On the $\mathcal{O}(1/k)$ Convergence of Distributed Gradient Methods Under Random Quantization

Amit Dutta, *Graduate Student Member, IEEE*, and Thinh T. Doan, *Senior Member, IEEE*

***Abstract*—We revisit the so-called distributed two-time-scale stochastic gradient method for solving a strongly convex optimization problem over a network of agents in a bandwidth-limited regime. In this setting, the agents can only exchange the quantized values of their local variables using a limited number of communication bits. Due to quantization errors, the existing best-known convergence results of this method can only achieve a suboptimal rate $\mathcal{O}(1/\sqrt{k})$, while the optimal rate is $\mathcal{O}(1/k)$ under no quantization, where k is the time iteration. The main contribution of this letter is to address this theoretical gap, where we study a sufficient condition and develop an innovative analysis and step-size selection to achieve the optimal convergence rate $\mathcal{O}(1/k)$ for the distributed gradient methods given any number of quantization bits. We provide numerical simulations to illustrate the effectiveness of our theoretical results.**

***Index Terms*—Distributed optimization, quantized communication, two-time-scale stochastic approximation.**

I. Introduction

In THIS letter, we focus on optimization problems defined over a network of N agents, where the goal is to solve

$$\min_{x\in\mathbb{R}^d} f(x) \triangleq \frac{1}{N}\sum_{i=1}^{N} f^i(x), \tag{1}$$

with $f^i : \mathbb{R}^d \to \mathbb{R}$ as the local objective function known only to agent i. We assume no central coordination and the agents communicate locally with neighbors over a graph to solve (1).

We are interested in studying distributed consensus stochastic gradient (**DCSG**) methods to solve problem (1), where each agent maintains a local estimate of the decision variable $x^\star$. Agents update their variables by communicating with their neighbors, averaging the received estimates, and then taking a gradient step of their local functions. A practical challenge when implementing this method is the so-called quantization error when the communication network has limited bandwidths, i.e., agents can only exchange limited information using a finite number of communication bits. This requires them to quantize their values before communicating with others, leading to "quantization errors" in their updates. These errors present a significant bottleneck in the design and analysis of distributed optimization algorithms [1].

In our previous work [2], [3], we propose a variant of **DCSG**, namely, distributed two-time-scale gradient methods, to solve problem (1) under quantized communication (see Algorithm 1 below). However, we showed that this method only achieves a suboptimal convergence rate due to quantization errors, i.e., the rate is $\mathcal{O}(1/\sqrt{k})$ when f is strongly convex [3]. This rate is known to be $\mathcal{O}(1/k)$ when there is no quantization.

Main Contribution: The main focus of this letter is to address the theoretical gap for the convergence complexity of **DCSG** under quantization. In particular, we study a sufficient condition and develop an innovative analysis and step-size selection to achieve the optimal convergence rate $\mathcal{O}(1/k)$ for the distributed two-time-scale stochastic gradient method given any number of quantization bits. To illustrate the effectiveness of our theoretical results, we simulate this method to solve an example of problem (1) and compare it with the performance of the classic **DCSG** without quantization.

A. Related Work

DCSG algorithm was first studied in [4] based on the classic work on distributed computation in [5]. Until now, **DCSG** have been well studied with many advanced theoretical results; see for example the recent survey in [1]. For example, **DCSG** achieves an optimal convergence rate $\mathcal{O}(1/k)$ when the objective function f is strongly convex, which is the same as the result of the centralized setting.

The practical challenge of quantized communication has motivated the existing literature to study the performance of **DCSG** under quantization errors. In [6], the authors provide the first convergence result for the convergence of **DCSG**, where this method only achieves an approximate convergence due to quantization errors. The following work in [2] studies the so-called two-time-scale **DCSG**, motivated by the special consensus algorithm with quantization [7], and shows that this method can find an exact solution of problem (1). However, this letter requires projecting the iterates to a compact set, introducing both projection and quantization errors, resulting

Received 16 September 2024; revised 22 November 2024; accepted 8 December 2024. Date of publication 16 December 2024; date of current version 26 December 2024. This work was supported in part by NSF-CAREER Grant under Grant 2339509, and in part by AFOSR YIP under Grant 420525. Recommended by Senior Editor S. Olaru. *(Corresponding author: Amit Dutta.)*

Amit Dutta is with the Electrical and Computer Engineering Department, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: amitdutta@vt.edu).

Thinh T. Doan is with the Aerospace Engineering and Engineering Mechanics Department, University of Texas at Austin, Austin, TX 78712 USA (e-mail: thinhdoan@utexas.edu).

Digital Object Identifier 10.1109/LCSYS.2024.3519013

in a suboptimal convergence rate of $\mathcal{O}(1/k^{1/3})$. This limitation was later addressed in [3], where the quantization bin sizes increased over time while keeping the number of communication bits constant, leading to a better, but still suboptimal, convergence rate of $\mathcal{O}(1/\sqrt{k})$. These works demonstrate that two-time-scale **DCSG** algorithms can handle quantization errors, but suffer a suboptimal convergence rate. Recent works in [8], [9] improve these results, where they propose a more complicated quantization scheme to obtain optimal convergence rates. These results, however, only apply to the deterministic setting and static communication graphs. Our focus in this letter is to improve the results in [2], [3], where we propose an innovative step size selection and analysis to achieve an optimal rate $\mathcal{O}(1/k)$ in the stochastic setting. For ease of exposition, we will consider static graphs. However, our result can be easily extended to the setting of time-varying graphs studied in [1].

Other approaches, for example, [10], [11], study communication compression using quantization. However, they require an impractical setting, where agents need to use (potentially) an infinite number of bits for quantization to exchange a real interval every iteration for decoding. Their results, therefore, are not applicable to the setting studied in this letter.

B. Notation

We denote $\|x\|$ and $\|\mathbf{X}\|$ the Euclidean norm and the Frobenius norm of the vector x and matrix $\mathbf{X}$, respectively. Let $\mathbf{1}$ be the vector whose entries are 1 and $\mathbf{I}$ the identity matrix. Next, we use superscript and subscript, e.g., x_k^i, to denote the agent indices and iterations, respectively.

Random Quantization. Given a real number $x \in [\ell, u]$, we partition the interval into B equal length bins with endpoints denoted by $\tau_m, m \in \{1, \ldots, B+1\}$, such that $\tau_1 = \ell$ and $\tau_{B+1} = u$. The length of each bin Δ, is defined as $\Delta = \frac{u-l}{B}$. The representation symbols for the quantizers are chosen from $\{\tau_m\}_{m=1}^{B+1}$, where each τ_m is mapped into a codeword of b bits. Thus, for a given number of bits b, the number of bins $B = 2^b - 1$, and $\Delta = (l-u)/(2^b-1)$.

Given $x \in [\tau_i, \tau_{i+1})$, we assign a probability based on its relative location within this interval, $p = (x-\tau_i)/\Delta$. We either choose τ_i or τ_{i+1} to represent x using the stochastic rule $\mathcal{Q}$ which follows the following

$$\mathcal{Q}(x) = \begin{cases} \tau_i & \text{with probability } 1-p, \\ \tau_{i+1} & \text{with probability } p. \end{cases} \tag{2}$$

The random variable $\mathcal{Q}(x)$ satisfies the following properties:

$$\mathbb{E}[\mathcal{Q}(x)|x] = x, \tag{3}$$

$$\mathbb{E}\Big[(\mathcal{Q}(x)-x)^2|x\Big] \leq \frac{\Delta^2}{4}, \tag{4}$$

$$\mathbf{P}(|\mathcal{Q}(x)-x| \leq \Delta) = 1. \tag{5}$$

Thus, the random quantizer is unbiased, has bounded variance, and ensures the quantized value is almost always within Δ of the true value x.

II. Distributed Consensus Stochastic Gradient With Random Quantization

The **DCSG** method with random quantization is formally presented in Algorithm 1, where each agent i maintains a local variable x_k^i to estimate for the optimal solution $x^\star$ of problem (1). At every iteration k, each agent i only exchanges a quantized value, $q_k^i = \mathcal{Q}(x_k^i)$, with its neighboring agents. Upon receiving the quantized values from its neighbors j, in (6) agent i first forms a $\beta-$convex combination of its local value x_k^i and the weighted average of these quantized values. The outcome of the first step is used to update x_{k+1}^i by using the sample of agent i's local gradient scaled by another step size α_k.

Algorithm 1 DCSG Under Random Quantization

Initialize: Each node i initializes $\{x_0^i, \alpha_k, \beta_k\}$

Iteration: For $k = 1, \ldots,$ node $i \in \mathcal{V}$ implements:

Compute quantization $q_k^i = \mathcal{Q}(x_k^i)$ and send to node $j \in \mathcal{N}^i$

Receive q_k^j from node $j \in \mathcal{N}^i$ and update

$$x_{k+1}^i = (1-\beta_k)x_k^i + \beta_k \sum_{j\in\mathcal{N}^i} a_{ij} q_k^j - \alpha_k \nabla f^i\big(x_k^i; \xi_k^i\big). \tag{6}$$

Here, the decentralized communication between agents is modeled by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = 1, \ldots, N$ represents the set of vertices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of edges. We denote by $\mathcal{N}^i = \{j \in \mathcal{V} | (i,j) \in \mathcal{E}\}$ the neighboring set of agent i. The matrix $\mathbf{A} = [a^{ij}]$ represents the communication structure associated with graph $\mathcal{G}$, i.e., $a^{ij} \in (0,1)$ if $j \in \mathcal{N}^i$ otherwise $a^{ij} = 0$. Note that when $\beta_k = 1$ and $q_k^i = x_k^i$, i.e., no quantization, Algorithm 1 reduces to the classic **DCSG** method introduced in [4]. However, in (6) β_k is chosen strictly smaller than 1 to remove the impact of quantization noise. In addition, β_k is chosen larger than α_k so that the quantization noise is addressed before the gradient updates. This is a distributed variant of the so-called two-time-scale stochastic approximation [12]. It turns out that by properly choosing α_k, β_k, Algorithm 1 can find an exact solution $x^\star$ of problem (1) even under random quantization [2]. However, as noted the existing results for the convergence complexity of Algorithm 1 are suboptimal, (e.g., $\mathcal{O}(1/\sqrt{k})$ when f is strongly convex). This rate is $\mathcal{O}(1/k)$ when there is no quantization. Our focus in this letter is, therefore, to close this gap, where we will provide a sufficient condition to achieve the optimal convergence rate $\mathcal{O}(1/k)$ of Algorithm 1.

III. Technical Assumptions and Preliminaries

We will consider the following assumptions, which we assume they always hold to the end of this letter.

Assumption 1 (Lipschitz Smoothness): For all i, the gradient of f_i is Lipschitz continuous with a positive constant L^i

$$\|\nabla f^i(x) - \nabla f^i(y)\| \leq L^i \|x-y\|, \ \ x, y \in \mathbb{R}^d. \tag{7}$$

Assumption 2 (Strong Convexity): The global objective function f is strongly convex with constant $\mu > 0$

$$(x-y)^T(\nabla f(x) - \nabla f(y)) \geq \mu \|x-y\|^2, \ \ x, y \in \mathbb{R}^d. \tag{8}$$

Assumption 3: The random variables ξ_k^i, $\forall i$ and $k \geq 0$, are i.i.d. and there exists a positive constant σ such that $\forall x \in \mathbb{R}^d$

$$\mathbb{E}\big[\nabla f^i\big(x, \xi_{k,t}^i\big) \mid \mathcal{F}_k\big] = \nabla f^i(x), \tag{9}$$

$$\mathbb{E}\Big[\|\nabla f^i\big(x, \xi_{k,t}^i\big) - \nabla f^i(x)\|^2 \mid \mathcal{F}_k\Big] \leq \sigma^2, \tag{10}$$

where $\mathcal{F}_k$ represents the filtration that contains the history of all variables generated by Algorithm 1 up to iteration k.

Assumption 4: The optimal solution $x^\star$ of (1) satisfies

$$x^\star \in \bigcap_i \arg\min_{x\in\mathbb{R}^d} f^i(x). \tag{11}$$

Remark 1: Assumption 2 implies that $x^\star$ is the unique solution of problem (1). Under Assumption 4, $x^\star$ is also a minimizer of each f^i. However, Assumption 4 does not imply that the set of minimizers of f^i is unique. Thus, under this assumption solving problem (1) is equivalent to searching for a point in the intersection of the minimizer sets of each f^i.

We consider the following example where Assumption 5 holds. Specifically, the global objective is given by $G(x) = \sum_{i=1}^N \|A_i x - b_i\|^2$, where $A_i \in \mathbb{R}^{d\times d}$ are rank-deficient local matrices. Despite the rank deficiency of each A_i, the aggregate matrix $\sum_{i=1}^N A_i^\top A_i$ is full rank. This ensures that $G(x)$ is strongly convex and admits a unique global minimizer $x^\star$. The rank deficiency of A_i implies that each local objective $\|A_i x - b_i\|^2$ is convex but can have multiple minimizers, resulting in local solution sets $S_i = \{x \mid A_i x = b_i\}$. By construction, it is possible to choose $x^\star$ such that it lies in the column space of A_i for all i, ensuring that $x^\star \in S_i$ for every i. Consequently, $x^\star$ resides in the intersection of the local solution sets $x^\star \in \bigcap_{i=1}^N S_i$. This setup satisfies Assumption 5, as $x^\star$ is the unique minimizer of the global objective $G(x)$ and lies in the intersection of the local solution sets. In Section V, we provide another example where this assumption holds.

The above assumption provides a sufficient condition to establish the optimal convergence rate $\mathcal{O}(1/k)$ for the two-time-scale distributed gradient descent method under random quantization. While this condition ensures the theoretical guarantees, it may not be necessary, which we leave for future studies. Compared to existing works [3], [13], this additional assumption is required to achieve the optimal rate for Algorithm 1. However, we do not require that each function f^i is Lipschitz continuous (or bounded gradients) as assumed in prior works.

Assumption 5: The matrix $\mathbf{A} = [a^{ij}]$ is doubly stochastic, i.e., $\sum_i a^{ij} = \sum_j a^{ij} = 1$ for all i, j.

We note that Assumptions 1–5 are standard in the literature of DCSG. Finally, our result is the same as the one in [8], where the authors use an adaptive quantization scheme. However, the result in [8] is for fixed graphs and requires a certain condition on the number of bits to control the quantization errors. On the other hand, our result is applicable to any value of B and can be extended to time-varying graphs using standard uniform connectivity assumption [6].

For convenience, we introduce the following notation. First, the quantization error at each agent i is defined as

$$e_k^i = x_k^i - q_k^i. \tag{12}$$

We denote $\mathbf{X}$ the matrix

$$\mathbf{X} = \begin{bmatrix} (x_k^1)^T \\ \vdots \\ (x_k^N)^T \end{bmatrix}, \tag{13}$$

whose i-th row is $(x^i)^T$ and $\bar{x} = \frac{1}{N}\sum_i x^i$ as the average for a given collection of vectors x^i, for iteration k. Let $\mathbf{W} = \mathbf{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ and $\mathbf{Y}$ be the consensus errors given as

$$\mathbf{Y}_k = \mathbf{X}_k - \mathbf{1}\bar{x}_k^T = \mathbf{W}\mathbf{X}_k, \tag{14}$$

Using the notation above, the matrix form of (6) is

$$\mathbf{X}_{k+1} = (1-\beta_k)\mathbf{X}_k + \beta_k \mathbf{A}\mathbf{Q}_k - \alpha_k \mathbf{G}_k(\mathbf{X}_k), \tag{15}$$

$$\bar{x}_{k+1} = (1-\beta_k)\bar{x}_k + \beta_k \bar{q}_k - \alpha_k \bar{g}_k, \tag{16}$$

where

$$\mathbf{G}_k(\mathbf{X}_k;\xi_k) = \begin{bmatrix} (\nabla f^1(x_k^1,\xi_k^1))^T \\ \vdots \\ (\nabla f^N(x_k^N,\xi_k^N))^T \end{bmatrix},$$

$$\bar{g}_k = \frac{1}{N}\sum_{i=1}^N \nabla f^i(x_k^i,\xi_k^i). \tag{17}$$

We will consider the following choice of step sizes

$$\alpha_k = \frac{C_\alpha}{1+h+k}, \quad \beta_k = \frac{C_\beta}{1+h+k}, \quad C_\alpha \le C_\beta, \tag{18}$$

where C_α, C_β and $h > 1$ are constants. The choice of these constants to guarantee an optimal convergence rate $\mathcal{O}(1/k)$ for Algorithm 1 will be given in Theorem 1. Finally, we consider the following lemmas to characterize the properties of the iterates generated by Algorithm 1. We present their proofs in the Appendix.

Lemma 1: For all $k \ge 0$ we have

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{Y}_{k+1}\|^2\Big] &\le \Big(1-(1-\sigma_2)\beta_k + \frac{8\alpha_k(L+1)^3N}{\mu}\Big)\mathbb{E}\Big[\|\mathbf{Y}_k\|^2\Big] + 2\beta_k^2 N\sigma^2 \\ &\quad + \Big(\frac{\mu\alpha_k}{4} + 2\alpha_k\beta_k(L+1)^2\Big)\mathbb{E}\Big[\|\bar{x}_k - x^\star\|^2\Big] + \frac{\beta_k^2 d^2\Delta^2 N}{2},\end{aligned} \tag{19}$$

where $L = \sum_{i=1}^N L^i$ and σ_2 is the second largest singular value adopted from the averaging matrix $\mathbf{A}$.

Lemma 2: For all $k \ge 0$ we have

$$\begin{aligned}\mathbb{E}\Big[\|\bar{x}_k - x^\star\|^2\Big] &\le \Big(1 - \frac{7\mu\alpha_k}{4} + 4\alpha_k\beta_k(L+1)^2\Big)\mathbb{E}\Big[\|\bar{x}_{k+1} - x^\star\|^2\Big] \\ &\quad + \frac{8\alpha_k(L+1)^3N}{\mu}\mathbb{E}\Big[\|\mathbf{Y}_k\|^2\Big] + \frac{\beta_k^2\Delta^2 d^2}{2} + \beta_k^2\sigma^2 N.\end{aligned} \tag{20}$$

IV. Main Results

The focus of this section is to study the convergence of Algorithm 1 when the global function f is strongly convex and each local function f_i has Lipschitz smooth gradients. Our main result shows that each iterate x_k^i converges to $x^\star$ at a rate $\mathcal{O}(1/k)$.

To demonstrate the convergence rate of Algorithm 1, we consider the following aggregate Lyapunov function:

$$V_k = \|\bar{x}_k - x^\star\|^2 + \|\mathbf{Y}_k\|^2. \tag{21}$$

Finally, we present our main result that establishes the convergence of Algorithm 1, achieved through the analysis of the aggregate Lyapunov function V_k defined in (21), as outlined in the following theorem.

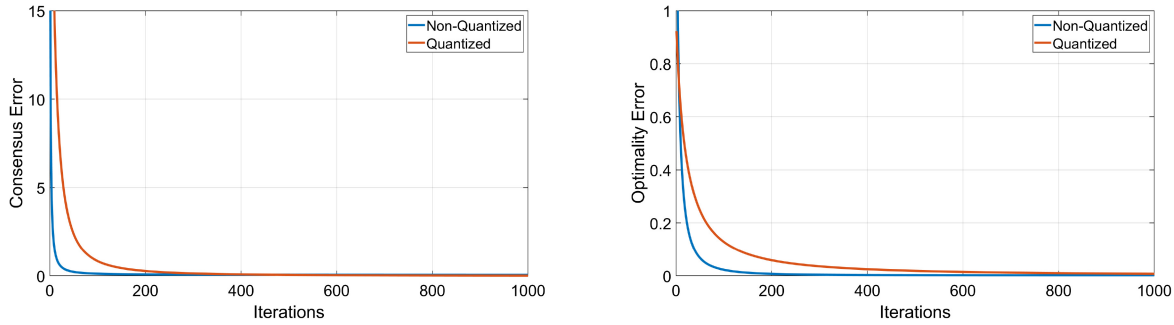

Fig. 1. Convergence of $\|\mathbf{Y}_k\|^2$ on the left and $\|\bar{x}_k - x^\star\|^2$ on the right under 16-bit quantization.

Theorem 1: Let C_α, C_β and $h > 1$ in (18) be chosen as

$$C_\alpha = \frac{16}{3\mu}, \quad C_\beta \geq \frac{17(L+1)^3 N C_\alpha}{\mu(1-\sigma_2)}, \quad h \geq \frac{48(L+1)^2 C_\beta}{3\mu}. \tag{22}$$

Then we obtain for all $k \geq 0$

$$\mathbb{E}\big[V_{k+1}\big] \leq \frac{h^2 \mathbb{E}[V_0]}{(k+h+1)^2} + \frac{d^2\Delta^2 C_\beta^2}{k+h+1} + \frac{3\sigma^2 N C_\beta^2}{k+h+1}. \tag{23}$$

Remark 2: Theorem 1 indicates that Algorithm 1 converges to the desired solution at a rate $\mathcal{O}(1/k)$, which is the same as DCSG when there is no quantization. Our result also shows that the complexity scales proportionally with d^2 and Δ^2. The term Δ^2 is expected as one can view quantization error as another source of noise with variance Δ^2. On the other hand, the dependence on the square of dimension, d^2, might be suboptimal. One would expect that the upper bound should depend linearly on d. At this point, we are uncertain whether this is an artifact from our analysis or a fundamental result. We leave this question for our future research.

Proof: Adding (19) into (20) and using (21) we have

$$\begin{aligned}\mathbb{E}\big[V_{k+1}\big] &\leq \left(1 - \frac{3\mu\alpha_k}{8}\right)\mathbb{E}[V_k] + d^2\Delta^2\beta_k^2 + 3\sigma^2 N\beta_k^2 \\ &\quad + \left(-\frac{3\mu\alpha_k}{8} + 6(L+1)^2\alpha_k\beta_k\right)\mathbb{E}\Big[\|\bar{x}_k - x^\star\|^2\Big] \\ &\quad + \left(-(1-\sigma_2)\beta_k + \frac{17(L+1)^3 N\alpha_k}{\mu}\right)\mathbb{E}\Big[\|\mathbf{Y}_k\|^2\Big]. \end{aligned} \tag{24}$$

Using the definition of step sizes α_k and β_k in (18) and (22), we have $\frac{3\mu\alpha_k}{8} - 6(L+1)^2\alpha_k\beta_k > 0$ and $(1-\sigma_2)\beta_k - \frac{17(L+1)^3 N}{\mu} > 0$ which when using into (24) gives

$$\begin{aligned}\mathbb{E}\big[V_{k+1}\big] &\leq \left(1 - \frac{3\mu\alpha_k}{8}\right)\mathbb{E}[V_k] + d^2\Delta^2\beta_k^2 + 3\sigma^2 N\beta_k^2 \\ &= \frac{k+h-1}{k+h+1}\mathbb{E}[V_k] + \frac{d^2\Delta^2 C_\beta^2}{(k+h+1)^2} + \frac{3\sigma^2 N C_\beta^2}{(k+k+1)^2}.\end{aligned}$$

Multiplying both sides of the preceding relation with $(k+h+1)^2$ and since $h > 1$ we have

$$\begin{aligned}&(k+h+1)^2\mathbb{E}\big[V_{k+1}\big] \\ &\leq (k+h+1)(k+h-1)\mathbb{E}[V_k] + d^2\Delta^2 C_\beta^2 + 3\sigma^2 N C_\beta^2 \\ &\leq h^2\mathbb{E}[V_0] + d^2\Delta^2 C_\beta^2(k+1) + 3\sigma^2 N C_\beta^2(k+1), \end{aligned} \tag{25}$$

which when dividing both sides by $(k+h+1)^2$ immediately yields (23). This concludes our proof. ■

V. Simulations

We simulate Algorithm 1 to solve a simple example of problem (1) when $N = 25$. Specifically, the agents aim to agree at an unknown point $x^\star$, where $x^\star$ is a 10-dimensional vector with all elements equal to unity. Each agent i makes 100 noisy observations of $x^\star$, denoted as $X^i = x^\star + Z^i$, where $Z^i \sim \mathcal{N}(0, \mathbf{I}_{10})$. Here, $f^i(x) = \frac{1}{2}\mathbb{E}[\|x - X^i\|^2]$ is the local objective for i^{th} agent with minimizer $x^\star$. We also note that $x^\star$ is also minimizer of the global function $\sum_{i=1}^N f^i(x)$. Thus our experimental setup satisfies Assumption 5. Here, note that this is a special case where the local objectives have a unique minimizer. For our implementation, we randomly generate a connected graph G and use $\mathbf{A}$ as the Metropolis adjacency matrix corresponding to G [2].

We implement two sets of simulations. In Figure 1, we compare the performance of the classic **DCSG** without quantization [4] and Algorithm 1. For the former, we set $\alpha_k = \frac{0.15}{1+k}$, while $\alpha_k = \frac{0.15}{6+k}$ and $\beta_k = \frac{0.5}{6+k}$ for the latter. Here we have $C_\alpha = 0.15$, $C_\beta = 0.5$ and $h = 5$, which satisfy $C_\alpha \leq C_\beta$ and $h > 1$ that meet the step size conditions for Algorithm 1. Furthermore, for the simulation of the Algorithm 1, each agent utilizes a 16-bit quantizer based on the random quantization scheme discussed in Section I-B. As shown in Figure 1 both methods have the same convergence rates to the desired values, which agrees with our theoretical result. Figure 2 demonstrates that as the number of quantization bits increases, the algorithm requires fewer iterations to reach the desired accuracy (e.g., $V_k \leq 0.045$), consistent with Theorem 1, where the upper bound scales proportionally with Δ^2 that gets smaller as the number of bits increases.

VI. Concluding Remarks

In this letter, we revisit the distributed two-time-scale stochastic gradient method under quantized communication. Our main contribution is to study a sufficient condition and develop an innovative analysis and step-size selection to achieve the optimal convergence rate $\mathcal{O}(1/k)$ for the distributed gradient methods given any number of quantization bits. One interesting question left by this letter is to study more general conditions to obtain an optimal convergence rate of **DCSG**. Another question is to understand whether one can achieve a linear dependence on the dimension d in our main results.

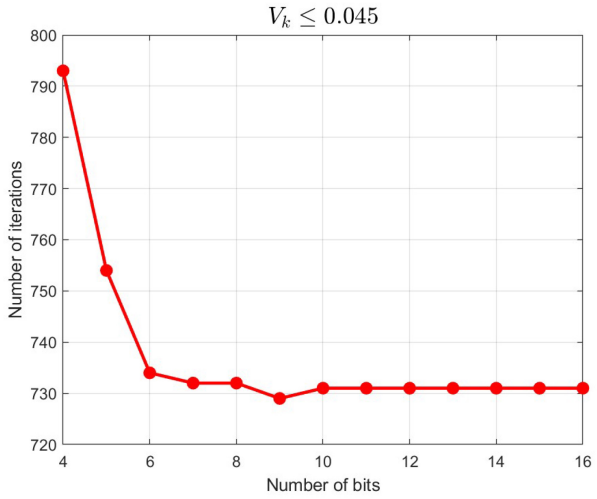

Fig. 2. Performance complexity as a function of bits.

APPENDIX

A. Proof of Lemma 1

Proof: From (6) and using the fact that $\mathbf{E}_k = \mathbf{Q}_k - \mathbf{X}_k$ and $\mathbf{AW} = \mathbf{WA}$ we have,

$$\begin{aligned}\|\mathbf{Y}_{k+1}\|^2 &= \|\mathbf{W}\mathbf{X}_{k+1}\|^2\\ &= \|(1-\beta_k)\mathbf{Y}_k + \beta_k\mathbf{W}\mathbf{A}\mathbf{E}_k + \beta_k\mathbf{A}\mathbf{Y}_k - \alpha_k\mathbf{W}\mathbf{G}_k\|^2\\ &= \|(\mathbf{I}-(\mathbf{I}-\mathbf{A})\beta_k)\mathbf{Y}_k + \beta_k\mathbf{W}\mathbf{A}\mathbf{E}_k\|^2 + \alpha_k^2\|\mathbf{W}\mathbf{G}_k(\mathbf{X}_k)\|^2\\ &\quad - 2\alpha_k((\mathbf{I}-(\mathbf{I}-\mathbf{A})\beta_k)\mathbf{Y}_k + \beta_k\mathbf{W}\mathbf{A}\mathbf{E}_k)^T(\mathbf{W}\mathbf{G}_k(\mathbf{X}_k))\\ &= A + B + C, \end{aligned} \tag{26}$$

where A, B and C are defined in that order. Taking conditional expectation with respect to $\mathcal{F}_k$ and using the unbiasedness of the random quantizers associated with each agent, term A can be analyzed as the following

$$\begin{aligned}\mathbb{E}[A|\mathcal{F}_k] &= \|(\mathbf{I}-(\mathbf{I}-\mathbf{A})\beta_k)\mathbf{Y}_k\|^2 + \beta_k^2\mathbb{E}\Big[\|\mathbf{W}\mathbf{A}\mathbf{E}_k\|^2|\mathcal{F}_k\Big]\\ &\leq (1-(1-\sigma_2)\beta_k)^2\|\mathbf{Y}_k\|^2 + \beta_k^2\sum_{i=1}^N\mathbb{E}\Big[\|e_k^i\|^2|\mathcal{F}_k\Big]\\ &\overset{(3)}{\leq} (1-(1-\sigma_2)\beta_k)^2\|\mathbf{Y}_k\|^2 + \frac{N\beta_k^2d^2\Delta^2}{4}. \end{aligned} \tag{27}$$

Next, the term B can be analyzed as the following

$$\begin{aligned}\mathbb{E}\big[B|\mathcal{F}_k\big] &\leq \alpha_k^2\mathbb{E}\Big[\|\mathbf{G}_k(\mathbf{X}_k)\|^2|\mathcal{F}_k\Big]\\ &= \alpha_k^2\sum_{i=1}^N\mathbb{E}\Big[\|\nabla f^i(x_k^i;\xi_k^i) - \nabla f^i(x_k^i)\|^2|\mathcal{F}_k\Big]\\ &\quad + \alpha_k^2\sum_{i=1}^N\|\nabla f^i(x_k^i)\|^2\\ &\leq \alpha_k^2\sum_{i=1}^N\|\big(\nabla f^i(x_k^i) - \nabla f^i(\bar{x}_k)\big) + \big(\nabla f^i(\bar{x}_k) - \nabla f^i(x^\star)\big)\|^2\\ &\quad + \alpha_k^2N\sigma^2 \leq 2\alpha_k^2L^2\|\mathbf{Y}_k\|^2 + 2\alpha_k^2L^2N\|\bar{x}_k - x^\star\|^2 + \beta_k^2N\sigma^2, \end{aligned} \tag{28}$$

where the last inequality is obtained using Cauchy-Schwarz inequality and due to $\alpha_k \leq \beta_k$ and $L^i \leq \sum_i L^i$. Finally taking expectation with respect to filtration $\mathcal{F}_k$, we analyze the term C as

$$\begin{aligned}&\mathbb{E}[C|\mathcal{F}_k]\\ &= -2\alpha_k\mathbb{E}\big[((\mathbf{I}-(\mathbf{I}-\mathbf{A})\beta_k)\mathbf{Y}_k)^T(\mathbf{W}\mathbf{G}_k(\mathbf{X}_k;\xi_k))|\mathcal{F}_k\big]\\ &\quad - 2\alpha_k\mathbb{E}\big[\beta_k(\mathbf{W}\mathbf{A}\mathbf{E}_k)^T(\mathbf{W}\mathbf{G}_k(\mathbf{X}_k;\xi_k))|\mathcal{F}_k\big]\\ &= C_1 + C_2, \end{aligned} \tag{29}$$

where C_1 and C_2 are defined in that order. Since $\beta_k \leq 1$ and $\sigma_2 \in (0,1)$ we have $1-(1-\sigma_2)\beta_k \leq 1$. By Assumption 3, term C_1 can be expressed as

$$\begin{aligned}C_1 &\leq 2\alpha_k\|(\mathbf{I}-(\mathbf{I}-\mathbf{A})\beta_k)\mathbf{Y}_k\|\|\mathbf{W}\mathbf{G}_k(\mathbf{X}_k)\|\\ &\leq 2\alpha_k(1-(1-\sigma_2)\beta_k)\|\mathbf{Y}_k\|\|\mathbf{G}_k(\mathbf{X}_k)\|\\ &\leq 2\alpha_k\|\mathbf{Y}_k\|\|\mathbf{G}_k(\mathbf{X}_k) - \mathbf{G}_k(\bar{x}_k)\|\\ &\quad + 2\alpha_k\|\mathbf{Y}_k\|\|\mathbf{G}_k(\bar{x}_k) - \mathbf{G}_k(x^\star)\|\\ &\leq 2\alpha_kL\|\mathbf{Y}_k\|^2 + 2\alpha_kL\sqrt{N}\|\mathbf{Y}_k\|\|\bar{x}_k - x^\star\|. \end{aligned} \tag{30}$$

Note that in the above inequality using the Frobenius norm of a matrix we have obtained,

$$\|\mathbf{G}_k(\mathbf{X}_k) - \mathbf{G}_k(\bar{x}_k)\| \leq L\sqrt{\sum_{i=1}^N\|x_k^i - \bar{x}_k\|^2} = L\|\mathbf{Y}_k\|.$$

Similarly, $\|\mathbf{G}_k(\bar{x}_k) - \mathbf{G}_k(x^\star)\| \leq L\sqrt{N}\|\bar{x}_k - x^\star\|$. Next, using the Cauchy-Schwarz inequality, term C_1 can be expressed as

$$C_1 \leq \left(2\alpha_kL + \frac{4\alpha_kL^2N}{\mu}\right)\|\mathbf{Y}_k\|^2 + \frac{\mu\alpha_k}{4}\|\bar{x}_k - x^\star\|^2. \tag{31}$$

Next using the unbiased property of both the stochastic gradients and the random quantizer we analyze term C_2 as

$$\begin{aligned}&\mathbb{E}[C_2|\mathcal{F}_k]\\ &= -2\alpha_k\beta_k\mathbb{E}\big[(\mathbf{W}\mathbf{A}\mathbf{E}_k)^T(\mathbf{W}(\mathbf{G}_k(\mathbf{X}_k;\xi_k) - \mathbf{G}_k(\mathbf{X}_k)))|\mathcal{F}_k\big]\\ &\quad - 2\alpha_k\beta_k\mathbb{E}\big[(\mathbf{W}\mathbf{A}\mathbf{E}_k)^T\mathbf{G}_k(\mathbf{X}_k)|\mathcal{F}_k\big]\\ &\leq \alpha_k\beta_k\mathbb{E}\Big[\|\mathbf{E}_k\|^2|\mathcal{F}_k\Big]\\ &\quad + \alpha_k\beta_k\mathbb{E}\Big[\|\mathbf{G}_k(\mathbf{X}_k;\xi_k) - \mathbf{G}_k(\mathbf{X}_k)\|^2|\mathcal{F}_k\Big]\\ &\leq \frac{\alpha_k\beta_kd^2\Delta^2N}{4} + \alpha_k\beta_kN\sigma^2 \leq \frac{\beta_k^2d^2\Delta^2N}{4} + \beta_k^2N\sigma^2. \end{aligned}$$

Substituting the above relation and (31) into (29), along with (27) and (28), and using $(1-(1-\sigma_2)\beta_k)^2 \leq 1-(1-\sigma_2)\beta_k$ (as $\beta_k \leq 1$, $\sigma_2 \in (0,1)$), $\mu \leq L$, and $\alpha_k \leq \beta_k$, we derive (19) from (26), completing the proof. ■

B. Proof of Lemma 2

Proof: From (16) we have the following,

$$\begin{aligned}\|\bar{x}_{k+1} - x^\star\|^2 &= \|\bar{x}_k - x^\star - \alpha_k\bar{g}_k + \beta_k\bar{e}_k\|^2\\ &= \|\bar{x}_k - x^\star - \alpha_k\bar{g}_k\|^2 + \beta_k^2\|\bar{e}_k\|^2\\ &\quad + 2\beta_k\big(\bar{x}_k - x^\star - \alpha_k\bar{g}_k\big)^T\bar{e}_k. \end{aligned} \tag{32}$$

The first term in the right-hand side of the above equation can be analyzed as the following,

$$\begin{aligned}&\|\bar{x}_k - x^\star - \alpha_k\bar{g}_k\|^2\\ &= \|\bar{x}_k - x^\star\|^2 + \alpha_k^2\|\bar{g}_k\|^2 - 2\alpha_k\big(\bar{x}_k - x^\star\big)^T\bar{g}_k. \end{aligned} \tag{33}$$

We analyze the last term in the right-hand side of the above equation in expectation. For this, using the unbiased gradient property, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[-2\alpha_k\big(\bar{x}_k - x^\star\big)^T \bar{g}_k|\mathcal{F}_k\Big] \\
&= -\frac{2\alpha_k}{N}\sum_{i=1}^N\big(\nabla f^i(x_k^i) - \nabla f^i(\bar{x}_k)\big)^T\big(\bar{x}_k - x^\star\big) \\
&\quad - \frac{2\alpha_k}{N}\sum_{i=1}^N \nabla f^i(\bar{x}_k)^T\big(\bar{x}_k - x^\star\big) \\
&\le \frac{2\alpha_k}{N}\sum_{i=1}^N L_i\|x_k^i - \bar{x}_k\|\|\bar{x}_k - x^\star\| - 2\alpha_k\nabla f(\bar{x}_k)^T\big(\bar{x}_k - x^\star\big) \\
&\overset{(8)}{\le} \frac{2\alpha_k L}{N}\sum_{i=1}^N\|x_k^i - \bar{x}_k\|\|\bar{x}_k - x^\star\| - 2\mu\alpha_k\|\bar{x}_k - x^\star\|^2 \\
&\le \frac{7\mu\alpha_k}{4}\|\bar{x}_k - x^\star\|^2 + \frac{4\alpha_k L^2}{\mu N}\|\mathbf{Y}_k\|^2, \qquad (34)
\end{aligned}$$

where the last inequality is obtained using the Cauchy-Schwarz inequality and using the definition of $\mathbf{Y}_k$. Next, we analyze the second term in the right hand side of (33) as,

$$\begin{aligned}
&\mathbb{E}\Big[\alpha_k^2\|\bar{g}_k\|^2|\mathcal{F}_k\Big] \\
&\le \frac{\alpha_k^2}{N}\sum_{i=1}^N\mathbb{E}\Big[\|\big(\nabla f^i(x_k^i;\xi_k^i) - \nabla f^i(x_k^i)\big) + \nabla f^i(x_k^i)\|^2|\mathcal{F}_k\Big] \\
&\le \frac{\alpha_k^2}{N}\sum_{i=1}^N\mathbb{E}\Big[\|\nabla f^i(x_k^i;\xi_k^i) - \nabla f^i(x_k^i)\|^2|\mathcal{F}_k\Big] \\
&\quad + \frac{\alpha_k^2}{N}\sum_{i=1}^N\|\big(\nabla f^i(x_k^i) - \nabla f^i(\bar{x}_k)\big) + \big(\nabla f^i(\bar{x}_k) - \nabla f^i(x^\star)\big)\|^2 \\
&\le \frac{2\alpha_k^2 L^2}{N}\|\mathbf{Y}_k\|^2 + 2\alpha_k^2 L^2\|\bar{x}_k - x^\star\|^2 + \alpha_k^2\sigma^2, \qquad (35)
\end{aligned}$$

where the last inequality is obtained using the Cauchy-Schwarz inequality and the lipschitz smoothness of the gradients. Putting the above result along with (34) back into (33) and using $\alpha_k \le 1$ along with $\mu \le L$ we get

$$\begin{aligned}
&\mathbb{E}\Big[\|\bar{x}_k - x^\star - \alpha_k\bar{g}_k\|^2\Big] \\
&\le \left(1 - \frac{7\mu\alpha_k}{4} + 2\alpha_k^2L^2\right)\mathbb{E}\Big[\|\bar{x}_k - x^\star\|^2\Big] \\
&\quad + \frac{6\alpha_k(L+1)^3N}{\mu}\mathbb{E}[\|\mathbf{Y}_k\|^2 + \alpha_k^2\sigma^2. \qquad (36)
\end{aligned}$$

The second term in the right hand side of (32) can be analyzed in expectation conditioned on $\mathcal{F}_k$ as $\beta_k^2\mathbb{E}[\|\bar{e}_k\|^2|\mathcal{F}_k] \le \frac{\beta_k^2}{N}\sum_{i=1}^N\mathbb{E}[\|e_k^i\|^2|\mathcal{F}_k] \le \frac{\beta_k^2 d^2\Delta^2}{4}$, which follows from the property of random quantization (3). Finally, since both the stochastic gradients and the random quantizer are unbiased, we analyze the last term from (32) as,

$$\begin{aligned}
&2\beta_k\mathbb{E}\Big[\big(\bar{x}_k - x^\star - \alpha_k\bar{g}_k\big)^T\bar{e}_k|\mathcal{F}_k\Big] \\
&= 2\beta_k\mathbb{E}\Big[\big(\bar{x}_k - x^\star\big)^T\bar{e}_k|\mathcal{F}_k\Big] - 2\alpha_k\beta_k\mathbb{E}\big[\bar{g}_k^T\bar{e}_k|\mathcal{F}_k\big] \\
&\le \alpha_k\beta_k\mathbb{E}\Big[\|\bar{g}_k\|^2|\mathcal{F}_k\Big] + \alpha_k\beta_k\mathbb{E}\Big[\|\bar{e}_k\|^2|\mathcal{F}_k\Big] \\
&\le \frac{\alpha_k\beta_k}{N}\sum_{i=1}^N\mathbb{E}\Big[\|\big(\nabla f^i(x_k^i;\xi_k^i) - \nabla f^i(x_k^i)\big) + \nabla f^i\big(x_k^i\big)\|^2\mathcal{F}_k\Big] \\
&\quad + \frac{\beta_k^2\Delta^2d^2N}{4} \\
&= \frac{\alpha_k\beta_k}{N}\sum_{i=1}^N\mathbb{E}\Big[\|\nabla f^i\big(x_k^i;\xi_k^i\big) - \nabla f^i\big(x_k^i\big)\|^2|\mathcal{F}_k\Big]^2 \\
&\quad + \frac{\alpha_k\beta_k}{N}\sum_{i=1}^N\|\nabla f^i\big(x_k^i\big)\|^2 + \frac{\beta_k^2\Delta^2d^2}{4} \\
&\le \alpha_k\beta_k\sigma^2 + \frac{\alpha_k\beta_k}{N}\sum_{i=1}^N\|\big(\nabla f^i(x_k^i) - \nabla f^i(\bar{x}_k)\big) \\
&\quad - \big(\nabla f^i(\bar{x}_k) - \nabla f^i(x^\star)\big)\|^2 + \frac{\beta_k^2\Delta^2d^2}{4} \\
&\le \frac{2\alpha_k(L+1)^3N}{\mu}\|\mathbf{Y}_k\|^2 + 2\alpha_k\beta_k(L+1)^2\|\bar{x}_k - x^\star\|^2 \\
&\quad + \frac{\beta_k^2\Delta^2d^2}{4} + \beta_k^2\sigma^2. \qquad (37)
\end{aligned}$$

Thus, taking expectation on both sides of (32) and using the above results we arrive at (20). This concludes our proof. ■

REFERENCES

[1] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proc. IEEE*, vol. 106, no. 5, pp. 953–976, May 2018.

[2] T. T. Doan, S. T. Maguluri, and J. Romberg, "Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach," *IEEE Trans. Autom. Control*, vol. 66, no. 10, pp. 4469–4484, Oct. 2021.

[3] M. M. Vasconcelos, T. T. Doan, and U. Mitra, "Improved convergence rate for a distributed two-time-scale gradient method under random quantization," in *Proc. 60th IEEE Conf. Decis. Control (CDC)*, 2021, pp. 3117–3122.

[4] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[5] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, Sep. 1986.

[6] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Trans. Autom. Control*, vol. 54, no. 11, pp. 2506–2517, Nov. 2009.

[7] A. I. Rikos and C. N. Hadjicostis, "Distributed average consensus under quantized communication via event- triggered mass summation," in *Proc. IEEE Conf. Decis. Control (CDC)*, 2018, pp. 894–899.

[8] T. T. Doan, S. T. Maguluri, and J. Romberg, "Fast convergence rates of distributed subgradient methods with adaptive quantization," *IEEE Trans. Autom. Control*, vol. 66, no. 5, pp. 2191–2205, May 2021.

[9] N. Michelusi, G. Scutari, and C.-S. Lee, "Finite-bit quantization for distributed algorithms with linear convergence," *IEEE Trans. Inf. Theory*, vol. 68, no. 11, pp. 7254–7280, Nov. 2022.

[10] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 4934–4947, Oct. 2019.

[11] D. Kovalev, A. Koloskova, M. Jaggi, P. Richtarik, and S. Stich, "A linearly convergent algorithm for decentralized optimization: Sending less bits for free!" in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 4087–4095.

[12] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ., 2008.

[13] T. T. Doan, "Nonlinear two-time-scale stochastic approximation: Convergence and finite-time performance," 2021, *arXiv:2011.01868*.