Accelerating the Convergence Rates of Distributed Subgradient Methods with Adaptive Quantization

Thinh T. Doan†⋆, Siva Theja Maguluri⋆, Justin Romberg†
† School of Electrical and Computer Engineering
⋆ H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology, GA, 30332, USA
{thinhdoan, siva.theja}@gatech.edu, jrom@ece.gatech.edu.

Abstract

We study distributed optimization problems over a network when the communication between the nodes is constrained, and so information that is exchanged between the nodes must be quantized. This imperfect communication poses a fundamental challenge, and this imperfect communication, if not properly accounted for, prevents the convergence of these algorithms. In this paper, our main contribution is to propose a novel quantization method, which we refer to as an adaptive quantization. The main idea of our approach is to quantize the nodes’ estimates based on the progress of the algorithm, which helps to eliminate the quantized errors. Under the adaptive quantization, we then derive the bounds on the convergence rates of the proposed method as a function of the bandwidths and the underlying network topology, for both convex and strongly convex objective functions. Our results shows that using the adaptive quantization, the rate of convergence of distributed consensus-based subgradient methods with and without quantization are the same, except for a factor which captures the number of quantization bits. Finally, we provide numerical simulations to compare the convergence properties of the distributed gradient methods with and without quantization for solving the well-known regression problems over networks, for both quadratic and absolute loss functions.

1 Introduction

In this paper, we consider optimization problems that are defined over a network of nodes\(^1\). The objective function is composed of a sum of local functions where each function is known by only one node. In addition, each node is only allowed to interact with its neighboring nodes that are connected to it through the network. We assume no central coordination between the nodes and since each node knows only its local function, they are required to cooperatively solve the problems. This necessitates the development of distributed algorithms, which can be done under communication and computation constraints.

Distributed algorithms for solving these problems have been received a surge of interests during the last decade. This is driven by a broad applications of distributed methods for solving many engineering problems, where centralized approaches may not be practical or sometimes impossible. A standard example is the problem of estimating the radio frequency in a wireless network of sensors. The goal is to cooperatively estimate the radio-frequency power spectrum density through minimizing a loss function defined over the entire measured data by the sensors, which are scattered across a large geographical area \(^1\). Another possible application is the problem of distributed

\(^1\)In this paper, nodes can be used to present for processors, robotics, or sensors.
information processing in edge (fog) computing, which has recently received a surge of interests due to the emergence of the Internet of Things [2].

Distributed algorithms for these problems have received wide attention during the last decade, mostly focusing on three classes of algorithms, namely, the alternating direction method of multipliers (ADMM) [1, 17, 33, 35], distributed dual methods (mirror descent/dual averaging) [6, 9, 14, 34, 37], and distributed gradient algorithms [3–5, 13, 16, 22, 23, 27, 29, 32, 38]. The focus in this paper will be on distributed (sub)gradient algorithms, as they have the benefits (in terms of convergence rates and simplicity) of both ADMM and dual methods. We refer interested readers to the recent survey paper [21] for a summary of existing results in this area.

Our focus in this paper is to study the convergence properties of distributed (sub)gradient methods when the nodes only communicate their quantized values, often referred to as quantized communication. Such quantization is necessary since in many practical applications the nodes often have to share a limited communication bandwidth. Different variants of distributed (sub)gradient methods under quantized communication have been studied in [7, 8, 15, 19, 26, 28, 36]. In [15, 19] the authors only show the convergence to a neighborhood around the optimal of the problem due to the quantized error. An asymptotic convergence to the optimal has been studied in [8, 26, 36]; however, a condition on the growing communication bandwidth is assumed in these works to remove the quantized error. Recently, the authors in [7, 28] show the asymptotic convergence of such methods under random quantization using only finite communication bandwidth. In particular, in [28] the rate of convergence in expectation is shown to occur at $O(1/k^{(1-\gamma)/2})$ for some $\gamma \in (0, 1)$, when the objective functions are strongly convex and smooth. On the other hand, the authors in [7] study distributed subgradient methods for nonsmooth problems and analyze their convergence rates by utilizing the techniques from stochastic approximation approach. Specifically, such algorithms asymptotically converge to the optimal value in expectation at a rate $O(\ln(k)/k^{1/4})$ and $O(\ln(k)/k^{1/3})$ for convex and strongly convex functions, respectively.

It is worth to note that without quantized communication distributed subgradient (DSG) methods converge to the optimal value at a rate $O(\ln(k)/k^{1/2})$ and $O(\ln(k)/k)$ for convex and strongly convex functions, respectively; see for example [21]. This is considered as an optimal rate of DSG without quantization. Thus, there is a gap in the rate of convergence of such methods between using quantization [7, 28] and no quantization. Our goal in this paper, therefore, is to close such convergence gap, where we show that DSG with quantization can perform as good as the case without quantization. In particular, we first provide a novel quantization scheme for distributed subgradient methods, which can alleviate the issues of finite communication bandwidth. Under our proposed quantization methods, we show that DSG with or without quantization achieves the same convergence rate, except for a constant factor which captures the communication capacity. The main contributions of our paper are stated as follows.

**Main Contributions.** The main contribution of this paper is to study the convergence properties of distributed subgradient methods under the impact of quantized communication. In particular, based on a simple observation in DSG we first propose a novel quantization method, which we will refer to as adaptive quantization. Such method, while only uses a finite number of bits for quantization, helps to asymptotically alleviate the quantized error. In addition, using this quantization method allows us to accelerate the rate of convergence of DSG, i.e., the convergence occurs at the rates $O(\eta \ln(k)/\sqrt{k})$ and $O(\eta \ln(k)/k)$ for convex and strongly convex functions, respectively. Here, $k$ is the number of iterations and $\eta$ is some constant depending on the network topology and the number of quantized bits $b$. Thus, our results show that under our proposed quantized scheme, the rates of convergence of DSG are unaffected by quantization, except for a factor which captures the number of quantization bits. Finally, we provide simulations to compare the convergence properties of DSG with and without quantization, for solving the well know least square problems over networks.
1.1 Notation and Definition

We first introduce here a set of notation and definition used throughout this paper. We use boldface to distinguish between vectors in $\mathbb{R}^d$ and scalars in $\mathbb{R}$. Given a collection of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$, we denote by $\mathbf{X}$ a matrix in $\mathbb{R}^{n \times d}$, whose $i$-th row is $\mathbf{x}_i^T$. We then denote by $\|\mathbf{x}\|$ and $\|\mathbf{X}\|$ the Euclidean norm and the Frobenius norm of $\mathbf{x}$ and $\mathbf{X}$, respectively. Let $\mathbf{1}$ be the vector whose entries are 1 and $\mathbf{I}$ the identity matrix. Given a closed convex set $\mathcal{X}$, we denote by $\mathcal{P}_\mathcal{X} [\mathbf{x}]$ the projection of $\mathbf{x}$ to $\mathcal{X}$. Given a nonsmooth convex function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\partial f(\mathbf{x})$ its subdifferential estimated at $\mathbf{x}$, i.e., $\partial f(\mathbf{x}) \triangleq \{ g \in \mathbb{R}^d | f(\mathbf{y}) \geq f(\mathbf{x}) + g^T (\mathbf{y} - \mathbf{x}), \forall \mathbf{y} \in \mathbb{R}^d \}$ is the set of subgradients of $f$ at $\mathbf{x}$. Since $f$ is convex, $\partial f(\cdot)$ is nonempty. The function $f$ is $L$-Lipschitz continuous if and only if
\[
| f(\mathbf{x}) - f(\mathbf{y}) | \leq L \| \mathbf{x} - \mathbf{y} \|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.
\] Note that the $L$-Lipschitz continuity of $f$ is equivalent to the subgradients of $f$ are uniformly bounded by $L$. A function $f$ is $\mu$-strongly convex if and only if $f$ satisfies $\forall \mathbf{x}, \mathbf{y} \[f(\mathbf{y}) - f(\mathbf{x}) - g(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \geq \frac{\mu}{2} (\mathbf{y} - \mathbf{x})].
\] Uniform Quantization: Given a finite interval $[c, d]$ we divide this interval into a $B \geq 2$ number of bins $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_B$ where $\tau_1 = c$ and $\tau_B = d$. In addition, we assume that these points $\tau_i$ are uniformly spaced with a distance $\Delta$, i.e., $\Delta = \tau_{i+1} - \tau_i$ for all $i = 0, \ldots, B - 1$ implying that $\Delta = (d - c) / (B - 1)$. Here, to present the points $\tau_i$ we need a finite $b$ bits where $b = \log_2 (B)$. Next, given a value $\mathbf{x} \in [c, d]$ we denote by $q = Q(\mathbf{x})$ its quantized value where
\[
Q(\mathbf{x}) \triangleq \min_i | \tau_i - \mathbf{x} |.
\] If $\tau_i$ and $\tau_{i+1}$ achieve the minimal value in Eq. (4), then we set $Q(\mathbf{x}) = \tau_i$. Also, by Eq. (3) the quantized error is given as
\[
| \mathbf{x} - Q(\mathbf{x}) | \leq \Delta = \frac{d - c}{B - 1}.
\] With some abuse of notation, given a vector $\mathbf{x} \in \mathcal{X}$ we denote by $\mathbf{q} = Q(\mathbf{x})$, where $q^i = Q(x^i)$, the quantization of $i$-th coordinate of $\mathbf{x}$, for $i = 1, 2, \ldots, d$. Here, each $q^i$ is defined by using Eq. (3) with a uniform distance $\Delta^i$ associated with each interval $[c^i, d^i]$, for all $i = 1 \ldots, d$.

2 Distributed Optimization over Networks

In this paper, we consider an optimization problem defined over a network of $n$ nodes. More precisely, associated with each node $i$ is a nonsmooth convex function $f_i : \mathbb{R}^d \to \mathbb{R}$ and a compact convex set $\mathcal{X} \subset \mathbb{R}^d$. The goal of the nodes is to solve the following minimization problem
\[
\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \triangleq \sum_{i=1}^{n} f_i(\mathbf{x}).
\] We assume no central coordination between the nodes and since each node $i$ knows only its local function $f_i$, the nodes are required to cooperatively solve the problem. We are interested in distributed consensus-based (sub)gradient methods implying that each node $i$ maintains its local copy $\mathbf{x}_i \in \mathbb{R}^d$ of $\mathbf{x}^*$, a solution of problem (5). Each node $i$ then exchanges its local copies with its neighboring nodes, where the goal of the nodes is to asymptotically drive the nodes’ estimates $\mathbf{x}_i$ to $\mathbf{x}^*$. 

3
Note that since the set $\mathcal{X}$ is compact, there exists a point $\mathbf{x}^*$ which solves problem (5). However, $\mathbf{x}^*$ may not be unique. We will use $\mathcal{X}^*$ to denote the set of optimal solutions to problem (5). Given a solution $\mathbf{x}^* \in \mathcal{X}^*$ we denote $f^* = \sum_{i=1}^{n} f_i(\mathbf{x}^*)$. Also, due to the compactness of $\mathcal{X}$, it is obvious that the subgradients of $f_i$ are uniformly bounded in $\mathcal{X}$. We state this observation formally in the following proposition.

**Proposition 1.** There exists a positive constant $L_i$, for all $i \in \mathcal{V}$, such that the 2-norm of subgradients $g_i(\cdot)$ of $f_i$ are uniformly bounded by $L_i$ in $\mathcal{X}$, i.e., the following condition holds

$$\|g_i(x)\| \leq L_i, \quad \text{for all } x \in \mathcal{X}. \tag{6}$$

Regarding the network topology and inter-node communications, we assume that each node is only allowed to interact with its neighbors that are directly connected to it through a connected and undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, n\}$ and $\mathcal{E} = (\mathcal{V} \times \mathcal{V})$ are the vertex and edge sets, respectively. In particular, node $i$ can communicate directly only with its neighbors $j \in \mathcal{N}_i$, where $\mathcal{N}_i := \{j \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}$ is the set of node $i$ is neighbors. Finally, for ease of exposition and notational convenience, we consider the problem (5) when the variable $\mathbf{x}$ is a scalar, i.e., $d = 1$. Extensions for the case $d > 1$ as well as time-varying graphs will be provided in the Appendix.

### 3 Distributed Subgradient Methods Under Quantization

For solving problem (5), we are interested in DSG methods, which were first studied and analyzed rigorously in [23,24]. In these methods each node $i$ iteratively updates $x_i$ as

$$x_i(k + 1) = \mathcal{P}_{\mathcal{X}} \left[ \sum_{j \in \mathcal{N}_i} a_{ij} x_j(k) - \alpha(k) g_i(x_i(k)) \right], \tag{7}$$

where $\alpha(k)$ is some sequence of stepsizes, $g_i(x_i(k)) \in \partial f_i(x_i(k))$, and $a_{ij}$ is some positive weight which node $i$ assigns for $x_j$ received from node $j$. Here, for simplicity we only consider the adjacency weights $a_{ij}$ corresponding to $\mathcal{G}$ are fixed. An extension to the case of time-varying weights will be provided in the Appendix. We make the following assumption about the weights $a_{ij}$, which captures the topology of $\mathcal{G}$.

**Assumption 1.** The matrix $\mathbf{A}$, whose $(i, j)$-th entries are $a_{ij}$, is doubly stochastic, i.e., $\sum_{i=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ij} = 1$. Moreover, $\mathbf{A}$ is irreducible and aperiodic. Finally, the weights $a_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}$ otherwise $a_{ij} = 0$.

This assumption also implies that $\mathbf{A}$ has 1 as the largest singular value and others are strictly less than 1; see for example, the Perron-Frobenius theorem [11]. Also, we denote by $\sigma_2 \in (0, 1)$ the second largest singular value of $\mathbf{A}$, which by the Courant-Fisher theorem [11] gives

$$\left\| \mathbf{A} \left( \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right) \right\| \leq \sigma_2 \left\| \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right\|. \tag{8}$$

Our focus in this paper is to study the impact of quantized communication between the nodes on the performance of DSG. In particular, at any iteration $k \geq 0$ the nodes are only allowed to send and receive the quantized values of their local copies to their neighboring nodes. Due to the quantization, we first modify the update in Eq. (7) to take into account the quantized error. That is, each node $i$, for all $i \in \mathcal{V}$, now considers the following update

$$x_i(k + 1) = \mathcal{P}_{\mathcal{X}} \left[ \underbrace{x_i(k) - q_i(k)}_{\text{quantized error}} + \sum_{j \in \mathcal{N}_i} a_{ij} q_j(k) - \alpha(k) g_i(x_i(k)) \right], \tag{9}$$

where $q_i(k)$ is the quantized value of $x_i(k)$.
where \( q_i(k) = Q_k(x_i(k)) \), for all \( i \in \mathcal{V} \), is the quantized value of \( x_i(k) \) at iteration \( k \). Second, unlike in Eq. (4) where the quantized interval \([c, d]\) is fixed, each node \( i \) uses the uniform quantized operator \( Q \) defined in Eq. (5) over finite time-varying intervals \([c_i(k), d_i(k)]\) for all \( k \geq 0 \). Thus, \( Q_k \) now depends on time, which we refer to as adaptive quantization. As will be seen shortly, such intervals are defined carefully at each node for each iteration to remove the quantized errors over the network. The main idea behind our proposed adaptive quantization method are explained in the next section.

3.1 Adaptive Quantization

Let \( w_i(k) \) be denoted as

\[
w_i(k) = P_X \left[ x_i(k) - q_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij} q_j(k) \right].
\]

Since the subgradients \( g_i(\cdot) \) are uniformly bounded by \( L_i \) in \( \mathcal{X} \) and by the nonexpansiveness of the projection operator, Eq. (8) gives

\[
|x_i(k + 1) - w_i(k)| \leq |\alpha(k) g_i(x_i(k))| \leq L_i \alpha(k),
\]

which implies that \( x_i(k + 1) \in [w_i(k) - L_i \alpha(k), w_i(k) - L_i \alpha(k)] \). Thus, at each iteration \( k \geq 0 \) each node \( i \) refines its quantized interval as follows

\[
c_i(k + 1) = w_i(k) - L_i \alpha(k) \quad \text{and} \quad d_i(k + 1) = w_i(k) + L_i \alpha(k).
\]

(10)

For each interval, we again use a finite number of bits \( b \) to divide \([c_i(k), d_i(k)]\) into a \( B = 2^b \) number of bins \( \tau_i^b(k) = c_i(k) \leq \tau_i^2(k) \leq \ldots \leq \tau_i^B(k) = d_i(k) \), for all \( i \in \mathcal{V} \) and \( k \geq 0 \). This implies that

\[
|\tau_i^\ell(k) - \tau_i^{\ell+1}(k)| = \Delta_i(k) = \frac{2L_i \alpha(k - 1)}{B - 1}, \quad \text{for all } \ell = 1, \ldots, B - 1 \text{ and } i \in \mathcal{V}.
\]

Since \( x_i(k) \in [c_i(k), d_i(k)] \) Eq. (8) gives \( q_i(k) \in [c_i(k), d_i(k)] \), for all \( k \geq 0 \). Thus, we obtain the following bound for the quantized error

\[
|x_i(k) - q_i(k)| \leq \frac{2L_i \alpha(k - 1)}{B - 1}, \quad \forall i \in \mathcal{V}, \ k \geq 0.
\]

(11)

Here, the refined interval \([\ell_i(k), u_i(k)]\) in Eq. (10) provides a more adaptive quantization scheme to the progress of the variables \( x_i \). Indeed, it is obvious to see that this scheme reduces the quantization error to zero as long as the stepsizes \( \alpha(k) \) decay to zero. This is opposed to uniform quantization over fixed interval, which usually causes a fixed error, preventing the algorithm converges to the optimal solution of problem (3).

Finally, we give a simple interpretation of the update in Eq. (9) with adaptive quantization: at any time \( k \geq 0 \), each node \( i \) first quantizes its estimate \( x_i(k) \) by using the quantization \( Q \) over interval \([c_i(k), d_i(k)]\) in Eq. (10), and then sends this quantized value to its neighbors \( j \in \mathcal{N}_i \). Such interval \([c_i(k), d_i(k)]\) can be defined and based only on local information available at node \( i \). Each node \( i \) then combines its quantized error \( x_i(k) - q_i(k) \) with the weighted quantized values received from its neighbors \( j \in \mathcal{N}_i \), with the goal of seeking consensus on their estimates. Finally, the nodes apply the subgradients of their respective objective functions to update their estimates, pushing the consensus point toward the optimal set \( \mathcal{X}^* \). Here \( x_i(k) - q_i(k) \), for all \( i \in \mathcal{V} \), plays the role of feedback signals to correct for the errors due to quantization across the network. The distributed subgradient algorithm with quantization in communication is formally formulated in Algorithm 1.
4 Convergence Analysis

The focus of this section is to analyze the performance of DSG methods under adaptive quantization given in Algorithm 1. In particular, we provide a rigorous analysis which establishes the convergence rate of the proposed algorithm for two cases, namely, when the functions $f_i$ are convex and strongly convex, respectively. For ease of exposition, we delay all the proofs of the results in this section to Section 6. The main steps of the analysis are as follows.

We first show that the distances between the estimates $x_i(k)$ to their average $\bar{x}(k)$ asymptotically converge to zero, implying the nodes achieve a consensus. The main idea of this step is to utilize the fact that quantization error in Eq. (11) is a function of $x_i(k)$ if the stepsize $\alpha(k)$ decays as $\alpha(k) = 1/\sqrt{k} + 1$, the objective function value $f$ estimated at each $z_i(k)$ converges to the optimal value with a rate $O(\eta \ln(k+1)/\sqrt{k} + 1)$, where $\eta$ is some constant depending on the network topology, function properties, and the number of quantization bits $b$. Next, under the strong convexity assumption on $f_i$ we show that the variables $z_i(k)$ converge to a solution $x^*$ of problem (5) with a rate $O(\eta \ln(k+1)/(k+1))$ under a proper choice of stepsizes $\alpha(k)$. Our results suggest that under the adaptive quantization, the rates of convergence of distributed subgradient methods are unaffected by quantization, except for a factor which captures the number of quantization bits.

We start our analysis by introducing more notation. Given a collection of scalars $x_1, \ldots, x_n$ in $\mathbb{R}$ we denote by $\mathbf{x}$ a vector whose $i$-th entries are $x_i$, i.e.,

$$
\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n.
$$

In addition, let $\bar{x}$ be the average of the vector $\mathbf{x}$, i.e., $\bar{x} = 1/n \sum_{i=1}^n x_i$. Without loss of generality, we consider the case $\mathcal{X} = [a, b]$ for some real numbers $a \leq b \in \mathbb{R}$, and note that the multi-dimensional case of $\mathcal{X}$ is presented in the Appendix. This simplification will allow us to write explicitly the projection on the set $\mathcal{X}$. In particular, Eq. (12) can now be rewritten as

$$
v_i(k) = \sum_{j \in \mathcal{N}_i} a_{ij} x_j(k) + x_i(k) - q_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij} (q_j(k) - x_j(k)) - \alpha(k) g_i(x_i(k))
\tag{13}
$$

$$
x_i(k+1) = \mathcal{P}_\mathcal{X} [v(k)] = \begin{cases} 
v_i(k) & \text{if } v_i(k) \in [a, b] \\ a & \text{if } v_i(k) < a \\ b & \text{otherwise.} \end{cases}
\tag{14}
$$

Given $v_i$ we denote by $\xi_i$ the error due to projection of $v_i$ to $\mathcal{X}$, i.e., $\xi(v_i) = v_i - \mathcal{P}_\mathcal{X}[v_i]$. Using this notation and the adjacency matrix $A$ in Assumption 1 Eqs. (13) and (14) can be written in matrix form as

$$
v(k) = Ax(k) + (A - I)(q(k) - x(k)) - \alpha(k) g(x(k))
\tag{15}
$$

$$
x(k+1) = \mathcal{P}_\mathcal{X} [v(k)] = v(k) - \xi(v(k)),
\tag{16}
$$
where $\mathcal{P}_X[v(k)]$ denotes the component-wise projection. Moreover, since $\mathbf{1}^T \mathbf{A} = \mathbf{1}$, Eqs. (15) and (16) give
\begin{equation}
\bar{v}(k) = \bar{x}(k) - \alpha(k) \frac{1}{n} \sum_{i=1}^{n} g_i(x_i(k)) \tag{17}
\end{equation}
\begin{equation}
\bar{x}(k+1) = \bar{v}(k) - \bar{\xi}(k), \tag{18}
\end{equation}
where $\bar{\xi}(k)$ is the average of $\xi(v_i(k))$. Let $L = \sum_{j=1}^{n} L_i$, where $L_i$ is given in Proposition 1. Finally, for notational convenience let $\mathbf{W} = \mathbf{I} - 1/n \mathbf{1} \mathbf{1}^T$ and we denote by $y(k)$ the consensus error and $r(k)$ the optimal error as
\begin{equation*}
y(k) = \mathbf{W} \mathbf{x}(k) = \mathbf{x}(k) - \bar{\mathbf{x}}(k) \mathbf{1}, \quad r(k) = \bar{x}(k) - x^*.
\end{equation*}

4.1 Preliminaries

We now consider the following sequence of lemmas, which provides fundamental preliminaries for our results given in the next sections. We first study some important properties of the projection errors $\xi_i$, which can be viewed as the one-dimensional version of Lemma 5 for the general convex set $\mathcal{X}$, stated in the Appendix. As will be seen in the analysis of this lemma, considering one dimensional case gives us more insight about the decreasing of the projection errors.

**Lemma 1.** Suppose that Assumption 1 holds. Let the sequence \{x_i(k)\}, for all $i \in \mathcal{V}$, be generated by Algorithm 2. Then for all $i \in \mathcal{V}$ we have
1. For all $k \geq 0$
\begin{equation}
|\xi_i(v_i(k))| \leq \frac{(B+1)L_i + \sum_{j \in \mathcal{N}_i} a_{ij} L_j}{B-1} \alpha(k-1). \tag{19}
\end{equation}
2. Given a point $w \in \mathcal{X} = [a, b]$, we have for all $k \geq 0$
\begin{equation}
(v_i(k) - w) \xi_i(v_i(k)) \geq \xi_i^2(v_i(k)), \quad \forall i \in \mathcal{V}. \tag{20}
\end{equation}

Next we provide an upper bound for the consensus error $\|y(k)\|$ in the following lemma.

**Lemma 2.** Suppose that Assumption 1 holds. Let the sequence \{x_i(k)\}, for all $i \in \mathcal{V}$, be generated by Algorithm 2. In addition, let $\{\alpha(k)\}$ be a nonnegative nonincreasing sequence of stepsizes. Then, we have
1. The consensus error $y(k)$ satisfies
\begin{equation}
\|y(k+1)\| \leq \sigma_2^{k+1} \|y(0)\| + \frac{L(2B + 3)}{B-1} \sum_{t=0}^{k} \sigma_2^{k-t} \alpha(t-1). \tag{21}
\end{equation}
2. If $\lim_{k \to \infty} \alpha(k) = 0$ then
\begin{equation}
\lim_{k \to \infty} x_i(k) = \lim_{k \to \infty} x_j(k), \quad \forall i, j \in \mathcal{V}. \tag{22}
\end{equation}

In addition, if $\alpha(k)$ is also square-summable, i.e.,
\begin{equation}
\sum_{k=0}^{\infty} \alpha^2(k) < \infty, \tag{23}
\end{equation}
then for all $k \geq 0$ we have
\begin{equation}
\sum_{t=0}^{k} \alpha(t) \|y(t)\| \leq \frac{\alpha(0) \|y(0)\|}{1 - \sigma_2} + \frac{L(2B + 3)}{(B-1)(1-\sigma_2)} \sum_{t=0}^{k-1} \alpha^2(t-1) < \infty. \tag{24}
\end{equation}
3. If \( \alpha(k) = 1/\sqrt{k + 1} \) then we have for all \( K \geq 0 \),
\[
\sum_{k=0}^{K} \alpha(k)\|y(k + 1)\| \leq \frac{\|y(0)\|}{1 - \sigma_2} + \frac{L(4B + 1)}{(2B - 1)(1 - \sigma_2)}(1 + \ln(k + 1)).
\]

Finally, we study an upper bound for the optimal distance \( r^2(k) = (\bar{x}(k) - x^*)^2 \) in the following lemma.

**Lemma 3.** Suppose that Assumption 2 holds. Let the sequence \( \{x_i(k)\} \), for all \( i \in V \), be generated by Algorithm 1. In addition, let \( x^* \in \mathcal{X}^* \) be a solution of problem (5). Then, we have
\[
r^2(k + 1) \leq r^2(k) + \frac{2(2B + 1)L}{n(B - 1)}\alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2}\alpha^2(k - 1)
- \frac{2\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*).
\]

### 4.2 Convergence Results of Convex Functions

We now present the first main result of this paper, which is the rate of convergence of Algorithm 1 to the optimal value of problem (5) when the local functions \( f_i \) are convex. Since the update of \( \bar{x}(k) \) in Eq. (18) can be viewed as a variant of a centralized projected subgradient methods used to solve problem (5), we utilize standard techniques in the analysis of these methods to derive the rate of Algorithm 1. Specifically, at any time \( k \geq 0 \) if each node \( i \in V \) maintains a variable \( z_i(k) \) to compute the time-weighted average of its estimate \( x_i(k) \) and if the stepsize \( \alpha(k) \) decays as \( \alpha(k) = 1/\sqrt{k + 1} \), the objective function value \( f \) in Eq. (5) estimated at each \( z_i(k) \) converges to the optimal value with a rate \( O(\eta \ln(k + 1)/\sqrt{k + 1}) \), where \( \eta \) is some constant depending on the algebraic connectivity \( 1 - \sigma_2 \) of the network, the number of quantized bits \( b \), and the Lipschitz constants \( L_i \) of \( f_i \). We also note that this condition on the stepsizes is also used to study the convergence rate of centralized subgradient methods [25]. The following theorem is used to show the convergence rate of Algorithm 1.

**Theorem 1.** Suppose that Assumption 2 holds. Let the sequence \( \{x_i(k)\} \), for all \( i \in V \), be generated by Algorithm 1. In addition, let \( \alpha(k) = 1/\sqrt{k + 1} \), a nonnegative nonincreasing sequence of stepsizes with \( \alpha(0) = 1 \). Moreover, suppose that each node \( i \), for all \( i \in V \), stores a variable \( z_i \in \mathbb{R} \) initiated arbitrarily in \( \mathcal{X} \) and updated as
\[
z_i(k) = \frac{\sum_{t=0}^{k} \alpha(t)x_i(t)}{\sum_{t=0}^{k} \alpha(t)}, \quad \forall i \in V.
\]

Let \( x^* \in \mathcal{X}^* \) be a solution of problem (5). Then for all \( i \in V \) and \( k \geq 0 \) we have
\[
f(z_i(k)) - f^* \leq \frac{m^2(0) + L(4B - 1)\|y(0)\|}{2\sqrt{k + 1}} - \frac{L(4B + 1)\|y(0)\|}{(B - 1)(1 - \sigma_2)}\sqrt{k + 1}
+ \frac{2B(5B + 7)L^2}{(B - 1)^2(1 - \sigma_2)^2}(1 + \ln(k + 1)).
\]

It is worth to mention that under the choice of \( \alpha(k) = 1/(k + 1) \), for all \( k \geq 0 \), one can show that \( x_i(k) \) asymptotically converges to \( x^* \) for all \( i \in V \). This is a consequence of Lemmas 2 and 3 and some standard analysis. The following lemma is to state this result, where its analysis is omitted and can be found in [8, Theorem 3].

**Lemma 4.** Suppose that Assumption 2 holds. Let the sequence \( \{x_i(k)\} \), for all \( i \in V \), be generated by Algorithm 1. In addition, let \( \alpha(k) = 1/k + 1 \) for \( k \geq 0 \). Then we obtain
\[
\lim_{k \to \infty} x_i(k) = x^*, \quad \text{for all } i \in V,
\]
for some \( x^* \) that is a solution of Problem (5).
4.3 Convergence Results of Strongly Convex Case

In this section, our goal is to study the convergence rate of Algorithm 1 when the local functions $f_i$ are strongly convex, that is, we make the following assumption on $f_i$

**Assumption 2.** Each function $f_i$ is strongly convex with some positive constant $\mu_i$, i.e., the condition holds.

Under this assumption, we show that if each node $i \in V$ maintains a variable $z_i(k)$ to compute the time average of its estimate $x_i(k)$ and if the stepsize $\alpha(k)$ decays as $\alpha(k) = a / k + 1$ for some proper constant $a$, the variable $z_i(k)$ converges to the optimal solution $x^*$ of problem (5) with a rate $O(\eta \ln(k+1)/k + 1)$, where $\eta$ is some constant depending on the algebraic connectivity $1 - \sigma_2$ of the network, the number of quantized bits $b$, and the constants $L_i$ and $\mu_i$ of $f_i$. The following theorem is used to show the convergence rate of Algorithm 1 under Assumption 2.

**Theorem 2.** Suppose that Assumptions 1 and 2 hold. Let the sequence \{$x_i(k)$\}, for all $i \in V$, be generated by Algorithm 1. We denote by $\mu = \min_{i \in V} \mu_i$. In addition, let $\{\alpha(k)\} = a / k + 1$ for some $a \geq 1 / \mu$. Moreover, suppose that each node $i$, for all $i \in V$, stores a variable $z_i \in \mathbb{R}$ initiated arbitrarily in $X$ and updated as

$$z_i(k) = \frac{\sum_{t=0}^{k} x_i(t)}{k+1}, \quad \forall i \in V.$$  \hspace{1cm} (30)

Let $x^* \in X^*$ be a solution of problem (5). Then for all $i \in V$ and $k \geq 0$ we have

$$(z_i(k) - x^*)^2 \leq \frac{4L(2B+1)\alpha^2(0)\|y(0)\|}{(B-1)(1-\sigma_2)} \frac{1}{k+1} + \frac{24L^2(B+1)^2\alpha^2(0)}{(B-1)^2(1-\sigma_2)} \frac{1 + \ln(k+1)}{k+1}. \hspace{1cm} (31)$$

5 Simulations

In this section, we apply Algorithm 1 for solving linear regression problems, the most popular technique for data fitting [10, 31] in statistical machine learning, over a network of processors under random quantization. The goal of this problem is to find a linear relationship between a set of variables and some real value outcome. That is, given a training set $S = \{(a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}\}$ for $i = 1, \ldots, n$, we want to learn a parameter $x$ that minimizes

$$\min_{x \in \mathcal{X}} \sum_{i=1}^{n} f_i(x; a_i, b_i),$$

where $\mathcal{X} = [-1, 1]^d$ and $d = 10$, i.e., $x, a_i \in \mathbb{R}^{10}$. Here, $f_i$ are the loss functions defined over the dataset. For the purpose of our simulation, we will consider two loss functions, namely, quadratic loss and absolute loss functions. While the quadratic loss is strongly convex, the absolute loss is only convex.

First, when $f_i$ are quadratic, we have the well-known least square problem given as

$$\min_{x \in \mathcal{X}} \sum_{i=1}^{n} (a_i^T x - b_i)^2.$$  

Second, regression problems with absolute loss functions (or L1 norm) is often referred to as robust regression, which is known to be robust to outliers [12], given as follows

$$\min_{x \in \mathcal{X}} \sum_{i=1}^{n} |a_i^T x - b_i|.$$
We consider simulated training data sets, i.e., \((a_i, b_i)\) are generated randomly with uniform distribution between \([0, 1]\). We consider the performance of the distributed subgradient methods on an undirected connected graph of 100 nodes, i.e., \(G = (V, E)\) and \(n = |V| = 100\). Our graph is generated as follows.

1. In each network, we first randomly generate the nodes’ coordinates in the plane with uniform distribution.

2. Then any two nodes are connected if their distance is less than a reference number \(r\), e.g., \(r = 0.4\) for our simulations.

3. Finally we check whether the network is connected. If not we return to step 1 and run the program again.

To implement our algorithm, the adjacency matrix \(A\) is chosen as a lazy Metropolis matrix corresponding to \(G\), i.e.,

\[
A = [a_{ij}] = \begin{cases} 
\frac{1}{2(\max(|N_i|,|N_j|))}, & \text{if } (i,j) \in E \\
0, & \text{if } (i,j) \notin E \text{ and } i \neq j \\
1 - \sum_{j \in N_i} a_{ij}, & \text{if } i = j
\end{cases}
\]

It is straightforward to verify that the lazy Metropolis matrix \(A\) satisfies Assumption 1.

5.1 Convergence of Function Values

In this simulation, we apply variants of distributed subgradient methods for solving the linear regression problems. In particular, we compare the performance of such methods for three different scenarios, namely, DSG with no quantization (a.k.a Eq. (7)), DSG with time-varying quantization in [8], distributed stochastic approximation under random quantization [7], and the proposed Algorithm 1 with adaptive quantization. The plots in Fig. 1 show the convergence of these four methods for both quadratic and absolute loss functions.

Note that, the work in [8] achieves the same rate of convergence as the one in [24], but requires that the nodes eventually exchange an infinite number of bits. On the other hand, the work
in \[7\] only requires a finite number of bits, but achieves a slow rate of convergence. The adaptive quantization in this paper achieves both benefits of \[8\] and \[7\], that is, achieving the same rate as the algorithm without quantization but only using a finite bit of quantization. In addition, as observed in Fig. 1a for quadratic loss and in Fig. 1b for absolute loss, Algorithm 1 performs almost as well as the one without quantization \[24\], and significantly better than the algorithms in \[7,8\].

5.2 Impacts of the Number of Bits \(b\)

Here, we consider the impacts of the number of communication bits \(b\) on the performance of Algorithm 1. In Fig. 2 we plots the number of iterations, needed to obtain the relative error \(f(z_i(k)) - f^* / f^* \leq 0.05\), as a function of \(b\). As we can see the more bits we use the faster the algorithm converges. Moreover, even we use very small number of bits, for example, \(b = 4\), the algorithm still works very well. Finally, these plots seem to describe the curve \((B + 1)^2 / (B - 1)^2\) up to some constant in the upper bound of convergence rates given in Theorems 1 and 2, where \(B = 2^b\) is the size of the bandwidths. This implies that this simulation seems to agree with our results.

6 Proofs of Results in Section 4

6.1 Proof of Lemma 1

**Proof.** 1. Recall that \(\xi_i(v_i(k)) = v_i(k) - \mathcal{P}_X[v_i(k)]\). By Eq. (14) we consider the following three cases.

(a) If \(v_i(k) \in X = [a, b]\) then \(\xi_i(v_i(k)) = v_i(k) - v_i(k) = 0\).

(b) If \(v_i(k) < a\) then \(\mathcal{P}_X[v_i(k)] = a\), implying \(\xi_i(v_i(k)) = v_i(k) - \mathcal{P}_X[v_i(k)] \leq 0\). On the other
2. Let $v_i(k) \in [a, b]$ for all $i \in \mathcal{V}$ and $k \geq 0$, Eq. (13) gives

$$\xi_i(v_i(k)) \geq v_i(k) - \mathcal{P}_X[v_i(k)]$$

$$\geq \sum_{j \in \mathcal{N}_i} a_{ij}x_j(k) - a + x_i(k) - q_i(k)$$

$$+ \sum_{j \in \mathcal{N}_i} a_{ij}(q_j(k) - x_j(k)) + \alpha(k)g_i(x_i(k))$$

$$\geq x_i(k) - q_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij}(q_j(k) - x_j(k)) + \alpha(k)g_i(x_i(k)),$$

which since $\xi_i(v_i(k)) < 0$ implies that

$$|\xi_i(v_i(k))| \leq |x_i(k) - q_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij}(q_j(k) - x_j(k)) + \alpha(k)g_i(x_i(k))|$$

$$\leq |x_i(k) - q_i(k)| + \sum_{j \in \mathcal{N}_i} |a_{ij}|q_j(k) - x_j(k)| + \alpha(k)|g_i(x_i(k))|$$

$$\leq \frac{2L_i\alpha(k - 1)}{B - 1} + \frac{\sum_{j \in \mathcal{N}_i} a_{ij}L_j\alpha(k - 1)}{B - 1} + L_i\alpha(k)$$

$$\leq \frac{(B + 1)L_i + \sum_{j \in \mathcal{N}_i} a_{ij}L_j}{B - 1}\alpha(k - 1).$$

(c) If $v_i(k) > b$ then $\mathcal{P}_X[v_i(k)] = b$, implying $\xi_i(v_i(k)) = v_i(k) - \mathcal{P}_X[v_i(k)] \geq 0$. Similarly to the case (b) above, since $x_i(k) \in [a, b]$ for all $i \in \mathcal{V}$ and $k \geq 0$, by Eq. (13) we have

$$\xi_i(v_i(k)) \leq x_i(k) - q_i(k) + \sum_{j \in \mathcal{N}_i} a_{ij}(q_j(k) - x_j(k)) + \alpha(k)g_i(x_i(k)),$$

which since $\xi_i(v_i(k)) > 0$ implies that

$$|\xi_i(v_i(k))| \leq \frac{(B + 1)L_i + \sum_{j \in \mathcal{N}_i} a_{ij}L_j}{B - 1}\alpha(k - 1).$$

From these three cases, we have Eq. (19).

2. Let $w \in \mathcal{X} = [a, b]$. Then we consider

$$(v_i(k) - w)\xi_i(v_i(k)) = (v_i(k) - \mathcal{P}_X[v_i(k)] + \mathcal{P}_X[v_i(k)] - w)\xi_i(v_i(k))$$

$$= \xi_i^2(v_i(k)) + \left(\mathcal{P}_X[v_i(k)] - w\right)\left(v_i(k) - \mathcal{P}_X[v_i(k)]\right).$$

We now investigate the second term of the previous relation for three cases

(a) If $v_i(k) \in [a, b]$ then $q_i = 0$.

(b) If $v_i(k) < a$ then $\mathcal{P}_X[v_i(k)] = a$. Thus, we have $\mathcal{P}_X[v_i(k)] - w \leq 0$ and $v_i(k) - \mathcal{P}_X[v_i(k)] \leq 0$, implying $q_i \geq 0$.

(c) If $v_i(k) > b$ then $\mathcal{P}_X[v_i(k)] = b$. Thus, we have $\mathcal{P}_X[v_i(k)] - w \geq 0$ and $v_i(k) - \mathcal{P}_X[v_i(k)] \geq 0$, implying $q_i \geq 0$.

Combining these three cases and by Eq. (32) we have Eq. (20).
6.2 Proof of Lemma 2

Proof. Recall that

\[ y(k) = Wx(k). \]

1. Since \( 1^T(A - I) = 0 \) and by Eqs. (15), (16) and (18) we have

\[
\|y(k+1)\| = \|x(k+1) - \bar{x}(k+1)\| \\
= \left\| \begin{array}{c}
Ax(k) + (A - I)(q(k) - x(k)) - \alpha(k)g(x(k)) - \xi(v(k)) \\
-\bar{x}(k)1 - \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))1 - \bar{x}(k)1
\end{array} \right\| \\
= \|Ay(k) + (A - I)(q(k) - x(k)) - \alpha(k)Wg(x(k)) - W\xi(v(k))\| \\
\leq \|Ay(k)\| + \|((A - I)(q(k) - x(k)))\| \\
+ \alpha(k)\|W(k)g(v(k))\| + \|W\xi(v(k))\|. \tag{33}
\]

We now analyze each component on the right-hand side of Eq. (33). First, by Eq. (8) we have

\[
\|Ay(k)\| \leq \sigma_2\|y(k)\|. \tag{34}
\]

Second, using \( L = \sum_{i=1}^{n} L_i \), the Cauchy-Schwarz inequality, and Eq. (11) gives

\[
\|((A - I)(q(k) - x(k)))\| \leq \frac{2La(k-1)}{B-1}. \tag{35}
\]

Third, by Eqs. (6) and (19), and the Cauchy-Schwarz inequality we have

\[
\|Wg(x(k))\| \leq \sum_{i=1}^{n} \frac{(B + 1)L_i + \sum_{j \in N_i} a_{ij}L_i}{B - 1} \alpha(k - 1) \\
= \frac{(B + 2)L}{B - 1} \alpha(k - 1). \tag{36}
\]

Substituting Eqs. (34), (36) into Eq. (33) to have Eq. (21), i.e.,

\[
\|y(k+1)\| \leq \sigma_2\|y(k)\| + \frac{L(2B + 3)}{B - 1} \alpha(k - 1) \\
\leq \sigma_2^k\|y(0)\| + \frac{L(2B + 3)}{B - 1} \sum_{t=0}^{k} \sigma_2^{k-t} \alpha(t - 1).
\]

2. Suppose that \( \lim_{k \to \infty} \alpha(k) = 0 \). Since \( \alpha(k) \) is nonnegative and nonincreasing we have

\[
\lim_{k \to \infty} \sum_{t=0}^{k} \sigma_2^{k-t} \alpha(t) = \lim_{k \to \infty} \left\{ \sum_{t=0}^{\lfloor k/2 \rfloor - 1} \sigma_2^{k-t} \alpha(t) + \sum_{t=\lfloor k/2 \rfloor}^{k} \sigma_2^{k-t} \alpha(t) \right\} \\
\leq \lim_{k \to \infty} \alpha(0) \left\{ \sum_{t=0}^{\lfloor k/2 \rfloor - 1} \sigma_2^{k-t} + \alpha([k/2]) \sum_{t=\lfloor k/2 \rfloor}^{k} \sigma_2^{k-t} \right\} \\
\leq \lim_{k \to \infty} \alpha(0) \left\{ \frac{\alpha(0)}{1 - \sigma_2} \sigma_2^{\lfloor k/2 \rfloor} + \frac{1}{1 - \sigma_2} \alpha([k/2]) \right\} = 0. \tag{37}
\]

Thus, by Eq. (21) and using the preceding relation, we have Eq. (22).
Suppose now that the condition (23) is held. Then, for some \( K \geq 1 \) we have Eq. (24), i.e.,

\[
\sum_{k=0}^{K} \alpha(k) \|y(k+1)\| \leq \sum_{k=0}^{K} \alpha(k) \left\{ \sigma_{2}^{k+1} \|y(0)\| + \frac{L(2B + 3)}{B-1} \sum_{t=0}^{k} \sigma_{2}^{k-t} \alpha(t-1) \right\} \]

\[
\leq \sum_{k=0}^{K} \sigma_{2}^{k+1} \alpha(0) \|y(0)\| + \frac{L(2B + 3)}{B-1} \sum_{k=0}^{K} \alpha(k) \sum_{t=0}^{k} \sigma_{2}^{k-t} \alpha(t-1) \]

\[
\leq \frac{\alpha(0) \|y(0)\|}{1 - \sigma_{2}} + \frac{L(2B + 3)}{B-1} \sum_{k=0}^{K} \sum_{t=0}^{k} \sigma_{2}^{k-t} \alpha^{2}(t-1) \]

\[
= \frac{\alpha(0) \|y(0)\|}{1 - \sigma_{2}} + \frac{L(2B + 3)}{B-1} \sum_{t=0}^{K} \alpha^{2}(t-1) \sum_{k=t}^{K} \sigma_{2}^{k} \]

\[
\leq \frac{\alpha(0) \|y(0)\|}{1 - \sigma_{2}} + \frac{L(2B + 3)}{(B-1)(1 - \sigma_{2})} \sum_{t=0}^{K} \alpha^{2}(t-1) < \infty. \]

3. Suppose now that \( \alpha(k) = 1/\sqrt{k+1} \) implying \( \alpha(0) = 1 \). Then by the inequality above we have Eq. (25), i.e.,

\[
\sum_{k=0}^{K} \alpha(k) \|y(k+1)\| \leq \frac{\|y(0)\|}{1 - \sigma_{2}} + \frac{L(2B + 3)}{(B-1)(1 - \sigma_{2})} \sum_{t=0}^{K} \alpha^{2}(t-1) \]

where we use the integral test in the last inequality to have

\[
\sum_{t=0}^{K} \frac{1}{t+1} \leq 1 + \int_{0}^{K} \frac{1}{t+1} dt \leq 1 + \ln(K+1). \]

6.3 Proof of Lemma 3

Proof. Let \( x^* \) be a solution of problem (5). Recall that \( y(k) = Wx(k) \) and \( r(k) = \bar{x}(k) - x^* \). By Eq. (18) we have

\[
r^2(k+1) = \left( \bar{x}(k) - x^* - \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) - \xi(k) \right)^2 \]

\[
= r^2(k) - 2(\bar{x}(k) - x^*)\xi(k) - \frac{2\alpha(k)}{n} (\bar{x}(k) - x^*) \sum_{i=1}^{n} g_i(x_i(k)) \]

\[
+ \left( \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) + \xi(k) \right)^2 \]

\[
= r^2(k) - 2(\bar{x}(k) - x^*)\xi(k) - 2\alpha(k) (\bar{x}(k) - x^*) \sum_{i=1}^{n} g_i(x_i(k)) \]

\[
+ \frac{\alpha^2(k)}{n^2} \sum_{i=1}^{n} g_i(x_i(k))^2 + \xi^2(k). \]

Note that

\[
\frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) + \xi(k) \]

is an affine function of \( x_i(k) \). Thus, it is sufficient to show that the expression

\[
\frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) + \xi(k) \]

is nonnegative. This follows from the fact that

\[
\sum_{i=1}^{n} g_i(x_i(k)) \]

is a nonnegative function of \( x_i(k) \) and

\[
\alpha(k) \]

is a nonnegative function of \( k \).
\[ r^2(k) - 2(\bar{x}(k) - x^*)\xi(k) - \frac{2\alpha(k)}{n}(\bar{x}(k) - x^*)\sum_{i=1}^{n} g_i(x_i(k)) \]
\[ + 2 \left( \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) \right)^2 + 2\bar{\xi}^2(k) \]
\[ \leq r^2(k) - 2(\bar{x}(k) - x^*)\xi(k) - \frac{2\alpha(k)}{n}(\bar{x}(k) - x^*)\sum_{i=1}^{n} g_i(x_i(k)) \]
\[ + \frac{2L^2\alpha^2(k)}{n^2} + \frac{2(B + 2)^2L^2}{n^2(B - 1)^2}\alpha^2(k - 1), \quad (38) \]

where the last inequality is due to Eq. (6) and (19), i.e.,
\[ \bar{\xi}(k) \leq \frac{1}{n} \sum_{i=1}^{n} \frac{(B + 1)\bar{L}_i + \sum_{j \in \mathcal{N}_i} a_{ij}\bar{L}_j}{B - 1} \alpha(k - 1) = \frac{(B + 2)\bar{L}\alpha(k - 1)}{n(B - 1)} \]

We now analyze the second term on the right-hand side of Eqs. (38)
\[ -2(\bar{x}(k) - x^*)\bar{\xi}(k) = -2 \frac{n}{n} \sum_{i=1}^{n} \xi_i(v_i(k))(\bar{x}(k) - x^*) \]
\[ = -2 \frac{n}{n} \sum_{i=1}^{n} \xi_i(v_i(k))(\bar{x}(k) - x_i(k) + x_i(k) - x^*) \]
\[ \leq 2 \frac{n}{n} \sum_{i=1}^{n} \xi_i(v_i(k)) |y_i(k)| - 2 \frac{n}{n} \sum_{i=1}^{n} \xi_i(v_i(k))(x_i(k) - x^*) \]
\[ \stackrel{(19)}{<} 2 \frac{n}{n} \sum_{i=1}^{n} \frac{(B + 1)\bar{L}_i + \sum_{j \in \mathcal{N}_i} a_{ij}\bar{L}_j}{B - 1} \alpha(k - 1) |y_i(k)| - 2 \frac{n}{n} \sum_{i=1}^{n} \xi_i^2(v_i(k)) \]
\[ \leq \frac{2(B + 2)L}{n(B - 1)} \alpha(k - 1)\|y(k)\|. \quad (39) \]

Next, we analyze the third term on the right-hand side of Eq. (38)
\[ -2\frac{\alpha(k)}{n}(\bar{x}(k) - x^*)\sum_{i=1}^{n} g_i(x_i(k)) \]
\[ = -2\frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(\bar{x}(k) - x_i(k)) - 2\frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*) \]
\[ \leq 2\frac{L\alpha(k)}{n}\|y(k)\| - 2\frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*). \quad (40) \]

Substituting Eqs. (39) and (40) into (38) we obtain Eq. (20), i.e.,
\[ r^2(k + 1) \leq r^2(k) + \frac{2(B + 2)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{2L^2}{n^2}\alpha^2(k) + \frac{2(B + 2)^2L^2}{n^2(B - 1)^2}\alpha^2(k - 1) \]
\[ + \frac{2L}{n}\alpha(k)\|y(k)\| - 2\frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*) \]
\[ \leq r^2(k) + \frac{2(B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2}\alpha^2(k - 1) \]
\[ - 2\frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*), \]

15
where the last inequality we use $\alpha(k)$ is nonincreasing and $B \geq 2$.

### 6.4 Proof of Theorem 1

*Proof.* By Eq. (26) we have

$$r(k + 1) \leq r(k) + \frac{2(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1)$$

$$- \frac{2\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*)$$

$$\leq r(k) + \frac{2(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1)$$

$$- \frac{2\alpha(k)}{n} \sum_{i=1}^{n} f_i(x_i(k)) - f_i(x^*),$$

(41)

where the last inequality is due to the convexity of $f_i$. We now analyze the last term on the right-hand side of Eq. (41). Indeed, by Eq. (6) and recall that $f = \sum_{i=1}^{n} f_i$ and $f^* = f(x^*)$, we have for a fixed $\ell \in \mathcal{V}$

$$- \sum_{i=1}^{n} f_i(x_i(k)) - f_i(x^*) = - \sum_{i=1}^{n} \left( f_i(x_i(k)) - f_i(x) + f_i(x) - f_i(x^*) \right)$$

$$\leq \sum_{i=1}^{n} L_i |y_i(k)| - \left( f(x) - f^* \right)$$

$$\leq L \|y(k)\| - \left( f(x) - f(x) + f(x) - f^* \right)$$

$$\leq 2L \|y(k)\| - \left( f(x) - f^* \right).$$

(42)

Substituting Eq. (42) into Eq. (41) yields

$$r(k + 1) \leq r(k) + \frac{2(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1)$$

$$+ \frac{4L}{n} \alpha(k)\|y(k)\| - \frac{2\alpha(k)}{n} \left( f(x) - f^* \right)$$

$$= r(k) + \frac{2L(4B - 1)}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1)$$

$$- \frac{2\alpha(k)}{n} \left( f(x) - f^* \right),$$

which when iteratively updating over $k = 0, \ldots, K$ for some $K \geq 0$ we have

$$r(K + 1) \leq r(0) + \frac{2L(4B - 1)}{n(2B - 1)} \sum_{k=0}^{K} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \sum_{k=0}^{K} \alpha^2(k - 1)$$

$$- \frac{2}{n} \sum_{k=0}^{K} \alpha(k) \left( f(x) - f^* \right).$$

16
Rearranging the preceding relation and dropping the nonnegative $r(K + 1)$ we obtain

$$\sum_{k=0}^{K} \alpha(k)\left(f(x_\ell(k)) - f^*\right) \leq \frac{nr(0)}{2} + \frac{2(B + 1)^2L^2}{n(B - 1)^2} \sum_{k=0}^{K} \alpha^2(k - 1)$$
$$+ \frac{L(4B - 1)}{(2B - 1)} \sum_{k=0}^{K} \alpha(k - 1)\|y(k)\|. \tag{43}$$

Since $\alpha(k) = 1 / \sqrt{k + 1}$, we use Eq. 25 to have

$$\sum_{k=0}^{K} \alpha(k)\left(f(x_\ell(k)) - f^*\right)$$
$$\leq \frac{nr(0)}{2} + \frac{2(B + 1)^2L^2(1 + \ln(K + 1))}{n(B - 1)^2}$$
$$+ \frac{L(4B - 1)\|y(0)\|}{(B - 1)(1 - \sigma_2)} + \frac{L^2(4B - 1)(2B + 1)(1 + \ln(K + 1))}{(B - 1)^2(1 - \sigma_2)}.$$

Dividing both sides of the equation above by $\sum_{k=0}^{K} \alpha(k)$ and using the Jensen’s inequality we obtain

$$f(z_\ell(K)) - f^* \leq \frac{nr(0)}{2} + \frac{L(4B - 1)\|y(0)\|}{(B - 1)(1 - \sigma_2)} \frac{1}{\sum_{k=0}^{K} \alpha(k)}$$
$$+ \frac{2B(5B + 7)L^2}{(B - 1)^2(1 - \sigma_2)} \frac{1}{\sum_{k=0}^{K} \alpha(k)}$$
$$\leq \frac{nr(0)}{2\sqrt{K + 1}} + \frac{L(4B - 1)\|y(0)\|}{(B - 1)(1 - \sigma_2) \sqrt{K + 1}}$$
$$+ \frac{2B(5B + 7)L^2}{(B - 1)^2(1 - \sigma_2)} \frac{1 + \ln(K + 1)}{\sqrt{K + 1}},$$

where in the last inequality we use the integral test for $K \geq 0$ to have

$$\sum_{k=0}^{K} \alpha(k) = \sum_{k=0}^{K} \frac{1}{k} \geq \int_{t=0}^{K+1} \frac{1}{\sqrt{t+1}} \, dt = 2(\sqrt{K + 2} - 1) \geq \sqrt{K + 1}.$$
6.5 Proof of Theorem 2

Proof. By Eq. (26) we have

\[ r(k + 1) \leq r(k) + \frac{2(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1) \]

\[-\frac{2\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*) \]

\[ \leq r(k) + \frac{2(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1) \]

\[-\frac{2\alpha(k)}{n} \sum_{i=1}^{n} \left( f_i(x_i(k)) - f_i(x^*) + \frac{\mu_i}{2}(x_i(k) - x^*)^2 \right) \],

where the last inequality is due to the strong convexity of \( f \), i.e., Eq. (2). First, using the Jensen’s inequality on quadratic function \((\cdot)^2\) we have

\[ -\frac{1}{n} \sum_{i=1}^{n} \mu_i(x_i(k) - x^*)^2 \leq -\mu \frac{1}{n} \sum_{i=1}^{n} (x_i(k) - x^*)^2 \leq -\mu(x(k) - x^*)^2 = -\mu r(k). \]

Fix some \( \ell \in \mathcal{V} \). Using the preceding relation and Eq. (12) in the proof of Theorem 1 into Eq. (44) yields

\[ r(k + 1) \leq r(k) + \frac{2(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1) \]

\[-\frac{2\alpha(k)}{n} \left( f(x_k(k)) - f^* \right) - \mu \alpha(k) r(k) \]

\[ \leq \left( 1 - \frac{1}{k + 1} \right) r(k) + \frac{2(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1) \]

\[-\frac{2\alpha(k)}{n} \left( f(x_k(k)) - f^* \right), \]

where we use \( \alpha(k) = a / (k + 1) \) with \( a \geq 1 / \mu \), implying \( \mu \alpha(k) \geq 1 / (k + 1) \). Note that \( \alpha(0) = a \).

Multiplying both sides of the preceding equation by \( k + 1 \) and using \( k + 1 / k \leq 2 \) we have

\[ (k + 1)r(k) \leq kr(k) + \frac{4(2B + 1)L\alpha(0)}{n(B - 1)} \|y(k)\| + \frac{8(B + 1)^2L^2\alpha(0)}{n^2(B - 1)^2} \alpha(k - 1) \]

\[-\frac{2\alpha(0)}{n} \left( f(x_k(k)) - f^* \right); \]

which when summing up both sides over \( k = 0, \ldots, K \) for some \( K \geq 0 \) and rearranging we obtain

\[ \frac{2\alpha(0)}{n} \sum_{k=0}^{K} \left( f(x_k(k)) - f^* \right) + (K + 1)r(K + 1) \]

\[ \leq \frac{4L(2B + 1)\alpha(0)}{n(B - 1)} \sum_{k=0}^{K} \|y(k)\| + \frac{8(B + 1)^2L^2\alpha(0)}{n^2(B - 1)^2} \sum_{k=0}^{K} \alpha(k - 1). \]

(45)
First, Eq. \[\text{21}\] yields

\[
\sum_{k=0}^{K} \|x(k) - \bar{x}(k)\| \leq \sum_{k=0}^{K} \sigma_{2k}^2 \alpha(0) \|y(0)\| + \frac{L(2B + 3)\alpha(0)}{B - 1} \sum_{k=0}^{K} \sum_{t=0}^{K-t} \sigma_{2t}^2 \frac{1}{t + 1}
\]

\[
\leq \frac{\alpha(0)\|y(0)\|}{1 - \sigma_2} + \frac{L(2B + 3)\alpha(0)}{B - 1} \sum_{t=0}^{K-t} \frac{1}{t + 1} \sum_{k=0}^{K} \sigma_{2t}^2
\]

\[
\leq \frac{\alpha(0)\|y(0)\|}{1 - \sigma_2} + \frac{L(2B + 3)\alpha(0)}{(B - 1)(1 - \sigma_2)} \sum_{t=0}^{K-t} \frac{1}{t + 1}
\]

\[
\leq \frac{\alpha(0)\|y(0)\|}{1 - \sigma_2} + \frac{L(2B + 3)\alpha(0)}{(B - 1)(1 - \sigma_2)} (1 + \ln(K + 1)),
\]

where in the last inequality since \(\alpha(k + 1) = \alpha(0)/(k + 1)\), we use

\[
\sum_{k=0}^{K} \alpha(k) \leq \alpha(0)(1 + \ln(K + 1)).
\]

Substituting the preceding two relations into Eq. \[\text{45}\] we obtain

\[
\frac{2\alpha(0)}{n} \sum_{k=0}^{K} \left( f(x_k(k)) - f^* \right) + (K + 1)r(K + 1)
\]

\[
\leq \frac{4L(2B + 1)\alpha(0)}{n(B - 1)} \left[ \frac{\alpha(0)\|y(0)\|}{1 - \sigma_2} + \frac{L(2B + 3)\alpha(0)}{(B - 1)(1 - \sigma_2)} (1 + \ln(K + 1)) \right]
\]

\[
+ \frac{8(B + 1)^2L^2\alpha(0)}{n^2(B - 1)} (1 + \ln(K + 1))
\]

\[
\leq \frac{4L(2B + 1)\alpha^2(0)\|y(0)\|}{n(B - 1)(1 - \sigma_2)} \frac{1}{K + 1} + \frac{24L^2(B + 1)^2\alpha^2(0)}{n(B - 1)^2(1 - \sigma_2)} (1 + \ln(K + 1)),
\]

which when dividing both sides by \((K + 1)/n\), and using the Jensen’s inequality and Eq. \[\text{30}\] we have

\[
2\alpha(0) \left[ f(z_k(K)) - F^* \right] + nr(K + 1)
\]

\[
\leq \frac{4L(2B + 1)\alpha^2(0)\|y(0)\|}{(B - 1)(1 - \sigma_2)} \frac{1}{K + 1} + \frac{24L^2(B + 1)^2\alpha^2(0)}{(B - 1)^2(1 - \sigma_2)} \frac{1 + \ln(K + 1)}{K + 1}.
\]

Since the functions \(f_i\) are strongly convex with constant \(\mu_i\), \(f\) is strongly convex with constant \(\mu\). Thus, the preceding equation and since \(\alpha(0) = a \geq 1/\mu\) implies Eq. \[\text{31}\], i.e.,

\[
(z_k(K) - x^*)^2 \leq \frac{2}{\mu} \left[ f(z_k(K)) - f^* + \frac{n}{2\alpha(0)} r(K + 1) \right]
\]

\[
\leq \frac{4L(2B + 1)\alpha^2(0)\|y(0)\|}{(B - 1)(1 - \sigma_2)} \frac{1}{K + 1} + \frac{24L^2(B + 1)^2\alpha^2(0)}{(B - 1)^2(1 - \sigma_2)} \frac{1 + \ln(K + 1)}{K + 1}.
\]

\[\square\]

**References**

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–22.
[2] M. Chiang and T. Zhang, *Fog and IoT: An overview of research opportunities*, IEEE Internet of Things Journal, 3 (2016), pp. 854–864.

[3] T. T. Doan, *Aggregating stochastic gradients in distributed optimization*, in Proceedings of American Control Conference (ACC), 2018.

[4] T. T. Doan, C. L. Beck, and R. Srikant, *On the convergence rate of distributed gradient methods for finite-sum optimization under communication delays*, Proceedings ACM Meas. Anal. Comput. Syst., 1 (2017), pp. 37:1–37:27.

[5] T. T. Doan, C. L. Beck, and R. Srikant, *Convergence rate of distributed subgradient methods under communication delays*, in Proceedings of American Control Conference (ACC), 2018.

[6] T. T. Doan, S. Bose, D. H. Nguyen, and C. L. Beck, *Convergence of the iterates in mirror descent methods*, IEEE Control Systems Letters, 3 (2019), pp. 114–119.

[7] T. T. Doan, S. T. Maguluri, and J. Romberg, *Distributed stochastic approximation for solving network optimization problems under random quantization*. Available at: [https://arxiv.org/abs/1810.11568](https://arxiv.org/abs/1810.11568) 2018.

[8] T. T. Doan, S. T. Maguluri, and J. Romberg, *On the convergence of distributed subgradient methods under quantization*. To appear in the Proceedings of Allerton Conference on Communication, Control, and Computing, Monticello, IL, 2018.

[9] J. Duchi, A. Agarwal, and M. Wainwright, *Dual averaging for distributed optimization: Convergence analysis and network scaling*, IEEE Transactions on Automatic Control, 57 (2012), pp. 592–606.

[10] T. Hastie, T. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York, 2nd ed., 2009.

[11] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge, U.K.: Cambridge Univ. Press, 1985.

[12] O. J. Karst, *Linear curve fitting using least deviations*, Journal of the American Statistical Association, 53 (1958), pp. 118–132.

[13] G. Lan, S. Lee, and Y. Zhou, *Communication-Efficient Algorithms for Decentralized and Stochastic Optimization*. Available at: [https://arxiv.org/abs/1708.03543](https://arxiv.org/abs/1708.03543) 2017.

[14] J. Li, G. Chen, Z. Dong, and Z. Wu, *Distributed mirror descent method for multi-agent optimization with delay*, Neurocomputing, 177 (2016), pp. 643 – 650.

[15] J. Li, G. Chen, Z. Wu, and X. He, *Distributed subgradient method for multiagent optimization with quantized communication*, Mathematical Methods in the Applied Sciences, 40 (2016), pp. 1201–1213.

[16] P. D. Lorenzo and G. Scutari, *NEXT: in-network nonconvex optimization*, IEEE Trans. Signal and Information Processing over Networks, 2 (2016), pp. 120–136.

[17] A. Makhdoumi and A. Ozdaglar, *Convergence rate of distributed admm over networks*, IEEE Transactions on Automatic Control, 62 (2017), pp. 5082–5095.

[18] G. Meteos, J. Bazerque, and G. Giannakis, *Distributed sparse linear regression*, IEEE Transactions on Signal Processing, 58 (2010), pp. 5262–5276.
[19] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *Distributed subgradient methods and quantization effects*, in 2008 47th IEEE Conference on Decision and Control, Dec 2008, pp. 4177–4184.

[20] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *On distributed averaging algorithms and quantization effect*, IEEE Transactions on Automatic Control, 54 (2009), pp. 2506–2517.

[21] A. Nedić, A. Olshevsky, and M. G. Rabbat, *Network topology and communication-computation tradeoffs in decentralized optimization*, Proceedings of the IEEE, 106 (2018), pp. 953–976.

[22] A. Nedić, A. Olshevsky, and W. Shi, *Achieving geometric convergence for distributed optimization over time-varying graphs*, SIAM Journal on Optimization, 27 (2017), pp. 2597–2633.

[23] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Transactions on Automatic Control, 54 (2009), pp. 48–61.

[24] A. Nedić, A. Ozdaglar, and P. A. Parrilo, *Constrained consensus and optimization in multi-agent networks*, IEEE Transactions on Automatic Control, 55 (2010), pp. 922–938.

[25] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Norwell, MA, 2004.

[26] Y. Pu, M. N. Zeilinger, and C. N. Jones, *Quantization design for distributed optimization*, IEEE Transactions on Automatic Control, 62 (2017), pp. 2107–2120.

[27] G. Qu and N. Li, *Harnessing smoothness to accelerate distributed optimization*, IEEE Transactions on Control of Network Systems, (2017).

[28] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, *Quantized Decentralized Consensus Optimization*. Available at: [https://arxiv.org/pdf/1806.11536.pdf](https://arxiv.org/pdf/1806.11536.pdf), 2018.

[29] S. M. Shah and V. S. Borkar, *Distributed stochastic approximation with local projections*. arXiv preprint: [https://arxiv.org/abs/1708.08246](https://arxiv.org/abs/1708.08246), 2018.

[30] S. Shalev-Shwartz, *Online learning and online convex optimization*, Foundations and Trends in Machine Learning, 4 (2012), pp. 107–194.

[31] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 1st ed., 2014.

[32] W. Shi, Q. Ling, G. Wu, and W. Yin, *EXTRA: an exact first-order algorithm for decentralized consensus optimization*, SIAM Journal on Optimization, 25 (2015), pp. 944–966.

[33] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, *On the linear convergence of the ADMM in decentralized consensus optimization*, IEEE Trans. Signal Processing, 62 (2014), pp. 1750–1761.

[34] K. Tsianos, S. Lawlor, and M. Rabbat, *Push-sum distributed dual averaging for convex optimization*, in Proceedings of the 51st IEEE Conference on Decision and Control (CDC), Hawaii, USA, Dec 2012.

[35] E. Wei and A. Ozdaglar, *Distributed alternating direction method of multipliers*, in 2012 IEEE 51st IEEE Conference on Decision and Control (CDC), Dec 2012, pp. 5445–5450.
Appendices

A Extensions to $\mathbb{R}^d$ and Time-Varying Graphs

A.1 Notation and Assumptions

In this section, we consider problem (5) over time-varying graphs when $x \in \mathbb{R}^d$ and $\mathcal{X} \subseteq \mathbb{R}^d$. In particular, we assume that each node is only allowed to interact with its neighbors that are directly connected to it through a sequence of time-varying undirected graphs. Specifically, we assume we are given a sequence of undirected graphs $G(k) = (V, E(k))$ with $V = \{1, \ldots, n\}$ is the vertex set and $E(k)$ is the edge set at time $k$. This implies that nodes $i$ and $j$ can exchange messages at time $k$ if and only if $(i, j) \in E(k)$.

Denote by $N_i(k)$ the neighboring set of node $i$ at time $k$. We make the following fairly standard assumption, which ensures the long-term connectivity of the network.

**Assumption 3.** There exists an integer $T$ such that the following graph is connected for all positive integers $\ell$

$$\left( V, E(\ell T) \cup E((\ell + 1)T - 1) \right).$$

The multi-dimensional variant of Eq. (9) over time-varying networks is given as

$$x_i(k + 1) = \mathcal{P}_X \left[ x_i(k) - q_i(k) + \sum_{j \in N_i(k)} a_{ij}(k)q_j(k) - \alpha(k)g_i(x_i(k)) \right], \quad (46)$$

where $a_{ij}(k)$ is the time-varying weight which node $i$ assigns for its neighbor $j \in N_i(k)$ at time $k$.

We consider the following assumption about the weights $a_{ij}(k)$, which is the time-varying version of Assumption 1.

**Assumption 4.** There exists a positive constant $\beta$ such that the matrix $A(k)$, whose $(i,j)$-th entries are $a_{ij}(k)$, satisfies the following condition for any $k \geq 0$

1. $a_{ii}(k) \geq \beta$, for all $i \in V$
2. $a_{ij}(k) \in [\beta, 1]$ if $(i, j) \in N_i(k)$ otherwise $a_{ij}(k) = 0$ for all $i, j \in V$
3. $\sum_{i=1}^{n} a_{ij}(k) = \sum_{j=1}^{n} a_{ij}(k) = 1$.

In Eq. (46), $q_i(k) = Q(x_i(k))$ is the quantized vector of $x_i(k)$ over $[c_i(k), d_i(k)]$ given as

$$w_i(k) = \mathcal{P}_X \left[ x_i(k) - q_i(k) + \sum_{j \in N_i(k)} a_{ij}(k)q_j(k) \right]$$

$$c_i(k + 1) = w_i(k) - L_i\alpha(k)1 \quad \text{and} \quad d_i(k + 1) = w_i(k) + L_i\alpha(k)1.$$
Similar to Eq. (11) we have the quantization error is upper bounded by
\[ \|x_i(k) - q_i(k)\| \leq \frac{2\sqrt{d}L_i\alpha(k-1)}{B-1}, \quad \forall i \in V, \; k \geq 0, \]
where recall that \( d \) is the dimension of \( x_i \). We consider the following notation
\[
X = \begin{pmatrix} x_1^T & \ldots & x_n^T \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad \bar{X} = \begin{pmatrix} \bar{x}^T \\ \ldots \\ \bar{x}^T \end{pmatrix} = 1\bar{x}^T \in \mathbb{R}^{n \times d}
\]
\[
G(X) = \begin{pmatrix} g_1^T(x_1) \\ \ldots \\ g_n^T(x_n) \end{pmatrix} \in \mathbb{R}^{n \times d}, \quad Y = WX = X - \bar{X}, \quad r = \bar{x} - x^*.
\]
Thus, by Eq. (46) we have
\[
V(k + 1) = A(k)X(k) + (A(k) - I)(Q(k) - X(k)) - \alpha(k)G(X(k + 1))
\]
\[
X(k + 1) = V(k) - \xi(V(k)),
\]
where \( \xi(V(k)) = V(k) - P_{\mathcal{X}}[V(k)] \). Since \( A(k) \) is doubly stochastic we obtain
\[
\bar{v}(k) = \bar{x}(k) - \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))
\]
\[
\bar{x}(k + 1) = \bar{v}(k) - \bar{\xi}(k),
\]
where \( \bar{\xi}(k) \) is the average of \( \xi(k) \). We denote by \( \sigma_2(A(k)) \) the second largest singular value of \( A(k) \). Furthermore, with some abuse of notation let
\[
A_T(k) = A(k)A(k + 1) \ldots A(k + T - 1),
\]
and \( \delta \) be a parameter representing the spectral properties of the sequence graph \( \{G(k)\} \) defined as
\[
\delta = \min \left\{ \left(1 - \frac{1}{2n^3}\right)^{1/T}, \sup_{k \geq 0} \sigma_2(A_T(k)) \right\} \in (0, 1).
\]
By Assumptions 3 and 4 the following condition holds 20 22
\[
\|A_T(k)X\| \leq \delta \|X\|.
\]
Finally, we use the following result studied in 24, which is a general version of Lemma 1 to analyze the impact of the projection.

**Lemma 5 (Lemma 1 24).** Let \( \mathcal{X} \) be a nonempty closed convex set in \( \mathbb{R}^d \). Then, we have for any \( x \in \mathbb{R}^d \)
\[ (a) \; (\mathcal{P}_\mathcal{X}[x] - x)^T(x - y) \leq -\|\mathcal{P}_\mathcal{X}[x] - x\|^2 \text{ for all } y \in \mathcal{X}. \]
\[ (b) \; \|\mathcal{P}_\mathcal{X}[x] - y\|^2 \leq \|x - y\|^2 - \|\mathcal{P}_\mathcal{X}[x] - x\|^2 \text{ for all } y \in \mathcal{X}. \]

**A.2 Convergence Results**

We now provide the analysis for the general versions of Lemmas 2 and 3 and Theorems 1 and 2

**Lemma 6.** Suppose that Assumptions 3 and 4 hold. Let the sequence \( \{x_i(k)\} \), for all \( i \in V \), be updated by Eq. (46). In addition, let \( \{\alpha(k)\} \) be a nonnegative nonincreasing sequence of stepsizes. Then, we have...
1. The consensus error $\mathbf{Y}(k)$ satisfies

$$
\|\mathbf{Y}(k+1)\| \leq \delta^{k+1}\|\mathbf{Y}(0)\| + \frac{\sqrt{dL}(2B+3)}{B-1} \sum_{t=0}^{k} \delta^{k-t} \alpha(t-1).
$$

(53)

2. If $\lim_{k \to \infty} \alpha(k) = 0$ then

$$
\lim_{k \to \infty} x_i(k) = \lim_{k \to \infty} x_j(k), \quad \forall i, j \in \mathcal{V}.
$$

(54)

In addition, if $\alpha(k)$ is also square-summable, i.e.,

$$
\sum_{k=0}^{\infty} \alpha^2(k) < \infty,
$$

(55)

then for all $k \geq 0$ we have

$$
\sum_{t=0}^{k} \alpha(t)\|\mathbf{Y}(t)\| \leq \frac{\|\mathbf{Y}(0)\|}{1-\delta} + \frac{\sqrt{dL}(2B+3)}{(B-1)(1-\delta)} \sum_{t=0}^{k-1} \alpha^2(t-1) < \infty.
$$

(56)

3. If $\alpha(k) = 1/\sqrt{k+1}$ then we have for all $K \geq 0$,

$$
\sum_{k=0}^{K} \alpha(k)\|\mathbf{Y}(k+1)\| \leq \frac{\|\mathbf{Y}(0)\|}{1-\delta} + \frac{\sqrt{dL}(4B+1)}{(2B-1)(1-\delta)} (1 + \ln(k+1)).
$$

(57)

**Sketch of Proof.** As can be seen in Section 6.2 the key step in the proof of Lemma 6 is to show Eq. (53). The analysis of Eqs. (54)–(57) are consequences of Eq. (53). Since $1^T(A(k) - I) = 0$ and by Eqs. (48) and (49) we have

$$
\mathbf{Y}(k+1) = A(k)\mathbf{Y}(k) + (A(k) - I)(Q(k) - \mathbf{X}(k)) + \alpha(k)\mathbf{W}\mathbf{G}(\mathbf{V}(k)) + \mathbf{W}\xi(\mathbf{V}(k))
$$

(58)

First, by Eq. (51) we have

$$
\prod_{t=0}^{k} \|A(t)\mathbf{Y}(0)\| \leq \delta^{k+1}\|\mathbf{Y}(0)\|.
$$

(59)

Second, using $L = \sum_{i=1}^{n} L_i$, the Cauchy-Schwarz inequality, and Eq. (47) gives

$$
\|(A(k) - I)(Q(k) - \mathbf{X}(k))\| \leq \frac{2\sqrt{d}\alpha(k-1)}{B-1}.
$$

(60)

Third, using Lemma 5(b) with $y = \sum_{j=1}^{n} a_{ij} x_j(k) \in \mathcal{X}$, for all $i \in \mathcal{V}$, we have

$$
\|\xi(\mathbf{V}(k))\| \leq \sqrt{d} \sum_{i=1}^{n} \frac{(B+1)L_i}{B-1} + \sum_{j \in \mathcal{X}_i} a_{ij} L_j \alpha(k-1)
$$

(61)

Thus, by applying the 2-norm on both sides of Eq. (58) and using Eqs. (59)–(61) give Eq. (53). □
Lemma 7. Suppose that Assumptions 3 and 4 hold. Let the sequence \(\{x_i(k)\}\), for all \(i \in \mathcal{V}\), be updated by Eq. 46. Then, given an optimal \(x^* \in X^*\) we have

\[
\|r(k+1)\|^2 \leq \|r(k)\|^2 + \frac{2\sqrt{d}(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4d(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1) - \frac{2\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))^T(x_i(k) - x^*). \tag{62}
\]

Sketch of Proof. The proof of this lemma can be derived from the one of Lemma 3 using Lemma 5 so we skip it.

Theorem 3. Suppose that Assumptions 3 and 4 hold. Let the sequence \(\{x_i(k)\}\), for all \(i \in \mathcal{V}\), be updated by Eq. 46. In addition, let \(\alpha(k) = 1/\sqrt{k + 1}\), a nonnegative nonincreasing sequence of stepsizes with \(\alpha(0) = 1\). Moreover, suppose that each node \(i\), for all \(i \in \mathcal{V}\), stores a variable \(z_i \in \mathbb{R}^d\) initiated arbitrarily in \(X\) and updated as

\[
z_i(k) = \sum_{t=0}^{k} \frac{\alpha(t) x_i(t)}{\sum_{t=0}^{k} \alpha(t)}, \quad \forall i \in \mathcal{V}. \tag{63}
\]

Let \(x^* \in X^*\) be a solution of problem 5. Then for all \(i \in \mathcal{V}\) and \(k \geq 0\) we have

\[
f(z_i(k)) - f^* \leq \frac{n\|r(0)\|^2}{2\sqrt{k + 1}} + \frac{\sqrt{d}(AB - 1)L\|y(0)\|}{(B - 1)(1 - \delta)} \frac{1}{\sqrt{k + 1}} + \frac{2dB(5B + 7)L^2}{(B - 1)^2(1 - \delta)} \ln(k + 1) \tag{64}
\]

Sketch of Proof. By Eq. 62 we have

\[
r(k + 1) \leq r(k) + \frac{2\sqrt{d}(2B + 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4d(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1) - \frac{2\alpha(k)}{n} \sum_{i=1}^{n} f_i(x_i(k)) - f_i(x^*). \tag{65}
\]

Similar to Eqs. 42 and 41 we have for some fixed \(\ell \in \mathcal{V}\)

\[
r(k + 1) \leq r(k) + \frac{2\sqrt{d}(4B - 1)L}{n(B - 1)} \alpha(k - 1)\|y(k)\| + \frac{4d(B + 1)^2L^2}{n^2(B - 1)^2} \alpha^2(k - 1) - \frac{2\alpha(k)}{n} \left(f(x_\ell(k)) - f^*\right),
\]

which when iteratively updating over \(k = 0, \ldots, K\) for some \(K \geq 0\) we have

\[
r(K + 1) \leq r(0) + \frac{2\sqrt{d}(4B - 1)L}{n(2B - 1)} \sum_{k=0}^{K} \alpha(k - 1)\|y(k)\| + \frac{4d(B + 1)^2L^2}{n^2(B - 1)^2} \sum_{k=0}^{K} \alpha^2(k - 1) - \frac{2}{n} \sum_{k=0}^{K} \alpha(k) \left(f(x_\ell(k)) - f^*\right).
\]

Rearranging the preceding relation and dropping the nonnegative \(r(K + 1)\) we obtain

\[
\sum_{k=0}^{K} \alpha(k) \left(f(x_\ell(k)) - f^*\right) \leq \frac{nr(0)}{2} + \frac{2d(B + 1)^2L^2}{n(2B - 1)^2} \sum_{k=0}^{K} \alpha^2(k - 1) + \frac{\sqrt{dL}(4B - 1)}{(2B - 1)} \sum_{k=0}^{K} \alpha(k - 1)\|y(k)\|. \tag{66}
\]
Since $\alpha(k) = 1 / \sqrt{k + 1}$, we use Eq. (57) to have
\[
\sum_{k=0}^{K} \alpha(k) \left( f(x_k(k)) - f^* \right) \leq \frac{nr(0)}{2} + \frac{2d(B + 1)^2L^2}{n(B - 1)^2} \sum_{k=0}^{K-1} \frac{1}{k + 1} + \frac{\sqrt{dL(4B - 1)}}{(B - 1)} \left( \frac{\|Y(0)\|}{1 - \delta} + \frac{\sqrt{dL(4B + 1)}}{(2B - 1)(1 - \delta)}(1 + \ln(K + 1)) \right).
\]
Dividing both sides of the equation above by $\sum_{k=0}^{K} \alpha(k)$ and using the Jensen’s inequality we obtain Eq. (64).

**Theorem 4.** Suppose that Assumptions 2, 3, and 4 hold. Let the sequence $\{x_i(k)\}$, for all $i \in \mathcal{V}$, be updated by Eq. (46). We denote by $\mu = \min_{i \in \mathcal{V}} \mu_i$. In addition, let $\{\alpha(k)\} = a / k + 1$ for some $a \geq 1 / \mu$. Moreover, suppose that each node $i$, for all $i \in \mathcal{V}$, stores a variable $z_i \in \mathbb{R}^d$ initiated arbitrarily in $\mathcal{X}$ and updated as
\[
z_i(k) = \frac{\sum_{t=0}^{k} x_i(t)}{k + 1}, \quad \forall i \in \mathcal{V}.
\]
Let $x^* \in \mathcal{X}^*$ be a solution of problem (5). Then for all $i \in \mathcal{V}$ and $k \geq 0$ we have
\[
\|z_i(k) - x^*\|^2 \leq \frac{4\sqrt{dL(2B + 1)a^2(0)}\|Y(0)\|}{(B - 1)(1 - \delta)} \frac{1}{k + 1} + \frac{24dL^2(B + 1)^2a^2(0)}{(B - 1)^2(1 - \delta)} \frac{1 + \ln(k + 1)}{k + 1}.
\]

**Sketch of Proof.** Similar to the analysis in Theorem 3, this proof can be derived from the analysis of Theorem 2 using Lemmas 6 and 7, so we skip it.