SIAM J. OPTIM.
Vol. 20, No. 2, pp. 691–717
© 2009 Society for Industrial and Applied Mathematics

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

INCREMENTAL STOCHASTIC SUBGRADIENT ALGORITHMS FOR CONVEX OPTIMIZATION*

S. SUNDHAR RAM†, A. NEDIĆ‡, AND V. V. VEERAVALLI†

Abstract. This paper studies the effect of stochastic errors on two constrained incremental subgradient algorithms. The incremental subgradient algorithms are viewed as decentralized network optimization algorithms as applied to minimize a sum of functions, when each component function is known only to a particular agent of a distributed network. First, the standard cyclic incremental subgradient algorithm is studied. In this, the agents form a ring structure and pass the iterate in a cycle. When there are stochastic errors in the subgradient evaluations, sufficient conditions on the moments of the stochastic errors are obtained that guarantee almost sure convergence when a diminishing step-size is used. In addition, almost sure bounds on the algorithm's performance with a constant step-size are also obtained. Next, the Markov randomized incremental subgradient method is studied. This is a noncyclic version of the incremental algorithm where the sequence of computing agents is modeled as a time nonhomogeneous Markov chain. Such a model is appropriate for mobile networks, as the network topology changes across time in these networks. Convergence results and error bounds for the Markov randomized method in the presence of stochastic errors for diminishing and constant step-sizes are obtained.

Key words. incremental optimization, convex optimization, network optimization, stochastic approximation, subgradient, random networks

AMS subject classifications. 90C25, 68W15

DOI. 10.1137/080726380

1. Introduction.

A problem of recent interest in distributed networks is the design of decentralized algorithms to minimize a sum of functions, where each component function is known only to a particular agent [5, 11, 22, 26, 31]. Such problems arise in many network applications, including in-network estimation, learning, signal processing, and resource allocation [8, 18, 30–33]. In these applications, there is no central coordinator that has access to all the information and, thus, decentralized algorithms are needed to solve the problems. In this paper, we consider decentralized subgradient methods for constrained minimization of a sum of convex functions, where each component function is only known partially (with stochastic errors) to a specific network agent. We study two incremental subgradient methods with stochastic errors: a cyclic and a (noncyclic) Markov randomized incremental method.

The cyclic incremental algorithm is a decentralized method in which the network agents form a ring and process the information cyclically. The incremental method was originally proposed by Kibardin [13] and has been extensively studied more recently in [5, 9, 19, 22, 35]. Incremental gradient algorithms were first used for optimizing the weights in neural network training [9, 19, 35], and most of the associated literature deals with differentiable nonconvex unconstrained optimization problems [4, 5, 9, 35, 36]. The incremental subgradient algorithm for nondifferentiable constrained convex optimization has been investigated in [21, 22] without errors, and in [15, 23, 31,

*Received by the editors June 5, 2008; accepted for publication (in revised form) February 12, 2009; published electronically June 10, 2009. This work is partially supported by Vodafone and the National Science Foundation under CAREER grant CMMI 07-42538.
http://www.siam.org/journals/siopt/20-2/72638.html
†Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (ssriniv5@illinois.edu, vvv@illinois.edu).
‡Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (angelia@illinois.edu).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

36] where the effects of deterministic errors are considered. The algorithm that we consider in this paper is stochastic and as such differs from the existing literature. For this algorithm, we establish convergence for diminishing step-size and provide an error bound for constant step-size.

The Markov randomized incremental algorithm is a decentralized method where the iterates are generated incrementally within the network by passing them from agent to agent. Unlike the cyclic incremental method, where the agent network has a ring structure and the information flow is along this ring (cycle), in the Markov randomized incremental method, the network can have arbitrary structure. However, similar to cyclic incremental method, in the Markov randomized incremental method, only one agent updates at any given time. In particular, an agent in the network updates the current iterate (by processing locally its own objective function), and either passes the new iterate to a randomly chosen neighbor, or processes it again. Thus the order in which the agents update the iterates is random. This class of incremental algorithms was first proposed in [22], where the agent that receives the iterate is chosen with uniform probability in each iteration (corresponding to the case of a fully connected agent network). Recently, this idea has been extended in [11] to the case where the sequence in which the agents process the information is a time *homogeneous Markov chain*. The rationale behind this model is that the agent updating the information at a given time is more likely to pass this information to a close neighbor rather than to an agent who is further away. In this paper, we consider a more general framework than that of [11] by allowing the sequence in which the agents process the information to be a time *nonhomogeneous Markov chain*.[1] We prove the algorithm convergence for diminishing step-size and establish an error bound for a constant step-size. This extends the results in [11], where an error bound for a homogeneous Markov randomized incremental subgradient method is discussed for a constant step-size and error-free case.

The Markov randomized incremental algorithm is also related to the decentralized computation model in [3, 6, 40] for stochastic optimization problems. However, the emphasis in these studies is on parallel processing where each agent completely knows the entire objective function to be minimized. More closely related is the work in studies in [24] that develops a "parallel" version of the unconstrained incremental subgradient algorithm. Also related is the constrained consensus problem studied in [27] where agents are interested in obtaining a solution to a feasibility problem, when different parts of the problem are known to different agents. At a much broader scale, the paper is also related to the literature on distributed averaging and consensus algorithms [3, 10, 12, 24, 27, 28, 37, 38, 40, 41].

Our main contributions in this paper are the development and analysis of the Markov randomized incremental method with stochastic subgradients and the use of a time nonhomogeneous Markov model for the sequence of computing agents. In addition, to the best of our knowledge, this is among the few attempts made at studying the effects of stochastic errors on the performance of decentralized optimization algorithms. The other studies are [17, 38, 39], but the algorithms considered are fundamentally different from the incremental algorithms studied in this paper.[2]

[1] The primary motivations for such models are mobile networks where the network connectivity structure is changing in time, and thus the set of the neighbors of an agent is time-varying.

[2] In that work, the components of the decision vector are distributed while the objective function is known to all agents. In contrast, in this paper, the objective function data are distributed, while each agent has an estimate of the entire decision vector.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

The paper is organized as follows. In section 2, we formulate the problem of interest, and introduce the cyclic incremental and Markov randomized incremental method with stochastic errors. We also discuss some applications that motivate our interest in these methods. In section 3, we analyze convergence properties of the cyclic incremental method. We establish convergence of the method under diminishing step-size and provide an error bound for the method with a constant step-size, both valid with probability 1. We establish analogous results for the Markov randomized incremental method in section 4. We give some concluding remarks in section 5.

2. Problem formulation and motivation. We consider a network of m agents, indexed by $i = 1, \ldots, m$. The network objective is to solve the following problem:

$$\begin{aligned} &\text{minimize} && f(x) = \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} && x \in X, \end{aligned} \tag{2.1}$$

where $x \in \Re^n$ is a decision or a parameter vector, X is a closed and convex subset of $\Re^n$, and each f_i is a convex function from $\Re^n$ to $\Re$ that is known only to agent i. Problems with the above structure arise in the context of estimation in sensor networks [31, 32], where x is an unknown parameter to be estimated and f_i is the cost function that is determined by the ith sensor's observations (for example, f_i could be the log-likelihood function in maximum likelihood estimation). Furthermore, problems with such structure also arise in resource allocation in data networks. In this context, x is the resource vector to be allocated among m agents and f_i is the utility function for agent i [8]. We discuss these examples in more detail later.

To solve the problem (2.1) in a network where agents are connected in a directed ring structure, we consider the cyclic *incremental subgradient method* [22]. Time is slotted, and in each time slot, the estimate is passed by an agent to the next agent along the ring. In particular, agent i receives the iterate from agent $i-1$, and updates the received estimate using a subgradient of its own "objective function f_i". The updated iterate is then communicated to the next agent in the cycle, which is agent $i+1$ when $i < m$, and agent 1 when $i = m$. We are interested in the case where the agent subgradient evaluations have random errors. Formally, the algorithm is given as follows:

$$\begin{aligned} &z_{0,k+1} = z_{m,k} = x_k, \\ &z_{i,k+1} = \mathcal{P}_X \left[z_{i-1,k+1} - \alpha_{k+1} \left(\nabla f_i(z_{i-1,k+1}) + \epsilon_{i,k+1} \right) \right], \end{aligned} \tag{2.2}$$

where the initial iterate $x_0 \in X$ is chosen at random. The vector x_k is the estimate at the end of cycle k, $z_{i,k+1}$ is the intermediate estimate obtained after agent i updates in $k+1$st cycle, $\nabla f_i(x)$ is the subgradient of f_i evaluated at x, and $\epsilon_{i,k+1}$ is a random error. The scalar α_{k+1} is a positive step-size and $\mathcal{P}_X$ denotes Euclidean projection onto the set X. We study the convergence properties of method (2.2) in section 3 for diminishing and constant step-size.

In addition, for a network of agents with arbitrary connectivity, we consider an incremental algorithm where the agent that updates is selected randomly according to a distribution depending on the agent that performed the most recent update. Formally, in this method the iterates are generated according to the following rule:

$$x_{k+1} = \mathcal{P}_X \left[x_k - \alpha_{k+1} \left(\nabla f_{s(k+1)}(x_k) + \epsilon_{s(k+1),k+1} \right) \right], \tag{2.3}$$

where the initial iterate $x_0 \in X$ is chosen at random and the agent $s(0)$ that initializes the method is also selected at random. The integer $s(k+1)$ is the index of the agent

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

that performs the update at time $k+1$, and the sequence $\{s(k)\}$ is modeled as a time nonhomogeneous Markov chain with state space $\{1,\dots,m\}$. In particular, if agent i was processing at time k, then the agent j will be selected to perform the update at time $k+1$ with probability $[P(k)]_{i,j}$. Formally, we have

$$\mathsf{Prob}\left\{s(k+1)=j \mid s(k)=i\right\} = [P(k)]_{i,j}.$$

When there are no errors ($\epsilon_{s(k+1),k+1}=0$) and the probabilities $[P(k)]_{i,j}$ are all equal to $\frac{1}{m}$, the method in (2.3) coincides with the incremental method with randomization that was first proposed by Ermoliev [7] (see Chapter 3, section 5), and later studied in [22].

Following [11], we refer to the method in (2.3) as the *Markov randomized incremental stochastic* algorithm. We analyze convergence properties of this method in section 4 for diminishing and constant step-sizes.

2.1. Motivation. As mentioned, we study the convergence properties of the incremental algorithms (2.2) and (2.3) for diminishing and constant step-size, and for zero and nonzero mean errors. Such errors may arise directly as computational round-off errors, which are of interest when the entire network is on a single chip and each agent is a processor on the chip [20]. In addition, stochastic errors also arise in the following context.

Let the function $f_i(x)$ have the following form

$$f_i(x) = \mathsf{E}[g_i(x,R_i)]\,,$$

where $\mathsf{E}[\cdot]$ denotes the expectation, $R_i \in \Re^d$ is a random vector, and $g_i : \Re^{n\times d} \to \Re$. Agent i does not know the statistics of R_i, and thus does not know its complete objective function f_i. However, agent i sequentially observes independent samples of R_i and uses these samples to determine an approximate subgradient using the Robbins–Monro approximation [34] or Kiefer–Wolfowitz approximation [14]. These approximate sub-gradients can be considered to be the actual sub-gradient corrupted by stochastic errors.

We next discuss some specific problems that fall within the framework that we consider and can be solved using the proposed methods.

Distributed regression. Consider m sensors that sense a time invariant spatial field. Let $r_{i,k}$ be the measurement made by i^{th} sensor in time slot k. Let ℓ_i be the location of the i^{th} sensor. For each i, let $h(\ell_i,x)$ be a model set, i.e., set of candidate models, for the spatial field at ℓ_i that is parameterized by x and selected based on a priori information. Thus, for each x, $h(\ell_i,x)$ is a model for the measurement $r_{i,k}$. The problem in regression is to choose the best model among the set of candidate models based on the collected measurements $r_{i,k}$, i.e., to determine the value for x that best describes the spatial field. In least squares regression, the parameter value x^* corresponding to the best model satisfies the following relation:

$$x^* \in \operatorname*{Argmin}_{x\in X} \lim_{N\to\infty} \sum_{i=1}^{m} \frac{1}{N} \sum_{k=1}^{N} \left(r_{i,k} - h(\ell_i,x)\right)^2.$$

If for each i, $\{r_{i,k}\}_{k\in\mathbb{N}}$ can be viewed as a sequence of i.i.d. samples of a random variable R_i, then the preceding limit exists and the problem simplifies to

$$x^* \in \operatorname*{Argmin}_{x\in X} \sum_{i=1}^{m} \mathsf{E}\Big[\left(R_i - h(\ell_i,x)\right)^2\Big].$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

In linear least squares regression, the models $h(\ell_i, x)$, $i = 1, \ldots, m$, are linear in x, so that each of the functions $f_i(x) = \mathsf{E}\left[(R_i - h(\ell_i, x))^2\right]$ is convex in x.

Distributed resource allocation. An important problem in wireless networks is the fair rate allocation problem [8]. Consider a wireless network represented by a graph with a set of directed (communication) links. Suppose that there are m flows indexed $1, \ldots, m$, whose rate can be adjusted, and let x_i be the rate of the ith flow. Each flow is characterized by a source node $b(i)$ and a destination node $e(i)$. The rate vector x must satisfy some constraints that are imposed by the individual link capacities of the network. For example, if there are multiple flows (or, parts of the flows) that use a link of total capacity c then the total sum of the rates of the flow routed through that link must not be greater than c. Apart from this there could also be some queuing delay constraints. Thus, only flow vectors that are constrained to a set X can be routed through the network. Associated with each flow, there is a reward function $U_i(x_i)$ depending only on the flow rate x_i and known only at the source node $b(i)$. The reward function is typically a concave and increasing function. In the *fair rate allocation problem*, the source nodes $\{b(\ell)\}$ need to determine the optimal flow rate x^* that maximizes the total network utility. Mathematically, the problem is to determine x^* such that

$$x^* \in \operatorname*{Argmax}_{x \in X} \sum_{i=1}^{m} U_i(x_i).$$

In some networks, the same flow can communicate different types of traffic that has different importance in different time slots. For example, in an intruder detection network, a "detected" message is more important (and is rewarded/weighted more) than a "not detected" message or some other system message. Thus, the reward function is also a function of the contents of the flow: If the type of flow i in time slot k is $r_{i,k}$, where $r_{i,k}$ takes values from the set of all possible types of flow data, then the reward is $U_i(x_i, r_{i,k})$ at time k. If the type of traffic on each flow across slots is i.i.d, then $\{r_{i,k}\}$ are i.i.d. samples of a random variable R_i and the *fair allocation rate* problem can be written as

$$\max_{x \in X} \sum_{i=1}^{m} \mathsf{E}[U_i(x_i, R_i)]\,.$$

The statistics of R_i may not be known since they may depend upon external factors such as the frequency of intruders in an intruder detection network.

2.2. Notation and basics. We view vectors as columns. We write $x^\top y$ to denote the inner product of two vectors x and y. We use $\|\cdot\|$ to denote the standard Euclidean norm. For a vector x, we use x_i to denote its ith component. For a matrix A, we use $[A]_{i,j}$ to denote its (i,j)th entry, $[A]_i$ its ith row, and $[A]^j$ its jth column. We use e to denote a vector with each entry equal to 1. Unless stated otherwise, all equalities and inequalities that involve random quantities are in an almost sure sense.

We use f^* to denote the optimal value of the problem (2.1), and we use X^* to denote its optimal set. Throughout the paper, we assume that the optimal value f^* is finite.

In our analysis, we use the subgradient defining property. Specifically, for a convex function $f : \Re^n \to \Re$, the vector $\nabla f(x)$ is a *subgradient of f at x* when the following

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

relation is satisfied:

$$\nabla f(x)^\top (y-x) \le f(y) - f(x) \qquad \text{for all } y \in \Re^n \tag{2.4}$$

(see, for example, [2]).

3. Cyclic incremental subgradient algorithm. Recall that the cyclic incremental stochastic subgradient algorithm is given by

$$\begin{aligned} z_{0,k+1} &= z_{m,k} = x_k, \\ z_{i,k+1} &= \mathcal{P}_X\left[z_{i-1,k+1} - \alpha_{k+1}\left(\nabla f_i(z_{i-1,k+1}) + \epsilon_{i,k+1}\right)\right], \end{aligned} \tag{3.1}$$

where $x_0 \in X$ is a random initial vector, $\nabla f_i(x)$ is a subgradient of f_i at x, $\epsilon_{i,k+1}$ is a random noise vector, and $\alpha_{k+1} > 0$ is a step-size.

The main difficulty in the study of the incremental stochastic subgradient algorithm is that the expected direction in which the iterate is adjusted in each subiteration is not necessarily a subgradient of the objective function f. For this reason, we cannot directly apply the classic stochastic approximation convergence results of [7, 16, 29] to study the convergence of method in (2.2). The key relation in our analysis is provided in Lemma 3.1 in section 3.1. Using this lemma and a standard super-martingale convergence result, we obtain results for diminishing step-size in Theorem 3.3. Furthermore, by considering a related "stopped" process to which we apply a standard supermartingale convergence result, we obtain the error bound results for a constant step-size in Theorem 3.5.

We make the following basic assumptions on the set X and the functions f_i.

ASSUMPTION 1. *The set $X \subseteq \Re^n$ is closed and convex. The function $f_i : \Re^n \to \Re$ is convex for each $i \in \{1, \ldots, m\}$.*

In our analysis, we assume that the first and the second moments of the subgradient noise $\epsilon_{i,k}$ are bounded uniformly over the agents, conditionally on the past realizations. In particular, we define F_k^i as the σ-algebra generated by x_0 and the subgradient errors $\epsilon_{1,1}, \ldots, \epsilon_{i,k}$, and assume the following.

ASSUMPTION 2. *There exist deterministic scalar sequences $\{\mu_k\}$ and $\{\nu_k\}$ such that*

$$\begin{aligned} \|\mathsf{E}\left[\epsilon_{i,k} \mid F_k^{i-1}\right]\| &\le \mu_k \qquad \text{for all } i \text{ and } k, \\ \mathsf{E}\left[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\right] &\le \nu_k^2 \qquad \text{for all } i \text{ and } k. \end{aligned}$$

Assumption 2 holds, for example, when the errors $\epsilon_{i,k}$ are independent across both i and k, and have finite moments. Note that under the assumption that the second moments are bounded, from Jensen's inequality we readily have

$$\left\|\mathsf{E}\left[\epsilon_{i,k} \mid F_k^{i-1}\right]\right\| \le \sqrt{\mathsf{E}\left[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\right]} \le \nu_k. \tag{3.2}$$

However, for a constant step-size, the terms $\left\|\mathsf{E}\left[\epsilon_{i,k} \mid F_k^{i-1}\right]\right\|$ and $\mathsf{E}\left[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\right]$ affect the error bounds on the performance of the incremental method differently, as seen in section 3.3. For this reason, we prefer to use different upper-bounds for the terms $\left\|\mathsf{E}\left[\epsilon_{i,k} \mid F_k^{i-1}\right]\right\|$ and $\mathsf{E}\left[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\right]$. We will also, without any loss of generality, assume that $\mu_k < \nu_k$.

We also assume that the subgradients $\nabla f_i(x)$ are uniformly bounded over the set X for each i. This assumption is commonly used in the convergence analysis of subgradient methods with a diminishing or a constant step-size.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

ASSUMPTION 3. *For every i, the subgradient set of the function f_i at $x \in X$ is nonempty and uniformly bounded over the set X by a constant C_i, i.e.,*

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

Assumption 3 holds, for example, when each f_i is a polyhedral function or when the set X is compact.

3.1. Preliminaries. In this section, we provide a lemma establishing a basic relation for the iterates generated by the incremental method (3.1) and any step-size rule. This relation plays a key role in our subsequent development.

LEMMA 3.1. *Let Assumptions* 1, 2, *and* 3 *hold. Then, the iterates generated by algorithm* (3.1) *are such that for any step-size rule and for any $y \in X$,*

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(y)\|^2 \mid F_k^m\right] \leq{}& \|d_k(y)\|^2 - 2\alpha_{k+1}\left(f(x_k) - f(y)\right)\\ &+ 2\alpha_{k+1}\mu_{k+1}\sum_{i=1}^{m}\mathsf{E}[\|d_{i-1,k+1}(y)\| \mid F_k^m]\\ &+ \alpha_{k+1}^2\left(m\nu_{k+1} + \sum_{i=1}^{m} C_i\right)^2, \end{aligned} \tag{3.3}$$

where $d_k(y) = x_k - y$ and $d_{i,k+1}(y) = z_{i,k+1} - y$ for all k.

Proof. Using the iterate update rule in (3.1) and the nonexpansive property of the Euclidean projection, we obtain for any $y \in X$,

$$\begin{aligned}\|d_{i,k+1}(y)\|^2 &= \|\mathcal{P}_X\left[z_{i-1,k+1} - \alpha_{k+1}\nabla f_i(z_{i-1,k+1}) - \alpha_{k+1}\epsilon_{i,k+1}\right] - y\|^2\\ &\leq \|z_{i-1,k+1} - \alpha_{k+1}\nabla f_i(z_{i-1,k+1}) - \alpha_{k+1}\epsilon_{i,k+1} - y\|^2\\ &= \|d_{i-1,k+1}(y)\|^2 - 2\alpha_{k+1}d_{i-1,k+1}(y)^\top\nabla f_i(z_{i-1,k+1})\\ &\quad - 2\alpha_{k+1}d_{i-1,k+1}(y)^\top\epsilon_{i,k+1} + \alpha_{k+1}^2\|\epsilon_{i,k} + \nabla f_i(z_{i-1,k+1})\|^2.\end{aligned}$$

Taking conditional expectations with respect to the σ-field F_{k+1}^{i-1}, we further obtain

$$\begin{aligned}\mathsf{E}\left[\|d_{i,k+1}(y)\|^2 \mid F_{k+1}^{i-1}\right] \leq{}& \|d_{i-1,k+1}(y)\|^2 - 2\alpha_{k+1}d_{i-1,k+1}(y)^\top\nabla f_i(z_{i-1,k+1})\\ &- 2\alpha_{k+1}d_{i-1,k+1}(y)^\top\mathsf{E}\left[\epsilon_{i,k+1} \mid F_{k+1}^{i-1}\right]\\ &+ \alpha_{k+1}^2\mathsf{E}\left[\|\epsilon_{i,k+1} + \nabla f_i(z_{i-1,k+1})\|^2 \mid F_{k+1}^{i-1}\right].\end{aligned}$$

By using Assumption 2 on the error moments, we have for all i and k,

$$-d_{i-1,k+1}(y)^\top\mathsf{E}\left[\epsilon_{i,k+1} \mid F_{k+1}^{i-1}\right] \leq \mu_{k+1}\|d_{i-1,k+1}(y)\|.$$

In addition, using Assumption 3 on the subgradient norms, we have for all i and k,

$$\mathsf{E}\left[\|\epsilon_{i,k+1} + \nabla f_i(z_{i-1,k+1})\|^2 \mid F_{k+1}^{i-1}\right] \leq (\nu_{k+1} + C_i)^2.$$

Combining the preceding three relations we obtain for all $y \in X$,

$$\begin{aligned}\mathsf{E}\left[\|d_{i,k+1}(y)\|^2 \mid F_{k+1}^{i-1}\right] \leq{}& \|d_{i-1,k+1}(y)\|^2 - 2\alpha_{k+1}d_{i-1,k+1}(y)^\top\nabla f_i(z_{i-1,k+1})\\ &+ 2\alpha_{k+1}\mu_{k+1}\|d_{i-1,k+1}(y)\| + \alpha_{k+1}^2(\nu_{k+1} + C_i)^2.\end{aligned} \tag{3.4}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

We now estimate the second term in the right-hand side of the preceding relation. By using the subgradient inequality in (2.4) and the subgradient norm bound from Assumption 3, we have

$$-d_{i-1,k+1}(y)^\top \nabla f_i(z_{i-1,k+1}) \leq -\left(f_i(x_k) - f_i(y)\right) + C_i \left\|z_{i-1,k+1} - x_k\right\|. \tag{3.5}$$

We next consider the term $\|z_{i-1,k+1} - x_k\|$. From (3.1) we have

$$\|z_{i-1,k+1} - x_k\| = \left\|\sum_{j=1}^{i-1} (z_{j,k+1} - z_{j-1,k+1})\right\| \leq \sum_{j=1}^{i-1} \|z_{j,k+1} - z_{j-1,k+1}\|.$$

By the nonexpansive property of the projection, we further have

$$\|z_{i-1,k+1} - x_k\| \leq \alpha_{k+1} \sum_{j=1}^{i-1} \left(\|\nabla f_j(z_{j-1,k+1})\| + \|\epsilon_{j,k+1}\|\right) \leq \alpha_{k+1} \sum_{j=1}^{i-1} \left(C_j + \|\epsilon_{j,k+1}\|\right). \tag{3.6}$$

By combining the preceding relation with (3.5), we have

$$-d_{i-1,k+1}(y)^\top \nabla f_i(z_{i-1,k+1}) \leq -\left(f_i(x_k) - f_i(y)\right) + \alpha_{k+1} C_i \sum_{j=1}^{i-1} \left(C_j + \|\epsilon_{j,k+1}\|\right).$$

By substituting the preceding estimate in the inequality in (3.4), we obtain for all $y \in X$,

$$\begin{aligned}\mathsf{E}\left[\|d_{i,k+1}(y)\|^2 \mid F_{k+1}^{i-1}\right] \leq{}& \|d_{i-1,k+1}(y)\|^2 - 2\alpha_{k+1}\left(f_i(x_k) - f_i(y)\right) \\ &+ 2\alpha_{k+1}^2 C_i \sum_{j=1}^{i-1} \left(C_j + \|\epsilon_{j,k+1}\|\right) \\ &+ 2\alpha_{k+1}\mu_{k+1}\|d_{i-1,k+1}(y)\| + \alpha_{k+1}^2 (C_i + \nu_{k+1})^2.\end{aligned}$$

Taking the expectation conditional on F_k^m, we obtain

$$\begin{aligned}\mathsf{E}\left[\|d_{i,k+1}(y)\|^2 \mid F_k^m\right] \leq{}& \mathsf{E}\left[\|d_{i-1,k+1}(y)\|^2 \mid F_k^m\right] - 2\alpha_{k+1}\left(f_i(x_k) - f_i(y)\right) \\ &+ 2\alpha_{k+1}\mu_{k+1}\mathsf{E}[\|d_{i-1,k+1}(y)\| \mid F_k^m] \\ &+ 2\alpha_{k+1}^2 C_i \sum_{j=1}^{i-1} (C_j + \nu_{k+1}) + \alpha_{k+1}^2 (C_i + \nu_{k+1})^2,\end{aligned}$$

where we have used Assumption 2 and Jensen's inequality to bound $\mathsf{E}[\|\epsilon_{j,k+1}\| \mid F_k^m]$ by ν_{k+1} [cf. (3.2)]. Summing over $i = 1, \ldots, m$, and noting that $d_{0,k+1}(y) = x_k - y$, we see that

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(y)\|^2 \mid F_k^m\right] \leq{}& \|d_k(y)\|^2 - 2\alpha_{k+1}\left(f(x_k) - f(y)\right) \\ &+ 2\alpha_{k+1}\mu_{k+1} \sum_{i=1}^{m} \mathsf{E}[\|d_{i-1,k+1}(y)\| \mid F_k^m] \\ &+ 2\alpha_{k+1}^2 \sum_{i=1}^{m} C_i \sum_{j=1}^{i-1} (C_j + \nu_{k+1}) + \sum_{i=1}^{m} \alpha_{k+1}^2 (C_i + \nu_{k+1})^2.\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Finally, by noting that

$$2\sum_{i=1}^{m} C_i \sum_{j=1}^{i-1} (C_j + \nu_{k+1}) + \sum_{i=1}^{m} (C_i + \nu_{k+1})^2 = \left(\sum_{i=1}^{m} C_i + m\nu_{k+1} \right)^2,$$

we obtain the desired relation. □

3.2. Convergence for diminishing step-size. We now study the convergence of the method in (3.1) for diminishing step-size rule. In our analysis, we use the following result due to Robbins and Siegmund (see Lemma 11, Chapter 2.2, [29]).

LEMMA 3.2. *Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space and let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots$ be a sequence of sub σ-fields of $\mathcal{F}$. Let u_k, v_k, and w_k, $k = 0, 1, 2 \dots$, be nonnegative $\mathcal{F}_k$-measurable random variables and let $\{q_k\}$ be a deterministic sequence. Assume that $\sum_{k=0}^{\infty} q_k < \infty$, and $\sum_{k=0} w_k < \infty$ 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 to a nonnegative random variable and $\sum_{k=0}^{\infty} v_k < \infty$.*

We next provide a convergence result for diminishing step-sizes.

THEOREM 3.3. *Let Assumptions* 1, 2, *and* 3 *hold. Assume that the step-size sequence $\{\alpha_k\}$ is positive and such that $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. In addition, assume that the bounds μ_k and ν_k on the moments of the error sequence $\{\epsilon_{i,k}\}$ are such that*

$$\sum_{k=1}^{\infty} \alpha_k \mu_k < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 \nu_k^2 < \infty.$$

Also, assume that the optimal set X^* is nonempty. Then, the iterate sequence $\{x_k\}$ generated by the method (3.1) *converges to an optimal solution with probability* 1.

Proof. First note that all the assumptions of Lemma 3.1 are satisfied. Let x^* be an arbitrary point in X^*. By letting $y = x^*$ in Lemma 3.1, we obtain for any $x^* \in X^*$,

$$\begin{aligned} \mathsf{E}\left[\|d_{k+1}(x^*)\|^2 \mid F_k^m\right] \leq & \|d_k(x^*)\|^2 - 2\alpha_{k+1}\left(f(x_k) - f^*\right) \\ & + 2\alpha_{k+1}\mu_{k+1} \sum_{i=1}^{m} \mathsf{E}[\|d_{i-1,k+1}(x^*)\| \mid F_k^m] \\ & + \alpha_{k+1}^2 \left(m\nu_{k+1} + \sum_{i=1}^{m} C_i \right)^2. \end{aligned} \tag{3.7}$$

We relate $\|d_{i-1,k+1}(x^*)\|$ to $\|d_k(x^*)\|$ by using the triangle inequality of norms,

$$\|d_{i-1,k+1}(x^*)\| = \|z_{i-1,k+1} - x_k + x_k - x^*\| \leq \|z_{i-1,k+1} - x_k\| + \|d_k(x^*)\|.$$

Substituting for $\|z_{i-1,k+1} - x_k\|$ from (3.6) we obtain

$$\|d_{i-1,k+1}(x^*)\| \leq \alpha_{k+1} \sum_{j=1}^{i-1} (C_j + \|\epsilon_{j,k+1}\|) + \|d_k(x^*)\|.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Taking conditional expectations and using the inequality $\|d_k(x^*)\| \leq (1+\|d_k(x^*)\|^2)/2$, we further obtain

$$\mathsf{E}[\|d_{i-1,k+1}(x^*)\| \mid F_k^m] \leq \frac{1}{2}(1+\|d_k(x^*)\|^2) + \alpha_{k+1}\sum_{j=1}^{i-1}(C_j+\nu_{k+1}),$$

where we have used Assumption 2 and Jensen's inequality to bound $\mathsf{E}[\|\epsilon_{j,k+1}\| \mid F_k^m]$ by ν_{k+1}. Using the preceding inequality in (3.7), we see that

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2 \mid F_k^m\right] \leq{}& (1+m\alpha_{k+1}\mu_{k+1})\,\|d_k(x^*)\|^2 - 2\alpha_{k+1}\,(f(x_k)-f^*)\\ &+ m\alpha_{k+1}\mu_{k+1} + 2\alpha_{k+1}^2\mu_{k+1}\sum_{i=1}^{m}\sum_{j=1}^{i-1}(C_j+\nu_{k+1})\\ &+ \alpha_{k+1}^2\left(m\nu_{k+1}+\sum_{i=1}^{m}C_i\right)^2. \qquad (3.8)\end{aligned}$$

By the assumptions on the step-size, and the sequences $\{\mu_k\}$ and $\{\nu_k\}$, we have

$$\begin{aligned}&\sum_{k=0}^{\infty} m\alpha_{k+1}\mu_{k+1} < \infty,\\ &\sum_{k=0}^{\infty} 2\alpha_{k+1}^2\mu_{k+1}\sum_{i=1}^{m}\sum_{j=1}^{i-1}(C_j+\nu_{k+1}) \leq 2\sum_{k=0}^{\infty}\sum_{i=1}^{m}\sum_{j=1}^{i-1}\left(\alpha_{k+1}^2\mu_{k+1}C_j+\alpha_{k+1}^2\nu_{k+1}^2\right) < \infty,\\ &\sum_{k=0}^{\infty}\alpha_{k+1}^2\left(m\nu_{k+1}+\sum_{i=1}^{m}C_i\right)^2 \leq 2\sum_{k=0}^{\infty}\alpha_{k+1}^2\left(m^2\nu_{k+1}^2+\left(\sum_{i=1}^{m}C_i\right)^2\right) < \infty,\end{aligned}$$

where in the second relation above, we have used $\mu_{k+1} \leq \nu_{k+1}$ [cf. (3.2)], while in the last inequality, we have used $(a+b)^2 \leq 2(a^2+b^2)$ valid for any scalars a and b. Thus, the conditions of Lemma 3.2 are satisfied with $u_k = \|d_k(x^*)\|^2$, $\mathcal{F}_k = F_k^m$, $q_k = m\alpha_{k+1}\mu_{k+1}$, $v_k = 2\alpha_{k+1}\,(f(x_k)-f^*)$, and

$$w_k = m\alpha_{k+1}\mu_{k+1} + 2\alpha_{k+1}^2\mu_{k+1}\sum_{i=1}^{m}\sum_{j=1}^{i-1}(C_j+\nu_{k+1}) + \alpha_{k+1}^2\left(m\nu_{k+1}+\sum_{i=1}^{m}C_i\right)^2.$$

Therefore, with probability 1, the scalar $\|d_{k+1}(x^*)\|^2$ converges to some nonnegative random variable for every $x^* \in X^*$. Also with probability 1, we have

$$\sum_{k=0}^{\infty}\alpha_{k+1}\,(f(x_k)-f^*) < \infty.$$

Since $\sum_{k=1}^{\infty}\alpha_k = \infty$, it follows that $\liminf_{k\to\infty} f(x_k) = f^*$ with probability 1. By considering a sample path for which $\liminf_{k\to\infty} f(x_k) = f^*$ and $\|d_{k+1}(x^*)\|^2$ converges for any x^*, we conclude that the sample sequence must converge to some vector in the optimal set X^* in view of continuity of f. Hence, the sequence $\{x_k\}$ converges to some vector in X^* with probability 1. □

Define $\mathsf{dist}\,(x, X^*)$ to be the distance between the point x and the set X^*. Note that under assumptions of Theorem 3.3, it can be seen that $\mathsf{E}\left[\mathsf{dist}\,(x_k, X^*)^2\right]$ also

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

converges to 0. In particular, since the solution set X^* is closed and convex, there exists a point $x_k^* \in X^*$ that is closest to x_k for every k. Letting $y = x_k^*$ in relation (3.8) and using the fact that $\mathsf{dist}\,(x_{k+1}, X^*) \leq d_{k+1}(x_k^*)$ with probability 1, we obtain for all k,

$$\begin{aligned}\mathsf{E}\Big[\mathsf{dist}\,(x_{k+1}, X^*)^2 \mid F_k^m\Big] \leq{}& (1 + m\alpha_{k+1}\mu_{k+1})\,\mathsf{dist}\,(x_k, X^*)^2 - 2\alpha_{k+1}\,(f(x_k) - f^*)\\ &+ m\alpha_{k+1}\mu_{k+1} + 2\alpha_{k+1}^2\mu_{k+1}\sum_{i=1}^{m}\sum_{j=1}^{i-1}(C_j + \nu_{k+1})\\ &+ \alpha_{k+1}^2\left(m\nu_{k+1} + \sum_{i=1}^{m} C_i\right)^2.\end{aligned}$$

Taking expectations, we obtain for all k,

$$\begin{aligned}\mathsf{E}\Big[\mathsf{dist}\,(x_{k+1}, X^*)^2\Big] \leq{}& (1 + m\alpha_{k+1}\mu_{k+1})\,\mathsf{E}\Big[\mathsf{dist}\,(x_k, X^*)^2\Big] - 2\alpha_{k+1}\,(f(x_k) - f^*)\\ &+ m\alpha_{k+1}\mu_{k+1} + 2\alpha_{k+1}^2\mu_{k+1}\sum_{i=1}^{m}\sum_{j=1}^{i-1}(C_j + \nu_{k+1})\\ &+ \alpha_{k+1}^2\left(m\nu_{k+1} + \sum_{i=1}^{m} C_i\right)^2.\end{aligned}$$

From the deterministic analog of Lemma 3.2, we can argue that $\mathsf{E}\Big[\mathsf{dist}\,(x_{k+1}, X^*)^2\Big]$ converges and $\liminf_{k\to\infty}\mathsf{E}[f(x_k)] = f^*$. Since $\{x_k\}$ converges to a point in X^* with probability 1, it follows that $\mathsf{E}\Big[\mathsf{dist}\,(x_{k+1}, X^*)^2\Big]$ converges to 0.

We next state a convergence result that requires weaker assumptions than Theorem 3.3. The result can be proved by combining the line of analysis used in the proof of Theorem 3.3 with the analysis used in [1].

THEOREM 3.4. *Let Assumption* 1 *hold. In addition, assume the following.*

(a) *For each $i \in \{1, \ldots, m\}$, the function $f_i(x)$ has linearly bounded subgradients, i.e., there exist $a_1 > 0$ and $a_2 > 0$ such that*

$$\|\nabla f_i(x)\| \leq a_1\|x\| + a_2 \qquad \text{for all subgradients } \nabla f_i(x) \text{ and for all } x \in X.$$

(b) *There exist two deterministic nonnegative sequences $\{\bar{\mu}_k\}$ and $\{\bar{\nu}_k\}$ such that for each $i \in \{1, \ldots, m\}$ and all k there is a subgradient $\nabla f_i(z_{i-1,k})$ satisfying*

$$\begin{aligned}\big\|\mathsf{E}\big[\epsilon_{i,k} \mid F_k^{i-1}\big]\big\| &\leq \bar{\mu}_k\,(1 + \|\nabla f_i(z_{i-1,k})\|)\,,\\ \mathsf{E}\big[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\big] &\leq A + \bar{\nu}_k^2\|\nabla f(z_{i-1,k})\|^2.\end{aligned}$$

(c) *The step-size sequence $\{\alpha_k\}$ is such that $\sum_{k=1}^{\infty}\alpha_k = \infty$ and $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$. The bounds $\bar{\mu}_k$ and $\bar{\nu}_k$ on the moments of the error sequence $\{\epsilon_{i,k}\}$ are such that*

$$\sum_{k=1}^{\infty}\alpha_k\bar{\mu}_k < \infty, \qquad \sum_{k=1}^{\infty}\alpha_k^2\bar{\nu}_k^2 < \infty.$$

(d) *The optimal set X^* is nonempty and bounded.*

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Then, the iterate sequence $\{x_k\}$ *generated by the method* (3.1) *converges to an optimal solution with probability* 1.

The condition in part (d) of Theorem 3.4 is satisfied, for example, when X is compact, or when the function $f(x)$ is coercive over X, i.e., $\lim_{\substack{\|x\|\to\infty \\ x\in X}} f(x) = \infty$.

Observe that the assumptions on the subgradients and the stochastic errors in Theorem 3.4 are weaker than in Theorem 3.3. For example, when each f_i is quadratic, then the conditions of Theorem 3.4 hold, while the conditions of Theorem 3.3 may not hold (since the gradients of f_i may not be bounded).

3.3. Error bound for constant step-size. Here, we study the behavior of the iterates $\{x_k\}$ generated by the method (3.1) with a constant step-size rule, i.e., $\alpha_k = \alpha$ for all k. In this case, we cannot guarantee the convergence of the iterates; however, we can provide bounds on the performance of the algorithm. In the following lemma, we provide an error bound for the expected values $\mathsf{E}[f(x_k)]$ and a bound for $\inf_k f(x_k)$ that holds with probability 1. The proofs of these results are similar to those used in [21].

THEOREM 3.5. *Let Assumptions* 1 *and* 2 *hold. Let the sequence* $\{x_k\}$ *be generated by the method* (3.1) *with a constant step-size rule, i.e.,* $\alpha_k = \alpha$ *for all* $k \geq 1$. *Also, assume that the set* X *is bounded, and*

$$\mu = \sup_{k\geq 1} \mu_k < \infty, \qquad \nu = \sup_{k\geq 1} \nu_k < \infty.$$

We then have

$$\liminf_{k\to\infty} \mathsf{E}[f(x_k)] \leq f^* + m\mu \max_{x,y\in X} \|x-y\| + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2, \tag{3.9}$$

and with probability 1,

$$\inf_{k\geq 0} f(x_k) \leq f^* + m\mu \max_{x,y\in X} \|x-y\| + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2. \tag{3.10}$$

Proof. Since X is compact and each f_i is convex over $\Re^n$, the subgradients of f_i are bounded over X for each i. Thus, all the assumptions of Lemma 3.1 are satisfied. Furthermore, the optimal set X^* is nonempty. Since $\mu_k \leq \mu$ and $\nu_k \leq \nu$ for all k, and $\|d_{i-1,k+1}(y)\| \leq \max_{x,y\in X} \|x-y\|$, according to the relation of Lemma 3.1, we have for $y = x^* \in X^*$,

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2 \mid F_k^m\right] \leq & \|d_k(x^*)\|^2 - 2\alpha\left(f(x_k) - f^*\right) + 2m\alpha\mu \max_{x,y} \|x-y\| \\ & + \alpha^2\left(\sum_{i=1}^m C_i + m\nu\right)^2. \end{aligned} \tag{3.11}$$

By taking the total expectation, we obtain for all $y \in X$ and all k,

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2\right] \leq & \mathsf{E}\left[\|d_k(x^*)\|^2\right] - 2\alpha\left(\mathsf{E}[f(x_k)] - f^*\right) + 2m\alpha\mu \max_{x,y} \|x-y\| \\ & + \alpha^2\left(\sum_{i=1}^m C_i + m\nu\right). \end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

If the relation (3.9) did not hold, then there would exist a $\gamma > 0$ and an index k_γ such that for all $k > k_\gamma$,

$$\mathsf{E}[f(x_k)] \geq f^* + \gamma + m\mu \max_{x,y\in X} \|x - y\| + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2.$$

This would imply that for all $k \geq k_\gamma$,

$$\mathsf{E}\big[\|d_{k+1}(x^*)\|^2\big] \leq \mathsf{E}\big[\|d_{k_\gamma}(x^*)\|^2\big] - 2\gamma\alpha(k - k_\gamma),$$

which evidently cannot hold for sufficiently large k. Thus, the relation (3.9) must hold.

We now prove the relation in (3.10). Define the set

$$L_N = \left\{x \in X : f(x) < f^* + \frac{1}{N} + m\mu \max_{x,y\in X} \|x - y\| + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2\right\}.$$

Let $x^* \in X^*$ and define the sequence $\hat{x}_k$ as follows:

$$\hat{x}_{k+1} = \begin{cases} x_{k+1} & \text{if } \hat{x}_k \notin L_N, \\ x^* & \text{if } \hat{x}_k \in L_N. \end{cases}$$

Thus, the process $\{\hat{x}_k\}$ is identical to the process $\{x_k\}$, until $\{x_k\}$ enters the set L_N. Define

$$\hat{d}_k(y) = \hat{x}_k - y.$$

Let us first consider the case when $\hat{x}_k \in L_N$. Since $\hat{x}_k = x^*$ and $\hat{x}_{k+1} = x^*$, we have $\hat{d}_k(x^*) = 0$ and $\hat{d}_{k+1}(x^*) = 0$, yielding

$$\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid F_k^m\Big] = \|\hat{d}_k(x^*)\|^2. \tag{3.12}$$

When $\hat{x}_k \notin L_N$, $\hat{x}_k = x_k$ and $\hat{x}_{k+1} = x_{k+1}$. Using relation (3.11), we conclude that

$$\begin{aligned}\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid F_k^m\Big] \leq& \|\hat{d}_k(x^*)\|^2 - 2\alpha\left(f(\hat{x}_k) - f(x^*)\right) + 2m\alpha\mu \max_{x,y\in X} \|x - y\| \\ &+ \alpha^2\left(\sum_{i=1}^m C_i + m\nu\right)^2.\end{aligned}$$

Observe that when $\hat{x}_k \notin L_N$,

$$f(\hat{x}_k) - f^* \geq \frac{1}{N} + m\mu \max_{x,y\in X} \|x - y\| + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2.$$

Therefore, by combining the preceding two relations, we obtain for $\hat{x}_k \notin L_N$,

$$\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid F_k^m\Big] \leq \|\hat{d}_k(x^*)\|^2 - \frac{2\alpha}{N}. \tag{3.13}$$

Therefore, from (3.12) and (3.13), we can write

$$\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid F_k^m\Big] \leq \|\hat{d}_k(x^*)\|^2 - \Delta_{k+1}, \tag{3.14}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where

$$\Delta_{k+1} = \begin{cases} 0 & \text{if } \hat{x}_k \in L_N, \\ \dfrac{2\alpha}{N} & \text{if } \hat{x}_k \notin L_N. \end{cases}$$

Observe that (3.14) satisfies the conditions of Lemma 3.2 with $u_k = \|\hat{d}_k(x^*)\|^2$, $\mathcal{F}_k = F_k^m$, $q_k = 0$, $v_k = \Delta_{k+1}$, and $w_k = 0$. Therefore, it follows that $\sum_{k=0}^{\infty} \Delta_{k+1} < \infty$ with probability 1. However, this is possible only if $\Delta_k = 0$ for all k sufficiently large. Therefore, with probability 1, we have $x_k \in L_N$ for all sufficiently large k. By letting $N \to \infty$, we obtain (3.10). ∎

As seen from relation (3.10) of Theorem 3.5, the error bound on the "best function" value $\inf_k f(x_k)$ depends on the step-size α, and the bounds μ and ν for the moments of the subgradient errors $\epsilon_{i,k}$. When the errors $\epsilon_{i,k}$ have zero mean, the results of Theorem 3.5 hold with $\mu = 0$. In this case[3], the error bound reduces to $\frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2$, which can be controlled with the step-size α. Furthermore, in the absence of any errors (i.e., $\mu = 0$ and $\nu = 0$), the error bound of Theorem 3.5 reduces to $f^* + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i\right)^2$, which coincides with the error bound for the cyclic incremental subgradient method (without errors) established in [22], Proposition 2.1.

4. Markov randomized incremental subgradient method. While the method of section 3 is implementable in networks with a ring structure (the agents form a cycle), the method of this section is implementable in networks with an arbitrary connectivity structure. The idea is to implement the incremental algorithm by allowing agents to communicate only with their neighbors. In particular, suppose at time k, an agent j updates and generates the estimate x_k. Then, agent j may pass this estimate to his neighboring agent i with probability $[P(k)]_{i,j}$. If agent j is not a neighbor of i, then this probability is 0. Formally, the update rule for this method is given by

$$x_{k+1} = \mathcal{P}_X\left[x_k - \alpha_{k+1}\left(\nabla f_{s(k+1)}(x_k) + \epsilon_{s(k+1),k+1}\right)\right], \tag{4.1}$$

where $x_0 \in X$ is some random initial vector, $\epsilon_{s(k+1),k+1}$ is a random noise vector, and $\alpha_{k+1} > 0$ is the step-size. The sequence of indices of agents updating in time evolves according to a time nonhomogeneous Markov chain with states $1, \ldots, m$. We let $P(k)$ denote the transition matrix of this chain at time k, i.e.,

$$[P(k)]_{i,j} = \mathsf{Prob}\left\{s(k+1) = j \mid s(k) = i\right\} \qquad \text{for all } i, j \in \{1, \ldots, m\}.$$

In the absence of subgradient errors ($\epsilon_{s(k),k} = 0$), when the probabilities $[P(k)]_{i,j}$ are all equal, i.e., $[P(k)]_{i,j} = \frac{1}{m}$ for all i, j and all k, the method reduces to the incremental method with randomization proposed in [22], which is applicable only to the agent networks that are fully connected.

We note here that the time nonhomogeneous Markov chain models networks where the set of neighbors of an agent may change in time (as the network may be mobile or for other reasons). We will also assume that the agents decide the probabilities with which they communicate with their neighbors, i.e., at time k, the agent j chooses the probabilities $[P(k)]_{i,j} \geq 0$ for his neighbors i.

[3] When $\mu = 0$, the result of Theorem 3.5 can be shown under the assumption that the subgradients are bounded, which is weaker than the assumption that X is bounded.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

The main difficulty in the analysis of the method in (4.1) comes from the dependence between the random agent index $s(k+1)$ and the iterate x_k. Assuming that the Markov chain is ergodic with the uniform steady-state distribution, in the absence of the errors $\epsilon_{i,k}$ (i.e., $\epsilon_{i,k} = 0$), it is intuitively possible that the method uses directions that approximate the subgradient $\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_k)$ in the steady state. This is the basic underlying idea that we exploit in our analysis.

For this idea to work, it is crucial not only that the Markov chain probabilities converge to a uniform distribution but also that the convergence rate estimate is available in an explicit form. The uniform steady state requirement is natural since it corresponds to each agent updating his objective f_i with the same steady state frequency, thus ensuring that the agents cooperatively minimize the overall network objective function $f(x) = \sum_{i=1}^{m} f_i(x)$, and not a weighted sum. We use the rate estimate of the convergence of the products $P(\ell)\cdots P(k)$ to determine the step-size choices that guarantee the convergence of the method in (4.1).

To ensure the desired limiting behavior of the Markov chain probabilities, we use the following two assumptions on the matrices $[P(k)]$.

ASSUMPTION 4. *Let $V = \{1, \ldots, m\}$. Let $E(k)$ be the set of edges (j, i) induced by the positive entries of the probability matrix $P(k)$, i.e.,*

$$E(k) = \{(i, j) \mid [P(k)]_{i,j} > 0\}.$$

There exists an integer $Q \geq 1$ such that the graph $\left(V, \cup_{l=k}^{k+Q-1} E(\ell)\right)$ is strongly connected for all k.

Generally speaking, Assumption 4 ensures that the agents are connected sufficiently often in time. To guarantee that each agent updates the estimate x_k with the same frequency in the long run, we impose the following assumption.

ASSUMPTION 5.

(a) *The diagonal entries of $P(k)$ are all positive for each k.*

(b) *All positive entries of $[P(k)]$ are uniformly bounded away from zero, i.e., there exists a scalar $\eta > 0$ such that for all $i, j \in \{1, \ldots, m\}$ and all k,*

$$\text{if} \quad [P(k)]_{i,j} > 0, \qquad \text{then} \quad [P(k)]_{i,j} > \eta.$$

(c) *The matrix $P(k)$ is doubly stochastic for each k; i.e., the sum of the entries in every row and every column is equal to* 1.

Assumptions 5(a) and 5(b) ensure that the information from each and every agent is persistent in time. Assumption 5(c) ensures that the limiting Markov chain probability distribution (if one exists) is uniform. Assumptions 4 and 5 together guarantee the existence of the uniform limiting distribution, as shown in [27]. We state this result in the next section.

Note that the cyclic incremental algorithm (2.2) does not satisfy Assumption 5. The transition probability matrix corresponding to the cyclic incremental method is a permutation matrix with the (i, i)th entry being zero when agent i updates at time k. Thus, Assumption 5(c) is violated.

We now provide some examples of transition matrices $[P(k)]$ satisfying Assumption 5. The second and third examples are variations of the Metropolis–Hasting weights [10, 41], defined in terms of the agent neighbors. We let $N_i(k) \subset \{1, \ldots, m\}$ be the set of neighbors of an agent i at time k, and let $|N_i(k)|$ be the cardinality of this set. Consider the following rules:

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

- *Equal probability scheme.* The probabilities that agent i uses at time k are

$$[P(k)]_{i,j} = \begin{cases} \dfrac{1}{m} & \text{if } j \neq i \text{ and } j \in N_i(k), \\ 1 - \dfrac{|N_i(k)|}{m} & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

- *Min-equal neighbor scheme.* The probabilities that agent i uses at time k are

$$[P(k)]_{i,j} = \begin{cases} \min\left\{\dfrac{1}{|N_i(k)|+1}, \dfrac{1}{|N_j(k)|+1}\right\} & \text{if } j \neq i \text{ and } j \in N_i(k), \\ 1 - \displaystyle\sum_{j \in N_i(k)} \min\left\{\dfrac{1}{|N_i(k)|+1}, \dfrac{1}{|N_j(k)|+1}\right\} & \text{if } j = i, \\ 0 & \text{otherwise.} \end{cases}$$

- *Weighted Metropolis–Hastings scheme.* The probabilities that agent i uses at time k are given by

$$[P(k)]_{i,j} = \begin{cases} \eta_i \min\left\{\dfrac{1}{|N_i(k)|}, \dfrac{1}{N_j(k)|}\right\} & \text{if } j \neq i \text{ and } j \in N_i(k), \\ 1 - \eta_i \displaystyle\sum_{j \in N_i(k)} \min\left\{\dfrac{1}{|N_i(k)|}, \dfrac{1}{|N_j(k)|}\right\} & \text{if } j = i, \\ 0 & \text{otherwise,} \end{cases}$$

  where the scalar $\eta_i > 0$ is known only to agent i.

In the first example, the parameter η can be defined as $\eta = \frac{1}{m}$. In the second example, η can be defined as

$$\eta = \min_{i,j}\left\{\frac{1}{|N_i(k)|+1}, \frac{1}{|N_j(k)|+1}\right\},$$

while in the third example, it can be defined as

$$\eta = \min_i\{\eta_i, 1-\eta_i\} \min_{i,j}\left\{\frac{1}{|N_i(k)|}, \frac{1}{|N_j(k)|}\right\}.$$

Furthermore, note that in the first example, each agent knows the size of the network and no additional coordination with the other agents is needed. In the other two examples, an agent must be aware of the number of the neighbors each of his neighbors has at any time.

4.1. Preliminaries. We first state a result from [25] for future reference. The result captures the convergence and the rate of convergence of the time nonhomogeneous Markov chain to its steady state. Define $\Phi(k,\ell)$, with $k > l$, to be the transition probability matrix for the Markov chain from time l to k, i.e., $\Phi(k,\ell) = P(\ell)\cdots P(k)$ with $k \geq l$. Then, we have the following convergence result for the transition matrices.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEMMA 4.1. *Assume the matrices $P(k)$ satisfy Assumptions* 4 *and* 5. *Then:*

1. $\lim_{k\to\infty} \Phi(k,s) = \frac{1}{m} e e^\top$ *for all* s.
2. *The convergence is geometric and the rate of convergence is given by*

$$\left| [\Phi(k,\ell)]_{i,j} - \frac{1}{m} \right| \le b\,\beta^{k-l} \qquad \text{for all } k \text{ and } l \text{ with } k \ge l \ge 0,$$

where

$$b = \left(1 - \frac{\eta}{4m^2}\right)^{-2} \quad \text{and} \quad \beta = \left(1 - \frac{\eta}{4m^2}\right)^{\frac{1}{Q}}.$$

We use the estimate of Lemma 4.1 to establish a key relation in Lemma 4.2, which is repeatedly invoked in our subsequent analysis. The idea behind Lemma 4.2 is the observation that when there are no errors ($\epsilon_{s(k),k} = 0$) and the Markov chain has a uniform steady state distribution, the directions $\nabla f_{s(k+1)}(x_k)$ used in (4.1) are approximate subgradients of the function $\frac{1}{m}\sum_{i=1}^m f_i(x)$ at points $x_{n(k)}$ far away from x_k in the past [i.e., $k >> n(k)$]. However, even though $x_n(k)$ are far away from x_k in time, their Euclidean distance $\|x_k - x_{n(k)}\|$ can be small when the step-size is selected appropriately. Overall, this means that each iterate of method in (4.1) can be viewed as an approximation of the iteration

$$x_{k+1} = P_X\left[x_k - \frac{\alpha_{k+1}}{m}\sum_{i=1}^m \nabla f_i(x_k) + \alpha_{k+1}\xi_k\right],$$

with correlated errors ξ_k depending on current and past iterates.

In the forthcoming lemma and thereafter, we let G_k denote the entire history of the method up to time k, i.e., the σ-field generated by the initial vector x_0 and $\{s(n), \epsilon_{s(n),n}; 0 \le n \le k\}$.

LEMMA 4.2. *Let Assumptions* 1–5 *hold. Then, the iterates generated by algorithm* (4.1) *are such that for any step-size rule, for any* $y \in X$, *and any nonnegative integer sequence* $\{n(k)\}$, $n(k) \le k$, *we have*

$$\begin{aligned}
\mathsf{E}\left[\|d_{k+1}(y)\|^2 \mid G_{n(k)}\right] \le{}& \mathsf{E}\left[\|d_k(y)\|^2 \mid G_{n(k)}\right] - \frac{2\alpha_{k+1}}{m}\left(f\left(x_{n(k)}\right) - f(y)\right) \\
&+ 2b\left(\sum_{i=1}^m C_i\right)\alpha_{k+1}\beta^{k+1-n(k)}\left\|d_{n(k)}(y)\right\| \\
&+ 2C\alpha_{k+1}\sum_{l=n(k)}^{k-1}\alpha_{l+1}\left(C + \nu_{l+1}\right) \\
&+ 2\alpha_{k+1}\mu_{k+1}\mathsf{E}\left[\|d_k(y)\| \mid G_{n(k)}\right] + \alpha_{k+1}^2(\nu_k + C)^2,
\end{aligned}$$

where $d_k(y) = x_k - y$ *and* $C = \max_{1\le i\le m} C_i$.

Proof. Using the iterate update rule in (4.1), the nonexpansive property of the Euclidean projection and the subgradient inequality in (2.4), we obtain for any $y \in X$ and $k \ge 0$,

$$\begin{aligned}
\|d_{k+1}(y)\|^2 \le{}& \|d_k(y)\|^2 - 2\alpha_{k+1}\left(f_{s(k+1)}(x_k) - f_{s(k+1)}(y)\right) \\
&- 2\alpha_{k+1}d_k(y)^\top\epsilon_{s(k+1),k+1} + \alpha_{k+1}^2\left\|\epsilon_{s(k+1),k+1} + \nabla f_{s(k+1)}(x_k)\right\|^2.
\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

By writing

$$f_{s(k+1)}(x_k) - f_{s(k+1)}(y) = \left(f_{s(k+1)}(x_k) - f_{s(k+1)}\left(x_{n(k)}\right)\right) + \left(f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}(y)\right),$$

and by taking conditional expectations with respect to the σ-field $G_{n(k)}$, we obtain

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(y)\|^2 \mid G_{n(k)}\right] \le\ & \mathsf{E}\left[\|d_k(y)\|^2 \mid G_{n(k)}\right] \\ & - 2\alpha_{k+1}\left(\mathsf{E}\left[f_{s(k+1)}(x_k) - f_{s(k+1)}\left(x_{n(k)}\right) \mid G_{n(k)}\right]\right) \\ & - 2\alpha_{k+1}\left(\mathsf{E}\left[f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}(y) \mid G_{n(k)}\right]\right) \\ & - 2\alpha_{k+1}\mathsf{E}\left[d_k(y)^\top \epsilon_{s(k+1),k+1} \mid G_{n(k)}\right] \\ & + \alpha_{k+1}^2 \mathsf{E}\left[\left\|\epsilon_{s(k+1),k+1} + \nabla f_{s(k+1)}(x_k)\right\|^2 \mid G_{n(k)}\right]. \end{aligned} \tag{4.2}$$

We use the subgradient inequality in (2.4) and the subgradient boundedness from Assumption 3 to estimate the second term in the preceding relation, as follows:

$$\mathsf{E}\left[f_{s(k+1)}(x_k) - f_{s(k+1)}\left(x_{n(k)}\right) \mid G_{n(k)}\right] \ge -C\mathsf{E}\left[\left\|x_{n(k)} - x_k\right\| \mid G_{n(k)}\right].$$

We next estimate $\mathsf{E}\left[\left\|x_{n(k)} - x_k\right\| \mid G_{n(k)}\right]$ from the iterate update rule (4.1) and the nonexpansive property of the Euclidean projection as follows:

$$\begin{aligned}\mathsf{E}\left[\left\|x_{n(k)} - x_k\right\| \mid G_{n(k)}\right] &\le \sum_{l=n(k)}^{k-1} \mathsf{E}\left[\|x_{l+1} - x_l\| \mid G_{n(k)}\right] \\ &\le \sum_{l=n(k)}^{k-1} \alpha_{l+1}\mathsf{E}\left[\left\|\nabla f_{s(\ell+1)}(x_l)\right\| + \left\|\epsilon_{s(\ell+1),l+1}\right\| \mid G_{n(k)}\right] \\ &\le \sum_{l=n(k)}^{k-1} \alpha_{l+1}(C + \nu_{l+1}),\end{aligned}$$

where in the last step we have used the law of iterated conditioning, and the boundedness of subgradients and the second moments of $\epsilon_{i,k}$ [cf. (3.2)]. From the preceding two relations, we obtain

$$\mathsf{E}\left[f_{s(k+1)}(x_k) - f_{s(k+1)}\left(x_{n(k)}\right) \mid G_{n(k)}\right] \ge -C\sum_{l=n(k)}^{k-1} \alpha_{l+1}(C + \nu_{l+1}).$$

For the last term in (4.2), by using the subgradient boundedness of Assumption 3 and the boundedness of the second moments of $\epsilon_{i,k}$ [cf. (3.2)], we have for all k,

$$\mathsf{E}\left[\left\|\epsilon_{s(k+1),k+1} + \nabla f_{s(k+1)}(x_k)\right\|^2 \mid G_{n(k)}\right] \le (\nu_k + C)^2.$$

We next estimate the term $\mathsf{E}\left[d_k(y)^\top \epsilon_{s(k+1),k+1} \mid G_{n(k)}\right]$ in (4.2). Since $G_{n(k)} \subset G_k$ and $d_k(y)$ is G_k-measurable, from the law of iterated conditioning it follows

$$\begin{aligned}\mathsf{E}\left[d_k(y)^\top \epsilon_{s(k+1),k+1} \mid G_{n(k)}\right] &= \mathsf{E}\left[\mathsf{E}\left[d_k(y)^\top \epsilon_{s(k+1),k+1} \mid G_k\right] \mid G_{n(k)}\right] \\ &= \mathsf{E}\left[d_k(y)^\top \mathsf{E}\left[\epsilon_{s(k+1),k+1} \mid G_k\right] \mid G_{n(k)}\right] \\ &\ge -\mathsf{E}\left[\|d_k(y)\| \left\|\mathsf{E}\left[\epsilon_{s(k+1),k+1} \mid G_k\right]\right\| \mid G_{n(k)}\right] \\ &\ge -\mu_{k+1}\mathsf{E}\left[\|d_k(y)\| \mid G_{n(k)}\right].\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

By substituting the preceding three estimates in relation (4.2), we obtain

$$\begin{aligned}\mathsf{E}\big[\|d_{k+1}(y)\|^2 \mid G_{n(k)}\big] \le{}& \mathsf{E}\big[\|d_k(y)\|^2 \mid G_{n(k)}\big] + 2C\alpha_{k+1}\sum_{l=n(k)}^{k-1}\alpha_{l+1}\left(C+\nu_{l+1}\right)\\ &- 2\alpha_{k+1}\left(\mathsf{E}\big[f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}\left(y\right) \mid G_{n(k)}\big]\right)\\ &+ 2\alpha_{k+1}\mu_{k+1}\mathsf{E}\big[\|d_k(y)\| \mid G_{n(k)}\big] + \alpha_{k+1}^2(\nu_k+C)^2.\end{aligned} \tag{4.3}$$

Finally, we consider the term $\mathsf{E}\big[f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}\left(y\right) \mid G_{n(k)}\big]$, and we use the fact that the probability transition matrix for the Markov chain $\{s(k)\}$ from time $n(k)$ to time k is $\Phi(k+1,n(k)) = P(n(k))\cdots P(k)$. We have

$$\begin{aligned}&\mathsf{E}\big[f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}\left(y\right) \mid G_{n(k)}\big]\\ &= \sum_{i=1}^m \left[\Phi(k+1,n(k))\right]_{s(n(k)),i}\left(f_i\left(x_{n(k)}\right) - f_i(y)\right)\\ &\ge \sum_{i=1}^m \frac{1}{m}\left(f_i\left(x_{n(k)}\right) - f_i(y)\right) - \sum_{i=1}^m \left|\left[\Phi(k+1,n(k))\right]_{s(n(k)),i} - \frac{1}{m}\right|\left|f_i\left(x_{n(k)}\right) - f_i(y)\right|\\ &\ge \frac{1}{m}\left(f\left(x_{n(k)}\right) - f(y)\right) - b\beta^{k+1-n(k)}\sum_{i=1}^m \left|f_i\left(x_{n(k)}\right) - f_i(y)\right|,\end{aligned} \tag{4.4}$$

where at the last step we have used Lemma 4.1. By the subgradient inequality (2.4), we further have

$$\left|f_i\left(x_{n(k)}\right) - f_i(y)\right| \le C_i\left\|x_{n(k)} - y\right\| = C_i\|d_{n(k)}(y)\|. \tag{4.5}$$

The result now follows by combining the relations in (4.3), (4.4), and (4.5). ∎

4.2. Convergence for diminishing step-size. In this section, we establish the convergence of the Markov randomized method in (4.1) for a diminishing step-size. Recall that in Theorem 3.3 for the cyclic incremental method, we showed an almost sure convergence result for a diminishing step-size α_k subject to some conditions that coordinate the choice of the step-size, and the bounds μ_k and ν_k on the moments of the errors $\epsilon_{i,k}$. To obtain an analogous result for the Markov randomized method, we use the boundedness of the set X and a more restricted step-size. In particular, we consider a step-size of the form $\alpha_k = \frac{a}{k^p}$ for a range of values of p, as seen in the following.

THEOREM 4.3. *Let Assumptions* 1, 2, 4, *and* 5 *hold. Assume that the step-size is* $\alpha_k = \frac{a}{k^p}$, *where* a *and* p *are positive scalars with* $\frac{2}{3} < p \le 1$. *In addition, assume that the bounds* μ_k *and* ν_k *on the error moments satisfy*

$$\sum_{k=1}^\infty \alpha_k\mu_k < \infty, \qquad \nu = \sup_{k\ge 1}\nu_k < \infty.$$

Furthermore, let the set X *be bounded. Then, with probability* 1, *we have*

$$\liminf_{k\to\infty} f(x_k) = f^*, \qquad \liminf_{k\to\infty}\operatorname{dist}(x_k, X^*) = 0.$$

Proof. Since the set X is compact and f_i is convex over $\Re^n$, it follows that the subgradients of f_i are bounded over X for each i. Thus, Assumption 3 is satisfied, and we can use Lemma 4.2.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Since X is compact and f is convex over $\Re^n$ (therefore, also continuous), the optimal set X^* is nonempty, closed, and convex. Let x_k^* be the projection of x_k on the set X^*. In Lemma 4.2, we let $y = x_k^*$ and let $n(k) = k + 1 - \lceil k^\gamma \rceil$, where $\gamma > 0$ (to be specified more precisely later on). Note that $n(k) \leq k$ for all $k \geq 1$. Using this and the relation $\mathsf{dist}\,(x_{k+1}, X^*) \leq \|x_{k+1} - x_k^*\|$, from Lemma 4.2, we obtain for all $k > 1$,

$$\begin{aligned}\mathsf{E}\Big[\mathsf{dist}\,(x_{k+1}, X^*)^2 \mid G_{n(k)}\Big] \leq \mathsf{E}\Big[\mathsf{dist}\,(x_k, X^*)^2 \mid G_{n(k)}\Big] - \frac{2\alpha_{k+1}}{m}\left(f\left(x_{n(k)}\right) - f^*\right)\\ + 2b\left(\sum_{i=1}^m C_i\right)\alpha_{k+1}\beta^{\lceil k^\gamma\rceil}\left\|d_{n(k)}(x_k^*)\right\|\\ + 2C\alpha_{k+1}\alpha_{n(k)+1}(\lceil k^\gamma\rceil - 2)\max_{n(k)\leq l\leq k}(C + \nu_{l+1})\\ + 2\alpha_{k+1}\mu_{k+1}\mathsf{E}\big[\mathsf{dist}\,(x_k, X^*) \mid G_{n(k)}\big] + \alpha_{k+1}^2(\nu_k + C)^2.\end{aligned}$$

Taking expectations and using $\sup_{k\geq 1}\nu_k = \nu$, we obtain for all $k > 1$,

$$\mathsf{E}\Big[\mathsf{dist}\,(x_{k+1}, X^*)^2\Big] \leq \mathsf{E}\Big[\mathsf{dist}\,(x_k, X^*)^2\Big] - \frac{2\alpha_{k+1}}{m}\left(\mathsf{E}\big[f\left(x_{n(k)}\right)\big] - f^*\right) + \tau_{k+1},$$

where

$$\begin{aligned}\tau_{k+1} = 2b\left(\sum_{i=1}^m C_i\right)\alpha_{k+1}\beta^{\lceil k^\gamma\rceil}\left\|d_{n(k)}(x_k^*)\right\|\\ + 2C(C+\nu)\alpha_{k+1}\alpha_{n(k)+1}(\lceil k^\gamma\rceil - 2)\\ + 2\alpha_{k+1}\mu_{k+1}\mathsf{E}[\mathsf{dist}\,(x_k, X^*)] + \alpha_{k+1}^2(\nu_k + C)^2.\end{aligned}$$

We next show that $\sum_{k=2}^\infty \tau_{k+1} < \infty$. Since $\alpha_k = \frac{a}{k^p}$, we have $\alpha_{k+1} < \alpha_k$ for all $k \geq 1$. Furthermore, since $\beta < 1$, we have $\beta^{\lceil k^\gamma\rceil} < \beta^{k^\gamma}$. Therefore, $\alpha_{k+1}\beta^{\lceil k^\gamma\rceil} < \frac{a\beta^{k^\gamma}}{k^p}$. By choosing $\gamma > 0$ such that $\gamma \geq 1 - p$, we see that $\frac{1}{k^p} \leq \frac{1}{k^{1-\gamma}}$ for all $k > 1$. Hence, for all $k > 1$,

$$\sum_{k=2}^\infty \alpha_{k+1}\beta^{\lceil k^\gamma\rceil} < \sum_{k=2}^\infty \frac{a\beta^{k^\gamma}}{k^p} \leq \sum_{k=2}^\infty \frac{a\beta^{k^\gamma}}{k^{1-\gamma}} \leq a\int_1^\infty \frac{\beta^{y^\gamma}}{y^{1-\gamma}}\,dy = -\frac{a\beta}{\gamma\ln(\beta)}.$$

Since the set X is bounded, it follows that

$$\sum_{k=2}^\infty 2b\left(\sum_{i=1}^m C_i\right)\alpha_{k+1}\beta^{\lceil k^\gamma\rceil}\left\|d_{n(k)}(x_k^*)\right\| < \infty. \tag{4.6}$$

Next, since $\lceil k^\gamma\rceil - 2 < k^\gamma$ for all $k \geq 2$, and since $\alpha_{k+1} < \alpha_k$, $\alpha_k = \frac{1}{k^p}$ and $n(k) = k + 1 - \lceil k^\gamma\rceil$, it follows that for all $k \geq 2$,

$$\alpha_{k+1}\alpha_{n(k)+1}(\lceil k^\gamma\rceil - 2) < \frac{a^2\,k^\gamma}{k^p(k + 2 - \lceil k^\gamma\rceil)^p} < \frac{a^2\,k^\gamma}{k^p(k - k^\gamma)^p} = \frac{a^2\,k^\gamma}{k^{2p}(1 - k^{\gamma-1})^p}.$$

By choosing $\gamma > 0$ such that it also satisfies $\gamma < 2p - 1$ (in addition to $\gamma \geq 1 - p$), we have $\gamma < 1$ (in view of $p \leq 1$). Therefore, for all $k \geq 2$,

$$\frac{k^\gamma}{k^{2p}(1 - k^{\gamma-1})^p} \leq \frac{1}{(1 - 2^{\gamma-1})^p}\frac{1}{k^{2p-\gamma}}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

By combining the preceding two relations, we have

$$\sum_{k=2}^{\infty} 2C(C+\nu)\alpha_{k+1}\alpha_{n(k)+1}(\lceil k^{\gamma}\rceil - 2) < 2C(C+\nu)\frac{a^2}{(1-2^{\gamma-1})^p}\sum_{k=2}^{\infty}\frac{1}{k^{2p-\gamma}} < \infty, \tag{4.7}$$

where the finiteness of the last sum follows from $2p-\gamma > 1$.

Finally, as a consequence of our assumptions, we also have

$$\sum_{k=2}^{\infty} 2\alpha_{k+1}\mu_{k+1}\mathsf{E}[\mathsf{dist}\,(x_k, X^*)] < \infty,$$
$$\sum_{k=2}^{\infty} \alpha_{k+1}^2(\nu_k + C)^2 < \infty.$$

From (4.6) and (4.7), and the preceding two relations, we see that $\sum_{k=2}^{\infty}\tau_{k+1} < \infty$.

From the deterministic analog of Lemma 3.2, we conclude that $\mathsf{E}\left[\mathsf{dist}\,(x_k, X^*)^2\right]$ converges to a nonnegative scalar and

$$\sum_{k=2}^{\infty}\frac{2\alpha_{k+1}}{m}\left(\mathsf{E}\left[f\left(x_{n(k)}\right)\right] - f^*\right) < \infty.$$

Since $p < 1$, we have $\sum_{k=2}^{\infty}\alpha_{k+1} = \infty$. Further, since $f\left(x_{n(k)}\right) \geq f^*$, it follows that

$$\liminf_{k\to\infty}\mathsf{E}\left[f\left(x_{n(k)}\right)\right] = f^*. \tag{4.8}$$

The function f is convex over $\Re^n$ and, hence, continuous. Since the set X is bounded, the function $f(x)$ is also bounded on X. Therefore, from Fatou's lemma it follows that

$$\mathsf{E}\left[\liminf_{k\to\infty} f\,(x_k)\right] \leq \liminf_{k\to\infty}\mathsf{E}[f\,(x_k)] = f^*,$$

implying that $\liminf_{k\to\infty} f\,(x_k) = f^*$ with probability 1. Moreover, from this relation, by the continuity of f and boundedness of X, it follows that $\liminf_{k\to\infty}\mathsf{dist}\,(x_k, X^*) = 0$ with probability 1. □

As seen in the proof of Theorem 4.3, $\mathsf{E}\left[\mathsf{dist}\,(x_k, X^*)^2\right]$ converges to a nonnegative scalar. Since $\liminf_{k\to\infty}\mathsf{dist}\,(x_k, X^*) = 0$ with probability 1, there is a subsequence of $\{\mathsf{dist}\,(x_k, X^*)^2\}$, which we denote by $\{\mathsf{dist}\,(x_{\ell_k}, X^*)^2\}$, which converges to 0 with probability 1. Since the set X is bounded, the sequence $\{\mathsf{dist}\,(x_{\ell_k}, X^*)^2\}$ is bounded. By the dominated convergence theorem, we conclude that $\mathsf{E}\left[\mathsf{dist}\,(x_{\ell_k}, X^*)^2\right]$ must also converge to 0. Therefore, the whole sequence $\left\{\mathsf{E}\left[\mathsf{dist}\,(x_k, X^*)^2\right]\right\}$ converges to 0.

4.3. Error bounds for constant step-size. We now establish error bounds when the Markov randomized incremental method is used with a constant step-size.

THEOREM 4.4. *Let Assumptions* 1, 2, 4, *and* 5 *hold. Let the sequence* $\{x_k\}$ *be generated by the method* (4.1) *with a constant step-size rule, i.e.,* $\alpha_k = \alpha$ *for all* k. *Also, assume that the set* X *is bounded, and*

$$\mu = \sup_{k\geq 1}\mu_k < \infty, \qquad \nu = \sup_{k\geq 1}\nu_k < \infty.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Then for any integer $T \geq 0$,

$$\begin{aligned}\liminf_k \mathsf{E}[f(x_k)] \leq f^* + \mu \max_{x,y\in X} \|x-y\| + \frac{1}{2}\alpha(\nu+C)^2 + \alpha TC\,(C+\nu)\\ + b\left(\sum_{i=1}^m C_i\right)\beta^{T+1}\max_{x,y\in X}\|x-y\|,\end{aligned} \tag{4.9}$$

where $\beta = \left(1-\frac{\eta}{4m^2}\right)^{\frac{1}{Q}}$ *and* $C = \max_{1\leq i\leq m} C_i$. *Furthermore, with probability* 1, *the same estimate holds for* $\inf_k f(x_k)$.

Proof. Since X is compact and each f_i is convex over $\Re^n$, the subgradients of f_i are bounded over X for each i. Thus, all the assumptions of Lemma 4.2 are satisfied. Let T be a nonnegative integer and let $n(k) = k - T$. Since $\mu_k \leq \mu$ and $\nu_k \leq \nu$ for all k, and $\|d_k(y)\| \leq \max_{x,y\in X}\|x-y\|$, according to Lemma 4.2, we have for $y = x^* \in X^*$ and $k \geq T$,

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2 \mid G_{n(k)}\right] \leq \mathsf{E}\left[\|d_k(x^*)\|^2 \mid G_{n(k)}\right] - \frac{2\alpha}{m}\left(f\left(x_{k-T}\right) - f^*\right)\\ + 2b\left(\sum_{i=1}^m C_i\right)\alpha\beta^{T+1}\max_{x,y\in X}\|x-y\|\\ + 2\alpha^2 TC\,(C+\nu)\\ + 2\alpha\mu\max_{x,y\in X}\|x-y\| + \alpha^2(\nu+C)^2.\end{aligned} \tag{4.10}$$

By taking the total expectation, we obtain for all $x^* \in X^*$ and all $k \geq T$,

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2\right] \leq \mathsf{E}\left[\|d_k(x^*)\|^2\right] - \frac{2\alpha}{m}\left(\mathsf{E}[f\left(x_{k-T}\right)] - f^*\right)\\ + 2b\left(\sum_{i=1}^m C_i\right)\alpha\beta^{T+1}\max_{x,y\in X}\|x-y\|\\ + 2\alpha^2 TC\,(C+\nu)\\ + 2\alpha\mu\max_{x,y\in X}\|x-y\| + \alpha^2(\nu+C)^2.\end{aligned}$$

Now assume that the relation (4.9) does not hold. Then, there will exist a $\gamma > 0$ and an index $k_\gamma \geq T$ such that for all $k \geq k_\gamma$,

$$\begin{aligned}\mathsf{E}[f(x_k)] \geq f^* + \gamma + \mu\max_{x,y\in X}\|x-y\| + \frac{1}{2}\alpha(\nu+C)^2 + \alpha TC\,(C+\nu)\\ + b\left(\sum_{i=1}^m C_i\right)\beta^{T+1}\max_{x,y\in X}\|x-y\|.\end{aligned}$$

Therefore, for $k \geq k_\gamma + T$, we have

$$\mathsf{E}\left[\|d_{k+1}(x^*)\|^2\right] \leq \mathsf{E}\left[\|d_k(x^*)\|^2\right] - 2\alpha\gamma \leq \cdots \leq \mathsf{E}\left[\|d_{k_\gamma}(x^*)\|^2\right] - 2\alpha\gamma(k-k_\gamma).$$

For sufficiently large k, the right-hand side of the preceding relation is negative, yielding a contradiction. Thus, the relation (4.9) must hold for all $T \geq 0$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

We next show that for any $T \geq 0$,

$$\begin{aligned}\inf_k f(x_k) \leq & f^* + \mu \max_{x,y\in X} \|x-y\| + \frac{1}{2}\alpha(\nu+C)^2 + \alpha TC\,(C+\nu) \\ & + b\left(\sum_{i=1}^m C_i\right)\beta^{T+1} \max_{x,y\in X}\|x-y\|, \end{aligned} \tag{4.11}$$

with probability 1. Define the set

$$\begin{aligned}L_N = \Bigg\{ x \in X : f(x) < f^* + \frac{1}{N} + \mu \max_{x,y\in X}\|x-y\| + \frac{1}{2}\alpha(\nu+C)^2 + \alpha TC\,(C+\nu) \\ + b\left(\sum_{i=1}^m C_i\right)\beta^{T+1}\max_{x,y\in X}\|x-y\|\Bigg\}. \end{aligned}$$

Let $x^* \in X^*$ and define the sequence $\hat{x}_k$ as follows:

$$\hat{x}_{k+1} = \begin{cases} x_{k+1} & \text{if } \hat{x}_k \notin L_N, \\ x^* & \text{otherwise.} \end{cases}$$

Thus, the process $\{\hat{x}_k\}$ is identical to the process $\{x_k\}$ until $\{x_k\}$ enters the set L_N. Define

$$\hat{d}_k(y) = \hat{x}_k - y \qquad \text{for any } y \in X.$$

Let $k \geq T$. Consider the case when $\hat{x}_k \in L_N$. Then, $\hat{x}_k = x^*$ and $\hat{x}_{k+1} = x^*$, so that $\hat{d}_k(x^*) = 0$ and $\hat{d}_{k+1}(x^*) = 0$, yielding

$$\mathsf{E}\left[\|\hat{d}_{k+1}(x^*)\|^2 \mid G_{n(k)}\right] = \mathsf{E}\left[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)}\right]. \tag{4.12}$$

Consider now the case when $\hat{x}_k \notin L_N$. Then, $\hat{x}_l = x_l$ and $x_l \notin L_N$ for all $l \leq k+1$. Therefore, by the definition of the set L_N, we have

$$\begin{aligned} f(x_{k-T}) - f^* \geq & \frac{1}{N} + \mu \max_{x,y\in X}\|x-y\| + \frac{1}{2}\alpha(\nu+C)^2 + \alpha TC\,(C+\nu) \\ & + b\left(\sum_{i=1}^m C_i\right)\beta^{T+1}\max_{x,y\in X}\|x-y\|. \end{aligned} \tag{4.13}$$

By using relations (4.10) and (4.13), we conclude that for $\hat{x}_k \notin L_N$,

$$\mathsf{E}\left[\|\hat{d}_{k+1}(x^*)\|^2 \mid G_{n(k)}\right] \leq \mathsf{E}\left[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)}\right] - \frac{2\alpha}{N}. \tag{4.14}$$

Therefore, from (4.12) and (4.14), we can write

$$\mathsf{E}\left[\|\hat{d}_{k+1}(x^*)\|^2 \mid G_{n(k)}\right] \leq \mathsf{E}\left[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)}\right] - \Delta_{k+1}, \tag{4.15}$$

where

$$\Delta_{k+1} = \begin{cases} 0 & \text{if } \hat{x}_k \in L_N, \\ \dfrac{2\alpha}{N} & \text{if } \hat{x}_k \notin L_N. \end{cases}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Observe that (4.15) satisfies the conditions of Lemma 3.2 with $u_k = \mathsf{E}[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)}]$, $\mathcal{F}_k = G_{n(k)}$, $q_k = 0$, $w_k = 2\Delta_{k+1}$, and $v_k = 0$. Thus, it follows that with probability 1,

$$\sum_{k=T}^{\infty} \Delta_{k+1} < \infty.$$

However, this is possible only if $\Delta_k = 0$ for all k sufficiently large. Therefore, with probability 1, we have $x_k \in L_N$ for all sufficiently large k. By letting $N \to \infty$, we obtain (4.11). ∎

Under Assumptions of Theorem 4.4, the function f is bounded over the set X, and by Fatou's lemma, we have

$$\mathsf{E}\left[\liminf_{k\to\infty} f(x_k)\right] \leq \liminf_{k\to\infty} \mathsf{E}[f(x_k)]\,.$$

It follows that the estimate of Theorem 4.4 also holds for $\mathsf{E}[\liminf_{k\to\infty} f(x_k)]$.

In the absence of errors ($\mu_k = 0$ and $\nu_k = 0$), the error bound in Theorem 4.4 reduces to

$$f^* + \frac{1}{2}\,\alpha C^2 + \alpha T C^2 + b\left(\sum_{i=1}^{m} C_i\right)\beta^{T+1} \max_{x,y\in X} \|x - y\|. \tag{4.16}$$

With respect to the parameter β, the error bound is obviously smallest when $\beta = 0$. This corresponds to uniform transition matrices $P(k)$, i.e., $P(k) = \frac{1}{m} ee^\top$ for all k (see Lemma 4.1). As mentioned, the Markov randomized method with uniform transition probability matrices $P(k)$ reduces to the incremental method with randomization in [22]. In this case, choosing $T = 0$ in (4.16) is optimal and the resulting bound is $f^* + \frac{\alpha}{2}C^2$, with $C = \max_{1\leq i\leq m} C_i$. We note that this bound is better by a factor of m than the corresponding bound for the incremental method with randomization given in Proposition 3.1 in [22].

When transition matrices are nonuniform ($\beta > 0$), and good estimates of the bounds C_i on subgradient norms and the diameter of the set X are available, one may optimize the error bound in (4.16) with respect to integer T for $T \geq 0$. In particular, one may optimize the term $\alpha T C^2 + b\,(\sum_{i=1}^m C_i)\,\beta^{T+1} \max_{x,y\in X}\|x - y\|$ over integers $T \geq 0$. It can be seen that the optimal integer T^* is given by

$$T^* = \begin{cases} 0 & \text{when } \dfrac{\alpha C^2}{C_0(-\ln\beta)} \geq 1, \\ \left\lceil (\ln\beta)^{-1} \ln\left(\dfrac{\alpha C^2}{C_0(-\ln\beta)}\right)\right\rceil - 1 & \text{when } \dfrac{\alpha C^2}{C_0(-\ln\beta)} < 1, \end{cases} \tag{4.17}$$

where $C_0 = b\,(\sum_{i=1}^m C_i) \max_{x,y\in X}\|x - y\|$.

A similar expression for optimal T^* in the presence of subgradient errors can be obtained, but it is rather cumbersome. Furthermore, such an expression (as well as the preceding one) may not be of practical importance when the bounds C_i, the diameter of the set X, and the bounds μ and ν on the error moments are "roughly" known. In this case, a simpler bound can be obtained by just comparing the values α and β, as given in the following.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

COROLLARY 4.5. *Let the conditions of Theorem* 4.4 *hold. Then,*

$$\liminf_{k\to\infty} \mathsf{E}[f(x_k)] \le f^* + \mu \max_{x,y\in X} \|x-y\| + \alpha\left[\frac{1}{2}(\nu+C)^2 + b\left(\sum_{i=1}^m C_i\right)\max_{x,y\in X}\|x-y\|\right] + \delta(\alpha,\beta),$$

where

$$\delta(\alpha,\beta) = \begin{cases} 0 & \text{if } \alpha \ge \beta, \\ \left\lceil \dfrac{\ln(\alpha)}{\ln(\beta)} \right\rceil - 1 & \text{if } \alpha < \beta. \end{cases}$$

Furthermore, with probability 1*, the same estimate holds for* $\inf_k f(x_k)$.

Proof. When $\alpha > \beta$ choose $T = 0$. In this case, from (Theorem 4.4) we get

$$\mathsf{E}[f(x_k)] \le f^* + \mu \max_{x,y\in X} \|x-y\| + \alpha\left(\frac{1}{2}(\nu+C)^2 + b\left(\sum_{i=1}^m C_i\right)\max_{x,y\in X}\|x-y\|\right).$$

When $\alpha < \beta$ we can choose $T = \left\lceil \frac{\ln(\alpha)}{\ln(\beta)} \right\rceil - 1$. Then, from (Theorem 4.4),

$$\mathsf{E}[f(x_k)] \ge f^* + \mu \max_{x,y\in X} \|x-y\| + \alpha\left[\frac{1}{2}(\nu+C)^2 + C\,(C+\nu)\left(\left\lceil \frac{\ln(\alpha)}{\ln(\beta)} \right\rceil - 1\right) + b\left(\sum_{i=1}^m C_i\right)\max_{x,y\in X}\|x-y\|\right]. \qquad \square$$

It can be seen that the error bounds in (4.17) and Corollary 4.5 converge to zero as $\alpha \to 0$. This is not surprising in view of the convergence of the method with a diminishing step-size.

As discussed earlier, the error bound in [11] is obtained assuming that there are no errors in subgradient evaluations and that the sequence of computing agents form a homogeneous Markov chain. Here, while we relax these assumptions, we make the additional assumption that the set X is bounded.

A direct comparison between the bound in Corollary 4.5 and the results in [11] is not possible. However, some qualitative comparisons on the nature of the bounds can be made. The bound in [11] is obtained for each individual agent's sequence of iterates (by sampling the iterates). This is a stronger result than our results in (4.17) and Corollary 4.5, which provide guarantees only on the entire iterate sequence (and not on the sequence of iterates at an individual agent). However, the bound in [11] depends on the entire network topology, through the probability transition matrix P of the Markov chain. Thus, the bound can be evaluated *only* when the complete network topology is available. In contrast, our bounds given in (4.17) and Corollary 4.5 can be evaluated without knowing the network topology. We require that the topology satisfies a connectivity assumption, as specified by Assumption 4, but we do not assume the knowledge of the exact network topology.

5. Discussion. Incremental algorithms form the middle ground between selfish agent behavior and complete network cooperation. Each agent can be viewed to be selfish, as it adjusts the iterate only using its own cost function. At the same time, the agents also cooperate by passing the iterate to a neighbor so that it may factor in its opinion by adjusting the iterate using its cost function. Theorems 3.3 and 4.3

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

show that a system level global optimum can still be obtained through some amount of cooperation.

The results we have obtained are asymptotic in nature. The key step in dealing with both the incremental algorithms was to obtain the basic iterate equation (Lemmas 3.1 and 4.2). This was then combined with standard stochastic analysis techniques to obtain asymptotic results. While we have restricted ourselves to establishing only convergence results, it is possible to combine the techniques in [21] with the basic iterate relation to obtain bounds on the expected rate of convergence of the algorithms. Finally, we have only listed a few possible applications for the results in this paper. The problem of aligning and coordinating mobile agents can also be cast in the optimization framework studied in this paper and the results obtained in this paper, especially the results on Markov stochastic subgradient algorithms, can be used to design suitable alignment algorithms.

An interesting extension that we plan to study in the future is an asynchronous version of the Markov randomized incremental algorithm. For the asynchronous version, dynamic step-size rules such as those studied in [6] are more appropriate than the off-line diminishing step-size rules considered in this paper.

Acknowledgments. We are very thankful to two anonymous referees for their valuable insights and suggestions that have improved this paper.

REFERENCES

[1] K. Barty, J.-S. Roy, and C. Strugarek, *Hilbert-valued perturbed subgradient algorithms*, Math. Oper. Res., 32 (2007), pp. 551–562.
[2] D. P. Bertsekas, A. Nedić, and A. Ozdalgar, *Convex Analysis and Optimization*, Athena Scientific, Belmont, MA, 2003.
[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Belmont, MA, 1997.
[4] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM J. Optim., 10 (2000), pp. 627–642.
[5] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with constant stepsize*, SIAM J. Optim., 18 (2007), pp. 29–51.
[6] V. Borkar, *Asynchronous stochastic approximations*, SIAM J. Control Optim., 36 (1998), pp. 840–851.
[7] Y. Ermoliev, *Stochastic Programming Methods*, Nauka, Moscow, 1976.
[8] A. Eryilmaz and R. Srikant, *Joint congestion control, routing, and MAC for stability and fairness in wireless networks*, IEEE J. Selected Areas Commun., 24 (2006), pp. 1514–1524.
[9] A. A. Gaivoronski, *Convergence properties of backpropogation for neural nets via theory of stochastic gradient methods. Part* 1, Optim. Methods Software, 4 (1994), pp. 117–134.
[10] A. Jadbabaie, J. Lin, and S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Trans. Automat. Control, 48 (2003), pp. 998–1001.
[11] 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.
[12] S. Kar and J. Moura, *Distributed consensus algorithms in sensor networks: Link and channel noise*, http://arxiv.org/abs/0711.3915, 2007.
[13] V. M. Kibardin, *Decomposition into functions in the minimization problem*, Autom. Remote Control, 40 (1980), pp. 1311–1323.
[14] J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist., 23 (1952), pp. 462–466.
[15] K. C. Kiwiel, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM J. Optim., 14 (2003), pp. 807–840.
[16] H. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.
[17] S. Li and T. Basar, *Asymptotic agreement and convergence of asynchronous stochastic algorithms*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 612–618.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[18] S. H. Low and D. E. Lapsley, *Optimization flow control* I: *Basic algorithm and convergence*, IEEE/ACM Transactions on Networking, 7 (1999), pp. 861–874.
[19] Z.-Q. Luo and P. Tseng, *Analysis of an approximate gradient projection method with applications to the backpropogation algorithm*, Optim. Methods Software, 4 (1994), pp. 85–101.
[20] S. Narayanan, G. Varatkar, D. Jones, and N. Shanbhag, *Computation as estimation: Estimation-theoretic IC design improves robustness and reduces power consumption*, in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2008.
[21] A. Nedić and D. P. Bertsekas, *Convergence rate of incremental algorithms*, Stochastic Optimization: Algorithms and Applications, (2001), pp. 223–264.
[22] A. Nedić and D. P. Bertsekas, *Incremental subgradient method for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.
[23] A. Nedić and D. P. Bertsekas, *The effect of deterministic noise in sub-gradient methods*, Technical report, Lab. for Information and Decision Systems, MIT, Boston, 2007.
[24] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *Distributed averaging algorithms and quantization effects*, http://arxiv.org/abs/0803.1202, 2008.
[25] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *Distributed subgradient algorithm and quantization effects*, in Proceedings of the Joint 47th IEEE Conference on Decision and Control (CDC '08), Cancun, Mexico, 2008.
[26] A. Nedić and A. Ozdaglar, *Distributed sub-gradient methods for multi-agent optimization*, Trans. Automat. Control, 64 (2008), pp. 48–61.
[27] A. Nedić, A. Ozdaglar, and P. A. Parrilo, *Constrained consensus*, Lab. for Information and Decision Systems, MIT, Report 2779, 2008.
[28] A. Olshevsky and J. N. Tsitsiklis, *Convergence speed in distributed consensus and averaging*, SIAM J. Control Optim., 48 (2009), pp. 33–55.
[29] B. T. Polyak, *Introduction to Optimization*, Optimization Software Inc., 1987.
[30] J. B. Predd, S. B. Kulkarni, and H. V. Poor, *Distributed learning in wireless sensor networks*, IEEE Signal Process. Mag., 23 (1999), pp. 56–69.
[31] M. G. Rabbat and R. D. Nowak, *Quantized incremental algorithms for distributed optimization*, IEEE Journal on Select Areas in Communications, 23 (2005), pp. 798–808.
[32] S. S. 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.
[33] S. S. Ram, V. V. Veeravalli, and A. Nedić, *Distributed recursive parameter estimation in parametrized linear state-space models*, 2008, http://arxiv.org/abs/0804.1607.
[34] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. of Math. Statist., 22 (1951), pp. 400–407.
[35] M. V. Solodov, *Incremental gradient algorithms with stepsizes bounded away from zero*, Computational Optimization and Algorithms, 11 (1998), pp. 23–35.
[36] M. V. Solodov and S. K. Zavriev, *Error stability properties of generalized gradient-type algorithms*, J. Optim. Theory Appl., 98 (1998), pp. 663–680.
[37] 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.
[38] J. N. Tsitsiklis, *Problems in decentralized decision making and computation*, Ph.D. thesis, Massachusetts Institute of Technology, Boston, 1984.
[39] J. N. Tsitsiklis and M. Athans, *Convergence and asymptotic agreement in distributed decision problems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 42–50.
[40] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Trans. Automat. Control, 31 (1986), pp. 803–812.
[41] L. Xiao, S. Boyd, and S.-J. Kim, *Distributed average consensus with least mean square deviation*, J. Parallel Distributed Computing, 67 (2007), pp. 33–46.

Downloaded 08/23/13 to 130.126.138.40. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.