Distributed Stochastic Approximation for Solving Network Optimization Problems Under Random 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. Our first contribution in this paper is to propose a modified consensus-based gradient method for solving such problems using random (dithered) quantization. This algorithm can be interpreted as a distributed variant of a well-known two-time-scale stochastic algorithm. We then study the convergence and derive upper bounds on the rates of convergence of the proposed method as a function of the bandwidths available between the nodes and the underlying network topology, for both convex and strongly convex objective functions. Our results complement for existing literature where such convergence and explicit formulas of the convergence rates are missing. 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 nodes1. 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.

We are motivated by various applications of such problems within engineering. A standard example is the problem of estimating the radio frequency in a wireless network of sensors where the goal is to cooperatively estimate the radio-frequency power spectrum density through solving a regression problem [24]. In this application, the objective function is the total loss over the entire measured data by the sensors, which are scattered across a large geographical area. Due to some privacy concerns, the sensors may not be willing to share their measurements, but only their own estimates, making distributed algorithms become necessary.

Another possible application is the problem of distributed information processing in edge (fog) computing, which has recently received a surge in interests [4]. This new technology, emerging from the rapid development of the Internet of Things, aims to reduce the burden of communication and computation at cloud or centralized servers by shifting the computing infrastructure closer to the source of data (e.g., smart devices, wireless sensors, or mobile robots). In this context, distributed algorithms provide a promising solution for coping with the large-scale complex networks while handling massive amounts of generated data.

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) [3, 22, 40, 44], distributed dual methods (mirror descent/dual averaging) [10, 12, 19, 42, 47], and distributed gradient algorithms [7–9, 18, 21, 27, 28, 33, 36, 39, 48]. 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 [26] for a summary of existing results in this area.

In distributed algorithms, the nodes are required to communicate and exchange information while cooperatively solving the problems. Thus, communication constraints, such as delays and finite bandwidth, are critical issues in distributed systems. For this reason, there are recent interests in studying the convergence of distributed gradient methods while taking into account these communication constraints. The convergence rates of such methods in the presence of communication delays have been studied in [8, 9, 41, 45], while some works presented in [18, 31] focus on reducing the number of communication rounds between nodes.

Our focus is on studying the convergence properties of distributed gradient methods when the nodes are only allowed to exchange their quantized values due to the finite bandwidths shared between them. Different variants of distributed gradient methods under quantized communication have been studied in [11, 20, 25, 32, 34, 46]. In [20, 25] the authors only show the convergence to a neighborhood around the optimal of the problem due to the quantized error. On the other hand, an asymptotic convergence to the optimal has been studied in [11, 32, 46]; however, a condition on the growing communication bandwidth is assumed in these works to remove the quantized error. Recently, the authors in [34] study distributed gradient methods with random quantization using finite

1In this paper, nodes can be used to present for processors, robotics, or sensors.
and show a convergence rate $O(1/k^{(1-\gamma)/2})$ for some $\gamma \in (0, 1)$ for unconstrained problems with strongly convex and smooth objective functions.

We consider in this paper a stochastic variant of distributed gradient methods with random quantization, which can be viewed as a distributed version of the well-known two-time-scale stochastic approximation. Similar to [34], we consider the problems where the nodes only share a finite communication bandwidth. However, unlike [34] we consider a constrained problem with nonsmooth objective functions. We derive explicit formulas for the rates of convergence of the algorithm, which show the dependence on the network topology and the communication capacity, for both convex and strongly convex objective functions. It is worth to note that the techniques used to derive the convergence rates in this paper are different from the ones in [34]. While the authors in [34] use a dual approach in their convergence analysis, we utilize the standard techniques from two-time-scale stochastic approximation studied in [2, 16, 17, 43]. This allows us to clearly show the impact of network topology and communication bandwidths on the convergence of the algorithm.

**Main Contributions.** The main contributions of this paper are two folds. We first propose a distributed variant of the well-known two-time-scale stochastic approximation for solving network optimization problems under random quantization. Second, we study the convergence and derive upper bounds on the rates of convergence of such methods. In particular, when the objective function is convex we first show the almost sure convergence of the decision variables to a central point that depends on the size of the communication bandwidths and the topology and communication bandwidths. We then derive upper bounds on the rates of convergence, which show the dependence on the size of the communication bandwidths. When the objective function is strongly convex, we further show that this rate occurs at $O\left(\frac{\ln(k+1)}{(1-\sigma_2)^3}k^{1/3}\right)$. We then conclude our paper with numerical experiments comparing the performance of distributed subgradient methods for solving the well-known convex optimization problems with and without quantization.

### 1.1 Notation And Definition

**Notation:** 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 $x_1, \ldots, x_n$ in $\mathbb{R}^d$, we denote by $X$ a matrix in $\mathbb{R}^{n \times d}$, whose $i$-th row is $x_i^T$. We then denote by $\|x\|$ and $\|X\|$ the Euclidean norm and the Frobenius norm of $x$ and $X$, respectively. Let $1$ be the vector whose entries are one and $I$ the identity matrix. Given a closed convex set $X$, we denote by $\mathcal{P}_X(x)$ the projection of $x$ to $X$.

Given a nonsmooth convex function $f : \mathbb{R}^d \to \mathbb{R}$, we denote by $\partial f(x)$ its subdifferential estimated at $x$, i.e., $\partial f(x) \triangleq \{ g \in \mathbb{R}^d \mid f(y) \geq f(x) + g^T(y-x), \forall y \in \mathbb{R}^d \}$ is the set of subgradients of $f$ at $x$. Since $f$ is convex, $\partial f(x)$ is nonempty. The function $f$ is $L$-Lipschitz continuous if and only if

$$||f(x) - f(y)|| \leq L||x - y||, \forall x, y \in \mathbb{R}^d.$$  \hspace{1cm} (1)

Note that the $L$-Lipschitz continuity of $f$ is equivalent to the subgradients of $f$ are uniformly bounded by $L$ [37]. A function $f$ is $\mu$-strongly convex if and only if $f$ satisfies $\forall x, y$

$$f(y) - f(x) - g(x)^T(y-x) \geq \frac{\mu}{2}(y-x).$$  \hspace{1cm} (2)

**Random Quantization:** We now present a brief review of random quantization adopted from [1], which is also equivalent to dithered quantization in signal processing. In particular, given a finite interval $[s, u]$ we divide this interval into a $B$ number of bins $\ell = \tau_1 \leq \tau_2 \leq \ldots \leq \tau_B = u$. We assume that the 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 $\Delta = (u-s)/(B-1)$. Thus, to present the points $\tau_i$ we need a finite $b$ bits where $b = \log_2(B)$.

Next given $x \in [\tau_i, \tau_{i+1})$ we denote by $p = (x - \tau_i)/\Delta$. Then the random quantization $q$ of $x$ is defined as

$$q = Q(x) \triangleq \left\{ \begin{array}{ll} \tau_i & \text{w/ prob. } 1 - p \\ \tau_{i+1} & \text{w/ prob. } p. \end{array} \right.$$  \hspace{1cm} (3)

As shown in [1] the random quantization Eq. (3) satisfies

$$\mathbb{E}[q] = x, \text{ and } \mathbb{E}((q - x)^2) \leq \frac{\Delta^2}{4} |x - q| \leq \Delta \text{ a.s.}$$  \hspace{1cm} (4)

In addition, we have $q = \tau_i$ a.s. if $x = \tau_i$ for some $i = 1, \ldots, B$. Thus, $q \in [s, u]$ a.s. if $x \in [s, u]$.

Finally, we consider the random quantization for the vector case. In particular, consider a compact set $X \subset \mathbb{R}^d$ defined as

$$X = [\ell^1, u^1] \times \ldots \times [\ell^d, u^d].$$

With some abuse of notation, given a vector $x \in X$ we denote by $q = Q(x)$, where $q^i = Q(x^i)$, the quantization of $i$-th coordinate of $x$, 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 $[\ell^i, u^i]$, for all $i = 1, \ldots, d$.

### 2 PROBLEM FORMULATION

We consider an optimization problem defined over a network of $n$ nodes. Associated with each node $i$ is a nonsmooth convex function $f_i : X \to \mathbb{R}$ over a convex compact set $X \subset \mathbb{R}^d$. The goal is to solve

$$\min_{x \in X} f(x) \triangleq \sum_{i=1}^{n} f_i(x).$$  \hspace{1cm} (5)

Each node $i$ knows only its local function $f_i$, and since there is no central coordination, the nodes are required to cooperatively solve the problem. We will use a distributed consensus-based (sub)gradient method where each node $i$ maintains its own version of the decision variables $x_i \in \mathbb{R}^d$, the goal is to have all the $x_i$ converge to $x^*$, a solution of problem (5). Each node can exchange a quantized version of $x_i$ with its neighbors, as defined through a connected and undirected graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ and $E = (V \times V)$ are the vertex and edge sets, respectively. We denote by $N_i := \{j \in V \mid (i, j) \in E\}$ the set of node $i$'s neighbors.

A concrete motivating example for this problem is distributed linear regression problems solved over a network of processors. Regression problems involving massive amounts of data are common in machine learning; see for example, [13, 38]. Each function $f_i$
is the empirical loss over the local data stored at processor $i$. The objective is to minimize the total loss over the entire dataset. Due to the difficulty of storing the enormous amount of data at a central location, the processors perform local computations over the local data, which are then exchanged to arrive at the globally optimal solution. Distributed gradient methods are a natural choice to solve such problems since they have been observed to be both fast and easily parallelizable in the case where the processors can exchange data instantaneously. The goal of this paper is to show that the algorithm continues to be convergent even the nodes only exchange the quantized values of their variables due to the finite bandwidths shared between them. In particular, we derive expressions for the convergence rate as a function of the communication bandwidths and the underlying network topology.

In the sequel, we will use $f^*$ to denote the optimal value of problem (5), i.e., $f^* = \sum_{i=1}^n f_i(x^*)$ where $x^*$ is a solution of problem (5). We denote by $X^*$ the solution set of problem (5), which is nonempty due to the compactness of $X$. In addition, since $X$ is compact it is obvious that each $f_i$ is Lipschitz continuous with some positive constant $L_i$, as stated in the following proposition.

**Proposition 1.** Each function $f_i$, for all $i \in V$, is $L_i$-Lipschitz continuous, i.e., Eq. (1) holds for some $L_i \geq 0$ for all $i \in V$.

### 3 DISTRIBUTED GRADIENT METHODS UNDER RANDOM QUANTIZATION

Distributed subgradient (DSG) methods, Eq. (6), for solving problem (5) were first studied and analyzed rigorously in [28, 29]. In these methods each node $i$ iteratively updates $x_i$ as

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

(6)

where $\alpha(k)$ is some sequence of stepsizes and $g_i(x_i(k)) \in \partial f_i(x_i(k))$. Here, $a_{ij}$ is some positive weight which node $i$ assigns for $x_j$. We assume that these weights, which capture the topology of $G$, satisfy the following condition.

**Assumption 1.** The matrix $A$, 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, $A$ is irreducible and aperiodic. Finally, the weights $a_{ij} > 0$ if and only if $(i, j) \in E$ otherwise $a_{ij} = 0$.

This assumption also implies that $A$ has 1 as the largest singular value and others are strictly less than 1; see for example, the Perron-Frobenius theorem [14]. Also, we denote by $\sigma_2 \in (0, 1)$ the second largest singular value of $A$, which by the Courant-Fisher theorem [14] gives

$$\left\| A \left( 1 - \frac{1}{n} I \right) \right\| \leq \sigma_2 \left\| 1 - \frac{1}{n} I \right\|. $$

(7)

Our focus in this section is to study DSG under random quantization in communication between nodes. 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 such quantized communication, we modify the update in

### Algorithm 1: Distributed Subgradient Algorithm Under Random Quantization

1. **Initialize:** Each node $i$ arbitrarily initializes $x_i(0) \in X$.

2. **Iteration:** For $k \geq 0$ each node $i$ implements

$$x_i(k + 1) = \mathcal{P}_X \left[ (1 - \beta(k))x_i(k) + \beta(k) \sum_{j \in \mathcal{N}_i} a_{ij} Q_j(k) \right] - \alpha(k)g_i(x_i(k)), $$

(8)

Eq. (6), that is, each node $i$ now considers the following update

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

where $q_j(k) = Q(x_j(k))$ given in Eq. (3). Here, in addition to $\alpha(k)$ we introduce a new stepsize $\beta(k)$ due to the random quantization exchanged between nodes.

This update has a simple interpretation. At any time $k \geq 0$, each node $i$ first obtains the quantized value $q_j(k)$ of its value $x_i(k)$. Each node $i$ then formulates a convex combination between its value $x_i(k)$ and the weighted quantized value received from its neighbors $j \in \mathcal{N}_i$, with the goal of seeking a consensus on their estimates. Each node then moves along the subgradients of its respective objective function to update its estimates, pushing the consensus point toward the optimal set $X^*$. The distributed subgradient algorithm under random quantization is formally stated in Algorithm 1.

#### 3.1 The Role of $\beta(k)$

We discuss in this section some aspects of the new stepsize $\beta(k)$. First, one can interpret Eq. (8) as a distributed two-time-scale stochastic algorithm [2, 16, 17, 43], where the first sum play the role of fast time scale while the gradient step is the slow time scale. Due to the random quantization, each node first uses the fast time scale to estimate the true average of their estimates. Each node then moves along the subgradients of its respective objective function to update its estimates, pushing the consensus point toward the optimal set $X^*$. The distributed subgradient algorithm under random quantization is formally stated in Algorithm 1.

Finally, one can view $\beta(k)$, in addition to $\alpha(k)$, is some weight which each node $i$ uses to indicate that it “trusts” its own value $x_i$ more than the value $x_j$ received from its neighbor $j$. As will be seen, to guarantee the convergence of the algorithm we will let $\beta(k)$ go to zero at some proper rate, implying eventually node $i$ only uses its own value.
4 CONVERGENCE RESULTS

The focus of this section is to study the convergence properties of Algorithm 1 for solving problem (5), when the objective functions are both convex and strongly convex. The key idea of our analysis is to utilize the standard techniques used in centralized subgradient methods and stochastic approximation approach. In particular, for convex objective functions we first show that \( x_i(k) \), for all \( i \in V \), converges almost surely to a solution \( x^* \) of problem (5) under some proper choice of stepsizes \( \{\alpha(k), \beta(k)\} \). We next show the convergence of the function \( f \) estimated at the time \( \alpha \)-weighted average of each \( x_i \) to the optimal value \( f^* \) in expectation at a rate \( O(n\Delta^2 L^2 \ln(k)/(1 - \sigma_2)^2 k^{1/4}) \), where 1 - \( \sigma_2 \) is the spectral gap of the network connectivity. Finally, when the objective functions are strongly convex, we derive the convergence of the time-weighted average of each \( x_i \) to an optimal solution \( x^* \) of problem (5) in expectation at a rate \( O(n\Delta^2 L^2 \ln(k)/(1 - \sigma_2)^3 k^{1/3}) \).

We start our analysis by introducing more notation. Given the nodes' estimates \( x_1, \ldots, x_n \) in \( \mathbb{R}^d \), we denote by \( X \in \mathbb{R}^{n \times d} \) a matrix whose \( i \)-th rows are \( x_i^T \), i.e.,

\[
X = \begin{bmatrix}
- x_1^T \\
- x_2^T \\
\vdots \\
- x_n^T
\end{bmatrix} \in \mathbb{R}^{n \times d}.
\]

Let \( \bar{x} \in \mathbb{R}^d \) be the average of \( x_i \), i.e.,

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = X^T 1 \in \mathbb{R}^d.
\]

For convenience, we use the following notation

\[
G(X) = \begin{bmatrix}
g_1(x_1) \\
g_2(x_2) \\
\vdots \\
g_n(x_n)
\end{bmatrix} \in \mathbb{R}^{n \times d}, \quad e_i(k) = q_i(k) - x_i(k)
\]

\[
\Delta = \sum_{i=1}^d \Delta_i, \quad L = \sum_{i=1}^n L_i, \quad W = I - \frac{1}{n} 1 1^T
\]

\[
r(k) = \|\bar{x}(k) - x^*\|, \quad Y(k) = X(k) - 1\bar{x}^T(k) = WX(k).
\]

Moreover, let \( \mathcal{F}_k \) be the filtration containing all the history generated by Eq. (8) up to time \( k \), i.e.,

\[
\mathcal{F}_k = \{X(0), Q(0), \ldots, X(k), Q(k)\}.
\]

Finally, given a vector \( v \in \mathbb{R}^d \) let \( \xi \) denote the error due to the projection of \( v \) to \( X \), i.e.,

\[
\xi = v - \mathcal{P}_X[v].
\]

Thus, Eq. (8) now can be rewritten as

\[
v_i(k) = (1 - \beta(k))x_i(k) + \beta(k) \sum_{j \in N_i} a_{ij}q_j(k)
\]

\[
- \alpha(k)g_i(x_i(k))
\]

\[
x_i(k + 1) = v_i(k) - \xi_i(k),
\]

which by using A the matrix form of Eq. (10) is given as

\[
V(k) = (1 - \beta(k))X(k) + \beta(k)AQ(k) - \alpha(k)G(X(k))
\]

\[
X(k + 1) = Y(k) - \Xi(k),
\]

where \( \Xi(k) \in \mathbb{R}^{n \times d} \) is the matrix whose \( i \)-th row is \( \xi_i^T(k) \). In addition, since A is doubly stochastic, we have

\[
v(k + 1) = (1 - \beta(k))\bar{x}(k) + \beta(k)q(k)
\]

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

\[
\bar{x}(k + 1) = \bar{x}(k) - \xi(k).
\]

4.1 Preliminaries

In this section, we consider some preliminary results, which are essential in our analysis given in the next section. For an ease of exposition, we delay the proofs of all results in this section to the appendix. However, we present a sketch of their proofs to explain some intuition behind our analysis.

We first provide an upper bound for the consensus error \( \|Y(k)\| = \|WX(k)\| \) in the following lemma.

**Lemma 1.** Suppose that Assumption 1 holds. Let the sequence \( \{x_i(k), \beta(k)\} \) for all \( i \in V \), be generated by Algorithm 1. In addition, let \( \{\alpha(k), \beta(k)\} \) be two sequences of nonnegative and nonincreasing stepsizes. Then we have

\[
\mathbb{E}[\|Y(k + 1)\|^2 | \mathcal{F}_k] 
\]

\[
\leq (1 - (1 - \sigma_2)\beta(k))\|Y(k)\|^2
\]

\[
+ n\sigma_2^2 L^2 \beta^2(k) + \frac{4L^2(\beta(0) + 1)^2}{(1 - \sigma_2)^2} \beta(k).
\]

Moreover, we also obtain

\[
\sum_{i=0}^k \beta(t)\mathbb{E}[\|Y(t)\|^2] 
\]

\[
\leq \frac{1}{1 - \sigma_2} \mathbb{E}\[\|Y(0)\|^2] + \sum_{i=0}^k \frac{n\sigma_2^2 L^2 \beta^2(t)}{1 - \sigma_2} + \frac{4L^2(\beta(0) + 1)^2}{(1 - \sigma_2)^2} \beta(t).
\]

**Sketch of Proof.** To show Eq. (13) we first use Eqs. (11) and (12) to have

\[
Y(k + 1) = WX(k + 1)
\]

\[
= (1 - \beta(k))Y(k) + \beta(k)AQ(k) - \alpha(k)WG(k) - WX(k).
\]

Next, we use the following Cauchy-Schwarz inequality with some \( \eta > 0 \) and \( a, b \in \mathbb{R} \)

\[
(a + b)^2 \leq (1 + \eta)a^2 + (1 + 1/\eta)b^2.
\]

Thus, by taking the 2-norm square of the first equation and using the preceding Cauchy-Schwarz inequality with \( \eta = 1 + (1 - \sigma_2)\beta(k) \) we have

\[
\|Y(k + 1)\|^2 
\]

\[
\leq (1 + (1 - \sigma_2)\beta(k))\|Y(k)^2 \| + (1 - \sigma_2)\beta(k)\|AQ(k)\|^2
\]

\[
+ \frac{2}{(1 - \sigma_2)\beta(k)} \|\alpha(k)WG(k)\|^2
\]

\[
+ \frac{2}{(1 - \sigma_2)\beta(k)} \|WX(k)\|^2, \quad (15)
\]

We next analyze each term on the right-hand side of Eq. (15). First, Proposition 1 gives

\[
\|WG(k)\|^2 \leq \|G(k)\|^2 \leq L^2.
\]
Second, using the projection lemma, Lemma 5(b) in the Appendix, one can show
\[ \|Wz(k)\|^2 \leq L^2 \alpha^2(k). \]

Third, by Eq. (4) we have
\[ \|E(k)\|^2 = \sum_{i=1}^{n} \|x_i(k) - q_i(k)\|^2 \leq \sum_{i=1}^{n} \Delta^2 = n\Delta^2. \]

Fourth, by Eq. (7) we have
\[ \|(1 - \beta(k))Y(k) + \beta(k)AY(k)\|^2 \leq \|(1 - (1 - \sigma_2)\beta(k))Y(k)\|. \]

Thus, taking the conditional expectation of Eq. (41) w.r.t. \( \mathcal{F}_k \) and using the last four inequalities we obtain Eq. (13).

Finally, taking the expectation on both sides of Eq. (13) and summing up over \( k = 0, \ldots, K \) for some \( K \) immediately give Eq. (14). \( \square \)

We next provide proper conditions on the stepsizes \( \{\alpha(k), \beta(k)\} \), which guarantees that the nodes achieve a consensus on their estimates \( x_i \). The analysis of this lemma is a consequence of Lemma 1 and Lemma 4 on the almost supermartingale convergence theorem given later.

**Lemma 2.** Suppose that Assumption 1 holds. Let the sequence \( \{x_i(k)\} \), for all \( i \in \mathcal{V} \), be generated by Algorithm 1. In addition, let \( \alpha(k) \) and \( \beta(k) \) satisfy
\[ \sum_{k=0}^{\infty} \alpha(k) = \sum_{k=0}^{\infty} \beta(k) = \infty, \quad \sum_{k=0}^{\infty} \alpha^2(k)/\beta(k) < \infty \]
\[ \sum_{k=0}^{\infty} \alpha^2(k) < \infty, \quad \sum_{k=0}^{\infty} \beta^2(k) < \infty. \] (16)

Then we have
\[ \lim_{k \to \infty} \|x_i(k) - \bar{x}(k)\| = 0 \quad \text{a.s.}, \quad \text{for all } i \in \mathcal{V}. \] (17)

Furthermore, the following condition holds
\[ \sum_{k=0}^{\infty} \beta(k)|Y(k)|^2 < \infty \quad \text{a.s.} \] (18)

**Remark 1.** One example of stepizes \( \{\alpha(k), \beta(k)\} \), which satisfies Eq. (16), can be chosen as follows
\[ \alpha(k) = \frac{1}{k + 2}, \quad \beta(k) = \frac{1}{(k + 2)^2}, \quad \forall s \in \left(\frac{1}{2}, 1\right). \] (19)

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

**Lemma 3.** Suppose that Assumption 1 holds. Let the sequence \( \{x_i(k)\} \), for all \( i \in \mathcal{V} \), be generated by Algorithm 1. Let \( \{\alpha(k), \beta(k)\} \) be two sequences of nonnegative and nonincreasing step sizes with \( \beta(0) < 1 \). Let \( x^* \) be a solution of problem (5). Then we have
\[ \mathbb{E}\left[ r(k + 1) \mid \mathcal{F}_k \right] \leq r(k) + \frac{6L^2 \alpha^2(k)}{n(1 - \beta(0))} + \frac{2L^2 \alpha^2(k)}{n \beta(k)} + \Delta^2 \beta^2(0) + \frac{2\beta(k)|Y(k)|^2}{n} \cdot \frac{2\alpha(k)}{n} \sum_{i=1}^{n} g_i^T(x_i(k)) \left(x_i(k) - x^*\right). \] (20)

**Sketch of Proof.** First, using Eq. (12) to have
\[ r(k + 1) = \|\bar{x}(k) - \bar{x}(k) - x^*\|^2 \]
\[ = \left\| (1 - \beta(k))x(k) - x^* - \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) + \beta(k)q(k) - \xi(k) \right\|^2, \]
which by expanding the right-hand side and taking the conditional expectation w.r.t. \( \mathcal{F}_k \) yields
\[ \mathbb{E}\left[ r(k + 1) \mid \mathcal{F}_k \right] \leq \left\| \bar{x}(k) - x^* - \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) \right\|^2 + \|\beta(k)e(k)\|^2 + \|\bar{x}(k) - x^*\|^2 - 2\bar{x}(k)^T \left(x^* - x(k) - \frac{\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)) \right), \]
\[ + \frac{2\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k)). \]
The next step is to provide an upper bound for each term on the right-hand side of the preceding equation to obtain Eq. (20). This step can be done in a similar concept of the one given in Lemma 1. \( \square \)

Finally, we utilize the result on almost supermartingale convergence studied in [35], stated as follows.

**Lemma 4 ([35]).** Let \( \{y(k)\}, \{z(k)\}, \{w(k)\}, \text{ and } \{\gamma(k)\} \) be non-negative sequences of random variables and satisfy
\[ \mathbb{E}\left[ y(k + 1) \mid \mathcal{F}_k \right] \leq (1 + \gamma(k))y(k) - z(k) + w(k) \]
\[ \sum_{k=0}^{\infty} y(k) < \infty \text{ a.s.}, \quad \sum_{k=0}^{\infty} w(k) < \infty \text{ a.s.}, \]
where \( \mathcal{F}_k = \{y(0), \ldots, y(k)\}, \) the history of \( y \) up to time \( k \). Then \( \{y(k)\} \) converges a.s., and \( \sum_{k=0}^{\infty} z(k) < \infty \text{ a.s.} \)

### 4.2 Convergence Results of Convex Functions

In this section, we study the convergence and the rate of convergence of Algorithm 1 when the objective functions \( f_i \) are convex. For an ease of explanation, we only provide a sketch of the proofs for all the main results in this section and the next section, where their details are presented in Section 6.

Our first main result is to show that if the stepizes \( \{\alpha(k), \beta(k)\} \) satisfy Eq. (16), then \( x_i(k) \), for all \( i \in \mathcal{V} \), converges almost surely to \( x^* \), a solution of problem (5). The following theorem is states this result.

**Theorem 1.** Suppose that Assumption 1 holds. Let the sequence \( \{x_i(k)\} \), for all \( i \in \mathcal{V} \), be generated by Algorithm 1. Let \( \{\alpha(k), \beta(k)\} \) be two sequences of nonnegative and nonincreasing step sizes satisfying Eq. (16) with \( \beta(0) < 1 \), e.g., Eq. (19) holds. Then we have
\[ \lim_{k \to \infty} x_i(k) = \bar{x} \quad \text{a.s.}, \quad \text{for all } i \in \mathcal{V}, \] (21)
for some \( \bar{x} \) that is a solution of Problem (5).
Proof Sketch. The main idea of this proof is first using the convexity of the functions $f_i$ into Eq. (20) in Lemma 3 to obtain
\[
\mathbb{E}[r(k + 1) | \mathcal{F}_k] \leq r(k) + O \left( a^2(k) + \beta^2(k) + \frac{\alpha^2(k)}{\beta(k)} \right) + \frac{3}{n} \beta(k) ||Y(k)||^2 - \frac{2\alpha(k)}{n} \left( f(\bar{x}(k)) - f^* \right).
\]
Second, since the stepizes $\{a(k), \beta(k)\}$ satisfy the conditions in Eq. (16), Eq. (18) holds. Thus, we can now apply Lemma 4 to the preceding relation to have
\[
\lim_{k \to \infty} r(k) \text{ exits a.s. for each } x^*.
\]
\[
\sum_{k=0}^{\infty} a(k) \left( f(\bar{x}(k)) - f^* \right) < \infty \quad \text{a.s.}
\]
Thus, using these relations and standard analysis on the convergence of subsequence of $\{\bar{x}(k)\}$ we can obtain Eq. (21). \hfill \Box

We now study the rate of convergence of Algorithm 1 to the optimal value in expectation, where we utilize a similar technique used to establish the convergence rate of centralized subgradient methods. In particular, we show that if each node $i$ maintains a variable $z_i$ used to estimate the time $\alpha$-weighted average of its local copy $x_i$, then the function value $f$ estimated at each $z_i$ converges in expectation to the optimal value $f^*$ with a rate $O \left( n^{\Delta^2} \ell^2 \ln(k) / (1 - \sigma_2)^3 \right)$. The dependence on the variance $\Delta^2$ of the quantized error in the upper bound of the rate is natural, as we often observe in stochastic gradient descent where such dependence is on the variance of the gradient noise. Such result is derived under different assumptions on the stepizes $\{a(k), \beta(k)\}$ as shown in the following theorem. Note that while the previous theorem studies almost sure convergence of the local copies, this theorem studies convergence in expectation of the functional value, and so it is not surprising that the stepizes are different.

**Theorem 2.** Suppose that Assumption 1 holds. Let the sequence $(x_i(k),)$ for all $i \in \mathcal{V}$, be generated by Algorithm 1. Let $\{a(k), \beta(k)\}$ be defined as
\[
a(k) = \frac{1}{(k + 2)^{3/4}}, \quad \beta(k) = \frac{1}{(k + 2)^{1/2}}.
\]
In addition, suppose that each node $i$ maintains a variable $z_{i}$ initialized arbitrarily in $X$ and updated as
\[
z_{i}(k) = \sum_{t=0}^{k} a(t)x_{i}(t).
\]
Then we have for all $i \in \mathcal{V}$ and $k \geq 0$
\[
\mathbb{E} \left[ f(z_{i}(k)) \right] - f^* \leq \frac{n \mathbb{E} \left[ r(0) \right] + n \mathbb{E} \left[ ||Y(0)||^2 \right] + 16L^2}{2(1 - \sigma_2)^{3/2}} + \frac{5n\Delta^2(1 + \ln(k + 2))}{2(1 - \sigma_2)^2} \times \frac{1}{(k + 1)^{1/4}}.
\]
\[
\sum_{k=0}^{\infty} \frac{\mathbb{E} \left[ ||z_{i}(k) - x^*||^2 \right]}{k + 1} \leq \frac{n \mathbb{E} \left[ r(0) \right] + n \mathbb{E} \left[ ||Y(0)||^2 \right]}{2(1 - \sigma_2)} + \frac{16L^2}{2(1 - \sigma_2)^2} \times \frac{1}{k + 2} + \frac{5n\Delta^2(1 + \ln(k + 2))}{2(1 - \sigma_2)^2} \times \frac{1}{k + 1}.
\]

\[\text{Remark 2. We note that } z_{i}(k) \text{ in Theorem 2 can be iteratively updated as follows:}
\]
\[
z_{i}(k + 1) = \frac{a(k)x_{i}(k) + S(k)z_{i}(k)}{S(k + 1)},
\]
where $S(0) = 0$ and $S(k + 1) = \sum_{t=0}^{k} a(t)$ for $k \geq 1$.

**4.3 Convergence Results of Strongly Convex Functions**

We study here the convergence rate of Algorithm 1 when $f_i$ are strongly convex, that is, we consider the following assumption.

**Assumption 2.** Each function $f_i$, for all $i \in \mathcal{V}$, is $\mu_i$-strongly convex, i.e., Eq. (2) holds for some $\mu_i \geq 0$.

Note that this assumption implies that $f$ is $\mu$-strongly convex where $\mu = \min_i \mu_i$. Under this assumption, we show the rate of convergence of Algorithm 1 to an optimal solution $x^*$ of problem (5) in expectation, stated in the following theorem.

**Theorem 3.** Suppose that Assumptions 1 and 2 hold. Let the sequence $(x_i(k),)$ for all $i \in \mathcal{V}$, be generated by Algorithm 1. Let $x^*$ be a solution of problem (5) and $\{a(k), \beta(k)\}$ be defined as
\[
a(k) = \frac{a}{k + 2}, \quad \beta(k) = \frac{b}{(k + 2)^{2/3}}, \quad \text{for } a \geq \frac{1}{\mu} \quad \text{and } b \geq \frac{1}{1 - \sigma_2}.
\]
In addition, suppose that each node $i$ maintains a variable $z_{i}$ initialized arbitrarily and updated as
\[
z_{i}(k) = \frac{\sum_{t=0}^{k} x_{i}(t)}{k + 1}.
\]
Then we have for all $i \in \mathcal{V}$ and $k \geq 0$
\[
\mathbb{E} \left[ ||z_{i}(k) - x^*||^2 \right] \leq \frac{n \mathbb{E} \left[ r(0) \right] + n \mathbb{E} \left[ ||Y(0)||^2 \right]}{2(1 - \sigma_2)} \times \frac{1}{k + 2} + \frac{5n\Delta^2(1 + \ln(k + 2))}{2(1 - \sigma_2)^2} \times \frac{1}{k + 1}.
\]
Proof Sketch. The first step in this analysis is using the strong convexity of the functions $f_i$ into Eq. (20) to have
\[
\mathbb{E}[r(k+1)|F_k] \leq (1 - \mu a(k)) r(k) + O \left(\frac{a^2(k) + \beta^2(k) + \frac{\alpha^2}{\bar{\beta}(k)} + \frac{3}{n}\bar{\beta}(k)\|Y(k)\|^2}{n} - \frac{2\epsilon(k)}{n}(\hat{f}(x(k)) - f^*)\right).
\]
The rest of this proof is similar to the one in Theorem 2.

5 SIMULATIONS

In this section, we apply Algorithm 1 for solving linear regression problems, the most popular technique for data fitting [13, 38] 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 $S = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}$ 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 $L_1$ norm) is often referred to as robust regression, which is known to be robust to outliers [15], 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 50 nodes, i.e., $G = (\mathcal{V}, \mathcal{E})$ and $n = |\mathcal{V}| = 50$. 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.,
\[
A = [a_{ij}] = \begin{cases} 
\frac{1}{2\max(|N_i|, |N_j|)}, & \text{if } (i, j) \in \mathcal{E} \\
0, & \text{if } (i, j) \notin \mathcal{E} \text{ and } i \neq j \\
1 - \sum_{j \in N_i} a_{ij}, & \text{if } i = j
\end{cases}
\]

It is straightforward to verify that the lazy Metropolis matrix $A$ satisfies Assumption 1.

![Figure 1: The convergence of function values using distributed subgradient methods without (——), with time-varying (-----), and with random (——) quantization for $n = 50$ and $d = 10$ are illustrated.](5)
nodes use a finite number of constant bits $b$ in their communication. However, as observed both in Fig. 1a for quadratic loss and in Fig. 1b for absolute loss, Algorithm 1 performs almost as well as the one in [11].

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 for the relative error $f(z_i(k)) - f^* / f^* \leq 0.2$, as a function of $b$. As we can see the more bits we use the faster the algorithm converges. Moreover, the number of iterations required by the algorithm seems to be the same when $b$ is larger than 11. This does make sense due to the numerical rounding required by the algorithm seems to be the same when $b$ is larger than 11. This does make sense due to the numerical rounding of the computer program. Finally, the curves in Fig. 2 seems to reflect the dependence of the rate of Algorithm 1 on the variance $\Lambda^2 = C/(2b - 1)$ within some constant factor $C$, which agrees with our results in Theorems 2 and 3.

6 PROOFS OF MAIN RESULTS

In this section, we present the proofs of our main results given in Section 4.

6.1 Proof of Theorem 1

By the convexity of $f_i$ we have

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

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

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

which by the $L_i$-Lipschitz continuity of $f_i$ in Proposition 1 yields

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

$$\leq -\frac{2}{n} \sum_{i=1}^{n} \alpha(k) L_i \|y_i(k)\| - \frac{2\alpha(k)}{n} \sum_{i=1}^{n} f_i(\bar{x}(k)) - f_i(x^*)$$

$$\leq \frac{L^2 \sigma^2(k)}{n\beta(k)} + \frac{\beta(k) \|y(k)\|^2}{n} - \frac{2\alpha(k)(f(\bar{x}(k)) - f^*)}{n}.$$

Substituting the preceding relation into Eq. (20) in Lemma 3 gives

$$\mathbb{E}[r(k + 1) | \mathcal{F}_k]$$

$$\leq r(k) + \frac{6L^2 \sigma^2(k)}{n(1 - \beta(0))} + \frac{3L^2 \sigma^2(k)}{n\beta(k)} + \frac{\Delta^2 \beta^2(k)}{n}$$

$$+ \frac{2\beta(k) \|y(k)\|^2}{n} + \frac{L^2 \sigma^2(k)}{n\beta(k)} + \frac{\beta(k) \|y(k)\|^2}{n}$$

$$- \frac{2\alpha(k)(f(\bar{x}(k)) - f^*)}{n}.$$

Figure 2: The number of iterations as a function of $b$ using distributed subgradient methods with random quantization for $n = 50$ and $d = 10$ are illustrated.

$$= r(k) + \frac{6L^2 \sigma^2(k)}{n(1 - \beta(0))} + \frac{3L^2 \sigma^2(k)}{n\beta(k)} + \frac{\Delta^2 \beta^2(k)}{n}$$

$$+ \frac{2\beta(k) \|y(k)\|^2}{n(1 - \beta(0))} + \frac{3\beta(k) \|y(k)\|^2}{n}$$

$$- \frac{2\alpha(k)(f(\bar{x}(k)) - f^*)}{n}.$$ (27)

Since $\{\alpha(k), \beta(k)\}$ satisfy Eq. (16), Eq. (18) also holds, which implies that

$$\sum_{k=0}^{\infty} \left( \alpha^2(k) + \beta^2(k) + \frac{\alpha^2(k)}{\beta(k)} + \beta(k) \|y(k)\|^2 \right) < \infty.$$
Thus, we can apply Lemma 4 to Eq. (27) to have

\[
\lim_{k \to \infty} r(k) \text{ exits a.s. for each } x^* \text{ if } \sum_{k=0}^{\infty} a(k) (f(\bar{x}(k)) - f^*) < \infty \quad \text{a.s.}
\]  

(28)

Consequently, since \( \sum_{k=0}^{\infty} a(k) = \infty \), Eq. (28) implies

\[
\lim_{k \to \infty} \inf_{\kappa} f(\bar{x}(k)) = f^* \quad \text{a.s.}
\]

Let \( (\bar{x}(k)) \) be a subsequence of \( (x(\kappa)) \) such that

\[
\lim_{\kappa \to \infty} f(\bar{x}(k)) = \lim \inf_{k} f(\bar{x}(k)) = f^* \quad \text{a.s.}
\]

Since \( \| x(\kappa) - x^* \| \) converges, the subsequence \( \{x(\kappa)\} \) is bounded. Hence, there is a convergent subsequence of \( \{x(\kappa)\} \), which converges to some minimizer \( \bar{x} \) of problem (5). Since \( \lim_{k \to \infty} f(\bar{x}(k)) = f^* \) a.s. In addition, since \( \| \bar{x}(k) - x^* \| \) converges a.s. for each \( x^* \), and in particular, \( \| \bar{x}(k) - \bar{x} \| \) converges a.s., we obtain

\[
\lim_{k \to \infty} f(\bar{x}(k)) = \bar{x} \quad \text{a.s.}
\]

which together with Eq. (17) implies Eq. (21).

6.2 Proof of Theorem 2

Taking the expectation on both sides of Eq. (27) yields

\[
\mathbb{E} \left[ r(k + 1) \right] \leq \mathbb{E} \left[ r(k) \right] + \frac{6L^2\sigma^2(k)}{n(1 - \beta(0))} + \frac{3L^2\sigma^2(k)}{n\beta(k)} + \frac{\alpha(0)}{n} \mathbb{E} \left[ f(\bar{x}(k)) - f^* \right].
\]

(29)

Fix some \( \ell \in \mathcal{W} \) and consider

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

\[
= - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f^* \right)
\]

\[
- \frac{2\alpha(k)}{n} \sum_{i=1}^{n} f_i(\bar{x}(k)) - f_i(x_\ell(k)),
\]

which by the \( L_i \)-Lipschitz continuity of \( f_i \) yields

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

\[
\leq - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f^* \right)
\]

\[
+ \frac{2}{n} \sum_{i=1}^{n} \alpha(k) L_i \left\| \bar{x}(k) - x_\ell(k) \right\|
\]

\[
\leq - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f^* \right)
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{L_i^2\sigma^2(k)}{\beta(k)} + \beta(k)\|\bar{x}(k) - x_\ell(k)\|^2 \right\}
\]

\[
\leq - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f^* \right) + \frac{L^2\sigma^2(k)}{n\beta(k)} + \beta(k)\|Y(k)\|^2,
\]

where the second inequality is due to the Cauchy-Schwarz inequality. Substituting the preceding relation into Eq. (27) gives

\[
\mathbb{E} \left[ r(k + 1) \right] \leq \mathbb{E} \left[ r(k) \right] + \frac{6L^2\sigma^2(k)}{n(1 - \beta(0))} + \frac{4L^2\sigma^2(k)}{n\beta(k)} + \alpha(k) \mathbb{E} \left[ f(x_\ell(k)) - f^* \right].
\]

(30)

Recall from Eq. (23) that

\[
\alpha(k) = \frac{1}{(k + 2)^{1/4}}, \quad \beta(k) = \frac{1}{(k + 2)^{1/2}},
\]

implying that \( 1 / (1 - \beta(0)) \leq 4 \). In addition, using the integral test we have

\[
\sum_{k=0}^{K} \alpha(k) = \sum_{k=0}^{K} \frac{1}{(k + 2)^{1/4}} \geq \int_{0}^{K} \frac{ds}{(s + 2)^{1/4}} \geq 4(K + 2)^{1/4}
\]

\[
\sum_{k=0}^{K} \alpha^2(k) \leq 1 + \int_{0}^{K} \frac{ds}{(s + 2)^{1/2}} \leq 5
\]

\[
\sum_{k=0}^{K} \beta^2(k) \leq 1 + \int_{0}^{K} \frac{ds}{s + 2} \leq 1 + \ln(K + 2)
\]

\[
\sum_{k=0}^{K} \alpha^2(k) \beta(k) \leq 1 + \int_{0}^{K} \frac{ds}{s + 2} \leq 1 + \ln(K + 2)
\]
Thus, dividing both sides of Eq. (30) by \(\sum_{k=0}^{K} \alpha(k)\) and by the Jensen’s inequality yields Eq. (24), i.e.,

\[
\mathbb{E} \left[ f \left( \frac{\sum_{k=0}^{K} \alpha(k)x_{r}(k)}{\sum_{k=0}^{K} \alpha(k)} \right) \right] - f^* \\
\leq \frac{\sum_{k=0}^{K} \alpha(k) \mathbb{E}[f(x_r(k))]}{\sum_{k=0}^{K} \alpha(k)} - f^* \\
\leq \frac{n\mathbb{E}[r(0)]}{8(K+2)^{1/4}} + \frac{n\mathbb{E}[\|Y(0)\|^2]}{2(1-\sigma_2)(K+2)^{1/4}} + \frac{16L^2}{(K+2)^{1/4}} + \frac{9nL^2(1+\ln(K+2))}{2(1-\sigma)^2(K+2)^{1/4}} + \frac{5n^2\Delta^2(1+\ln(K+2))}{8(K+2)^{1/4}}.
\]

### 6.3 Proof of Theorem 3

Fix some \(\ell \in \mathcal{V}\). First, the strong convexity of \(f_i\) gives

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

which by the \(L_1\)-Lipschitz continuity of \(f_i\) yields

\[
- \frac{2\alpha(k)}{n} \sum_{i=1}^{n} g_i(x_i(k))(x_i(k) - x^*) \\
\leq -\frac{2}{n} \sum_{i=1}^{n} \alpha(k) L_i \|y_i(1)\| - \frac{2\alpha(k)}{n} \sum_{i=1}^{n} f_i(x_i(k)) - f_i(x^*) \\
= -\mu \alpha(k) \frac{n}{n} \sum_{i=1}^{n} \|x_i(k) - x^*\|^2 \\
\leq \frac{L^2\alpha^2(k)}{n\beta(k)} + \frac{\beta(k)\|Y(0)\|^2}{n} - \mu \alpha(k) r(k) \\
= -\frac{2\alpha(k)}{n} \sum_{i=1}^{n} f_i(x_i(k)) - f_i(x^*) + f_i(x_i(k)) - f_i(x^*) \\
\leq \frac{2L^2\alpha^2(k)}{n\beta(k)} + 2\beta(k)\|Y(0)\|^2 - \mu \alpha(k) r(k) \\
= -\frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f(x^*) \right).
\]

Thus, substituting Eq. (31) into Eq. (20) we obtain

\[
\mathbb{E} \left[ \left[ r(k+1) \mid \mathcal{F}_k \right] \right] \\
\leq r(k) + \frac{6L^2\alpha^2(k)}{n(1-\beta(0))} + \frac{4L^2\alpha^2(k)}{n\beta(k)} + \frac{\Delta^2\beta^2(k)}{n\beta(k)} \\
+ \frac{2\beta(k)\|Y(0)\|^2}{n} + \frac{2L^2\alpha^2(k)}{n\beta(k)} + 2\beta(k)\|Y(0)\|^2 \\
- \mu \alpha(k) r(k) - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f(x^*) \right) \\
= (1 - \mu \alpha(k)) r(k) + \frac{6L^2\alpha^2(k)}{n(1-\beta(0))} + \frac{4L^2\alpha^2(k)}{n\beta(k)} + \frac{\Delta^2\beta^2(k)}{n\beta(k)} \\
+ \frac{4\beta(k)\|Y(0)\|^2}{n} - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f(x^*) \right).
\]

Note that since \(\alpha(k)\) satisfies Eq. (25), we have

\[
1 - \mu \alpha(k) \leq 1 - \frac{1}{k + 2}.
\]

Then, using Eq. (25) into Eq. (32) gives

\[
\mathbb{E} \left[ \left[ r(k+1) \mid \mathcal{F}_k \right] \right] \\
\leq \left( 1 - \frac{1}{k + 2} \right) r(k) + \frac{6L^2\alpha^2(k)}{n(1-\beta(0))} + \frac{4L^2\alpha^2(k)}{n\beta(k)} + \frac{\Delta^2\beta^2(k)}{n\beta(k)} \\
+ \frac{2\beta(k)\|Y(0)\|^2}{n} - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f(x^*) \right) \\
\leq \frac{k + 1}{k + 2} r(k) + \frac{6L^2\alpha^2(k)}{n(1-\beta(0))} + \frac{4L^2\alpha^2(k)}{n\beta(k)} + \frac{\Delta^2\beta^2(k)}{n\beta(k)} \\
+ \frac{4\beta(k)\|Y(0)\|^2}{n} - \frac{2\alpha(k)}{n} \left( f(x_\ell(k)) - f(x^*) \right) \\
\leq \frac{k + 1}{k + 2} r(k) + \frac{6L^2\alpha^2(k)}{n(1-\beta(0))(k + 2)^{1/3}} + \frac{4L^2\alpha^2(k)}{n\beta(k)(k + 2)^{1/3}} \\
+ \frac{\Delta^2\beta^2(k)}{(k + 2)^{1/3}} + \frac{4\beta(k)\|Y(0)\|^2}{n} - \frac{2\alpha(k)}{n(k + 2)} \left( f(x_\ell(k)) - f(x^*) \right),
\]

which when multiplying both sides by \((k + 2)\) yields

\[
(k + 2) \mathbb{E} \left[ \left[ r(k+1) \mid \mathcal{F}_k \right] \right] \\
\leq (k + 1) r(k) + \frac{6L^2\alpha^2(0)}{n(1-\beta(0))(k + 2)} + \frac{4L^2\alpha^2(0)}{n\beta(0)(k + 2)^{1/3}} \\
+ \frac{\Delta^2\beta^2(0)}{(k + 2)^{1/3}} + 4\beta(k)\|Y(0)\|^2 \\
- \frac{2\alpha(0)}{n} \left( f(x_\ell(k)) - f(x^*) \right).
\]
By iteratively updating over \( k \) of the preceding relation we have

\[
(k + 2) \mathbb{E} \left[ r(k + 1) \mid \mathcal{F}_k \right] 
\leq r(0) + \sum_{t=0}^{k} \frac{6L^2 \alpha^2(0)}{n(1 - \beta(0))(t + 2)^2} (t + 2) 
+ \sum_{t=0}^{k} \frac{\Delta^2 \beta^2(0)}{(t + 2)^{1/3}} + \frac{4L^2 \alpha^2(0)}{n \beta(0)(t + 2)^{1/3}} (t + 2) 
+ 4 \sum_{t=0}^{k} (t + 2) \beta(t) \|Y(t)\|^2 - \frac{2 \alpha(0)}{n} \sum_{t=0}^{k} \left( f(x_t(t)) - f(x^*) \right) 
\leq r(0) + \frac{6L^2 \alpha^2(0)}{n(1 - \beta(0))} \sum_{t=0}^{k} \frac{1}{(t + 2)} 
+ \sum_{t=0}^{k} \frac{\Delta^2 \beta^2(0)}{(t + 2)^{1/3}} + \frac{4L^2 \alpha^2(0)}{n \beta(0)(t + 2)^{1/3}} 
+ 4 \sum_{t=0}^{k} (t + 2) \beta(t) \mathbb{E} \left[ \|Y(t)\|^2 \right],
\]

Taking the expectation of the preceding inequality and rearranging gives

\[
\frac{2 \alpha(0)}{n} \sum_{t=0}^{k} \mathbb{E} \left[ f(x_t(t)) - f(x^*) \right] + (k + 2) \mathbb{E} \left[ r(k + 1) \right] 
\leq \mathbb{E} \left[ r(0) \right] + \frac{6L^2 \alpha^2(0)}{n(1 - \beta(0))} \sum_{t=0}^{k} \frac{1}{(t + 2)} 
+ \sum_{t=0}^{k} \frac{\Delta^2 \beta^2(0)}{(t + 2)^{1/3}} + \frac{4L^2 \alpha^2(0)}{n \beta(0)(t + 2)^{1/3}} 
+ 4 \sum_{t=0}^{k} (t + 2) \beta(t) \mathbb{E} \left[ \|Y(t)\|^2 \right],
\]

which when dropping the nonnegative term \( \mathbb{E} \left[ r(k + 1) \right] \) and dividing both sides by \( (k + 2) / n \) yields

\[
\frac{2 \alpha(0)}{n} \sum_{t=0}^{k} \mathbb{E} \left[ f(x_t(t)) - f(x^*) \right] 
\leq \frac{n \mathbb{E} \left[ r(0) \right]}{k + 2} + \frac{6L^2 \alpha^2(0)}{(1 - \beta(0))(k + 2)} \sum_{t=0}^{k} \frac{1}{(t + 2)} 
+ \frac{n}{k + 2} \sum_{t=0}^{k} \frac{\Delta^2 \beta^2(0)}{(t + 2)^{1/3}} + \frac{4L^2 \alpha^2(0)}{n \beta(0)(t + 2)^{1/3}} 
+ \frac{4n}{(k + 2)} \sum_{t=0}^{k} (t + 2) \beta(t) \mathbb{E} \left[ \|Y(t)\|^2 \right].
\]
Thus, using the equation above into Eq. (34) we have
\[
\sum_{t=0}^{k} \left( (t+2)\beta(t)\mathbb{E} \left[ \|Y(t)\|^2 \right] \right) \leq \mathbb{E} \left[ \|Y(0)\| \left( 1 + \ln(k+2) \right) \right] + \sum_{t=0}^{k} \frac{3n\sigma^2\Delta^2\beta^2(0)}{(1-\sigma_2)^2(k+2)^{1/3}} + \sum_{t=0}^{k} \frac{20L^2\sigma^2(0)}{\beta(0)(1-\sigma_2)^3(k+2)^{1/3}}.
\]
which when substituting into Eq. (33) yields
\[
\frac{2\alpha(0)}{k+2} \sum_{t=0}^{k} \mathbb{E} \left[ f(x_t(t)) - f(x^*) \right] \leq \frac{n\mathbb{E} \left[ r(0) \right]}{k+2} + \frac{4n\mathbb{E} \left[ \|Y(0)\| \left( 1 + \ln(k+2) \right) \right]}{(1-\sigma_2)(k+2)} + \frac{nL^2\sigma^2(0)}{\beta(0)(k+2)^{1/3}} + \sum_{t=0}^{k} \frac{27nL^2\sigma^2(0)}{(1-\sigma_2)^2(k+2)^{1/3}}.
\]
Moreover, note that $\mu(0) > 1$. Thus, using (37) and (38) gives Eq. (26), i.e.
\[
\mathbb{E} \left[ \|z_t(0) - x^*\|^2 \right] \leq \frac{n\mathbb{E} \left[ r(0) \right]}{(k+2)} + \frac{4n\mathbb{E} \left[ \|Y(0)\| \left( 1 + \ln(k+2) \right) \right]}{(1-\sigma_2)(k+2)} + \frac{6L^2\sigma^2(0)(1+\ln(k+2))}{(1-\beta(0))(k+2)} + \frac{n\Delta^2\beta^2(0)}{(k+2)^{1/3}} + \frac{4L^2\sigma^2(0)}{\beta(0)(k+2)^{1/3}} + \frac{8n^2\sigma^2\Delta^2\beta^2(0)}{(1-\sigma_2)^2(k+2)^{1/3}} + \frac{27nL^2\sigma^2(0)}{(1-\sigma_2)^2(k+2)^{1/3}}.
\]

7 CONCLUDING REMARKS

In this paper, we consider distributed optimization over networks of nodes under quantized communication. For solving such problems, we propose a distributed variant of the popular stochastic approximation. Our main contribution is to establish the convergence of the distributed stochastic approximation. In addition, we provide an explicit formula for the rate of convergence of our proposed method as a function of the underlying network topology and the number of quantized bits.

As mentioned, the distributed stochastic approximation considered in this paper can be viewed as a distributed two-time-scale algorithm. Thus, we believe that the proposed algorithm can be extended for solving other problems, such as, distributed reinforcement learning over multi-agent systems. This is an interesting topic which we leave for our future studies.

REFERENCES

[1] T. C. Aysal, M. J. Coates, and M. G. Rabbat. 2008. Distributed Average Consensus With Dithered Quantization. IEEE Transactions on Signal Processing 56, 10 (2008), 4985–4998.

[2] V.S. Borkar. 2008. Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. 2011. Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. Foundations and Trends in Machine Learning 3 (2011), 1–122.

[4] M. Chiang and T. Zhang. 2016. Fog and IoT: An Overview of Research Opportunities. IEEE Internet of Things Journal 3, 6 (Dec 2016), 854–864.

[5] J. Cortes, S. Martinez, T. Karatas, and F. Bullo. 2004. Coverage control for mobile sensing networks. IEEE Transactions on Robotics and Automation 20, 4 (2004), 243–255.

[6] G. Dalal, Gugan Thopppe, Balázs Szörényi, and Shie Mannor. 2018. Finite Sample Analysis of Two-Time-scale Stochastic Approximation with Applications to Reinforcement Learning. In Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6–9 July 2018. 1199–1233.

[7] T. T. Doan. 2018. Aggregating Stochastic Gradients in Distributed Optimization. In Proceedings of American Control Conference (ACC).

[8] T. T. Doan, C. L. Beck, and R. Srikant. 2017. On the Convergence Rate of Distributed Gradient Methods for Finite-Sum Optimization Under Communication Delays. Proceedings ACM Meas. Anal. Comput. Syst. 1, 2 (Dec 2017), 37:1–37:27.

[9] T. T. Doan, C. L. Beck, and R. Srikant. 2018. Convergence Rate of Distributed Subgradient Methods under Communication Delays. In Proceedings of American Control Conference (ACC).

[10] T. T. Doan, S. Bose, D. H. Nguyen, and C. L. Beck. 2019. Convergence of the Iterates in Mirror Descent Methods. IEEE Control Systems Letters 3, 1 (Jan 2019), 114–119.

[11] T. T. Doan, S. T. Maguluri, and J. Romberg. 2018. On the Convergence of Distributed Subgradient Methods under Quantization. To appear in the Proceedings of Allerton Conference on Communication, Control, and Computing, Monticello, IL.

[12] J.C. Duchi, A. Agarwal, and M.J. Wainwright. 2012. Dual averaging for distributed optimization: Convergence analysis and network scaling. IEEE Transactions on Automatic Control 57, 3 (2012), 592–606.
A PROOFS OF RESULTS IN SECTION 4.1

In this section, we provide the analysis for the results stated in Section 4.1. To do that, we utilize the result on the properties of the projection studied in [29], stated as follows.

Lemma 5 (Lemma 1 [29]). Let \( X \) be a nonempty closed convex set in \( \mathbb{R}^d \). Then, we have for any \( x \in \mathbb{R}^d \)

\[
(a) \quad \langle P_X(x) - x, y - x \rangle \leq -\|P_X(x) - x\|^2 \quad \text{for all} \quad y \in X.
\]

\[
(b) \quad \|P_X(x) - u\|^2 \leq \|x - u\|^2 - \|P_X(x) - x\|^2 \quad \text{for all} \quad u \in X.
\]

A.1 Proof of Lemma 1

Since \( Y(k) = X(k) - 1^T(k) = WX(k) \), Eqs. (11) and (12) give

\[
Y(k + 1) = WX(k + 1) = (1 - \beta(k)Y(k) + \beta(k)AWQ(k) - \alpha(k)WG(k) - W\Xi(k)).
\]

By the Cauchy-Schwarz inequality we have for any \( \eta > 0 \) and \( a, b \in \mathbb{R} \)

\[
(a + b)^2 \leq (1 + \eta)a^2 + (1 + 1/\eta)b^2.
\]

Thus, by taking the 2-norm square of Eq. (39) and using Eq. (40) with \( \eta = 1 + (1 - \sigma_2)\beta(k) \) we have

\[
\|Y(k + 1)\|^2 \leq (1 + (1 - \sigma_2)\beta(k)) \left\| (1 - \beta(k)Y(k) + \beta(k)AWQ(k) - \alpha(k)WG(k) - W\Xi(k)) \right\|^2
\]

\[
\leq 1 + (1 - \sigma_2)\beta(k) \left( (1 + \beta(k))Y(k) + \beta(k)AWQ(k) \right)^2 + \frac{2}{\sigma_2\beta(k)} \left\| (\alpha(k)WG(k) + W\Xi(k)) \right\|^2
\]

\[
\leq 1 + (1 - \sigma_2)\beta(k) \left( (1 + \beta(k))Y(k) + \beta(k)AWQ(k) \right)^2 + \frac{2}{\sigma_2\beta(k)} \left\| (\alpha(k)WG(k)) \right\|^2 + \frac{2}{\sigma_2\beta(k)} \left\| W\Xi(k) \right\|^2.
\]
since \( x_i(k), q_i(k) \in X \), for all \( i \in \mathcal{V} \) and \( k \geq 0 \). By Lemma 5(b) we have

\[
\|W \Xi(k)\|^2 \leq \|\Xi(k)\|^2 = \sum_{i=1}^{n} \|\xi_i(k)\|^2
\]

\[
= \sum_{i=1}^{n} \|v_i(k) - x_i(k + 1)\|^2 = \sum_{i=1}^{n} \|v_i(k) - \mathcal{P}_X[v_i(k)]\|^2
\]

\[
\leq \sum_{i=1}^{n} \|v_i(k) - u_i(k)\|^2 = \sum_{i=1}^{n} \|\alpha(k)g_i(k)\|^2
\]

\[
\leq \sum_{i=1}^{n} \alpha^2(k)\ell_i^2 \leq L^2 \alpha^2(k),
\]

where the last inequality is due Proposition 1. Third, by Eq. (4) we have

\[
\|E(k)\|^2 = \sum_{i=1}^{n} \|x_i(k) - q_i(k)\|^2 \leq \sum_{i=1}^{n} \|\xi_i(k)\|^2 \leq \sum_{i=1}^{n} \sum_{\ell=1}^{d} (\Delta^\ell)^2
\]

\[
\leq \sum_{i=1}^{n} \Delta^2 = n\Delta^2,
\]

where recall that \( \Delta^\ell \) is the distance of \( \ell \)-coordinate of \( x_i - q_i \). In addition, by Eq. (7) we have

\[
\|(1 - \beta(k))Y(k) + \beta(k)AY(k)\|^2
\]

\[
\leq \|(1 - (1 - \sigma_2)\beta(k))Y(k)\|.
\]

Thus, using Eqs. (44) and (45) we consider

\[
\|(1 - \beta(k))Y(k) + \beta(k)AWQ(k)\|^2
\]

\[
= \|(1 - (1 - \beta(k))Y(k) + \beta(k)AY(k) + \beta(k)AWE(k)\|^2
\]

\[
\leq \|(1 - \beta(k))Y(k) + \beta(k)AY(k)\|^2 + \|\beta(k)AWE(k)\|^2
\]

\[
+ 2\beta(k)\left(1 - \beta(k)\right)Y(k) + \beta(k)AY(k)\right)^T AWE(k)
\]

\[
\leq \|(1 - \beta(k))Y(k) + \beta(k)AY(k)\|^2 + \beta(k)\left(1 - \beta(k)\right)Y(k) + \beta(k)AY(k)\right)^T AWE(k)
\]

\[
\leq \|(1 - (1 - \sigma_2)\beta(k))Y(k)\|^2 + \frac{n\sigma_2^2\Delta^2\beta^2(k)}{2}
\]

\[
+ 2\beta(k)\left(1 - \beta(k)\right)Y(k) + \beta(k)AY(k)\right)^T AWE(k).
\]

Note that by Eq. (4) we have \( \mathbb{E}[\xi_i(k)] = 0 \). Thus, taking the conditional expectation of Eq. (41) w.r.t. \( \mathcal{F}_k \) and using Eqs. (42), (43), and (46) yields Eq. (13), i.e.,

\[
\mathbb{E}[\|Y(k + 1)\|^2 | \mathcal{F}_k]
\]

\[
\leq (1 + (1 - \sigma_2)\beta(k)(1 - (1 - \sigma_2)\beta(k))^2\|Y(k)\|^2
\]

\[
+ \frac{n\sigma_2^2\Delta^2\beta^2(k)}{2} + \frac{L^2}{2} \left(\frac{1}{1 - \sigma_2}\right)\beta(k)^2
\]

\[
\leq (1 - (1 - \sigma_2)\beta(k))\|Y(k)\|^2
\]

\[
+ \frac{n\sigma_2^2\Delta^2\beta^2(k)}{2} + \frac{L^2}{2} \left(\frac{1}{1 - \sigma_2}\right)\beta(k)^2,
\]

where the last inequality is because \( \beta(k) \) is nonincreasing.

Finally, taking the expectation on both sides of the preceding relation and summing up over \( k = 0, \ldots, K \) for some \( K \) immediately give Eq. (14).

### A.2 Proof of Lemma 2

Recall from Eq. (13) that

\[
\mathbb{E}[\|Y(k + 1)\|^2 | \mathcal{F}_k]
\]

\[
\leq (1 - \sigma_2)\beta(k)\|Y(k)\|^2
\]

\[
+ \frac{n\sigma_2^2\Delta^2\beta^2(k)}{2} + \frac{L^2}{2} \left(\frac{1}{1 - \sigma_2}\right)\beta(k)^2.
\]

In addition, since \( \alpha(k) \) and \( \beta(k) \) satisfy Eq. (16) we can apply Lemma 4 to the preceding equation. Thus, we obtain \( \{Y(k)\} \) converges a.s. and \( \sum_{k=0}^{\infty} \beta(k)|Y(k)|^2 < \infty \). Since \( \sum_{k=0}^{\infty} \beta(k) = \infty \), the preceding relation immediately gives

\[
\lim_{k \to \infty} \|x_i(k) - \bar{x}(k)\| = 0 \quad \text{a.s.}, \quad \text{for all } i \in \mathcal{V}.
\]

### A.3 Proof of Lemma 3

First, Eq. (12) gives

\[
r(k + 1) = \|v(k) - \bar{x}(k) - x^*\|^2
\]

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

\[
+ \beta(k)q(k) - \bar{x}(k)
\]

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

\[
+ \frac{\|\beta(k)\bar{e}(k)\|^2}{2} + \|\bar{e}(k)\|^2 - 2\bar{e}(k)^T \bar{x}(k)
\]

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

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

which by taking the conditional expectation w.r.t. \( \mathcal{F}_k \) yields

\[
\mathbb{E}[\|r(k + 1) | \mathcal{F}_k\]

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

\[
+ \frac{\|\beta(k)\bar{e}(k)\|^2}{2} + \|\bar{e}(k)\|^2 - 2\bar{e}^T(k) \bar{x}(k) - x^*
\]

\[
+ 2\alpha(k) - \bar{e}^T(k) \sum_{i=1}^{n} g_i(x_i(k)).
\]

First, we use Eq. (43) to have \( \|\xi_i(k)\| \leq L_1\alpha(k) \). Thus, by Eq. (4) we obtain

\[
\|\beta(k)\bar{e}(k)\|^2 + \|\bar{e}(k)\|^2
\]

\[
= \beta^2(k) \left(\frac{1}{n} \sum_{i=1}^{n} e_i(k)\right)^2 + \left(\frac{1}{n} \sum_{i=1}^{n} \xi_i(k)\right)^2
\]

\[
\leq \frac{\beta^2(k)}{n} \sum_{i=1}^{n} |e_i|^2 + \frac{\alpha^2(k)}{n} \sum_{i=1}^{n} \xi_i^2 \leq \Delta^2 \beta^2(k) + \frac{L^2\alpha^2(k)}{n}.
\]
Second, using Proposition 1 gives

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

(49)

Third, for convenience let \( r_i(k), h_i(k) \) be defined as

\[ r_i(k) = \frac{\beta(k)}{1 - \beta(k)} v_i(k) - \frac{\beta^2(k)}{1 - \beta(k)} \sum_{j=1}^{n} a_{ij} q_j(k) \]

\[ h_i(k) = (1 - \beta(k))x^* + \beta(k) \sum_{j=1}^{n} a_{ij} q_j(k) \in \mathcal{X} \text{ a.s.} \]

Using \( r_i(k), h_i(k) \) and by Lemma 5(a) we consider

\[ = \frac{-2}{n\beta(k)} \sum_{i=1}^{n} \xi_i^T(k) \left( r_i(k) - \beta(k)x^* \right) \]

\[ = \frac{-2}{n(1 - \beta(k))} \sum_{i=1}^{n} \xi_i^T(k) \left( v_i(k) - h_i(k) \right) \]

\[ = \frac{2}{n(1 - \beta(k))} \sum_{i=1}^{n} \left( \mathcal{P}_\mathcal{X}[v_i(k)] - v_i(k) \right) \left( v_i(k) - h_i(k) \right) \]

\[ \leq - \frac{2}{n(1 - \beta(k))} \sum_{i=1}^{n} \| \xi_i(k) \|^2 \leq 0. \]

(50)

Thus, we have

\[ - 2 \xi^T(k) (\bar{x}(k) - x^*) \]

\[ = - \frac{2}{n\beta(k)} \sum_{i=1}^{n} \xi_i^T(k) \left( \beta(k)\bar{x}(k) - \beta(k)x^* \right) \]

\[ = - \frac{2}{n\beta(k)} \sum_{i=1}^{n} \xi_i^T(k) \left( \beta(k)x(k) - r_i(k) \right) \]

\[ - \frac{2}{n\beta(k)} \sum_{i=1}^{n} \xi_i^T(k) \left( r_i(k) - \beta(k)x^* \right) \]

\[ \leq - \frac{2}{n\beta(k)} \sum_{i=1}^{n} \xi_i^T(k) \beta(k) (\bar{x}(k) - x_i(k)) \]

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

\[ \leq \frac{2}{n} \sum_{i=1}^{n} \| \xi_i^T(k) \| \| y_i(k) \| + \frac{2\alpha(k)}{n(1 - \beta(k))} \sum_{i=1}^{n} L_i \| \xi_i(k) \| \]

\[ \leq \frac{2L^2\alpha^2(k)}{n\beta(k)} + \frac{\beta(k)\|Y(k)\|^2}{n} + \frac{2L^2\alpha^2(k)}{n(1 - \beta(k))} \]

(43)

\[ \leq \frac{L^2\alpha^2(k)}{n\beta(k)} + \frac{\beta(k)\|Y(k)\|^2}{n} + \frac{2L^2\alpha^2(k)}{n(1 - \beta(k))}. \]

(51)

where the last inequality is due to the Cauchy-Schwarz inequality and \( 1 - \beta(k) \geq 1 - \beta(0) \). Next, consider the following

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

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

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

\[ \leq \| \bar{x}(k) - x^* \|^2 + \frac{2L^2\alpha^2(k)}{n^2} \]

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

\[ \leq r(k) + \frac{L^2\alpha^2(k)}{n^2} + \frac{L^2\alpha^2(k)}{n\beta(k)} + \frac{\beta(k)\|Y(k)\|^2}{n} \]

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

(52)

We now substitute Eqs. (48), (49), (51), and (52) into Eq. (47) to have Eq. (20), i.e.,

\[ \mathbb{E} \left\{ r(k + 1) \mid F_k \right\} \]

\[ \leq r(k) + \frac{L^2\alpha^2(k)}{n^2} + \frac{L^2\alpha^2(k)}{n\beta(k)} + \frac{\beta(k)\|Y(k)\|^2}{n} \]

\[ + \Delta^2 \beta^2(k) \]

\[ + \frac{L^2\alpha^2(k)}{n^2} + \frac{2L^2\alpha^2(k)}{n} \]

\[ + \frac{L^2\alpha^2(k)}{n\beta(k)} + \frac{\beta(k)\|Y(k)\|^2}{n} + \frac{2L^2\alpha^2(k)}{n(1 - \beta(0))} \]

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

\[ \leq r(k) + \frac{2L^2\alpha^2(k)}{n\beta(k)} + \Delta^2 \beta^2(k) \]

\[ + \frac{6L^2\alpha^2(k)}{n(1 - \beta(0))} + \frac{2\beta(k)\|Y(k)\|^2}{n} \]

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