Asynchronous Gossip Algorithms for Stochastic Optimization

S. Sundhar Ram
ECE Dept.
University of Illinois
Urbana, IL 61801
ssrini5@illinois.edu

A. Nedić
IESE Dept.
University of Illinois
Urbana, IL 61801
angelia@illinois.edu

V. V. Veeravalli
ECE Dept.
University of Illinois
Urbana, IL 61801
vvv@illinois.edu

***Abstract*— We consider a distributed multi-agent network system where the goal is to minimize an objective function that can be written as the sum of component functions, each of which is known partially (with stochastic errors) to a specific network agent. We propose an asynchronous algorithm that is motivated by random gossip schemes where each agent has a local Poisson clock. At each tick of its local clock, the agent averages its estimate with a randomly chosen neighbor and adjusts the average using the gradient of its local function that is computed with stochastic errors. We investigate the convergence properties of the algorithm for two different classes of functions. First, we consider differentiable, but not necessarily convex functions, and prove that the gradients converge to zero with probability 1. Then, we consider convex, but not necessarily differentiable functions, and show that the iterates converge to an optimal solution almost surely.**

I. Introduction

The problem of minimizing of a sum of functions when each component function is available partially (with stochastic errors) to a specific network agent is an important problem in the context of wired and wireless networks [15], [24], [27], [28]. These problems require the design of optimization algorithms that are distributed, (i.e. without a central coordinator) and local in the sense that each agent can only use its local objective function and can exchange some limited information with its immediate neighbors.

In this paper, we propose an asynchronous distributed algorithm that is inspired by the random gossip averaging scheme of [7]. Each agent has a local Poisson clock and maintains an iterate sequence. At each tick of its local clock, the agent first randomly selects a neighbor, and computes the average of its current iterate and the iterate received from the selected neighbor. Then, the agent adjusts the computed average using the gradient of its local function, which is known only with stochastic errors. We investigate the convergence properties of the algorithm under two different assumptions on the objective functions: (a) differentiable but not necessarily convex, and (b) convex but not necessarily differentiable.

The algorithm in this paper is related to the distributed consensus-based optimization algorithm proposed in [22] and further studied in [14], [16], [18], [21], [25], [27], [28]. In consensus-based algorithms, each agent maintains an iterate sequence and updates using its local function gradient information. These algorithms are synchronous and require the agents to update simultaneously, which is in contrast with the asynchronous algorithm proposed in this paper. A different distributed model has been proposed in [31] and also studied in [2], [5], [32], where the complete objective function information is available to each agent, with the aim of distributing the processing by allowing an agent to update only a part of the decision vector. Related to the algorithm of this paper is also the literature on incremental algorithms [4], [12], [14], [15], [17], [19], [20], [24], [26], [27], [30], where the network agents sequentially update a single iterate sequence and only one agent updates at any given time in a cyclic or a random order. While being local, the incremental algorithms differ fundamentally from the algorithm studied in this paper (where all agents maintain and update their own iterate sequence). In addition, the work in this paper is related to a much broader class of gossip algorithms used for averaging [1], [8]. Since we are interested in the effect of stochastic errors, our work is also related to the stochastic (sub)gradient methods [3], [10], [11].

The novelty of our work is in several directions. First, our gossip-based asynchronous algorithms allow the agents to use the stepsize based on the number of their local updates; thus *the stepsize is not coordinated among the agents*. Second, we study the convergence of the algorithm when *the functions are non-convex*, which is unlike the recent trend in the distributed network optimization where typically convex functions[1] are considered (see e.g., [16], [18], [21], [22], [25], [27], [28]). Third, we are dealing with the general case where the agents compute their *(sub)gradients with stochastic errors*. Due to agent information exchange, the stochastic errors propagate across agents and time, which together with the stochastic nature of the agent stepsizes, highly complicates the convergence analysis. Our analysis combines the ideas used to study the basic gossip-averaging algorithm [7] with the tools that are generally used to study the convergence of the stochastic gradient schemes.

The rest of the paper is organized in the following manner. In the next section, we describe the problem of our interest,

This research is supported by a Vodafone Graduate Fellowship, and by NSF Awards CNS-0831670 and CMMI-0742538.

[1]There are papers that discuss the convergence of cyclic incremental algorithms when the functions are nonconvex (e.g., [29]). However, cyclic incremental algorithms are not distributed since the agents have to be organized in a cycle by a central coordinator.

present our algorithm and assumptions. In Section III, among other preliminaries, we investigate the asymptotic properties of the agent disagreements. In Section IV, the convergence properties of the algorithm are studied. We conclude with a discussion in Section V.

II. Problem, algorithm and assumptions

We consider a network of m agents that are indexed by $1, \ldots, m$; when convenient, we will use $V = \{1, \ldots, m\}$. The network has a static topology that is represented by the bidirectional graph (V, E), where E is the set of links in the network. We have $\{i, j\} \in E$ if agent i and agent j can communicate with each other. We assume that the network [i.e., the graph (V, E)] is connected. The network objective is to solve the following optimization problem [2]:

$$\begin{aligned} &\text{minimize} \quad f(x) := \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in \Re, \end{aligned} \tag{1}$$

where $f_i : \Re \to \Re$ for all i. The function f_i is only known to agent i that can compute the gradient $\nabla f_i(x)$ with stochastic errors [3] The goal is to solve problem (1) using an algorithm that is distributed and local.

A. Asynchronous Gossip Optimization Algorithm

Let $N(i)$ be the set of neighbors of agent i, i.e. $N(i) = \{j \in V : \{i, j\} \in E\}$. Each agent has a local clock that ticks at a Poisson rate[4] of 1. At each tick of its clock, agent i averages its iterate with a randomly selected neighbor $j \in N(i)$, where each neighbor has an equal chance of being selected. Agents i and j then adjust their averages along the negative direction of ∇f_i and ∇f_j, respectively, which are computed with stochastic errors.

As in [7] we will find it easier to study the gossip algorithms in terms of a single virtual clock that ticks whenever any of the local Poisson clock ticks. Thus, the virtual clock ticks according to a Poisson process with rate m. Let Z_k denote the k-th tick of the virtual clock and let I_k denote the index of the agent whose local clock actually ticked at that instant. The fact that the Poisson clocks at each agent are independent imply that I_k is uniformly distributed in the set V. In addition, the memoryless property of the Poisson arrival process ensure that the process $\{I_k\}$ is i.i.d. Let J_k denote the random index of the agent communicating with agent I_k. Observe that J_k, conditioned on I_k, is uniformly distributed in the set $N(I_k)$. Let $x_{i,k-1}$ denote agent i iterate at time immediately before Z_k. The iterates evolve according to

$$x_{i,k} = \begin{cases} \bar{x}_{I_k,J_k} - \frac{1}{\Gamma_k(i)} \left(\nabla f_i \left(\bar{x}_{I_k,J_k} \right) + \epsilon_{i,k} \right) & \text{if } i \in \{I_k, J_k\} \\ x_{i,k-1} & \text{otherwise,} \end{cases} \tag{2}$$

where $x_{i,0},\ i \in V$ are initial iterates of the agents,

$$\bar{x}_{I_k,J_k} = \frac{1}{2} \left(x_{I_k,k-1} + x_{J_k,k-1} \right),$$

$\nabla f_i(x)$ denotes the gradient[5] of f_i at x, $\epsilon_{i,k}$ is the stochastic error and $\Gamma_k(i)$ denotes the total number of agent i updates up to the time Z_k.

B. Assumptions

We make the following assumption on the functions.

Assumption 1: The gradients are uniformly bounded, i.e., $\sup_{x \in \Re} |\nabla f_i(x)| \leq C$ for some $C > 0$ and for all $i \in V$.

In addition to this, we will use two complimentary sets of assumptions on the functions f_i, as discussed later.

Let F_{k-1} be the σ-algebra generated by the entire history of the algorithm up to time Z_k, i.e., $F_{k-1} = \{I_\ell, J_\ell, \epsilon_{I_\ell,\ell}, \epsilon_{J_\ell,\ell}; 0 \leq \ell \leq k-1\}$. We make the following assumptions on the stochastic errors.

Assumption 2: With probability 1, we have:

(a) $\mathsf{E}\left[|\epsilon_{i,k}|^2 \mid F_{k-1}\right] \leq \nu^2$ for all k and $i \in V$, and some ν.

(b) $\mathsf{E}[\epsilon_{I_k,k} \mid F_{k-1}, I_k, J_k] = 0$, $\mathsf{E}[\epsilon_{J_k,k} \mid F_{k-1}, I_k, J_k] = 0$.

The assumption is satisfied, for example, when the errors are zero mean, independent across time and have bounded second moments.

III. Preliminaries

All vectors are column vectors, x_i to denotes the i-th component of a vector x, and $\|x\|$ denotes the Euclidean norm of a vector x. We use $\mathbb{1}$ to denote a vector with all components equal to 1. In our analysis, we frequently invoke the following result due to Robbins and Siegmund (see Lemma 11, Chapter 2.2, [23]).

Lemma 1: Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots$ be a sequence of sub σ-fields of $\mathcal{F}$. Let $\{u_k\}$, $\{v_k\}$, $\{q_k\}$ and $\{w_k\}$ be $\mathcal{F}_k$-measurable random variables, where $\{u_k\}$ is uniformly bounded below, and $\{v_k\}$, $\{q_k\}$ and $\{w_k\}$ are non-negative. Let $\sum_{k=0}^{\infty} w_k < \infty$, $\sum_{k=0}^{\infty} q_k < \infty$ and

$$\mathsf{E}[u_{k+1} \mid \mathcal{F}_k] \leq (1 + q_k) u_k - v_k + w_k$$

hold with probability 1. Then, with probability 1, the sequence $\{u_k\}$ converges and $\sum_{k=0}^{\infty} v_k < \infty$.

A. Relative Frequency of Agents Updates

We characterize the number $\Gamma_i(k)$ of times agent i updates its iterate until time Z_k inclusively (see Eq. (2)). Define the event $E_{i,k} = \{I_k = i\} \cup \{J_k = i\}$. This is essentially the event that agent i updates its iterate at time Z_k. It is easy to see that $\{E_{i,k}\}$ are independent events with the same (time invariant) probability distribution. Define γ_i to be the probability of event $E_{i,k}$. Since I_k is uniformly distributed on the set V and J_k, conditioned on $I_k = j$, is uniformly distributed on the set $N(j)$, it follows that $\gamma_i = \frac{1}{m} \left(1 + \sum_{j \in N(i)} \frac{1}{|N(j)|} \right)$.

[2] By componentwise application, our results and proofs can be extended to the case when x is a finite-dimension vector.

[3] See [26] for the motivation for studying stochastic errors.

[4] The model and the analysis can be easily extended to the case when the clocks have different rates.

[5] If the function is not differentiable but convex then $\nabla f_i(x)$ denotes a subgradient. We will discuss this later.

Define χ_A to be the indicator function of an event A, and note that $\Gamma_i(k) = \sum_{\ell=1}^{k} \chi_{E_{i,k}}$. Since the events $\{\chi_{E_{i,k}}\}$ are i.i.d., from the law of iterated logarithms [9], we can conclude that for any $p, q > 0$, with probability 1,

$$\lim_{k\to\infty} \frac{|\Gamma_i(k) - k\gamma_i|}{k^{\frac{1}{2}+q}} \leq p \qquad \text{for all } i \in V.$$

We can therefore conclude that with probability 1, for all $i \in V$ and for all sufficiently large k,

$$\frac{1}{\Gamma_i(k)} \leq \frac{m}{k}, \tag{3}$$

$$\left|\frac{1}{\Gamma_i(k)} - \frac{1}{\gamma_i k}\right| \leq \frac{p}{k^{\frac{3}{2}+q}}. \tag{4}$$

B. Alternative Representation of the Algorithm

We next give the algorithm (2) in a more convenient form for our analysis. Let e_i denote the unit vector with only its i-th component being non-zero. Define

$$W_k = I - \frac{1}{2}(e_{I_k} - e_{J_k})^\top (e_{I_k} - e_{J_k}).$$

Since $\{I_k\}$, $\{J_k\}$ are i.i.d. sequences, $\{W_k\}$ is also an i.i.d. sequence. Define $\bar{W} = \mathsf{E}[W_k]$. Since each W_k is symmetric and doubly stochastic with probability 1, $\bar{W}$ is also symmetric and doubly stochastic. Further, the maximum eigenvalue of $\bar{W}$ is 1, and 1 is not a repeated eigenvalue when the network is connected[6]. We also have $\mathsf{E}\big[W_k^2\big] = \bar{W}$ (see [7]).

Let x_k be the vector with components $x_{i,k}$, $i = 1, \dots, m$. Then, from the definition of the method in (2), we have

$$x_k = W_k x_{k-1} + p_k \qquad \text{for } k \geq 1, \tag{5}$$

where

$$p_k = -\sum_{i\in\{I_k,J_k\}} \frac{1}{\Gamma_k(i)} \left(\nabla f_i\left(\bar{x}_{I_k,J_k}\right) + \epsilon_{i,k}\right) e_i,$$

and $x_{I_k,J_k} = (x_{I_k,k-1} + x_{J_k,k-1})/2$. Define $y_k = \frac{\mathbb{1}^\top x_k}{m}$. We then have

$$y_k = \frac{\mathbb{1}^\top x_k}{m} = \frac{\mathbb{1}^\top \left(W_k x_{k-1} + p_k\right)}{m}.$$

By the doubly stochasticity of W_k, with probability 1, it follows

$$y_k = \frac{\mathbb{1}^\top x_{k-1} + \mathbb{1}^\top p_k}{m} = y_{k-1} + \frac{\mathbb{1}^\top p_k}{m}. \tag{6}$$

C. Agent Consensus

We use $\|x_k - y_k\mathbb{1}\|$ to quantify the disagreement between the agents, and we show that the disagreements converge to 0.

Lemma 2: Let Assumptions 1 and 2(a) hold. Then, with probability 1, we have $\sum_{k=1}^{\infty} \frac{\|x_k - y_k\mathbb{1}\|}{k} < \infty$ and $\lim_{k\to\infty} \|x_k - y_k\mathbb{1}\| = 0$.

[6]In this case, $\bar{W}$ is a stochastic irreducible matrix and $\lambda = 1$ is its largest real eigenvalue with a unique right eigenvector, see e.g. [13], Corollary 3, page 116.

Proof: From (5) and (6) it follows

$$\begin{aligned}
&\mathsf{E}[\|x_k - y_k\mathbb{1}\| \mid F_{k-1}] \\
&= \mathsf{E}\left[\left\|W_k x_{k-1} + p_k - y_{k-1}\mathbb{1} - \frac{\mathbb{1}^\top p_k}{m}\mathbb{1}\right\| \mid F_{k-1}\right] \\
&\leq \mathsf{E}[\|W_k(x_{k-1} - y_{k-1}\mathbb{1})\| \mid F_{k-1}] + 2\mathsf{E}[\|p_k\| \mid F_{k-1}],
\end{aligned} \tag{7}$$

where the inequality follows from the triangle inequality of norms and the doubly stochasticity of W_k. The first term can be estimated using the relation $\mathsf{E}\big[W_k^\top W_k\big] = \mathsf{E}[W_k] = \bar{W}$ (implying that $\bar{W}$ is positive semi-definite) as follows:

$$\begin{aligned}
&\mathsf{E}\big[\|W_k(x_{k-1} - y_{k-1}\mathbb{1})\|^2 \mid F_{k-1}\big] \\
&= (x_{k-1} - y_{k-1}\mathbb{1})^\top \mathsf{E}\big[W_k^\top W_k\big] (x_{k-1} - y_{k-1}\mathbb{1}) \\
&= (x_{k-1} - y_{k-1}\mathbb{1})^\top \bar{W} (x_{k-1} - y_{k-1}\mathbb{1}) \\
&= \sum_{i=1}^{m} \lambda_i (v_i^\top (x_{k-1} - y_{k-1}\mathbb{1}))^2,
\end{aligned}$$

where λ_i is the i-th largest eigenvalue and v_i is the corresponding eigenvector of $\bar{W}$. The last step follows from the eigenvector decomposition of the symmetric positive semi-definite matrix $\bar{W}$. Recall that $\lambda_1 = 1$ (the largest value of $\bar{W}$) and the corresponding eigenvector is $\mathbb{1}$. Hence,

$$\mathsf{E}\big[\|W_k(x_{k-1} - y_{k-1}\mathbb{1})\|^2 \mid F_{k-1}\big] \leq \lambda_2 \|x_{k-1} - y_{k-1}\mathbb{1}\|^2. \tag{8}$$

We next estimate the second term in (7). Using (3) and the boundedness of the gradients (Assumption 1), we can conclude that for sufficiently large k, we have

$$\begin{aligned}
&\mathsf{E}\big[\|p_k\|^2 \mid F_{k-1}\big] \\
&\leq 2\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \frac{1}{\Gamma_i(k)^2} \left|\nabla f_i\left(\bar{x}_{I_k,J_k}\right) + \epsilon_{i,k}\right|^2 \mid F_{k-1}\right] \\
&\leq 4\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \frac{1}{\Gamma_i(k)^2} \left(\left(\nabla f_i\left(\bar{x}_{I_k,J_k}\right)\right)^2 + \epsilon_{i,k}^2\right) \mid F_{k-1}\right] \\
&\leq \frac{4m^2(C+\nu)^2}{k^2}
\end{aligned} \tag{9}$$

From (7), (8), (9), and the Jensen's inequality we can see that

$$\mathsf{E}[\|x_k - y_k\mathbb{1}\|] \leq \sqrt{\lambda_2}\mathsf{E}[\|x_{k-1} - y_{k-1}\mathbb{1}\|] + \frac{4m(C+\nu)}{k},$$

where $\lambda_2 < 1$. Therefore, we have for sufficiently large k,

$$\begin{aligned}
\frac{1}{k}\mathsf{E}[\|x_k - y_k\mathbb{1}\|] &\leq \frac{1}{k-1}\mathsf{E}[\|x_{k-1} - y_{k-1}\mathbb{1}\|] \\
&\quad - \frac{1-\sqrt{\lambda_2}}{k-1}\mathsf{E}[\|x_{k-1} - y_{k-1}\mathbb{1}\|] + \frac{4m(C+\nu)}{(k-1)^2}.
\end{aligned}$$

Using the deterministic analog of Lemma 1, we see that $\sum_k \frac{\mathsf{E}[\|x_{k-1} - y_{k-1}\mathbb{1}\|]}{k-1} < \infty$, which implies $\sum_k \frac{\|x_{k-1} - y_{k-1}\mathbb{1}\|}{k-1} < \infty$ with probability 1.

We next prove the second part of the statement. As a consequence of the preceding result, it follows $\liminf_{k\to\infty} \|x_{k-1} - y_{k-1}\mathbb{1}\| = 0$. We only need to prove almost sure convergence of $\|x_{k-1} - y_{k-1}\mathbb{1}\|$ to complete the proof. From the definitions of x_k and y_k in (5) and (6), we obtain

$$\mathsf{E}\big[\|x_k - y_k\mathbb{1}\|^2 \mid F_{k-1}\big]$$

$$
\begin{aligned}
&= \mathsf{E}\left[\left\|W_k x_{k-1} + p_k - y_{k-1}\mathbb{1} - \frac{\mathbb{1}^\top p_k}{m}\mathbb{1}\right\|^2 \mid F_{k-1}\right] \\
&\leq \mathsf{E}\left[\|W_k(x_{k-1} - y_{k-1}\mathbb{1})\|^2 \mid F_{k-1}\right] \\
&\quad +2\sqrt{\mathsf{E}[\|W_k(x_{k-1} - y_{k-1}\mathbb{1})\|^2 \mid F_{k-1}]} \\
&\quad \times \sqrt{\mathsf{E}\left[\left\|p_k - \frac{\mathbb{1}^\top p_k}{m}\mathbb{1}\right\|^2 \mid F_{k-1}\right]} \\
&\quad +\mathsf{E}\left[\left\|p_k - \frac{\mathbb{1}^\top p_k}{m}\mathbb{1}\right\|^2 \mid F_{k-1}\right], \qquad (10)
\end{aligned}
$$

where in the last step we use Cauchy-Schwartz inequality. We next estimate the last term in (10), as follows

$$
\begin{aligned}
&\mathsf{E}\left[\left\|p_k - \frac{\mathbb{1}^\top p_k}{m}\mathbb{1}\right\|^2 \mid F_{k-1}\right] \\
&\leq 2\mathsf{E}\left[\|p_k\|^2 \mid F_{k-1}\right] + 2\mathsf{E}\left[\left\|\frac{\mathbb{1}^\top p_k}{m}\mathbb{1}\right\|^2 \mid F_{k-1}\right] \\
&\leq 2\left(1 + \frac{2}{m}\right)\mathsf{E}\left[\|p_k\|^2 \mid F_{k-1}\right] \\
&\leq 4\mathsf{E}\left[\|p_k\|^2 \mid F_{k-1}\right].
\end{aligned}
$$

In the last step we use the fact that only two components of p_k are non-zero. Using this in (10), substituting from (8) and (9), and taking into account $\lambda_2 < 1$, we obtain

$$
\begin{aligned}
&\mathsf{E}\left[\|x_k - y_k\mathbb{1}\|^2 \mid F_{k-1}\right] \leq \|x_{k-1} - y_{k-1}\mathbb{1}\|^2 \\
&+8m\sqrt{\lambda_2}(C+\nu)\frac{\|x_{k-1} - y_{k-1}\mathbb{1}\|}{k} + \frac{16m^2(C+\nu)^2}{k^2}.
\end{aligned}
$$

As shown earlier, we have $\sum_k \frac{\|x_{k-1} - y_{k-1}\mathbb{1}\|}{k} < \infty$ with probability 1. We can invoke Lemma 1 to conclude that $\|x_k - y_k\mathbb{1}\|$ converges with probability 1. ■

IV. Convergence Analysis

We here study the convergence of the algorithms under two different sets of conditions. The first requires the function to be differentiable with Lipschitz continuous gradient, i.e.,

$$
\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \qquad (11)
$$

A point $x^* \in \Re$ is a stationary point of $f(x)$ if $\nabla f(x^*) = 0$. A global minimum of $f(x)$ is also a stationary point of $f(x)$. Typically, when the objective function is non-convex and iterative methods are employed, the iterates may converge to a stationary point.

Theorem 1: Let Assumptions 1 and 2 hold, and let the function $f(x)$ be bounded below with Lipschitz derivatives. Then, with probability 1, we have $\lim_{k\to\infty} |x_{i,k} - y_k| = 0$ for all $i \in V$, $\{f(x_{i,k})\}$ converges, and $\liminf_{k\to\infty} \nabla f(x_{i,k}) = 0$.

Proof: Lemma 2 asserts that $\lim_{k\to\infty} |x_{i,k} - y_k| = 0$. Next, from the definition of p_k in (5) we obtain

$$
\begin{aligned}
p_k^\top \mathbb{1} &= -\sum_{i\in\{I_k,J_k\}} \frac{1}{\Gamma_k(i)}\left(\nabla f_i\left(\bar{x}_{I_k,J_k}\right) + \epsilon_{i,k}\right) \\
&= -\sum_{i\in\{I_k,J_k\}} \frac{1}{k\gamma_i}\nabla f_i\left(y_{k-1}\right) - \sum_{i\in\{I_k,J_k\}} \frac{\epsilon_{i,k}}{\Gamma_k(i)} \\
&\quad -\left(\sum_{i\in\{I_k,J_k\}} \frac{1}{\Gamma_k(i)}\nabla f_i\left(\bar{x}_{I_k,J_k}\right)\right. \\
&\qquad \left. -\sum_{i\in\{I_k,J_k\}} \frac{1}{\Gamma_k(i)}\nabla f_i\left(y_{k-1}\right)\right) \\
&\quad -\sum_{i\in\{I_k,J_k\}} \left(\frac{1}{\Gamma_k(i)} - \frac{1}{k\gamma_i}\right)\nabla f_i\left(y_{k-1}\right).
\end{aligned}
$$

Taking conditional expectations, and using (3), (11) and the boundedness of the gradient we obtain

$$
\begin{aligned}
&\left|\mathsf{E}\left[p_k^\top \mid F_{k-1}\right]\mathbb{1} + \frac{\nabla f(y_{k-1})}{k}\right| \\
&\leq \left|\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \frac{\nabla f_i\left(y_{k-1}\right)}{k\gamma_i} \mid F_{k-1}\right] - \frac{\nabla f(y_{k-1})}{k}\right| \\
&+\sum_{i=1}^m \frac{m\left|\mathsf{E}[\epsilon_{i,k} \mid F_{k-1}]\right|}{k} + \sum_{i=1}^m \frac{mL}{k}\mathsf{E}[|x_{i,k-1} - y_{k-1}| \mid F_{k-1}] \\
&+C\sum_{i=1}^m \mathsf{E}\left[\left|\frac{1}{\Gamma_k(i)} - \frac{1}{k\gamma_i}\right| \mid F_{k-1}\right]. \qquad (12)
\end{aligned}
$$

Since γ_i is the probability that agent i updates at time Z_k it follows that

$$
\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \frac{1}{\gamma_i}\nabla f_i\left(y_{k-1}\right) \mid F_{k-1}\right] = \nabla f(y_{k-1}),
$$

so that the first term in (12) is equal to 0. Using Assumption 2, we can see that the second term is 0. Further, note from (4) and Lemma 2 it follows that the last three terms in (12) are summable. Thus, from (6) we obtain

$$
y_k = y_{k-1} - \frac{\nabla f(y_{k-1})}{k} + a_k,
$$

where $\sum_k |\mathsf{E}[a_k \mid F_{k-1}]| < \infty$. Additionally, from Assumption 2(a), Lemma 2 (showing that $|x_{i,k-1} - y_{k-1}| \to 0$) and relation (4) it follows that $\mathsf{E}\left[|a_k|^2 \mid F_{k-1}\right] < \infty$. The result now follows from classic stochastic optimization theory (see [3], or Chapter 2 of [6]). ■

Observe that, in view of Lipschitz continuity of the gradient, the assumption that the gradients are bounded is equivalent to the following standard assumption.

Assumption 3: The sequences $\{x_{i,k}\}$, $i \in V$, are bounded with probability 1.

This assumption is implicit and not very easy to establish. We refer the reader to Chapter 3 of [6] for some discussions on techniques to verify this assumption.

We will next investigate the convergence when the functions are convex, but not necessarily differentiable. At points where the gradient does not exist, we use the notion of subgradient. A vector $\nabla g(x)$ is a *subgradient* of a function g at a point $x \in \operatorname{dom} g$ if the following relation holds

$$\nabla g(x)^\top (y-x) \leq g(y) - g(x) \qquad \text{for all } y \in \operatorname{dom} g. \quad (13)$$

We next discuss the convergence of the algorithms.

Theorem 2: Let Assumptions 1 and 2 hold. Assume that $X^* = \mathsf{Argmin}_{x\in\Re} f(x)$ is non-empty, and $f_i(x)$ is convex for each $i \in V$. Then, with probability 1, the sequences $\{x_{i,k}\}$, $i \in V$, converge to the same point in X^*.

Proof: Let x^* be an arbitrary point in X^*. Using (6) we obtain

$$\begin{aligned}|y_k - x^*|^2 =& |y_{k-1} + \frac{p_k^\top \mathbb{1}}{m} - x^*|^2 \leq |y_{k-1} - x^*|^2 \\ &+ \frac{2(p_k^\top \mathbb{1})(y_{k-1}-x^*)}{m} + \frac{(p_k^\top \mathbb{1})^2}{m^2} \\ \leq& |y_{k-1} - x^*|^2 + \frac{2(p_k^\top \mathbb{1})(\frac{x_{I_k,k-1}+x_{J_k,k-1}}{2} - x^*)}{m} \\ &+ \frac{2|p_k^\top \mathbb{1}| \sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{m} + \frac{2\|p_k\|^2}{m^2}.\end{aligned}$$

From the definition of p_k in (5) and the subgradient inequality in (13) we can write

$$\begin{aligned}|y_k - x^*|^2 \leq& |y_{k-1} - x^*|^2 - 2\sum_{i\in\{I_k,J_k\}} \frac{f_i\left(\bar{x}_{I_k,J_k}\right) - f_i(x^*)}{m\Gamma_i(k)} \\ &+ \frac{2\left(\epsilon_{I_k,k} + \epsilon_{J_k,k}\right)\left(\frac{x_{I_k,k-1}+x_{J_k,k-1}}{2} - x^*\right)}{m\Gamma_i(k)} \\ &+ \frac{2|p_k^\top \mathbb{1}| \sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{m} + \frac{2\|p_k\|^2}{m^2} \\ \leq& |y_{k-1} - x^*|^2 - 2\sum_{i\in\{I_k,J_k\}} \frac{f_i(y_{k-1}) - f_i(x^*)}{m\Gamma_i(k)} \\ &+ \frac{2\left(\epsilon_{I_k,k} + \epsilon_{J_k,k}\right)\left(\frac{x_{I_k,k-1}+x_{J_k,k-1}}{2} - x^*\right)}{m\Gamma_i(k)} \\ &+ 2\sum_{i\in\{I_k,J_k\}} \frac{f_i\left(\bar{x}_{I_k,J_k}\right) - f_i(y_{k-1})}{m\Gamma_i(k)} \\ &+ \frac{2|p_k^\top \mathbb{1}| \sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{m} + \frac{2\|p_k\|^2}{m^2}.\end{aligned}$$

Using the subgradient inequality (13) and subgradient boundedness (Assumption 1) to bound the fourth term, we get

$$\begin{aligned}|y_k - x^*|^2 \leq& |y_{k-1} - x^*|^2 - 2\sum_{i\in\{I_k,J_k\}} \frac{f_i(y_{k-1}) - f_i(x^*)}{m\Gamma_i(k)} \\ &+ \frac{2\left(\epsilon_{I_k,k} + \epsilon_{J_k,k}\right)\left(\frac{x_{I_k,k-1}+x_{J_k,k-1}}{2} - x^*\right)}{m\Gamma_i(k)} \\ &+ 2C\sum_{i=1}^m \frac{|y_{k-1} - x_{i,k-1}|}{m\Gamma_i(k)} \\ &+ \frac{2|p_k| \sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{m} + \frac{2\|p_k\|^2}{m^2}.\end{aligned}$$

Taking conditional expectations and using (3), we obtain

$$\begin{aligned}&\mathsf{E}\left[|y_k - x^*|^2 \mid F_{k-1}\right] \\ &\leq |y_{k-1} - x^*|^2 - 2\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \frac{f_i(y_{k-1}) - f_i(x^*)}{m\Gamma_i(k)} \mid F_{k-1}\right] \\ &+ \mathsf{E}\left[\frac{2\left(\epsilon_{I_k,k} + \epsilon_{J_k,k}\right)\left(\frac{x_{I_k,k-1}+x_{J_k,k-1}}{2} - x^*\right)}{m\Gamma_i(k)} \mid F_{k-1}\right] \\ &+ \frac{2C}{k}\sum_{i=1}^m |y_{k-1} - x_{i,k-1}| \\ &+ \frac{2\mathsf{E}[|p_k| \mid F_{k-1}] \sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{m} + \frac{2\mathsf{E}\left[\|p_k\|^2 \mid F_{k-1}\right]}{m^2}.\end{aligned}$$

Using the bounds in (9), we obtain for sufficiently large k,

$$\begin{aligned}&\mathsf{E}\left[|y_k - x^*|^2 \mid F_{k-1}\right] \\ &\leq |y_{k-1} - x^*|^2 - 2\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \frac{f_i(y_{k-1}) - f_i(x^*)}{m\Gamma_i(k)} \mid F_{k-1}\right] \\ &+ \mathsf{E}\left[\frac{2\left(\epsilon_{I_k,k} + \epsilon_{J_k,k}\right)\left(\frac{x_{I_k,k-1}+x_{J_k,k-1}}{2} - x^*\right)}{m\Gamma_i(k)} \mid F_{k-1}\right] \\ &+ \frac{2C}{k}\sum_{i=1}^m |y_{k-1} - x_{i,k-1}| \\ &+ \frac{4(C+\nu)\sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{k} + \frac{8(C+\nu)^2}{k^2} \\ &\leq |y_{k-1} - x^*|^2 - 2\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \frac{f_i(y_{k-1}) - f_i(x^*)}{m\gamma_i k} \mid F_{k-1}\right] \\ &+ \mathsf{E}\left[\frac{2\left(\epsilon_{I_k,k} + \epsilon_{J_k,k}\right)\left(\frac{x_{I_k,k-1}+x_{J_k,k-1}}{2} - x^*\right)}{m\gamma_i k} \mid F_{k-1}\right] \\ &+ \mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} |f_i(y_{k-1}) - f_i(x^*)| \left|\frac{1}{m\gamma_i k} - \frac{1}{m\Gamma_i(k)}\right| \mid F_{k-1}\right] \\ &+ \frac{(6C+4\nu)\sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{k} + \frac{8(C+\nu)^2}{k^2}.\end{aligned}$$

Note from Assumption 2(b) that the third term is 0. Since γ_i is the probability that agent i updates at time Z_k, we have

$$\begin{aligned}&\mathsf{E}\left[|y_k - x^*|^2 \mid F_{k-1}\right] \\ &\leq |y_{k-1} - x^*|^2 - 2\frac{f(y_{k-1}) - f(x^*)}{mk} \\ &+ 2\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} |f_i(y_{k-1}) - f_i(x^*)| \left|\frac{1}{m\gamma_i k} - \frac{1}{m\Gamma_i(k)}\right| \mid F_{k-1}\right] \\ &+ \frac{(6C+4\nu)\sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{k} + \frac{8(C+\nu)^2}{k^2}.\end{aligned}$$

Using the subgradient inequality (13) and the inequality $2a < 1 + a^2$, we can bound the third term as follows

$$\begin{aligned}&\sum_{i\in\{I_k,J_k\}} |f_i(y_{k-1}) - f_i(x^*)| \left|\frac{1}{m\gamma_i k} - \frac{1}{m\Gamma_i(k)}\right| \\ &\leq 2C|y_{k-1} - x^*| \left|\frac{1}{m\gamma_i k} - \frac{1}{m\Gamma_i(k)}\right| \\ &\leq C\left|\frac{1}{m\gamma_i k} - \frac{1}{m\Gamma_i(k)}\right| (1 + |y_{k-1} - x^*|^2).\end{aligned}$$

Combining the two preceding relations we obtain

$$\begin{aligned}
&\mathsf{E}\left[|y_k - x^*|^2 \mid F_{k-1}\right] \\
&\le \left(1 + 2C\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \left|\frac{1}{m\gamma_i k} - \frac{1}{m\Gamma_i(k)}\right| \mid F_{k-1}\right]\right) \\
&\quad \times |y_{k-1} - x^*|^2 - \frac{2(f(y_{k-1}) - f(x^*))}{mk} \\
&\quad + 2C\mathsf{E}\left[\sum_{i\in\{I_k,J_k\}} \left|\frac{1}{m\gamma_i k} - \frac{1}{m\Gamma_i(k)}\right| \mid F_{k-1}\right] \\
&\quad + \frac{(6C + 4\nu)\sum_{i=1}^m |y_{k-1} - x_{i,k-1}|}{k} + \frac{8(C+\nu)^2}{k^2}.
\end{aligned}$$

Using (4), we can see that the conditions of Lemma 1 are satisfied. Therefore $\{|y_k - x^*|\}$ converges and $\sum_k \frac{f(y_{k-1})-f(x^*)}{k} < \infty$ with probability 1, which implies that $\{y_k\}$ converges to a point in the set X^* with probability 1. This and the fact $\lim_{k\to\infty} |x_{i,k} - y_k| = 0$ for all $i \in V$, with probability 1, (shown in Lemma 2) imply that $\{x_{i,k}\}$ converge to the same point in X^*, with probability 1. ■

V. Discussion

Using very similar ideas the algorithm and the proof of convergence can be extended to the case when x is a finite dimensional vector. When the problem in (1) is a constrained optimization problem where x is restricted to a convex and closed set X, then the algorithm in (2) can be extended by projecting onto the set X at each iteration. It is easy to obtain a convergence result similar to Theorem 2 for this case using Euclidean projection inequalities. As a part of our future work, we plan to investigate optimization algorithms based on different gossip schemes.

References

[1] T. Aysal, M. Yildiz, A. Sarwate, and A. Scaglione, *Broadcast gossip algorithms: Design and analysis for consensus*, Proceedings of the 47th IEEE Conference on Decision and Control, 2008.

[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), no. 3, 627–642.

[4] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with constant stepsize*, SIAM Journal of Optimization **18** (2007), no. 1, 29–51.

[5] V. Borkar, *Asynchronous stochastic approximations*, SIAM Journal on Control and Optimization **36** (1998), no. 3, 840–851.

[6] ______, *Stochastic approximation: A dynamical viewpoint*, Cambridge University Press, 2008.

[7] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Randomized gossip algorithms*, IEEE Transactions on Information Theory **52** (2006), no. 6, 2508–2530.

[8] A. Dimakis, A. Sarwate, and M. Wainwright, *Geographic gossip: Efficient averaging for sensor networks*, IEEE Transactions on Signal Processing **56** (2008), no. 3, 1205–1216.

[9] R. Dudley, *Real analysis and probability*, Cambridge University Press, 2002.

[10] Y. Ermoliev, *Stochastic programming methods*, Nauka, Moscow, 1976.

[11] ______, *Stochastic quasi-gradient methods and their application to system optimization*, Stochastics **9** (1983), no. 1, 1–36.

[12] A. A. Gaivoronski, *Convergence properties of backpropogation for neural nets via theory of stochastic gradient methods. Part 1.*, Optimization Methods and Software **4** (1994), no. 2, 117–134.

[13] R. G.Gallager, *Discrete stochastic processes*, Kluwer Academic Publishers, Norwell, Massachusetts, USA, 1996.

[14] B. Johansson, *On distributed optimization in networked systems*, Ph.D. thesis, Royal Institute of Technology, Stockholm, Sweden, 2008.

[15] B. Johansson, M. Rabi, and M. Johansson, *A simple peer-to-peer algorithm for distributed optimization in sensor networks*, Proceedings of the 46th IEEE Conference on Decision and Control, 2007, pp. 4705–4710.

[16] B. Johanssson, T. Keviczsky, M. Johansson, and K. Johansson, *Subgradient methods and consensus algorithms for solving convex optimization problems*, Proceedings of the 47th IEEE Conference on Decision and Control, 2008, pp. 4185—4190.

[17] K. C. Kiwiel, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM Journal on Optimization **14** (2003), no. 3, 807–840.

[18] I. Lobel and A. Ozdaglar, *Distributed subgradient methods over random networks*, Lab. for Information and Decision Systems, MIT, Report 2800, 2008.

[19] A. Nedić and D. P. Bertsekas, *Incremental subgradient method for nondifferentiable optimization*, SIAM Journal of Optimization **12** (2001), 109–138.

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

[21] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *Distributed subgradient algorithms and quantization effects*, Proceedings of the 47th IEEE Conference on Decision and Control, 2008.

[22] A. Nedić and A. Ozdaglar, *On the rate of convergence of distributed asynchronous subgradient methods for multi-agent optimization*, Proceedings of the 46th IEEE Conference on Decision and Control, 2007, pp. 4711–4716.

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

[24] M. G. Rabbat and R. D. Nowak, *Quantized incremental algorithms for distributed optimization*, IEEE Journal on Select Areas in Communications **23** (2005), no. 4, 798–808.

[25] S. Sundhar Ram, A. Nedić, and V. V. Veeravalli, *Distributed stochastic subgradient algorithm for convex optimization*, Available at http://arxiv.org/abs/0811.2595, 2008.

[26] 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.

[27] S. Sundhar Ram, V. V. Veeravali, and A. Nedić, *Sensor networks: When theory meets practice*, ch. Distributed and recursive estimation, Springer, 2009.

[28] S. Sundhar Ram, V. V. Veeravalli, and A. Nedić, *Distributed and non-autonomous power control through distributed convex optimization*, IEEE INFOCOM, 2009.

[29] M. V. Solodov, *Incremental gradient algorithms with stepsizes bounded away from zero*, Computational Optimization and Algorithms **11** (1998), no. 1, 23–35.

[30] M. V. Solodov and S. K. Zavriev, *Error stability properties of generalized gradient-type algorithms*, Journal of Optimization Theory and Applications **98** (1998), no. 3, 663–680.

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

[32] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Transactions on Automatic Control **31** (1986), no. 9, 803–812.