DISTRIBUTED STOCHASTIC SUBGRADIENT PROJECTION ALGORITHMS FOR CONVEX OPTIMIZATION
S. SUNDHAR RAM, A. NEDIĆ, AND V. V. VEERAVALLI *

Abstract. We consider a distributed multi-agent network system where the goal is to minimize a sum of agent objective functions subject to a common set of constraints. For this problem, we propose a distributed subgradient algorithm in which each agent maintains an iterate sequence and in each iteration the latest iterate is communicated by the agent to its neighbors. Each agent then averages the received iterates with its own iterate, and then adjusts the iterate using subgradient information (known with stochastic errors) of its own function and by projecting onto the constraint set. The focus of this paper is to explore the effects of stochastic subgradient errors on the convergence of the algorithm. We consider general stochastic errors that have uniformly bounded second moments and obtain bounds on the limiting performance of the algorithm for diminishing and non-diminishing stepsizes. Under the additional condition that the mean of the errors diminish, we prove that with diminishing stepsizes there is mean consensus between the agents and mean convergence to the optimum function value. When the convergence of the mean errors is sufficiently quick, we strengthen the result to consensus and convergence to the optimal with probability 1.

Key words. Distributed algorithm, convex optimization, subgradient methods, stochastic approximation.

AMS subject classifications. 90C15

1. Introduction. A number of problems that arise in the context of wired and wireless networks can be posed as the minimization of a sum of functions, when each component function is available only to a specific agent [22, 24, 25]. Often, it is not efficient, or not possible, for the network agents to share their objective functions with each other or with a central coordinator. In such scenarios, distributed algorithms that only require the agents to locally exchange limited and high level information are preferable. For example, in a large wireless network, energy is a scarce resource and it might not be efficient for a central coordinator to learn the individual objective functions from each and every agent [22]. In a network of databases from which information is to be mined, privacy considerations may not allow the sharing of the objective functions [33]. In a distributed network on a single chip, for the chip to be fault tolerant, it is desirable to perform the processing in a distributed manner to account for the statistical process variations [31].

We consider constrained minimization of a sum of convex functions, where each component function is known partially (with stochastic errors) to a specific network agent. The algorithm proposed builds on the distributed algorithm proposed in [18] for the unconstrained minimization problem. Each agent maintains an iterate sequence and in each iteration the latest iterate is communicated by the agent to its neighbors. Each agent then averages the received iterates with its own iterate, and then adjusts the iterate using subgradient information (known with stochastic errors) of its own function and by projecting onto the constraint set. The inter-agent information exchange model is a synchronous and delayless version of the computational model proposed by Tsitsiklis [29]. The algorithm is distributed since there is no central coor-
The algorithm is local since each agent uses only locally available information (its objective function) and communicates locally with its immediate neighbors.

Related to this work are the distributed incremental algorithms, where the network agents sequentially update a single iterate sequence in a cyclic or a random order \cite{5,12,15,22,24}. The effect of stochastic errors on these algorithms have been investigated in \cite{3,9,14,16,22,24,27}. In an incremental algorithm, there is a single iterate sequence and only one agent updates the iterate at a given time. Thus, while distributed and local, incremental algorithms differ fundamentally from the algorithm studied in this paper (where all agents update simultaneously). Also related are the optimization algorithms in \cite{2,30}. However, these algorithms are not local as the complete objective function information is available to each and every agent.\footnote{There, the goal is to distribute the processing.}

The work in this paper is also related at a much broader level to the distributed consensus algorithms \cite{2,11,13,17,19,20,28–30,32}. In these algorithms, each agent starts with a different value and through local information exchange, an agreement on the same value is reached by all agents. The effect of random errors on consensus algorithms have been investigated in \cite{10,13,32}. In addition, since we are interested in the effect of stochastic errors, the paper is also related to the literature on stochastic subgradient methods \cite{6–8}.

This paper adds to the multi-agent distributed optimization framework studied in \cite{18}. The novelties are: 1) the study of the effects of stochastic errors in subgradient evaluations; 2) the consideration of constrained optimization problem within the distributed multi-agent setting. The practical importance of studying both constrained and stochastic optimization problems are well understood and require no motivation. The presence of the constraint set complicates the analysis as it introduces non-linearities in the system dynamics. The non-linearity issues that we face have some similarities to those in the constrained consensus problem investigated in \cite{19}, though the problems are fundamentally different. The presence of subgradient stochastic errors adds another layer of complexity to the analysis as the errors made by each agent propagate through the network to every other agent and also across time, making the iterates statistically dependent across time and agents.

We consider general stochastic errors that have uniformly bounded second moments and obtain bounds on the limiting performance of the algorithm for diminishing and non-diminishing stepsizes. Under the additional condition that the mean of the errors diminish, we prove that with diminishing stepsizes there is mean consensus between the agents and mean convergence to the optimum function value. When the convergence of the mean errors is sufficiently quick, we strengthen the result to consensus and convergence to the optimal with probability 1.

The rest of the paper is organized as follows. In Section 2, we formulate the problem, describe the algorithm and state our basic assumptions. In Section 3, we state some results from literature that will be used in the analyses. In Section 4, we derive two important lemmas that will form the backbone of the analysis. In Section 5, we consider general stochastic errors and obtain bounds on their performance and in Section 6, we study errors with diminishing means. Finally, we provide some concluding remarks in Section 8.

\section*{2. Problem, algorithm and an overview of results} In this section, we formulate the problem of interest and describe the algorithm that we propose. We also state and discuss our assumptions on the agent connectivity and information.
exchange. At the end, we provide some basic relations that we use later on when investigating the algorithm’s behavior.

2.1. Problem. We consider a network of $m$ agents that are indexed by $1, \ldots, m$. Often, when convenient, we index the agents by using set $V = \{1, \ldots, m\}$. The network objective is to solve the following constrained optimization problem:

$$\begin{align*}
&\text{minimize} \quad \sum_{i=1}^{m} f_i(x) \\
&\text{subject to} \quad x \in X,
\end{align*}$$

(2.1)

where $X \subseteq \mathbb{R}^n$ is a constraint set and $f_i : \mathbb{R}^n \to \mathbb{R}$ for all $i$. We are interested in the following case.

Assumption 1. The functions $f_i$ and set $X$ are such that

(a) The set $X \subseteq \mathbb{R}^n$ is closed and convex.

(b) The functions $f_i$, $i \in V$, are convex over an open set that contains the set $X$.

The function $f_i$ is known only partially to agent $i$ and it can only obtain a noisy estimate of the function subgradient. The goal is to solve problem (2.1) using an algorithm that is distributed and local. We refer the reader to [24, 25] for wireless network applications that can be cast in this framework.

Note that we make no assumption on the differentiability of the functions $f_i$. At points where the gradient does not exist, we use the notion of subgradients. A vector $\nabla f_i$ is a subgradient of $f_i$ at a point $x$ if

$$\nabla f_i(x)^T(y - x) \leq f_i(y) - f_i(x) \quad \text{for all } y \in X. \quad (2.2)$$

Since the set $X$ is contained in an open set over which the functions are convex, a subgradient of $f_i$ exists at any point of the set $X$ (see [26] or [1]). Related to problem (2.1), we will also use the following notation

$$f(x) = \sum_{i=1}^{m} f_i(x), \quad f^* = \min_{x \in X} f(x), \quad X^* = \{x \in X : f(x) = f^*\}.$$ 

2.2. Algorithm. Let $w_{i,k}$ be the iterate with agent $i$ at the end of iteration $k$. At the beginning of iteration $k + 1$, agent $i$ receives the current iterate of a subset of the agents. Then, agent $i$ computes a weighted average of these iterates and adjusts this average along the negative subgradient direction of $f_i$, which is computed with stochastic errors. The adjusted iterate is then projected onto the constraint set $X$. Mathematically, each agent $i$ generates its iterate sequence $\{w_{i,0}\}$ according to the following relation:

$$w_{i,k+1} = P_X \left[ v_{i,k} - \alpha_{k+1} (\nabla f_i (v_{i,k}) + \epsilon_{i,k+1}) \right], \quad (2.3)$$

starting with some initial iterate $w_{0,k} \in X$. Here, $\nabla f_i(v_{i,k})$ denotes the subgradient of $f_i$ at $v_{i,k}$ and $\epsilon_{i,k+1}$ is the stochastic error in the subgradient evaluation, $\alpha_{k+1}$ is the stepsize, and $P_X$ denotes the Euclidean projection onto the set $X$. The vector $v_{i,k}$ is the weighted average computed by agent $i$ and is given by

$$v_{i,k} = \sum_{j \in N_i(k+1)} a_{i,j}(k+1)w_{j,k}, \quad (2.4)$$
where $N_i(k+1)$ denotes the set of agents whose current iterates are available to agent $i$ in the $(k+1)$-th iteration. We assume that $i \in N_i(k+1)$ for all agents and at all times $k$. The scalars $a_{i,j}(k+1)$ are the non-negative weights that agent $i$ assigns to agent $j$’s iterate. We will find it convenient to define $a_{i,j}(k+1)$ as 0 for $j \notin N_i(k+1)$ and rewrite (2.4) as

$$v_{i,k} = \sum_{j=1}^{m} a_{i,j}(k+1)w_{j,k}.$$  (2.5)

2.3. Additional assumptions. In addition to Assumption 1, we make the following assumptions on the inter-agent exchange model and the weights. The first assumption requires the agents to communicate sufficiently often so that all the component functions, directly or indirectly, influence the iterate sequence of any agent. Recall that we defined $N_i(k+1)$ as the set of agents that agent $i$ communicates with in iteration $k+1$. Define $(V,E_{k+1})$ to be the graph with edges $E_{k+1} = \{(j,i) : j \in N_i(k+1), i \in V\}$.

Assumption 2. There exists a scalar $Q$ such that the graph $(V, \cup_{l=1}^{Q} E_{k+l})$ is strongly connected for all $k$.

It is also important that the influence of the functions $f_i$ is “equal” in a long run so that the sum of the component functions is minimized rather than a weighted sum of them. The influence of a component $f_j$ on the iterates of agent $i$ depends on the weights that agent $i$ uses. To ensure equal influence, we make the following assumption on the weights.

Assumption 3. For $i \in V$ and all $k$,

(a) $a_{i,j}(k+1) \geq 0$, and $a_{i,j}(k+1) = 0$ when $j \notin N_i(k+1)$,

(b) $\sum_{j=1}^{m} a_{i,j}(k+1) = 1$,

(c) There exists a scalar $\eta$, $0 < \eta < 1$, such that $a_{i,j}(k+1) \geq \eta$ when $j \in N_i(k+1)$,

(d) $\sum_{i=1}^{n} a_{i,j}(k+1) = 1$.

Assumptions 2 and 3 state that each agent calculates a weighted average of all the iterates it has access to. Assumption 3 ensures that each agent gives a sufficient weight to its current iterate and all the iterates it receives. Assumption 3, together with Assumption 2 as we will see later, ensures that all the agents are equally influential in the long run. To satisfy Assumption 2, the agents need to coordinate their weights. Some coordination schemes are discussed in [18, 24].

3. Preliminaries. In this section, we state some results for future reference.

3.1. Euclidean norm inequalities. For $M \in \mathbb{N}$, and vectors $v_1, \ldots, v_M \in \mathbb{R}^n$,

$$\sum_{i=1}^{M} \left\| v_i - \frac{1}{M} \sum_{j=1}^{M} v_j \right\|^2 \leq \sum_{i=1}^{M} \left\| v_i - x \right\|^2 \quad \text{for any } x \in \mathbb{R}^n.$$  (3.1)

The preceding relation states that the average of a finite set of vectors minimizes the sum of distances between each vector and any vector in $\mathbb{R}^n$. It is easy to verify the inequality by using the first derivative test to observe that $x = \frac{1}{M} \sum_{j=1}^{M} v_j$ minimizes the function $\sum_{i=1}^{M} \left\| v_i - x \right\|^2$.

---

$^2$Agents need not be aware of the common bound $\eta$. 

---
Both the Euclidean norm and its square are convex functions. Therefore, for nonnegative scalars $\beta_1, \ldots, \beta_M$ such that $\sum_{i=1}^{M} \beta_i = 1$, and vectors $v_1, \ldots, v_M \in \mathbb{R}^n$, 

$$\left\| \sum_{i=1}^{M} \beta_i v_i \right\| \leq \sum_{i=1}^{M} \beta_i \|v_i\|,$$ 

(3.2)

and 

$$\left\| \sum_{i=1}^{M} \beta_i v_i \right\|^2 \leq \sum_{i=1}^{M} \beta_i \|v_i\|^2.$$ 

(3.3)

The following inequality is related to the Euclidean projection onto a nonempty, closed and convex set $X$.

$$\|\mathcal{P}_X[x] - \mathcal{P}_X[y]\| \leq \|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n.$$ 

(3.4)

This property is known as the non-expansive property of the Euclidean projection (see [1], Proposition 2.2.1).

### 3.2. Scalar sequence

For a scalar $\beta$ and a scalar sequence $\{\gamma_k\}$, we consider the "$k$-length convolution" sequence $\sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} = \beta^k \gamma_0 + \beta^{k-1} \gamma_1 + \ldots + \beta \gamma_{k-1} + \gamma_k$. We have the following result.

**Lemma 3.1.** Let $0 < \beta < 1$ and let $\{\gamma_k\}$ be a positive scalar sequence.

(a) If $\lim_{k \to \infty} \gamma_k = 0$, then $\lim_{k \to \infty} \sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} = 0$.

(b) If $\sum_{k=0}^{\infty} \gamma_k < \infty$, then $\sum_{k=0}^{\infty} (\sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell}) < \infty$.

(c) If $\{\zeta_k\}$ is a non-negative scalar sequence such that $\sum_{k=1}^{\infty} \zeta_k = \infty$ and $\gamma_k \to 0$, then $\sum_{k=1}^{k} \gamma_k \zeta_k = 0$.

**Proof.** (a) Let $\epsilon > 0$ be arbitrary. Since $\gamma_k \to 0$, there is an index $K$ such that $\gamma_k \leq \epsilon$ for all $k \geq K$. For all $k \geq K + 1$, we have

$$\sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} = \sum_{\ell=0}^{K} \beta^{k-\ell} \gamma_{\ell} + \sum_{\ell=K+1}^{k} \beta^{k-\ell} \gamma_{\ell} \leq \max_{0 \leq \ell \leq K} \gamma_{\ell} \sum_{\ell=0}^{K} \beta^{k-\ell} + \epsilon \sum_{\ell=K+1}^{k} \beta^{k-\ell}.$$

Since $\sum_{\ell=K+1}^{k} \beta^{k-\ell} \leq \frac{1}{\beta}$ and

$$\sum_{\ell=0}^{K} \beta^{k-\ell} = \beta^k + \cdots + \beta^K = \beta^k (1 + \cdots + \beta^{K-1}) \leq \frac{\beta^k}{1 - \beta},$$

it follows that for all $k \geq K + 1$,

$$\sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} \leq \max_{0 \leq \ell \leq K} \gamma_{\ell} \frac{\beta^k}{1 - \beta} + \frac{\epsilon}{1 - \beta}.$$

Therefore,

$$\limsup_{k \to \infty} \sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} \leq \frac{\epsilon}{1 - \beta}.$$

Since $\epsilon$ is arbitrary, we conclude that $\limsup_{k \to \infty} \sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} = 0$, implying

$$\lim_{k \to \infty} \sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} = 0.$$
(b) Let $\sum_{k=0}^{\infty} \gamma_k < \infty$. For any integer $M \geq 1$, we have
\[
\sum_{k=0}^{M} \left( \sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} \right) = \sum_{\ell=0}^{M-1} \gamma_{\ell} \sum_{k=\ell+1}^{M} \beta^{\ell} \leq \sum_{\ell=0}^{M-1} \gamma_{\ell} \frac{1}{1-\beta},
\]
implying that
\[
\sum_{k=0}^{\infty} \left( \sum_{\ell=0}^{k} \beta^{k-\ell} \gamma_{\ell} \right) \leq \frac{1}{1-\beta} \sum_{\ell=0}^{\infty} \gamma_{\ell} < \infty.
\]

(c) Since $\gamma_k \to 0$, for every $\epsilon > 0$ there will exist a $K$ such that $\gamma_k < \epsilon$ for all $k \geq K$. Thus,
\[
\sum_{k=K+1}^{\infty} \gamma_k \zeta_k \leq \sum_{k=1}^{K} \gamma_k \zeta_k + \sum_{k=K+1}^{\infty} \gamma_k \zeta_k 
\leq \sum_{k=1}^{K} \gamma_k \zeta_k + \epsilon \sum_{k=1}^{\infty} \zeta_k
\leq \epsilon.
\]
since $\zeta_k$ is nonnegative and $\sum_k \zeta_k = \infty$. Since $\epsilon$ arbitrary we can conclude the result.

3.3. Matrix convergence. Let $A(k)$ be the matrix with $(i,j)$-th entry equal to $a_{i,j}(k)$. As a consequence of Assumptions 3a, 3b and 3d, the matrix $A(k)$ is doubly stochastic. Define, for all $k, s$ with $k \geq s$,
\[
\Phi(k,s) = A(k)A(k-1) \cdots A(s+1).
\]

We next state a result from [17] (Corollary 1) on the convergence properties of the matrix $\Phi(k,s)$. Let $[\Phi(k,s)]_{i,j}$ denote the $(i,j)$-th entry of the matrix $\Phi(k,s)$, and let $e \in \mathbb{R}^m$ be the column vector with all entries equal to 1.

**Lemma 3.2.** Let Assumptions 2 and 3 hold. Then
1. $\lim_{k \to \infty} \Phi(k,s) = \frac{1}{m} e e^T$ for all $s$.
2. Further, the convergence is geometric and the rate of convergence is given by
\[
|[\Phi(k,s)]_{i,j} - \frac{1}{m}| \leq \theta \beta^{k-s},
\]
where
\[
\theta = \left( 1 - \frac{\eta}{4m^2} \right)^{-2} \quad \beta = \left( 1 - \frac{\eta}{4m^2} \right)^{\frac{1}{2}}.
\]

3.4. Stochastic convergence. We next state some results that deal with the convergence of a sequence of random vectors. The first theorem is the well known Fatou’s lemma [4].

**Lemma 3.3.** Let $\{X_i\}$ be a sequence of non-negative random variables. Then
\[
\mathbb{E}\left[ \liminf_{n \to \infty} X_n \right] \leq \liminf_{n \to \infty} \mathbb{E}[X_n].
\]

\[3\] The sum of its entries in every row and in every column is equal to 1.
The next theorem concerns the summability of a random sequence.

**Lemma 3.4.** Let \( \{X_i\} \) be a sequence non-negative random variables such that \( \sum_{k=1}^{\infty} E[X_k] < \infty \). Then \( \sum_{k=1}^{\infty} X_k < \infty \) with probability 1.

To see that the result is true, first note from the monotone convergence theorem [4] that \( E[\sum_{k=1}^{\infty} X_k] = \sum_{k=1}^{\infty} E[X_k] \), and is hence finite. If the expected value of a random variable is finite then it has to be finite with probability 1 and so \( \sum_{k=1}^{\infty} X_k < \infty \) with probability 1. The final theorem we state is due to Robbins and Siegmund (Lemma 11, Chapter 2.2, [21]).

**Theorem 3.5.** Let \( \{B_k\}, \{D_k\}, \) and \( \{H_k\} \) be non-negative random sequences and let \( \{\zeta_k\} \) be a deterministic sequence. Let \( G_k \) be the \( \sigma \)-algebra generated by \( B_1, \ldots, B_k, D_1, \ldots, D_k, H_1, \ldots, H_k \). Suppose that \( \sum_k \zeta_k < \infty \),

\[
E[B_{k+1} \mid G_k] \leq (1 + \zeta_k)B_k - D_k + H_k \quad \text{for all } k,
\]

(3.6)

and \( \sum_k H_k < \infty \) with probability 1. Then, the sequence \( \{B_k\} \) converges to a non-negative random variable and \( \sum_k D_k < \infty \) with probability 1 and in mean square sense.

4. Basic iterate relations. In this section, we derive two basic relations that form the basis for the analysis in this paper. The first relation describes the evolution of the disagreements across the agents. In particular, we study the behavior of \( \|y_k - w_{i,k}\| \), where \( \{y_k\} \) is the auxiliary vector sequence defined by

\[
y_k = \frac{1}{m} \sum_{i=1}^{m} w_{i,k} \quad \text{for all } k.
\]

(4.1)

We first obtain a basic relation for the agent iterate sequences \( \{w_{i,k}\} \) generated by the distributed method in (2.3)–(2.4).

**Lemma 4.1.** Let Assumptions 1a, 2, and 3 hold. Assume that the subgradients of \( f_i \) are uniformly bounded over the set \( X \), i.e., there are scalars \( C_i \) such that

\[
\|\nabla f_i(x)\| \leq C_i \quad \text{for all } x \in X \text{ and all } i \in V.
\]

Then, for all \( j \in V \) and \( k \geq 0 \),

\[
\|y_{k+1} - w_{j,k+1}\| \leq \theta m^2 \beta^k \max_{i \in V} \|w_{i,0}\| + \theta [\sum_{\ell=1}^{k} \alpha_{\ell} \beta^{k+1-\ell} \sum_{i=1}^{m} (C_i + \|\epsilon_{i,\ell}\|) + \frac{\alpha_{k+1}}{m} \sum_{i=1}^{m} (C_i + \|\epsilon_{i,k+1}\|)] + \alpha_{k+1} (C_j + \|\epsilon_{j,k+1}\|).
\]

Proof. Define for all \( i \in V \) and all \( k \),

\[
p_{i,k+1} = w_{i,k+1} - \sum_{j=1}^{m} a_{i,j}(k+1)w_{j,k}.
\]

(4.2)

Using the matrices \( \Phi(k,s) \) defined in (3.5) we can write (as in [16])

\[
w_{j,k+1} = \sum_{i=1}^{m} [\Phi(k+1,0)]_{j,i}w_{i,0} + p_{j,k+1} + \sum_{\ell=1}^{k} \left( \sum_{i=1}^{m} [\Phi(k+1,\ell)]_{j,i}p_{i,\ell} \right).
\]

(4.3)
Using (4.2), we can also rewrite $y_k$, defined in (4.1), as follows

$$y_{k+1} = \frac{1}{m} \left( \sum_{i=1}^{m} \sum_{j=1}^{m} a_{i,j} (k + 1) w_{j,k} + \sum_{i=1}^{m} p_{i,k+1} \right)$$

$$= \frac{1}{m} \left( \sum_{j=1}^{m} \left( \sum_{i=1}^{m} a_{i,j} (k + 1) \right) w_{j,k} + \sum_{i=1}^{m} p_{i,k+1} \right).$$

In the view of the double stochasticity of the weights, we have $\sum_{i=1}^{m} a_{i,j} (k + 1) = 1$, implying that

$$y_{k+1} = \frac{1}{m} \left( \sum_{j=1}^{m} w_{j,k} + \sum_{i=1}^{m} p_{i,k+1} \right) = y_k + \frac{1}{m} \sum_{i=1}^{m} p_{i,k+1}.$$

Therefore

$$y_{k+1} = y_0 + \frac{1}{m} \sum_{\ell=1}^{k+1} \sum_{i=1}^{m} p_{i,\ell} = \frac{1}{m} \sum_{i=1}^{m} w_{i,0} + \frac{1}{m} \sum_{\ell=1}^{k+1} \sum_{i=1}^{m} p_{i,\ell}. \quad (4.4)$$

Substituting for $y_{k+1}$ from (4.1) and for $w_{j,k+1}$ from (4.3), we obtain

$$\|y_{k+1} - w_{j,k+1}\| = \left\| \frac{1}{m} \sum_{i=1}^{m} w_{i,0} + \frac{1}{m} \sum_{\ell=1}^{k+1} \sum_{i=1}^{m} p_{i,\ell} - \sum_{i=1}^{m} [\Phi(k+1,0)]_{j,i} w_{i,0} + \sum_{\ell=1}^{k+1} \sum_{i=1}^{m} [\Phi(k+1,\ell)]_{j,i} p_{i,\ell} \right\|$$

$$= \left\| \sum_{i=1}^{m} \left( \frac{1}{m} - [\Phi(k+1,0)]_{j,i} \right) w_{i,0} + \sum_{\ell=1}^{k+1} \sum_{i=1}^{m} \left( \frac{1}{m} - [\Phi(k+1,\ell)]_{j,i} \right) p_{i,\ell} + \left( \frac{1}{m} \sum_{i=1}^{m} p_{i,k+1} - p_{j,k+1} \right) \right\|. $$

Therefore, for all $j \in V$ and all $k$,

$$\|y_{k+1} - w_{j,k+1}\| \leq \sum_{\ell=1}^{k} \left\| \frac{1}{m} - [\Phi(k+1,0)]_{j,i} \right\| \|w_{i,0}\|$$

$$+ \sum_{\ell=1}^{k+1} \sum_{i=1}^{m} \left\| \frac{1}{m} - [\Phi(k+1,\ell)]_{j,i} \right\| \|p_{i,\ell}\| + \frac{1}{m} \sum_{i=1}^{m} \|p_{i,k+1}\| + \|p_{j,k+1}\|. $$

We can bound $\|w_{i,0}\| \leq \max_{i \in V} \|w_{i,0}\|$. Further, we can use the rate of convergence result from Lemma 3.2 to bound $\frac{1}{m} - [\Phi(k,\ell)]_{j,i}$. We obtain

$$\|y_{k+1} - w_{j,k+1}\| \leq m \theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + \theta \sum_{\ell=1}^{k} \beta^{k+1-\ell} \sum_{i=1}^{m} \|p_{i,\ell}\|$$

$$+ \frac{1}{m} \sum_{i=1}^{m} \|p_{i,k+1}\| + \|p_{j,k+1}\|. \quad (4.5)$$
We next estimate the norms of the vectors $\|p_{i,k}\|$ for any $k$. From the definition of $p_{i,k+1}$ in (4.2) and the definition of the vector $v_{i,k}$ in (2.4), we have $p_{i,k+1} = w_{i,k+1} - v_{i,k}$. Note that, being a convex combination of vectors $w_{j,k}$ in the convex set $X$, the vector $v_{i,k}$ is in the set $X$. By the definition of the iterate $w_{i,k+1}$ in (2.3) and the non-expansive property of the Euclidean projection in (3.4), we have

$$
\|p_{i,k+1}\| = \|P_X [v_{i,k} - \alpha_{k+1} (\nabla f_i(v_{i,k}) + \epsilon_{i,k+1})] - v_{i,k}\|
\leq \alpha_{k+1} \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\|
\leq \alpha_{k+1} (C_i + \|\epsilon_{i,k+1}\|).
$$

In the last step we have used the subgradient boundedness. By substituting the preceding relation in (4.5), we obtain the desired relation.

We next provide a relation that relates the iterate to the value of the objective function. Recall that $f = \sum_{i=1}^m f_i$.

**Lemma 4.2.** Let Assumptions 1, 2, 3 hold. Assume that the subgradients of $f_i$ are uniformly bounded over the set $X$, i.e., there are scalars $C_i$ such that

$$
\|\nabla f_i(x)\| \leq C_i \quad \text{for all } x \in X \text{ and all } i \in V.
$$

Then, for any $z \in X$ and all $k$,

$$
\sum_{i=1}^m \|v_{i,k+1} - z\|^2 \leq \sum_{i=1}^m \|v_{i,k} - z\|^2 - 2\alpha_{k+1} (f(y_k) - f(z))
+ 2\alpha_{k+1} \left(\max_{i \in V} C_i\right) \sum_{j=1}^m y_k - w_{j,k}\|
- 2\alpha_{k+1} \sum_{i=1}^m \epsilon_{i,k+1}(v_{i,k} - z) + \alpha_{k+1}^2 \sum_{i=1}^m (C_i + \|\epsilon_{i,k+1}\|)^2.
$$

**Proof.** Using the Euclidean projection property in (3.4), from the definition of the iterate $w_{i,k+1}$ in (2.3), we have for any $z \in X$ and all $k$,

$$
\|w_{i,k+1} - z\|^2 = \|P_X [v_{i,k} - \alpha_{k+1} (\nabla f_i(v_{i,k}) + \epsilon_{i,k+1})] - z\|^2
\leq \|v_{i,k} - z\|^2 - 2\alpha_{k+1} \nabla f_i(v_{i,k})^T (v_{i,k} - z) - 2\alpha_{k+1} \epsilon_{i,k+1}^T (v_{i,k} - z)
+ \alpha_{k+1}^2 \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\|^2.
$$

By using the subgradient inequality in (2.2) to bound the second term, we obtain

$$
\|w_{i,k+1} - z\|^2 \leq \|v_{i,k} - z\|^2 - 2\alpha_{k+1} (f_i(v_{i,k}) - f_i(z))
- 2\alpha_{k+1} \epsilon_{i,k+1}^T (v_{i,k} - z) + \alpha_{k+1}^2 \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\|^2. \quad (4.6)
$$

Note that by the convexity of the squared norm [cf. Eq. (3.3)], we have

$$
\sum_{i=1}^m \|v_{i,k+1} - z\|^2 = \sum_{i=1}^m \sum_{j=1}^m a_{i,j} (k + 2) (w_{j,k+1} - z)^2 \leq \sum_{i=1}^m \sum_{j=1}^m a_{i,j} (k + 2) \|w_{j,k+1} - z\|^2.
$$

In view of Assumption 3, we have $\sum_{i=1}^m a_{i,j} (k + 2) = 1$ for all $j$ and $k$, implying that

$$
\sum_{i=1}^m \|v_{i,k+1} - z\|^2 \leq \sum_{j=1}^m \|w_{j,k+1} - z\|^2.
$$
By summing the relations in (4.6) over all \( i \in V \) and by using the preceding relation, we obtain
\[
\sum_{i=1}^{m} \|v_{i,k+1} - z\|^2 \leq \sum_{i=1}^{m} \|v_{i,k} - z\|^2 - 2\alpha_{k+1} \sum_{i=1}^{m} (f_i(v_{i,k}) - f_i(z)) - 2\alpha_{k+1} \sum_{i=1}^{m} \epsilon_{i,k+1}^T (v_{i,k} - z) + \alpha_{k+1}^2 \sum_{i=1}^{m} \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\|^2.
\]

From (2.2) we have
\[
f_i(v_{i,k}) - f_i(z) \geq (f_i(v_{i,k}) - f_i(y_k)) + (f_i(y_k) - f_i(z)) \geq -\|\nabla f_i(v_{i,k})\| \|y_k - v_{i,k}\| + (f_i(y_k) - f_i(z)).
\] (4.7)

Recall that \( v_{i,k} = \sum_{j=1}^{m} a_{i,j}(k+1) w_{j,k} \) [cf. (2.6)]. Substituting for \( v_{i,k} \) and using the convexity of the norm [cf. (2.5)], from (4.7) we obtain
\[
\sum_{i=1}^{m} f_i(v_{i,k}) - f_i(z) \geq - \sum_{i=1}^{m} \|\nabla f_i(v_{i,k})\| \sum_{j=1}^{m} a_{i,j}(k+1) w_{j,k} + (f_i(y_k) - f_i(z)) \geq - \sum_{i=1}^{m} \|\nabla f_i(v_{i,k})\| \sum_{j=1}^{m} a_{i,j}(k+1) \|y_k - w_{j,k}\| + (f_i(y_k) - f_i(z)) \geq - \left( \max_{i \in V} \|\nabla f_i(v_{i,k})\| \right) \sum_{j=1}^{m} \|y_k - w_{j,k}\| + (f_i(y_k) - f_i(z)).
\] (4.8)

By combining the relations in (4.7) and (4.8), we have
\[
\sum_{i=1}^{m} \|v_{i,k+1} - z\|^2 \leq \sum_{i=1}^{m} \|v_{i,k} - z\|^2 - 2\alpha_{k+1} (f_i(y_k) - f_i(z)) + 2\alpha_{k+1} \left( \max_{i \in V} \|\nabla f_i(v_{i,k})\| \right) \sum_{j=1}^{m} \|y_k - w_{j,k}\| - 2\alpha_{k+1} \sum_{i=1}^{m} \epsilon_{i,k+1}^T (v_{i,k} - z) + \alpha_{k+1}^2 \sum_{i=1}^{m} \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\|^2.
\]

The result follows immediately by bounding the subgradient term \( \nabla f_i(v_{i,k}) \) with \( C_i \).

5. General errors. In this section, we study the performance of the algorithm under the following assumption on the errors.

Assumption 4. The second moments of the errors are uniformly bounded, i.e., there are scalars \( \tilde{\nu}_i \) such that \( \mathbb{E}[\|\epsilon_{i,k+1}\|^2] \leq \tilde{\nu}_i \) for all \( i \in V \) and all \( k \).
This condition is very general and holds in all practical cases of interests. We first study $E[\|w_{i,k} - y_k\|]$ and subsequently use it to characterize the performance the method.

**Theorem 5.1.** Let Assumptions 1a, 1b, and 1c hold. Suppose that the subgradients of each $f_i$ are uniformly bounded over $X$, i.e., for each $i \in V$ there is $C_i$ such that

$$\|\nabla f_i(x)\| \leq C_i \quad \text{for all } x \in X.$$ 

If the stepsize $\{\alpha_k\}$ is non-negative and such that $\alpha_k \to \alpha$, $\alpha \geq 0$, then for all $j \in V$,

$$\limsup_{k \to \infty} E[\|y_{k+1} - w_{j,k+1}\|] \leq \alpha \max_{i \in V} \{ C_i + \bar{\nu}_i \} \left( 2 + \frac{m \theta \beta}{1 - \beta} \right).$$

**Proof.** The conditions of Lemma 4.1 are satisfied. Taking the expectation and using the relation $E[\|\epsilon_{i,k}\|] \leq \sqrt{E[\|\epsilon_{i,k}\|^2]} = \bar{\nu}_i$, we obtain for all $j \in V$, we obtain

$$E[\|y_{k+1} - w_{j,k+1}\|] \leq m \theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + m \theta \beta \max_{i \in V} \{ C_i + \bar{\nu}_i \} \sum_{\ell=1}^{k} \beta^{k-\ell} \alpha_{\ell}$$

$$+ 2 \alpha_{k+1} \max_{i \in V} \{ C_i + \bar{\nu}_i \}$$

$$\leq m \theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + m \theta \beta \max_{i \in V} \{ C_i + \bar{\nu}_i \} \sum_{\ell=1}^{k} \beta^{k-\ell} (\alpha_{\ell} - \alpha)$$

$$+ \theta \beta m \alpha \max_{i \in V} \{ C_i + \bar{\nu}_i \} \sum_{\ell=1}^{k} \beta^{k-\ell} + 2 \alpha_{k+1} \max_{i \in V} \{ C_i + \bar{\nu}_i \}. \quad (5.1)$$

Note that when $\lim_{k \to \infty} \alpha_k = \alpha$ for all $i$, then by Lemma 3.1(a), we have

$$\lim_{k \to \infty} \sum_{\ell=1}^{k} \beta^{k-\ell} (\alpha_{\ell} - \alpha) = 0.$$

By taking the limit superior in (5.1) as $k \to \infty$ and using the preceding relation, we obtain the result.

Observe that when the stepsize is diminishing, i.e., $\alpha = 0$, $E[\|y_{k+1} - w_{j,k+1}\|]$ converges to 0, and there is consensus in mean. We formally state this as a corollary.

**Corollary 5.2.** Let the conditions of Theorem 5.1 hold with $\alpha = 0$. Then $E[\|w_{j,k} - y_k\|] \to 0$ for all $j \in V$.

We next obtain bounds on the performance of the algorithm. We will make the additional assumption that the set $X$ is bounded. Thus the subgradients of each $f_i$ are also bounded (see [1], Proposition 4.2.3). We will continue to use $C_i$ to denote an upperbound on the subgradient of $f_i$ over the set $X$. We will also use the notation

$$\|E[\epsilon_{i,k+1}]\| = \bar{\mu}_{i,k}, \quad \limsup_{k \to \infty} \bar{\mu}_{i,k} = \bar{\mu}_i.$$

Note from Jensen’s inequality that $\bar{\mu}_{i,k} \leq \bar{\nu}_i$. In our analysis, we often use the preceding relation to simplify certain terms. However, we also use the bounds $\bar{\mu}_{i,k}$ explicitly to obtain special results for the case when the errors are zero mean, i.e., $\bar{\mu}_{i,k} = 0$, or they diminish to zero, i.e., $\bar{\mu}_i = 0$. 
Theorem 5.3. Let Assumptions 1, 2, 3 and 4 hold. Assume that the set $X$ is bounded. Let $\alpha_k$ be a non-negative sequence that converges to $\alpha$, where $\alpha \geq 0$. If $\alpha = 0$, also assume that $\sum \alpha_k = \infty$. Then, for all $j \in V$,
\[
\liminf_{k \to \infty} E[f(w_{j,k})] \leq f^* + \max_{x, y \in X} \|x - y\| \sum_{i=1}^{m} \tilde{\mu}_i + m\alpha \left( \max_{i \in V} \{C_i + \bar{\nu}_i\} \right)^2 \left( \frac{9}{2} + \frac{2m\beta}{1 - \beta} \right),
\]
where $C_i$ is an upperbound on the subgradient of $f_i$ over the set $X$.

Proof. Since the set $X$ is bounded the subgradient of $f_i$ over the set $X$ is bounded and hence $C_i$ is finite. Thus the conditions of Lemma 4.2 are satisfied. Thus, we have for any $z \in X$ and all $k$,
\[
\sum_{i=1}^{m} \|v_{i,k+1} - z\|^2 \leq \sum_{i=1}^{m} \|v_{i,k} - z\|^2 - 2\alpha_{k+1} (f(y_k) - f(z)) \]
\[
+ 2\alpha_{k+1} \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \|y_k - w_{j,k}\| \]
\[
+ 2\alpha_{k+1} \max_{i \in V} \|v_{i,k} - z\|^2 + \alpha_{k+1}^2 \sum_{i=1}^{m} (C_i + \|\epsilon_{i,k+1}\|)^2. \]

Since $X$ is bounded, by using $\|v_{i,k} - z\| \leq \max_{x, y \in X} \|x - y\|$, taking the expectation and bounding the error moments, we obtain
\[
\sum_{i=1}^{m} E[\|v_{i,k+1} - z\|^2] \leq \sum_{i=1}^{m} E[\|v_{i,k} - z\|^2] - 2\alpha_{k+1} (E[f(y_k)] - f(z)) \]
\[
+ 2\alpha_{k+1} \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} E[\|y_k - w_{j,k}\|] \]
\[
+ 2\alpha_{k+1} \max_{x, y \in X} \|x - y\| \sum_{i=1}^{m} \tilde{\mu}_{i,k+1} + m\alpha_{k+1} \left( \max_{i \in V} \{C_i + \bar{\nu}_i\} \right)^2. \]

Therefore, for any $z \in X$ and for all $k$,
\[
\sum_{i=1}^{m} E[\|v_{i,k+1} - z\|^2] \leq \sum_{i=1}^{m} E[\|v_{i,k} - z\|^2] \]
\[
- 2\alpha_{k+1} \left( E[f(y_k)] - f(z) \right) - \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} E[\|y_k - w_{j,k}\|] \]
\[
- \max_{x, y \in X} \|x - y\| \sum_{i=1}^{m} \tilde{\mu}_{i,k+1} - \frac{m\alpha_{k+1}}{2} \left( \max_{i \in V} \{C_i + \bar{\nu}_i\} \right)^2. \]

By rearranging the terms and summing over $k = 1, \ldots, K$, for an arbitrary $K$, we
obtain
\[
2 \sum_{k=1}^{K} \alpha_{k+1} \left( \mathbb{E}[f(y_k)] - f(z) \right) - \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \mathbb{E}[\|y_k - w_{j,k}\|]
\]
\[
- \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \bar{\mu}_{i,k+1} - \frac{m \alpha_{k+1}}{2} \left( \max_{i \in V} \{C_i + \bar{\nu}_i\} \right)^2
\]
\[
\leq \sum_{i=1}^{m} \mathbb{E}[\|v_{i,1} - z\|^2] - \sum_{i=1}^{m} \mathbb{E}[\|v_{i,K+1} - z\|^2] \leq m \max_{x,y \in X} \|x - y\|^2.
\]

Note that when \( \alpha_{k+1} \to \alpha, \alpha > 0 \), then \( \sum_{k} \alpha_{k} = \infty \). When \( \alpha = 0 \), we have assumed that the convergence is sufficiently slow so that \( \sum_{k} \alpha_{k} = \infty \). Therefore, by letting \( K \to \infty \), we have for any \( z \in X \),
\[
\liminf_{k \to \infty} \left( \mathbb{E}[f(y_k)] - \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \mathbb{E}[\|y_k - w_{j,k}\|]
\right)
\]
\[
- \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \bar{\mu}_{i,k+1} - \frac{m \alpha_{k+1}}{2} \left( \max_{i \in V} \{C_i + \bar{\nu}_i\} \right)^2 \leq f(z).
\]

Since this is true for any \( z \), by taking infimum over \( z \in X \), and using \( \liminf_{k \to \infty} \bar{\mu}_{i,k} = \bar{\mu}_i \) and \( \liminf_{k \to \infty} \alpha_{k} = \alpha \) we obtain
\[
\liminf_{k \to \infty} \mathbb{E}[f(y_k)] \leq f^* + \frac{m \alpha_{k+1}}{2} \left( \max_{i \in V} \{C_i + \bar{\nu}_i\} \right)^2
\]
\[
+ \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \limsup_{k \to \infty} \mathbb{E}[\|y_k - w_{j,k}\|] + \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \bar{\mu}_i.
\]

Next from the convexity inequality in (2.2) and the boundedness of the subgradients it follows that for all \( k \) and \( j \in V \),
\[
\mathbb{E}[f(w_{j,k}) - f(y_k)] \leq \left( \sum_{i=1}^{m} C_i \right) \mathbb{E}[\|y_k - w_{j,k}\|],
\]

implying
\[
\liminf_{k \to \infty} \mathbb{E}[f(w_{j,k})] \leq f^* + \frac{m \alpha_{k+1}}{2} \left( \max_{i \in V} \{C_i + \bar{\nu}_i\} \right)^2
\]
\[
+ \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \limsup_{k \to \infty} \mathbb{E}[\|y_k - w_{j,k}\|] + \left( \sum_{i=1}^{m} C_i \right) \liminf_{k \to \infty} \mathbb{E}[\|y_k - w_{j,k}\|] + \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \bar{\mu}_i.
\]

By Theorem 5.1 we have for all \( j \in V \),
\[
\limsup_{k \to \infty} \mathbb{E}[\|y_{k+1} - w_{j,k+1}\|] \leq \alpha \max_{i \in V} \{C_i + \bar{\nu}_i\} \left( 2 + \frac{m \theta\beta}{1 - \beta} \right) . \tag{5.2}
\]
By using the preceding relation, we see that

\[
\liminf_{k \to \infty} \mathbb{E}[f(w_{i,k})] \leq f^* + \frac{m \alpha_{k+1}}{2} \left( \max_{i \in V} \{C_i + \tilde{v}_i\} \right)^2 + \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \tilde{\mu}_i \\
+ m \alpha \left( \max_{i \in V} C_i \right) \max_{i \in V} \{C_i + \tilde{v}_i\} \left( 2 + \frac{m \theta \beta}{1 - \beta} \right) \\
+ \alpha \left( \sum_{i=1}^{m} C_i \right) \max_{i \in V} \{C_i + \tilde{v}_i\} \left( 2 + \frac{m \theta \beta}{1 - \beta} \right) \\
\leq f^* + \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \tilde{\mu}_i + m \alpha \max_{i \in V} \{C_i + \tilde{v}_i\} \left( \frac{9}{2} + \frac{2m \theta \beta}{1 - \beta} \right).
\]

When all error moments \(\tilde{\mu}_{i,k}\) are zero, the result of Theorem 5.3 holds even when the boundedness of \(X\) is replaced by the weaker assumption that the subgradients of each \(f_i\) are bounded over the set \(X\). The network topology influences the error only through the term \(\frac{\theta \beta}{1 - \beta}\) and can hence be used as a figure of merit for comparing topologies. For a network that is fully connected at every time, [i.e., \(Q = 1\) in Assumption 2] and when \(\eta\) in Assumption 4 does not depend on the number \(m\) of agents, the term \(\frac{\theta \beta}{1 - \beta}\) is of the order \(m^2\) and the error bound scales as \(m^2\).

We next show that stronger bounds can be obtained for a specific weighted time averages of the iterates \(w_{i,k}\). In particular, we will investigate the limiting behavior of \(\{f(z_{i,t})\}\), where \(z_{i,t} = \sum_{k=1}^{t} \frac{\alpha_{k+1} w_{i,k}}{\sum_{k=1}^{t} \alpha_{k+1}}\). Note that agent \(i\) can locally and recursively evaluate \(z_{i,t+1}\) from \(z_{i,t}\) and \(w_{i,t+1}\).

**Theorem 5.4.** Let the conditions of Theorem 5.3 hold. Then

\[
\limsup_{t \to \infty} \mathbb{E}[f(z_{i,t})] \leq f^* + \sum_{i=1}^{m} \tilde{\mu}_i \max_{x,y \in X} \|x - y\| + m \alpha \left( \max_{i \in V} C_i + \tilde{v}_i\right)^2 \left( \frac{9}{2} + \frac{2m \theta \beta}{1 - \beta} \right).
\]

**Note to Self:** Changed the proof. Also added a Lemma 3.1c

**Proof.** Taking expectations in Lemma 5.3, bounding \(||w_{i,k} - z||\) by the diameter of the set and bounding the moments of the errors we obtain

\[
\sum_{i=1}^{m} \mathbb{E}[||v_{i,k+1} - x^*||^2] \leq \sum_{i=1}^{m} \mathbb{E}[||v_{i,k} - x^*||^2] - 2 \alpha_{k+1} \left( \mathbb{E}[f(y_k)] - f^* \right) \\
+ 2 \alpha_{k+1} \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \mathbb{E}[||y_k - w_{j,k}||] \\
+ 2 \alpha_{k+1} \sum_{i=1}^{m} \tilde{\mu}_{i,k+1} \max_{x,y \in X} \|x - y\| + \alpha_{k+1}^2 \sum_{i=1}^{m} (C_i + \tilde{v}_i)^2.
\]

From the boundedness of the subgradient and 2.2 we have

\[
\mathbb{E}[f(y_k)] - \mathbb{E}[f(w_{j,k})] \geq - \left( \sum_{i=1}^{m} C_i \right) \mathbb{E}[||w_{j,k} - y_k||] \geq - m \left( \max_{i \in V} C_i \right) \mathbb{E}[||w_{j,k} - y_k||].
\]
Therefore, we obtain

\[
\sum_{i=1}^{m} E[\|v_{i,k+1} - x^*\|^2] \leq \sum_{i=1}^{m} E[\|v_{i,k} - x^*\|^2] - 2\alpha_{k+1} \left( E[f(w_{j,k})] - f^* \right) + 2\alpha_{k+1} \left( \max_{i \in V} C_i \right) \left( mE[\|y_{k} - w_{j,k}\|] + \sum_{i=1}^{m} E[\|y_{k} - w_{i,k}\|] \right) + 2\alpha_{k+1} \sum_{i=1}^{m} \bar{\mu}_{i,k+1} \max_{x,y \in X} \|x - y\| + \alpha_{k+1}^2 \sum_{i=1}^{m} (C_i + \bar{\nu}_i)^2.
\]

By re-arranging these terms, summing over \( k = 1, \ldots, t \) and dividing by \( \sum_{k=1}^{t} \alpha_{k+1} \) we obtain

\[
\sum_{k=1}^{t} \alpha_{k+1} E[f(w_{j,k})] \leq f^* + \frac{1}{2 \sum_{k=1}^{t} \alpha_{k+1}} \sum_{i=1}^{m} E[\|v_{i,1} - x^*\|^2] + \sum_{k=1}^{t} \alpha_{k+1} \left( \max_{i \in V} C_i \right) \left( mE[\|y_{k} - w_{j,k}\|] + \sum_{i=1}^{m} E[\|y_{k} - w_{i,k}\|] \right) + \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \frac{\sum_{k=1}^{t} \bar{\mu}_{i,k+1}}{\sum_{k=1}^{t} \alpha_{k+1}} \sum_{i=1}^{m} (C_i + \bar{\nu}_i)^2.
\]

Next by the convexity of \( f \) note that

\[
f(z_{i,t}) = f \left( \frac{\sum_{k=1}^{t} \alpha_{k+1} w_{j,k}}{\sum_{k=1}^{t} \alpha_{k+1}} \right) \leq \sum_{k=1}^{t} \alpha_{k+1} f(w_{j,k}).
\]

From the preceding two relations we obtain

\[
E[f(z_{i,t})] \leq f^* + \frac{1}{2 \sum_{k=1}^{t} \alpha_{k+1}} \sum_{i=1}^{m} E[\|v_{i,1} - x^*\|^2] + \sum_{k=1}^{t} \alpha_{k+1} \left( \max_{i \in V} C_i \right) \left( mE[\|y_{k} - w_{j,k}\|] + \sum_{i=1}^{m} E[\|y_{k} - w_{i,k}\|] \right) + \max_{x,y \in X} \|x - y\| \sum_{i=1}^{m} \frac{\sum_{k=1}^{t} \bar{\mu}_{i,k+1}}{\sum_{k=1}^{t} \alpha_{k+1}} \sum_{i=1}^{m} (C_i + \bar{\nu}_i)^2. \tag{5.3}
\]

First note that in the limit as \( t \to \infty \), boundedness of the set \( X \) implies that the first term converges to 0 since \( \sum_{k=1}^{t} \alpha_{k+1} \to \infty \). The limiting value of the third term can be bounded by \( \sum_{i=1}^{m} \bar{\mu}_{i} \max_{x,y \in X} \|x - y\| \), where \( \bar{\mu}_{i} \) was defined as \( \limsup_{k \to \infty} \mu_{i,k+1} \). The convergence of \( \alpha_{k} \to \alpha \) implies that the last term converges to \( \alpha \sum_{i=1}^{m} (C_i + \bar{\nu}_i)^2 \).

If we prove that the the limit of the second term can be bounded by \( \alpha \left( \frac{\beta}{1-\beta} + 2 \right) \), then we can obtain the result

\[
\limsup_{t \to \infty} E[f(z_{i,t})] \leq f^* + 4m \alpha \left( \max_{i \in V} C_i \right) \left( \max_{i \in V} \{ C_i + \bar{\nu}_i \} \left( \frac{\beta}{1-\beta} + 2 \right) \right) + 2 \sum_{i=1}^{m} \bar{\mu}_{i} \max_{x,y \in X} \|x - y\| + \frac{\alpha}{2} \sum_{i=1}^{m} (C_i + \bar{\nu}_i)^2 \leq f^* + \sum_{i=1}^{m} \bar{\mu}_{i} \max_{x,y \in X} \|x - y\| + m \alpha \left( \max_{i \in V} \{ C_i + \bar{\nu}_i \} \right)^2 \left( \frac{\beta}{1-\beta} + 2 \right) \left( \frac{9}{2} + \frac{2m\alpha\beta}{1-\beta} \right).
\]
We complete the proof by proving that
\[
\limsup_{t \to \infty} \frac{\sum_{k=1}^{t} \alpha_{k+1} E[\|y_{k+1} - w_{j,k+1}\|]}{\sum_{k=1}^{t} \alpha_{k+1}} \leq \alpha \max_{i \in V} \{C_i + \tilde{\nu}_i\} \left( \frac{\beta}{1 - \beta} + 2 \right).
\]

Taking expectations in Lemma 4.1 we obtain
\[
E[\|y_{k+1} - w_{j,k+1}\|] \leq m\theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + m\theta \beta \max_{i \in V} \{C_i + \tilde{\nu}_i\} \sum_{\ell=1}^{k} \beta^{k-\ell} \alpha_{\ell}
+ 2\alpha_{k+1} \max_{i \in V} \{C_i + \tilde{\nu}_i\}.
\]

Therefore,
\[
\frac{\sum_{k=1}^{t} \alpha_{k+1} E[\|y_{k+1} - w_{j,k+1}\|]}{\sum_{k=1}^{t} \alpha_{k+1}} \leq m\theta \max_{i \in V} \|w_{i,0}\| \frac{\sum_{k=1}^{t} \alpha_{k+1} \beta^{k+1}}{\sum_{k=1}^{t} \alpha_{k+1}} + m\theta \beta \max_{i \in V} \{C_i + \tilde{\nu}_i\} \frac{\sum_{k=1}^{t} \alpha_{k+1} \sum_{\ell=1}^{k} \beta^{k-\ell} \alpha_{\ell}}{\sum_{k=1}^{t} \alpha_{k+1}}
+ 2\sum_{k=1}^{t} \alpha_{k+1} \max_{i \in V} \{C_i + \tilde{\nu}_i\}.
\]

Since the stepsizes are bounded note that in the limit as \(t \to \infty\),
\[
\limsup_{t \to \infty} \frac{\sum_{k=1}^{t} \alpha_{k+1} \beta^{k+1}}{\sum_{k=1}^{t} \alpha_{k+1}} \leq \limsup_{t \to \infty} \frac{(\max_{k \in \mathbb{N}} \alpha_{k+1}) \beta}{(\sum_{k=1}^{t} \alpha_{k+1}) (1 - \beta)} \to 0
\]

since \(\sum_{k=1}^{\infty} \alpha_{k+1} = \infty\). Next,
\[
\frac{\sum_{k=1}^{t} \alpha_{k+1} \sum_{\ell=1}^{k} \beta^{k-\ell} \alpha_{\ell}}{\sum_{k=1}^{t} \alpha_{k+1}} = \sum_{k=1}^{t} \alpha_{k+1} \frac{\sum_{\ell=1}^{k} \beta^{k-\ell} \alpha_{\ell}}{\sum_{k=1}^{t} \alpha_{k+1}}
+ \sum_{k=1}^{t} \alpha_{k+1} \frac{\sum_{\ell=1}^{k} \beta^{k-\ell} (\alpha_{\ell} - \alpha)}{\sum_{k=1}^{t} \alpha_{k+1}}
\leq \frac{\alpha \beta}{1 - \beta} + \frac{\sum_{k=1}^{t} \alpha_{k+1} \sum_{\ell=1}^{k} \beta^{k-\ell} (\alpha_{\ell} - \alpha)}{\sum_{k=1}^{t} \alpha_{k+1}}.
\]

Since \((\alpha_{\ell} - \alpha) \to 0\) it follows from Lemma 3.1 that \(\sum_{\ell=1}^{k} \beta^{k-\ell} (\alpha_{\ell} - \alpha) < \infty\). Therefore, from Lemma 3.1 with \(\sum_{\ell=1}^{k} \beta^{k-\ell} (\alpha_{\ell} - \alpha)\) as \(\gamma_k\) and \(\alpha_{k+1}\) as \(\zeta_k\) we can conclude that
\[
\lim_{t \to \infty} \frac{\sum_{k=1}^{t} \alpha_{k+1} \sum_{\ell=1}^{k} \beta^{k-\ell} (\alpha_{\ell} - \alpha)}{\sum_{k=1}^{t} \alpha_{k+1}} = 0,
\]
and
\[
\limsup_{t \to \infty} \frac{\sum_{k=1}^{t} \alpha_{k+1} \sum_{\ell=1}^{k} \beta^{k-\ell} \alpha_{\ell}}{\sum_{k=1}^{t} \alpha_{k+1}} \leq \frac{\alpha \beta}{1 - \beta}.
\]

(5.6)
Finally, note that since \( \alpha_k \to \alpha \),
\[
\sum_{k=1}^{t} \frac{\alpha_k^2}{\alpha_{k+1}} \to \alpha.
\]  
(5.7)

From (5.4) – (5.7) we obtain
\[
\limsup_{t \to \infty} \sum_{k=1}^{t} \frac{\alpha_{k+1}E[\|y_{k+1} - w_{j,k+1}\|]}{\alpha_{k+1}} \leq \alpha \max_{i \in V} \left\{ C_i + \bar{\nu}_i \right\} \left( \frac{\beta}{1 - \beta} + 2 \right).
\]

Note that the bounds in Theorems 5.3 and 5.4 are identical. However, there is an important difference. Theorem 5.4 guarantees that all subsequences of \( E[f(z_{i,k})] \) will satisfy the bound. In contrast, Theorem 5.3 only guarantees that there is a subsequence that satisfies the bound. It is not clear if \( z_{i,k} \) has fundamentally better properties than \( w_{i,k} \), or whether this is an artifact of our analysis.

When a constant stepsize \( \alpha \) is used \( z_{j,t} \) is simply the average of all the iterates at agent \( j \) till time \( t \), i.e., \( z_{j,t} = \frac{\sum_{k=1}^{t} w_{j,k}}{t} \). For this case, with zero mean errors, (5.3) simplifies to
\[
E[f(z_{i,t})] \leq f^* + \frac{1}{2t\alpha} \sum_{i=1}^{m} E[\|v_{i,1} - x^*\|^2] + \left( \max_{i \in V} C_i \right) \frac{1}{t} \sum_{k=1}^{t} \left( mE[\|y_k - w_{j,k}\|] + \sum_{i=1}^{m} E[\|y_k - w_{i,k}\|] \right) + \alpha \sum_{i=1}^{m} (C_i + \bar{\nu}_i)^2.
\]  
(5.8)

Taking expectations in Lemma 4.1 we obtain
\[
E[\|y_{k+1} - w_{j,k+1}\|] \leq m\theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + m\alpha \theta \beta \max_{i \in V} \left\{ C_i + \bar{\nu}_i \right\} \sum_{\ell=1}^{k} \beta^{k-\ell} + 2\alpha \max_{i \in V} \left\{ C_i + \bar{\nu}_i \right\} \leq m\theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + \frac{m\alpha \theta \beta \max_{i \in V} \left\{ C_i + \bar{\nu}_i \right\}}{1 - \beta} + 2\alpha \max_{i \in V} \left\{ C_i + \bar{\nu}_i \right\}
\]

Combining this with (5.8) we obtain
\[
E[f(z_{i,t})] \leq f^* + \frac{1}{2t\alpha} \sum_{i=1}^{m} E[\|v_{i,1} - x^*\|^2] + \frac{2m^2 \theta \beta}{t(1 - \beta)} \left( \max_{i \in V} C_i \right) \left( \max_{i \in V} \|w_{i,0}\| \right) + m\alpha (C_i + \bar{\nu}_i)^2 \left( \frac{9}{2} + \frac{2m\theta \beta}{1 - \beta} \right).
\]  
(5.9)

The preceding equation is important as it provides a bound on the algorithms performance after a finite number of iteration. The bound can be used in obtaining stopping rules for the algorithm. For example, consider the error free case \( (\bar{\nu}_i = 0) \) and suppose
that the goal is to determine the number of iterations required so that each agent can determine a point in the $\epsilon$-optimal set, i.e., in the set $X_\epsilon = \{ x \in X : f(x) < f^* + \epsilon \}$. Minimizing the bound in (5.9) over different stepsize values $\alpha$, we can show that $\epsilon$-optimality can be achieved in $\lceil \frac{1}{\psi_{\epsilon}^2} \rceil$ iterations with a stepsize $\alpha_\epsilon = \sqrt{\frac{A\psi_{\epsilon}}{C}}$, where $\psi_{\epsilon}$ is the positive root of the quadratic equation

$$Bx^2 + 2\sqrt{AC}x - \epsilon = 0,$$

and $A, B$ and $C$ are

$$A = \sum_{i=1}^{m} \| v_{i,1} - x^* \|^2_2, \quad B = \frac{2m^2\theta\beta}{(1 - \beta)} \left( \max_{i \in V} C_i \right) \left( \max_{i \in V} \| w_{i,0} \| \right),$$

$$C = m(C_i + \bar{v}_i)^2 \left( \frac{9}{2} + \frac{2m\theta\beta}{1 - \beta} \right).$$

Since $\psi_{\epsilon}$ scales as $\sqrt{\epsilon}$, we can conclude that $N_\epsilon$ scales as $\frac{1}{\epsilon^2}$. Equivalently, we can state that the distance to optimality diminishes inversely with the square of the number of iterations.

6. Mean Diminishing stochastic errors. In the previous section we made minimal assumptions on the errors. In this section, we will study the special case when the expected values of the errors diminish. While it is possible to study non-diminishing stepsize, we will focus on diminishing stepsizes and obtain sufficient conditions for the convergence of the algorithm to the optimal values. We will first consider the following case.

**Assumption 5.** The expected errors converge to 0, i.e., $\lim_{k \to \infty} \| E[\epsilon_{i,k+1}] \| = 0$ for all $i \in V$.

With very little effort, we can draw the following conclusions from Theorem 5.3 and Theorem 5.4.

**Corollary 6.1.** Let the conditions of Theorem 5.3 hold with $\alpha = 0$. Further, let Assumption 5 hold. Then $\liminf_{k \to \infty} E[f(w_{i,k})] = f^*$ and $\lim_{k \to \infty} E[f(z_{i,k})] = f^*$.

Corollary 6.1 establish convergence in the mean. This does not guarantee convergence of the algorithm to an optimal point for every realization of the error sequences. We next investigate the additional assumptions on the errors that are sufficient to guarantee convergence of the iterate to the optimal set with probability 1. Towards this, define $F_k$ to be the $\sigma$ algebra $\sigma (\epsilon_{i,\ell}; \ i \in V, 0 \leq \ell \leq k)$. Thus, $F_k$ captures the history of the algorithm until the end of iteration $k$. We will strengthen Assumptions 4 and 5 to the following assumption on the errors.

**Assumption 6.** For all $i \in V$, $\sum_{k=0}^{\infty} \| E[\epsilon_{i,k} | F_{k-1}] \|^2 < \infty$ with probability 1. Further, there are scalars $\nu_i$ such that $E[\| \epsilon_{i,k+1} \|^2 | F_k] \leq \nu_i$ for all $k$ with probability 1.

The above assumption is stronger than Assumptions 4 and 5 in two ways. First, note that the conditions are imposed on the conditional moments of the errors, rather than on the absolute moments as in Assumptions 4 and 5. When the errors across iterations are independent this simplifies to assumptions on the absolute error moments. Imposing assumptions on the conditional moments, rather than the absolute moments, is quite standard in stochastic optimization literature [3]. Second, Assumptions 6 requires the conditional mean to not just converge to 0, but converge at a sufficiently fast rate so that it is summable.
For notational brevity we will denote \( \|E[\epsilon_{i,k} \mid F_{k-1}]\| \) by \( \mu_{i,k} \). As in the previous section, we will first investigate the nature of the disagreements between the agents in light of the new assumption on the errors.

**Theorem 6.2.** Let Assumptions 1a, 2, and 3 hold. Suppose that the subgradients of each \( f_i \) are uniformly bounded over \( X \), i.e., for each \( i \in V \) there is \( C_i \) such that

\[
\|\nabla f_i(x)\| \leq C_i \quad \text{for all } x \in X.
\]

Also, assume that, for each \( i \in V \), there is a scalar \( \nu_i \) such that

\[
E[\|\epsilon_{i,k+1}\|^2 \mid F_k] \leq \nu_i^2 \quad \text{for all } k.
\]

If \( \sum_{k=0}^{\infty} \alpha_{k+1}^2 < \infty \) for all \( i \in V \), then with probability 1 for all \( j \in V \),

\[
\sum_{k=1}^{\infty} \alpha_{k+2} \|y_{k+1} - w_{j,k+1}\| < \infty
\]

and \( \|y_{k+1} - w_{j,k+1}\| \) converges to 0 with probability 1 and in mean square.

**Proof.** By Lemma 4.11 and the subgradient boundedness, we have for all \( j \in V \),

\[
\|y_{k+1} - w_{j,k+1}\| \leq m \theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + \theta \sum_{\ell=1}^{k} \beta^{k+1-\ell} \sum_{i=1}^{m} \alpha_{\ell} (C_i + \|\epsilon_{i,\ell}\|)
\]

\[
+ \frac{1}{m} \sum_{i=1}^{m} \alpha_{k+1} (C_i + \|\epsilon_{i,k+1}\|) + \alpha_{k+1} (C_j + \|\epsilon_{j,k+1}\|).
\]

Using the inequalities

\[
\alpha_{k+2} \alpha_{\ell} (C_i + \|\epsilon_{i,\ell}\|) \leq \frac{1}{2} \left( \alpha_{k+2}^2 + \alpha_{\ell}^2 (C_i + \|\epsilon_{i,\ell}\|)^2 \right)
\]

and \( (C_i + \|\epsilon_{i,\ell}\|)^2 \leq 2C_i^2 + 2\|\epsilon_{i,\ell}\|^2 \), we obtain

\[
\alpha_{k+2} \|y_{k+1} - w_{j,k+1}\| \leq \alpha_{k+2} \theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\|
\]

\[
+ \theta \sum_{\ell=1}^{k} \beta^{k+1-\ell} \sum_{i=1}^{m} \left( \frac{1}{2} \alpha_{k+2}^2 + \alpha_{\ell}^2 (C_i^2 + \|\epsilon_{i,\ell}\|^2) \right)
\]

\[
+ \frac{1}{m} \sum_{i=1}^{m} \left( \frac{1}{2} \alpha_{k+2}^2 + \alpha_{k+1}^2 (C_i^2 + \|\epsilon_{i,k+1}\|^2) \right)
\]

\[
+ \frac{1}{2} \alpha_{k+2}^2 + \alpha_{k+1}^2 (C_j^2 + \|\epsilon_{j,k+1}\|^2).
\]

By using the inequalities \( \sum_{\ell=1}^{k} \beta^{k+1-\ell} \leq \frac{1}{1-\theta} \) for all \( k \geq 1 \) and \( \frac{1}{2m} + \frac{1}{2} \leq 1 \), and by grouping the terms accordingly, from the preceding relation we have

\[
\alpha_{k+2} \|y_{k+1} - w_{j,k+1}\| \leq \alpha_{k+2} \theta \beta^{k+1} \max_{i \in V} \|w_{i,0}\| + \left( 1 + \frac{m \theta \beta}{2(1-\beta)} \right) \alpha_{k+2}^2
\]

\[
+ \theta \sum_{\ell=1}^{k} \alpha_{\ell}^2 \beta^{k+1-\ell} \sum_{i=1}^{m} \left( C_i^2 + \|\epsilon_{i,\ell}\|^2 \right)
\]

\[
+ \frac{1}{m} \sum_{i=1}^{m} \left( C_i^2 + \|\epsilon_{i,k+1}\|^2 \right) + \alpha_{k+1}^2 (C_j^2 + \|\epsilon_{j,k+1}\|^2).
\]
Taking the conditional expectation and using $E[\|e_i,\ell\|^2 | F_k] \leq \nu_i^2$, and then taking the expectation again, we obtain

$$E[\alpha_{k+2} \|y_{k+1} - w_{j,k+1}\|] \leq \alpha_{k+2} m \theta \beta^{k+1} \max_{i \in V} \|w_i,0\| + \left(1 + \frac{m \theta \beta}{2(1-\beta)}\right) \alpha_{k+2}$$

$$+ \theta \left(\sum_{i=1}^{m} (C_i^2 + \nu_i^2)\right) \sum_{\ell=1}^{k-1} \alpha_{\ell}^2 \beta^{k+1-\ell}$$

$$+ \frac{1}{m} \alpha_{k+1}^2 \sum_{i=1}^{m} (C_i^2 + \nu_i^2) + \alpha_{k+1}^2 \left(C_j^2 + \nu_j^2\right).$$

(6.1)

Since $\sum_k \alpha_k^2 < \infty$ we have that $\{\alpha_k\}$ is bounded for all $i \in V$. Thus, the first two terms and the last two terms are summable. Furthermore, in view of Lemma 3.4, we have

$$\sum_{k=1}^{\infty} \sum_{\ell=1}^{k} \alpha_{\ell} \beta^{k+1-\ell} \alpha_{k}^2 < \infty.$$

Thus, the third term is also summable. Hence $\sum_{k=1}^{\infty} E[\alpha_{k+2} \|y_{k+1} - w_{j,k+1}\|] < \infty$, and from Lemma 3.4 we see that with probability 1

$$\sum_{k=1}^{\infty} \alpha_{k+2} \|y_{k+1} - w_{j,k+1}\| < \infty. \quad (6.2)$$

We now show that $\lim_{k \to \infty} \|y_k - w_{j,k}\| = 0$ with probability 1. Note that the conditions of Theorem 5.1 are satisfied with $\bar{\nu}_i = \nu_i$ and $\alpha = 0$. Therefore, $\|y_k - w_{j,k}\|$ converges to 0 in the mean and from Lemma 3.3 it follows that

$$0 \leq E\left[\liminf_{k \to \infty} \|y_k - w_{j,k}\|\right] \leq \liminf_{k \to \infty} E[\|y_k - w_{j,k}\|] = 0,$$

and hence $E[\liminf_{k \to \infty} \|y_k - w_{j,k}\|] = 0$. Therefore, with probability 1,

$$\liminf_{k \to \infty} \|y_k - w_{j,k}\| = 0. \quad (6.3)$$

To complete the proof, it suffices to show that $\|y_k - w_{j,k}\|$ converges with probability 1. Define

$$r_{i,k+1} = \sum_{j=1}^{m} a_{i,j} (k+1) w_j(k) - \alpha_{k+1} (\nabla f_i(v_{i,k}) + \epsilon_{i,k+1})$$

and note that $P_X[r_{i,k+1}] = w_{i,k+1}$. Since $y_k = \frac{1}{m} \sum_{i=1}^{m} w_{i,k}$ and the set $X$ is convex, it follows that $y_k \in X$ for all $k$. Therefore, by the non-expansive property of the Euclidean projection in (3.4), we have $\|w_{i,k+1} - y_k\|^2 \leq \|r_{i,k+1} - y_k\|^2$ for all $i \in V$ and all $k$. Summing these relations over all $i$, we obtain

$$\sum_{i=1}^{m} \|w_{i,k+1} - y_k\|^2 \leq \sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2 \quad \text{for all } k.$$
From $y_{k+1} = \frac{1}{m} \sum_{i=1}^{m} w_{i,k+1}$ and the fact that the average of vectors minimizes the sum of distances between each vector and arbitrary vector in $\mathbb{R}^n$ [cf. Eq (3.1)], we further obtain

$$\sum_{i=1}^{m} \|w_{i,k+1} - y_{k+1}\|^2 \leq \sum_{i=1}^{m} \|w_{i,k+1} - y_k\|^2.$$ 

Therefore, for all $k$,

$$\sum_{i=1}^{m} \|w_{i,k+1} - y_{k+1}\|^2 \leq \sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2.$$  \hspace{1cm} (6.4)

We next relate $\sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2$ to $\sum_{i=1}^{m} \|w_{i,k} - y_k\|^2$. From the definition of $r_{i,k+1}$ and the equality $\sum_{j=1}^{m} a_{i,j}(k+1) = 1$ [cf. Assumption 3.2], we have

$$r_{i,k+1} - y_k = \sum_{j=1}^{m} a_{i,j}(k+1) (w_{j,k} - y_k) - \alpha_{k+1} (\nabla f_i(v_{i,k}) + \epsilon_{i,k+1})$$

By Assumption 3.1 and 3.2, we have that the weights $a_{i,j}(k+1), j \in V$ yield a convex combination. Thus, by the convexity of the norm (3.2 and 3.3) and by the subgradient boundedness, we have

$$\|r_{i,k+1} - y_k\|^2 \leq \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\|^2 + 2\alpha_{k+1} \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\|^2$$

$$+ 2\alpha_{k+1} \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\| \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\|$$

$$\leq \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\|^2 + 2\alpha_{k+1}^2 (C_i^2 + \|\epsilon_{i,k+1}\|^2)$$

$$+ 2\alpha_{k+1} (C_i + \|\epsilon_{i,k+1}\|) \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\|.$$ 

Summing over all $i$ and using $\sum_{i=1}^{m} a_{i,j}(k+1) = 1$ [cf. Assumption 3.4], we obtain

$$\sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2 \leq \sum_{j=1}^{m} \|w_{j,k} - y_k\|^2 + 2\alpha_{k+1}^2 \sum_{i=1}^{m} (C_i^2 + \|\epsilon_{i,k+1}\|^2)$$

$$+ 2\alpha_{k+1} \sum_{i=1}^{m} (C_i + \|\epsilon_{i,k+1}\|) \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\|.$$ 

Using this in (6.4) and taking the conditional expectation, we see that for all $k$,

$$\sum_{i=1}^{m} \mathbb{E}[\|w_{i,k+1} - y_{k+1}\|^2 | F_k] \leq \sum_{i=1}^{m} \|w_{i,k} - y_k\|^2 + 2\alpha_{k+1}^2 \sum_{i=1}^{m} (C_i^2 + \nu_i^2)$$

$$+ 2\alpha_{k+1} \sum_{i=1}^{m} (C_i + \nu_i) \sum_{j=1}^{m} \|w_{j,k} - y_k\|,$$  \hspace{1cm} (6.5)
where we also use \( E[\|e_{i,k}\|^2 \mid F_k] \leq \nu_i^2, E[\|e_{i,k}\| \mid F_k] \leq \nu_i \) and \( a_{i,j}(k + 1) \leq 1 \).

We now apply Theorem 6.3 to the relation in (6.5). To verify that the conditions of Theorem 6.3 are satisfied, note that the stepsize satisfies \( \sum_{k=1}^{\infty} \alpha_{k+1}^2 < \infty \) for all \( i \in V \). We also have \( \sum_{k=1}^{\infty} \alpha_{k+1} \|w_{j,k} - y_k\| < \infty \) with probability 1 [cf. (6.2)]. Therefore, the relation in (6.5) satisfies the conditions of Theorem 6.3 with \( \zeta_k = D_k = 0 \), thus implying that \( \|w_{j,k} - y_k\| \) converges with probability 1.

Let us compare Theorem 6.2 and Corollary 3.2. Corollary 3.2 provided sufficient conditions for the different agents to have consensus in the mean. Theorem 6.2 strengthens this to consensus with probability 1 and in mean square sense, for a smaller class of stepsize sequences under a stricter assumption. We next show that the consensus vector is actually in the optimal set.

**Theorem 6.3.** Let Assumptions 1, 2, 3 and 6 hold. Suppose that the subgradients of each \( f_i \) are uniformly bounded over \( X \), i.e., for each \( i \in V \) there is \( C_i \) such that
\[
\|\nabla f_i(x)\| \leq C_i \quad \text{for all } x \in X.
\]
Further, let \( \alpha_k \) be a positive scalar stepsize sequence such that \( \sum_k \alpha_k = \infty \) and \( \sum_k \alpha_k^2 < \infty \). If the optimal set \( X^* \) is nonempty, then the iterate sequence \( \{w_{i,k}\} \) at each agent \( i \in V \) converges to the same optimal point in \( X^* \) with probability 1 and in mean square.

**Proof.** Observe that the conditions of Lemma 4.2 are satisfied. Fixing \( z = x^* \) for \( x^* \in X^* \), taking conditional expectations and using the bounds on the error moments, we obtain for any \( x^* \in X^* \) and any \( k \),
\[
\sum_{i=1}^{m} E[\|v_{i,k+1} - x^*\|^2 \mid F_k] \leq \sum_{i=1}^{m} \|v_{i,k} - x^*\|^2 - 2\alpha_{k+1} \|f(y_k) - f^*\|
+ 2\alpha_{k+1} \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \|y_k - w_{j,k}\|
+ 2\alpha_{k+1} \sum_{i=1}^{m} \mu_{i,k+1} \|v_{i,k} - x^*\| + \alpha_{k+1}^2 \sum_{i=1}^{m} (C_i + \nu_i)^2,
\]
where \( f^* = f(x^*) \). Using the inequality
\[
2\alpha_{k+1} \mu_{i,k+1} \|v_{i,k} - x^*\| \leq \alpha_{k+1}^2 \|v_{i,k} - x^*\|^2 + \mu_{i,k+1}^2
\]
we obtain
\[
\sum_{i=1}^{m} E[\|v_{i,k+1} - x^*\|^2 \mid F_k] \leq \sum_{i=1}^{m} (1 + \alpha_{k+1}^2) \|v_{i,k} - x^*\|^2
- 2\alpha_{k+1} \left( f(y_k) - f^* \right) - \left( \max_{i \in V} C_i \right) \sum_{j=1}^{m} \|y_k - w_{j,k}\|
+ \sum_{i=1}^{m} \mu_{i,k+1}^2 - \frac{1}{2} \alpha_{k+1} \sum_{i=1}^{m} (C_i + \nu_i)^2.
\]
By Theorem 6.2, we have with probability 1,
\[
\sum_{k} \alpha_{k+1} \|w_{j,k} - y_k\| < \infty.
\]
Further, since \( \sum_k \mu_k^2 < \infty \) and \( \sum_k \alpha_k^2 < \infty \) with probability 1, the relation in \((6.6)\) satisfies the conditions of Theorem \[6.5\]. We therefore have
\[
\sum_k \alpha_k (f(y_k) - f^*) < \infty,
\]
and \( \|v_i,k - x^*\| \) converges with probability 1 and in mean square. In addition, by Theorem \[6.2\] we have \( \lim_{k \to \infty} \|w_{i,k} - y_k\| = 0 \) for all \( i \), with probability 1. Hence, \( \lim_{k \to \infty} \|v_{i,k} - y_k\| \to 0 \) for all \( i \), with probability 1. Therefore, \( \|y_k - x^*\| \) converges with probability 1 for any \( x^* \in X^* \). Moreover, from \((6.7)\) and the fact that \( \sum_k \alpha_k = \infty \), by continuity of \( f \), it follows that \( y_k \), and hence \( w_{i,k} \), must converge to a vector in \( X^* \) with probability 1 and in mean square. \( \square \)

Note that Corollary \[6.1\] required boundedness of the constraint set \( X \) while this has been relaxed in Theorem \[6.3\].

7. Implications. The primary source of stochastic errors in the subgradient evaluation is when the objective function is not completely known and has some randomness in it. Such settings arise in sensor network applications that involve distributed and recursive estimation [23].

Let the function \( f_i(x) = \mathbb{E}[g_i(x, R_i)] \), where \( R_i \) is a random variable whose statistics do not depend on \( x \). The statistics of \( R_i \) are not available to agent \( i \) and hence the function \( f_i \) is not known to agent \( i \). Instead, agent \( i \) observes samples of \( R_i \). Thus, in any optimization algorithm that uses the subgradient function, the subgradient must be suitably approximated using the observed samples. In the Robbins-Monro stochastic approximation [21], the subgradient \( \nabla f(x) \) is approximated by \( \nabla g(x, r_i) \), where \( r_i \) denotes a sample of \( R_i \). The associated distributed Robbins-Monro stochastic optimization algorithm is
\[
w_{i,k+1} = \mathcal{P}_X (v_{i,k} - \alpha_{k+1} \nabla g_i (v_{i,k}, r_{i,k+1})) ,
\]
where \( r_{i,k+1} \) are samples of \( R_i \). The expression for the errors is
\[
\epsilon_{i,k+1} = \nabla g_i (v_{i,k}, r_{i,k+1}) - \mathbb{E}\left[ \nabla g_i (v_{i,k}, R_i) \right].
\]
If the samples obtained across iterations are independent then
\[
\mathbb{E}[\epsilon_{i,k+1} | F_k] = \mathbb{E}[\epsilon_{i,k+1} | v_{i,k}] = 0.
\]
If in addition, \( \text{Var}[\nabla g_i(x, R_i)] \) is bounded for all \( x \in X \) then the conditions of Theorems \[5.3\] and \[6.3\] are satisfied.

Let us next consider the case when \( f_i(x) = \mathbb{E}[g_i(x, R_i(x))] \), where \( R_i(x) \) is a random variable that is parameterized by \( x \). To keep the discussion simple let us also assume that \( x \in \mathbb{R} \). As in the previous case, the statistics of \( R_i(x) \) are not known to agent \( i \) but it can obtain samples of \( R_i(x) \) at any level of \( x \). In the Kiefer-Wolfowitz approximation [21],
\[
\nabla f(x) \approx g_i(x, r_i(x + \beta)) - g_i(x, r_i(x)),
\]
where \( r_i(x) \) denotes a sample of the random variable \( R_i(x) \). The corresponding distributed optimization algorithm is
\[
w_{i,k+1} = \mathcal{P}_X \left[ v_{i,k} - \alpha_{k+1} \frac{g_i(v_{i,k}, r_i(v_{i,k} + \beta_{i,k+1})) - g_i(v_{i,k}, r_i(v_{i,k}))}{\beta_{i,k+1}} \right],
\]
where $\beta_{i,k+1}$ is a scalar. In this case, the errors are

$$\epsilon_{i,k} = \frac{g_i(v_{i,k}, r_i(v_{i,k} + \beta_{i,k+1})) - g_i(v_{i,k}, r_i(v_{i,k}))}{\beta_{i,k+1}} - \nabla f(v_{i,k}).$$

If the function $g_i$ is differentiable then $E[\epsilon_{i,k} | v_{i,k}]$ is of the order $\Theta(\beta_{i,k})$. Thus the conditions on the mean value of the errors can be controlled through the sequence $\{\beta_{i,k}\}$ and the conditions in Theorems 5.3, 5.4 and 6.3 can be met by suitably choosing the sequence $\{\beta_{i,k}\}$.

In our analysis we have assumed that the agents can share their iterates with their neighbours in a perfect manner. This may not be a realistic assumption and usually the iterates are quantized before they are communicated. The effect of quantization can also be dealt with in the framework studied in this paper. With quantization, the algorithm (2.3)–(2.4) will be modified to

$$\tilde{v}_{i,k} = \sum_{j \in N_i(k+1)} a_{i,j}(k+1)Q[\tilde{w}_{j,k}],$$

$$\tilde{w}_{i,k+1} = P_X [\tilde{v}_{i,k} - \alpha_{k+1} \nabla f_i(\tilde{v}_{i,k})], \tag{7.2}$$

Here $Q[x]$ denotes the quantized value of $x$. Thus, we can rewrite (7.2) as follows

$$v_{i,k} = \sum_{j \in N_i(k+1)} a_{i,j}(k+1)w_{j,k},$$

$$w_{i,k+1} = P_X [v_{i,k} - \alpha_{k+1} (\nabla f_i(v_{i,k}) + \epsilon_{i,k+1})],$$

with

$$||\epsilon_{i,k+1}|| \leq ||v_{i,k} - \tilde{v}_{i,k}|| + ||f_i(v_{i,k}) - f_i(\tilde{v}_{i,k})||.$$ 

If the function $f_i$ are differentiable and are gradient Lipschitz with constant $L_i$, then

$$||\epsilon_{i,k+1}|| \leq \left(\frac{1}{\alpha_{k+1}} + L\right) ||v_{i,k} - \tilde{v}_{i,k}||.$$ 

Let us now suppose that the quantization granularity along each dimensions is $2d$ units. Then $||Q[x] - x|| = nd$, where $n$ is the dimensions of $x$. Therefore, it follows that $||v_{i,k} - \tilde{v}_{i,k}|| \leq nd$ and

$$||\epsilon_{i,k+1}|| \leq \left(\frac{1}{\alpha} + L\right) nd.$$ 

Thus we have captured the effect of quantization through a bounded and deterministic error term.

8. Discussion. We summarize the main results. We first considered very general errors that have bounded second moments and obtained an explicit bound on the expected deviation from the optimal of the limiting function value as a function of the network properties, objective function and the error moments. For networks that are connected at all times, the bound scaled as $\alpha \nu_i^2 m^4$, where $m$ is the number of agents in the network, $\nu_i^2$ is the bound on the second moment of the errors and $\alpha$ was the stepsize limit. For the constant stepsize case, we obtain a bound on the performance of the algorithm after a finite number of iterations. Using this we argued that the deviation
from optimality diminishes at rates faster than $\frac{1}{k^2}$, where $k$ is the number of iterations. We then strengthened the assumption on the errors and assumed that the expected error converges to 0. For this case we could show that under diminishing stepsizes the expected function value converges to the minimum of the objective function. Finally, we proved that when the expected error converges to 0 sufficiently quickly and the stepsize is also then the iterate sequence converges to an optimal point with probability 1.

We make the following remarks. First, it can be shown that the results in Corollary 5.2 and Theorem 6.2 hold even when the agents use non-identical stepsizes. However, with non-identical agent stepsizes there is no guarantee that the sum of the objectives rather than a weighted sum, is minimized. If it is known that the errors have zero mean then the boundedness assumption on the constraint set $X$ in Theorems 5.3 and 5.4 can also be relaxed. Alternatively, if the set $X$ is bounded, the condition on the error moments in Theorem 6.2 can be weakened from $\sum_{k=1}^{\infty} \mu_{i,k}^2 < \infty$ to $\sum_{k} \alpha_k \mu_{i,k} < \infty$.

In this paper we have considered the class of convex functions. This restricts the number of possible applications for the algorithm. An avenue for further research is to develop distributed algorithms when the functions $f_i$ are not convex, but instead gradient Lipschitz continuous. We have assumed no communication delays between the agents and synchronous processing. Another possible extension is to understand the properties of the algorithm in asynchronous networks with communication delays, as in [29].

REFERENCES

[1] D. P. Bertsekas, A. Nedić, and A. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, 2003.
[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and distributed computation: Numerical methods*, Athena Scientific, 1997.
[3] ———, *Gradient convergence in gradient methods with errors*, SIAM Journal of Optimization, 10 (2000), pp. 627–642.
[4] P. Billingsley, *Probability and Measure*, John Wiley and Sons, 1979.
[5] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with constant stepsize*, SIAM Journal of Optimization, 18 (2007), pp. 29–51.
[6] Y. Ermoliev, *Stochastic programming methods*, Nauka, Moscow, 1976.
[7] ———, *Stochastic quasi-gradient methods and their application to system optimization*, Stochastics, 9 (1983), pp. 1–36.
[8] ———, *Stochastic quazigradient methods*, Numerical Techniques for Stochastic Optimization, (Eds.) Y. Ermoliev and R. J-B Wets, Springer-Verlag, N.Y., (Eds.) Y. Ermoliev and R. J-B Wets, Springer-Verlag, N.Y. (1988), pp. 141–186.
[9] A. A. Gaivoronski, *Convergence properties of backpropogation for neural nets via theory of stochastic gradient methods. Part 1.*, Optimization Methods and Software, 4 (1994), pp. 117–134.
[10] M. Huang and J. H. Manton, *Stochastic approximation for consensus seeking: Mean square and almost sure convergence*, in Proceedings of the 46th IEEE Conference on Decision and Control, 2007.
[11] A. Jadbabaie, J. Lin, and S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Transactions on Automatic Control, 48 (2003), pp. 998–1001.
[12] B. Johansson, M. Rabi, and M. Johansson, *A simple peer-to-peer algorithm for distributed optimization in sensor networks*, in Proceedings of the 46th IEEE Conference on Decision and Control, 2007, pp. 4705–4710.
[13] S. Kar and J. Moura, *Distributed consensus algorithms in sensor networks: Link and channel noise*, http://arxiv.org/abs/0711.3915, 2007.
[14] K. C. Kiwiel, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM Journal on Optimization, 14 (2003), pp. 807–840.
[15] A. Nedić and D. P. Bertsekas, *Incremental subgradient method for nondifferentiable optimization*, SIAM Journal of Optimization, 12 (2001).

[16] ———, *The effect of deterministic noise in sub-gradient methods*, tech. report, Lab. for Information and Decision Systems, MIT, 2007.

[17] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *Distributed subgradient algorithms and quantization effects*, http://arxiv.org/abs/0803.1202, 2008.

[18] A. Nedić and A. Ozdaglar, *Distributed sub-gradient methods for multi-agent optimization*, To appear in Transactions on Automatic Control, (2008).

[19] A. Nedić, A. Ozdaglar, and P. A. Parrilo, *Constrained consensus*. Lab. for Information and Decision Systems, MIT, Report 2779, 2008.

[20] A. Olshevsky and J. N. Tsitsiklis, *Convergence speed in distributed consensus and control*, To appear in SIAM Journal on Control and Optimization, (2008).

[21] B. T. Polyak, *Introduction to optimization*, Optimization Software Inc., 1987.

[22] M. G. Rahmat and R. D. Nowak, *Quantized incremental algorithms for distributed optimization*, IEEE Journal on Select Areas in Communications, 23 (2005), pp. 798–808.

[23] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli, *Stochastic incremental gradient descent for estimation in sensor networks*, in Proceedings of the Asilomar Conference on Signals, Systems, and Computers, 2007, pp. 582–586.

[24] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli, *Incremental stochastic sub-gradient algorithms for convex optimization*. Available at http://arxiv.org/abs/0806.1092, 2008.

[25] S. Sundhar Ram, V. Veeravalli, and A. Nedić, *Distributed and non-autonomous power control through distributed convex optimization*, Submitted, 2008.

[26] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[27] M. V. Solodov and S. K. Zavriev, *Error stability properties of generalized gradient-type algorithms*, Journal of Optimization Theory and Applications, 98 (1998), pp. 663–680.

[28] D. P. Spanos, R. Olfati-Saber, and R. M. Murray, *Approximate distributed Kalman filtering in sensor networks with quantifiable performance*, in Proceedings of IEEE International Conference on Information Processing in Sensor Networks, 2005, pp. 133–139.

[29] J. N. Tsitsiklis, *Problems in decentralized decision making and computation*, PhD thesis, Massachusetts Institute of Technology, 1984.

[30] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Transactions on Automatic Control, 31 (1986), pp. 803–812.

[31] G. Varatkar, S. Narayanan, N. Shanbhag, and D. Jones, *Sensor network-on-chip*, in International Symposium on Systems-on-Chip, 2007, pp. 1–4.

[32] L. Xiao, S. Boyd, and S.-J. Kim, *Distributed average consensus with least mean square deviation*, Journal of Parallel and Distributed Computing, 67 (2007), pp. 33–46.

[33] N. Ye, ed., *Data mining handbook*, Lawrence Erlbaum Associates, 2002, ch. Distributed data mining: Algorithms, systems, and applications. Authored by B. Park and H. Kargupta.