1756–1764.
[64] 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.
[65] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson, "Distributed learning with compressed gradients," *arXiv: 1806.06573*, June 2018.
[66] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, Dec. 2014.
[67] B. T. Polyak, *Introduction to optimization*, ser. Translations Series in Mathematics and Engineering. Optimization Software, May 1987.
[68] C. A. Rogers, "Covering a sphere with spheres," *Mathematika*, vol. 10, no. 2, pp. 157—164, Dec. 1963.
[69] 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.
[70] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O\left(1/k^2\right)$," in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1982, pp. 372–376.
[71] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," *arXiv: 1407.1537*, July 2014.
[72] 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.
[73] R. Zamir, *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. Cambridge University Press, Sep. 2014.
[74] 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.
[75] 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.
[76] 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.
[77] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[78] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning From Data*. AMLBook, 2012.
[79] M. Schmidt and N. L. Roux, "Fast convergence of stochastic gradient descent under a strong growth condition," Aug. 2013.
[80] 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.
[81] 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.
[82] 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.

Chung-Yi Lin received the B.Sc. degree in electrical engineering and the M.Sc. degree in communication engineering from the National Taiwan University in 2015 and 2018, respectively. He was a Ph.D. student in electrical engineering at the California Institute of Technology between 2018 and 2020. He has been a quantitative researcher at the Kronos Research based in Taiwan since 2021.

Victoria Kostina (S'12–M'14) received the bachelor's degree from Moscow Institute of Physics and Technology (MIPT) in 2004, the master's degree from University of Ottawa in 2006, and the Ph.D. degree from Princeton University in 2013. During her studies at MIPT, she was affiliated with the Institute for Information Transmission Problems of the Russian Academy of Sciences.

She is currently a Professor of electrical engineering and computing and mathematical sciences at California Institute of Technology. Her research interests include information theory, coding, control, learning, and communications. She received the Natural Sciences and Engineering Research Council of Canada postgraduate scholarship during 2009–2012, Princeton Electrical Engineering Best Dissertation Award in 2013, Simons-Berkeley research fellowship in 2015 and the NSF CAREER award in 2017.

She is a Guest Editor for the IEEE Journal on Selected Areas in Information Theory 2022 special issue on Modern Compression.

Babak Hassibi (Member, IEEE) was born in Tehran, Iran, in 1967. He received the B.S. degree from the University of Tehran in 1989, and the M.S. and Ph.D. degrees from Stanford University in 1993 and 1996, respectively, all in electrical engineering.

He has been with the California Institute of Technology since January 2001, where he is currently the Mose and Lilian S. Bohn Professor of Electrical Engineering. From 2013-2016 he was the Gordon M. Binder/Amgen Professor of Electrical Engineering and from 2008-2015 he was Executive Officer of Electrical Engineering, as well as Associate Director of Information Science and Technology. From October 1996 to October 1998 he was a research associate at the Information Systems Laboratory, Stanford University, and from November 1998 to December 2000 he was a Member of the Technical Staff in the Mathematical Sciences Research Center at Bell Laboratories, Murray Hill, NJ. He has also held short-term appointments at Ricoh California Research Center, the Indian Institute of Science, and Linkoping University, Sweden. His research interests include communications and information theory, control and network science, and signal processing and machine learning. He is the coauthor of the books (both with A.H. Sayed and T. Kailath) *Indefinite Quadratic Estimation and Control: A Unified Approach to* H^2 *and* H^∞ *Theories* (New York: SIAM, 1999) and *Linear Estimation* (Englewood Cliffs, NJ: Prentice Hall, 2000). He is a recipient of an Alborz Foundation Fellowship, the 1999 O. Hugo Schuck best paper award of the American Automatic Control Council (with H. Hindi and S.P. Boyd), the 2002 National Science Foundation Career Award, the 2002 Okawa Foundation Research Grant for Information and Telecommunications, the 2003 David and Lucille Packard Fellowship for Science and Engineering, the 2003 Presidential Early Career Award for Scientists and Engineers (PECASE), and the 2009 Al-Marai Award for Innovative Research in Communications, and was a participant in the 2004 National Academy of Engineering "Frontiers in Engineering" program.

He has been a Guest Editor for the IEEE Transactions on Information Theory special issue on "space-time transmission, reception, coding and signal processing" was an Associate Editor for Communications of the IEEE Transactions on Information Theory during 2004-2006, and is currently an Editor for the Journal "Foundations and Trends in Information and Communication" and for the IEEE Transactions on Network Science and Engineering. He is an IEEE Information Theory Society Distinguished Lecturer for 2016-2017 and was General Co-Chair if the 2020 IEEE International Symposium on Information Theory (ISIT 2020).