

A RANDOMIZED INCREMENTAL SUBGRADIENT METHOD FOR DISTRIBUTED OPTIMIZATION IN NETWORKED SYSTEMS*

BJÖRN JOHANSSON, MABEN RABI, AND MIKAEL JOHANSSON[†]

Abstract. We present an algorithm that generalizes the randomized incremental subgradient method with fixed stepsize due to Nedić and Bertsekas. Our novel algorithm is particularly suitable for distributed implementation and execution, and possible applications include distributed optimization, e.g., parameter estimation in networks of tiny wireless sensors. The stochastic component in the algorithm is described by a Markov chain, which can be constructed in a distributed fashion using local information only. We provide a detailed convergence analysis of the proposed algorithm and compare it with existing, both deterministic and randomized, incremental subgradient methods.

Key words. convex programming, subgradient optimization, distributed optimization, markov chain

AMS subject classifications. 65K05, 90C25

1. Introduction. We consider the following convex optimization problem

$$\begin{aligned} & \text{minimize}_x && \sum_{n=1}^N f_n(x) \\ & \text{subject to} && x \in \mathcal{X}, \end{aligned} \tag{1.1}$$

where $f_n(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex functions and $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set. Let $f(x) = \sum_{n=1}^N f_n(x)$, and let f^* and x^* denote the optimal value and the optimizer of (1.1), respectively. To the problem we associate a connected N -node network, specified by the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The problem can now be interpreted as a *networked system*, where each node incurs a loss $f_n(x)$ of operating at x and nodes cooperate to find the optimal operating point (the optimizer x^* of (1.1)). In other words, each component in the objective function corresponds to a node in a network; see Fig. 1.1 for an example setup. The goal of this paper is to devise and analyze a novel distributed algorithm that iteratively solves (1.1) by passing an estimate of the optimizer between *neighboring nodes* in the network. There is a substantial interest in such algorithms, since centralized algorithms scale poorly with the number of nodes and are less resilient to failure of the central node. Moreover, peer-to-peer algorithms, that only exchange data between immediate neighbors, are attractive, since they make minimal assumptions on the networking support required for message passing between nodes. Application examples include estimation in sensor networks, coordination in multi-agent systems, and resource allocation in wireless systems; see, e.g., [6, 12].

One popular way of solving (1.1), is to use subgradient methods, which were pioneered by Shor [13] and recently unified in [8]. The methods' popularity stems from their ease of implementation and their capability of handling non-differentiable objective functions. Another key property is that subgradient methods often can be executed in a distributed fashion. The prototype subgradient method iteration for constrained convex minimization is

$$x_{k+1} = \mathcal{P}_{\mathcal{X}}\{x_k - \alpha_k h_k\},$$

*Parts of this research have previously been published in the proceedings of the IEEE Conference on Decision and Control 2007 [5]. The research was partially funded by the Swedish Research Council, the Swedish Foundation for Innovation Systems, and the European Commission.

[†]The authors are with the School of Electrical Engineering, Royal Institute of Technology (KTH), 100 44 Stockholm, Sweden. Email: `firstname.lastname@ee.kth.se`

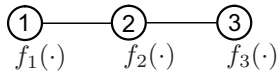


FIG. 1.1. *Example setup with three nodes. A line between nodes implies that they are connected.*

where $\mathcal{P}_{\mathcal{X}}\{\cdot\}$ denotes Euclidean projection on the feasible set \mathcal{X} , α_k is a stepsize, and h_k is a subgradient of the objective function at x_k ; there exist quite a few variations and extensions, but none of them fit our needs.

Naturally, the structure of the problem can be exploited and tailored algorithms, so called *incremental subgradient methods*, can be used. These algorithms are based on the iteration

$$x_{k+1} = \mathcal{P}_{\mathcal{X}} \{x_k - \alpha_k g_{n_k}(x_k)\}, \quad (1.2)$$

where $g_{n_k}(x_k)$ is a subgradient of the function $f_{n_k}(\cdot)$ at x_k , defined by

$$g_{n_k}(x_k) \in \partial f_{n_k}(x_k) = \{y \in \mathbb{R}^n \mid f_{n_k}(z) \geq f_{n_k}(x_k) + y^T(z - x_k), \forall z \in \mathcal{X}\}. \quad (1.3)$$

The set $\partial f_{n_k}(x_k)$ is called the subdifferential. Depending on how n_k and α_k are chosen, the resulting algorithms have rather diverse properties, and the stepsize, α_k , typically needs to be diminishing to insure asymptotic convergence of the iterates to an optimizer. To the authors' knowledge, most results on deterministic incremental subgradient methods are covered and unified in [8]. Although these methods were originally devised to boost convergence speed, they can also be used as decentralized mechanisms for optimization. A simple decentralized algorithm, proposed and analyzed in [10], is to use (1.2) with a fixed stepsize and let n_k cycle deterministically over the set $\{1, \dots, N\}$ in a round-robin fashion. We call this algorithm the *deterministic incremental subgradient method* (DISM). In [10], another variant, also suitable for distributed implementation, is proposed: it is a randomized algorithm where n_k is a sequence of independent and identically distributed random variables which take on values from the set $\{1, \dots, N\}$ with equal probability. We call this algorithm the *randomized incremental subgradient method* (RISM). If the problem setup permits, it is also possible to use incremental gradient methods [1]. In all of these methods, the iterate can be interpreted as being passed around between the nodes in the network. Finally, (1.2) is similar to the iterations used in stochastic approximation [9]. However, in stochastic approximation algorithms, the stepsize is typically diminishing and *not* fixed as it is in the algorithm we propose and analyze in this paper.

In the remainder of this paper, we will develop an algorithm that is based on (1.2), but where the sequence n_k is constructed such that *only neighbors* need to communicate with each other (in techspeak, this means that the network does not need to provide any multi-hop routing mechanism for the message passing). This is in contrast with both RISM and DISM, where nodes far apart need to communicate with each other. The outline is as follows: in Section 2, we present the novel algorithm, and in Section 3 we analyze its convergence properties. In Section 4, we compare, in the sense of performance bounds, the novel algorithm with existing algorithms. Finally, Section 5 concludes the paper with a discussion.

2. Algorithm. We associate an N -state Markov chain, MC , with the optimization problem (1.1). We make the following assumptions.

ASSUMPTION 1. *i) The functions $f_n : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex and the set \mathcal{X} is convex and non-empty. ii) The Markov chain MC is irreducible, aperiodic, and its stationary*

distribution is uniform. *iii) The subgradients are bounded,*

$$\sup \{ \|z\|_2 \mid z \in \partial f_n(x), n \in 1, \dots, N, x \in \mathcal{X} \} \leq C.$$

Remark: In general, the subgradients of a convex function are not bounded and the last assumption may seem to be rather restrictive. However, in several important cases this assumption is true, e.g., the functions $f_n(\cdot)$ are the pointwise maximum of a finite set of linear functions or the set \mathcal{X} is compact.

We are now ready to define our novel algorithm, which we denote the *Markov incremental subgradient method* (MISM). The iterations are defined as follows

$$x_{k+1} = \mathcal{P}_{\mathcal{X}} \{x_k - \alpha g_{w_k}(x_k)\}, k \geq 0, \quad (2.1)$$

where w_k is the state of MC.

Remark: The iterations (2.1) are interesting in its own right, and they generalize DISM and RISM. As we will see in Section 4, MISM reduces to DISM or RISM by choosing MC appropriately. However, MISM is particularly interesting in the context of distributed implementation. As we will see in the next section, by choosing MC in a special way, we can interpret the iterations as an estimate of the optimizer that is passed around between neighboring nodes and thereby iteratively improved.

2.1. Markov Chain for Distributed Execution. In this section we investigate what we need to be able to implement (2.1) in a decentralized fashion. We make the following assumptions on the graph associated with the optimization problem (1.1).

ASSUMPTION 2. *The undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is composed of N nodes and is connected.*

The assumption guarantees that there is a path between all nodes. The question is: how do we to construct, using only local information, the transition matrix of MC, P , such that the iterate, x_k , only jumps to an adjacent node? If the sparsity constraints $[P]_{ij} = 0$ when $(i, j) \notin \mathcal{E}$ are fulfilled, then the state of MC can only jump from state i to state j if $(i, j) \in \mathcal{E}$. The sparsity constraints therefore imply that the iterate in (2.1) is passed to a node that is adjacent to the current node.

It turns out there is a simple way to find such a Markov chain using the so called Metropolis-Hastings scheme, and we have the following lemma (see, e.g., [2]).

LEMMA 1. *Under Assumption 2, MC fulfills Assumption 1 and the sparsity constraints $[P]_{ij} = 0, (i, j) \notin \mathcal{E}$, if the elements of the transmission probability matrix of MC are set to*

$$[P]_{ij} = \begin{cases} \min\{\frac{1}{d_i}, \frac{1}{d_j}\} & \text{if } (i, j) \in \mathcal{E} \text{ and } i \neq j \\ \sum_{(i,k) \in \mathcal{E}} \max\{0, \frac{1}{d_i} - \frac{1}{d_k}\} & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (2.2)$$

where d_i is the number of edges of node i .

Note that each node only needs to know its neighbors' number of edges in order to be able to construct its part of the Markov chain.

3. Convergence Analysis. To show convergence we need some notation and three lemmas.

3.1. Technical Preliminaries. Denote the starting state of MC by i . Under Assumption 1 and with probability 1, all states in MC are visited equally often. Thus, we can form the subsequence $\{\bar{x}_k\}_{k=0}^\infty$ of $\{x_l\}_{l=0}^\infty$ by sampling it whenever the Markov chain visits the state i , i.e., whenever $w_l = i$. For example, if $w_0 = i$, $w_3 = i$, and $w_5 = i$, then

$$\begin{array}{ccccccc} w_0, & w_1, w_2, w_3, & w_4, w_5, & \dots & & & \\ \boxed{x_0}, & \underbrace{x_1, x_2, \boxed{x_3}}_{R_0^i}, & \underbrace{x_4, \boxed{x_5}}_{R_1^i}, & \dots & & & \\ \bar{x}_0, & \bar{x}_1, & \bar{x}_2, & \dots, & & & \end{array} \quad (3.1)$$

where $\bar{x}_0 = x_0$, $\bar{x}_1 = x_3$, and $\bar{x}_2 = x_5$. In addition, let R_k^i be the recurrence time for state i ,

$$R_k^i = \begin{cases} \inf \left\{ t - \sum_{m=0}^{k-1} R_m^i \mid w_t = i, t \geq \sum_{m=0}^{k-1} R_m^i + 1, t \in \mathbb{N} \right\}, & k > 0 \\ \inf \{ t \mid w_t = i, t \geq 1, t \in \mathbb{N} \}, & k = 0 \\ 0, & k < 0. \end{cases} \quad (3.2)$$

Successive recurrence times to a state form an independent and identically distributed sequence of random variables, due to the strong Markov property [11, Theorem 1.4.2], and we note that the statistics of R_k^i will not depend on k . Also note that R_k^i is independent of $\bar{x}_k, \bar{x}_{k-1}, \bar{x}_{k-2}, \dots$. Furthermore, we let $v_k^{i,j}$ be the random number of visits to state j during the time interval $[\sum_{m=0}^{k-1} R_m^i + 1, \sum_{m=0}^k R_m^i]$.

The first lemma concerns the average number of visits to other states over a recurrence time.

LEMMA 2. *Under Assumption 1, we have that*

$$(\mathbb{E}[v_k^{i,1}] \quad \dots \quad \mathbb{E}[v_k^{i,N}]) = \mathbf{1}_N^T \text{ and } \mathbb{E}[R_k^i] = N, \text{ for all } i = 1, \dots, N, k \geq 0,$$

where $\mathbf{1}_N$ denotes the N column vector with all entries equal to one.

Proof. Without loss of generality, we assume that $i = 1$ (we can always permute the states). Note that the statistics of $v_k^{i,1}$ and R_k^i do not depend on k due to the strong Markov property. Let MC have the following transition matrix

$$P = \begin{pmatrix} p_{11} & P_{12} \\ P_{21} & Q \end{pmatrix}.$$

Due to the definitions of R_k^i and $v_k^{i,j}$, it follows that $v_k^{1,1} = 1$. Now consider the Markov chain with transition matrix

$$P' = \begin{pmatrix} 1 & 0 \\ P_{21} & Q \end{pmatrix}.$$

This new chain has the same state space as the original and has the first row of the transition modified from $(p_{11} \ P_{12})$ to $(1 \ 0 \ \dots \ 0)$. Hence, state 1 is absorbing and all other states are transient. If the absorbing chain is started in a transient state $m + 1$, then the expected number of visits to a transient state $n + 1$ (including the starting position) is given by $[Z]_{mn}$, with $Z = (I - Q)^{-1} \in \mathbb{R}^{(N-1) \times (N-1)}$ [7, Theorem 3.2.4]. If we now consider the original Markov chain with transition matrix P , then if we start in state i , the expected number of visits to the other states before

returning to state 1 are given by $(\mathbb{E}[v_k^{1,2}] \dots \mathbb{E}[v_k^{1,N}]) = P_{12}Z$. Thus, we have $(\mathbb{E}[v_k^{1,1}] \dots \mathbb{E}[v_k^{1,N}]) = (1 \ P_{12}Z)$. It turns out that the vector of expected visits is an eigenvector to P since

$$(1 \ P_{12}Z)P = (p_{11} + P_{12}ZP_{21} \ P_{12}Z(I - Q + Q)) = (1 \ P_{12}Z),$$

where the last step follows from $ZP_{21} = \mathbf{1}_{N-1}$ [7, Theorem 3.3.7]. The transition matrix P has only one eigenvector with eigenvalue 1, namely the invariant distribution [7, Theorem 4.1.6]. Since P is assumed to have a uniform stationary distribution, we have that $(1 \ P_{12}Z) = \mathbf{1}_N^T$, which is the desired result. Finally, $\mathbb{E}[R_k^i] = \sum_{n=1}^N \mathbb{E}[v_k^{1,n}] = N$. \square

The second lemma concerns the second moment of the recurrence times, $\mathbb{E}[(R_k^i)^2]$.

LEMMA 3 ([7, Theorem 4.5.2]). *Under Assumption 1, the second moment of the recurrence time R_k^i is finite and given as*

$$\mathbb{E}[(R_k^i)^2] = 2[\Gamma]_{ii}N^2 - N,$$

with $\Gamma = (I - P + \lim_{n \rightarrow \infty} P^n)^{-1}$.

The last lemma establishes a bounding inequality that we will use in the convergence proof.

LEMMA 4. *Under Assumption 1, the sequence $\{\bar{x}_k\}_{k=0}^\infty$, formed by sampling the sequence $\{x_l\}_{l=0}^\infty$ whenever $w_l = i$ generated by (2.1), fulfills*

$$\mathbb{E}[\|\bar{x}_{k+1} - y\|_2^2 | \bar{x}_k] \leq \|\bar{x}_k - y\|_2^2 - 2\alpha(f(\bar{x}_k) - f(y)) + \alpha^2 C^2 K, \quad (3.3)$$

with $K = \max_i \mathbb{E}[(R_k^i)^2] < \infty$.

Proof. In this proof, we need to use both sequences $\{\bar{x}_k\}_{k=0}^\infty$ and $\{x_l\}_{l=0}^\infty$, and we need to keep track of which elements correspond to each other. For this purpose, let $l = \sum_{m=0}^{k-1} R_m^i$, so that $x_l = \bar{x}_k$ and $x_{l+R_k^i} = \bar{x}_{k+1}$. Using the non-expansion property of Euclidean projection, the definition of a subgradient, and the assumption that the subgradients are bounded, we have that, for any $y \in \mathcal{X}$,

$$\begin{aligned} \|x_{l+1} - y\|_2^2 &\leq \|x_l - y\|_2^2 - 2\alpha(g_{w_l}(x_l))^T(x_l - y) + \alpha^2 C^2 \\ &\leq \|x_l - y\|_2^2 - 2\alpha(f_{w_l}(x_l) - f_{w_l}(y)) + \alpha^2 C^2. \end{aligned}$$

Along the same lines of reasoning, we get the family of inequalities

$$\begin{aligned} \|x_{l+1} - y\|_2^2 &\leq \|x_l - y\|_2^2 - 2\alpha(f_{w_l}(x_l) - f_{w_l}(y)) + \alpha^2 C^2, \\ \|x_{l+2} - y\|_2^2 &\leq \|x_{l+1} - y\|_2^2 - 2\alpha(f_{w_{l+1}}(x_{l+1}) - f_{w_{l+1}}(y)) + \alpha^2 C^2, \\ &\vdots \\ \|x_{l+R_k^i} - y\|_2^2 &\leq \|x_{l+R_k^i-1} - y\|_2^2 - 2\alpha(f_{w_{l+R_k^i-1}}(x_{l+R_k^i-1}) - f_{w_{l+R_k^i-1}}(y)) + \alpha^2 C^2. \end{aligned}$$

Combining all of them together we get

$$\|x_{l+R_k^i} - y\|_2^2 \leq \|x_l - y\|_2^2 - 2\alpha \sum_{j=0}^{R_k^i-1} (f_{w_{l+j}}(x_{l+j}) - f_{w_{l+j}}(y)) + R_k^i \alpha^2 C^2,$$

which can be rewritten as

$$\begin{aligned} \left\| x_{l+R_k^i} - y \right\|_2^2 &\leq \|x_l - y\|_2^2 - 2\alpha \sum_{j=0}^{R_k^i-1} (f_{w_{l+j}}(x_{l+j}) - f_{w_{l+j}}(x_l)) \\ &\quad - 2\alpha \sum_{j=0}^{R_k^i-1} (f_{w_{l+j}}(x_l) - f_{w_{l+j}}(y)) + R_k^i \alpha^2 C^2. \end{aligned} \quad (3.4)$$

Notice that

$$f_{w_{l+j}}(x_l) - f_{w_{l+j}}(x_{l+j}) \leq \|g_{w_{l+j}}(x_l)\|_2 \|x_{l+j} - x_l\|_2 \leq C \|x_{l+j} - x_l\|_2 \leq \alpha j C^2.$$

This enables us to rewrite inequality (3.4) as follows

$$\left\| x_{l+R_k^i} - y \right\|_2^2 \leq \|x_l - y\|_2^2 - 2\alpha \sum_{j=0}^{R_k^i-1} (f_{w_{l+j}}(x_l) - f_{w_{l+j}}(y)) + \alpha^2 C^2 (R_k^i)^2.$$

Using $v_k^{i,j}$ as defined in Lemma 2, we express (3.4) as

$$\left\| x_{l+R_k^i} - y \right\|_2^2 \leq \|x_l - y\|_2^2 - 2\alpha \sum_{j=1}^N v_k^{i,j} (f_j(x_l) - f_j(y)) + \alpha^2 C^2 (R_k^i)^2. \quad (3.5)$$

Now, due to the Markov property and Lemma 2, we have

$$\mathbb{E} \left[\sum_{j=1}^N v_k^{i,j} (f_j(x_l) - f_j(y)) \middle| x_l, w_l \right] = \sum_{j=1}^N \mathbb{E}[v_k^{i,j}] (f_j(x_l) - f_j(y)) = f(x_l) - f(y). \quad (3.6)$$

Define

$$K = \max_{j \in \{1, \dots, N\}} \mathbb{E} \left[(R_k^j)^2 \right], \quad (3.7)$$

which is known to be finite and easily computable from Lemma 3. Note that $K \geq \mathbb{E} \left[(R_k^j)^2 \right] = \mathbb{E} \left[(R_k^j)^2 \mid x_l, w_l \right]$ for any $j \in \{1, \dots, N\}$. By taking the conditional expectation of (3.5) with respect to x_l and w_l , and using the equations (3.6) and (3.7), we obtain the desired result. \square

3.2. Proof of Convergence. Now we are ready for the main result of this paper.

THEOREM 1. *Let $\{x_l\}_{l=0}^\infty$ be generated by (2.1). Under Assumption 1 and with probability 1, we have the following:*

a) *The sequence $\{x_l\}_{l=0}^\infty$ fulfills*

$$\begin{cases} \liminf_{l \rightarrow \infty} f(x_l) = f^*, & \text{if } f^* = -\infty \\ \liminf_{l \rightarrow \infty} f(x_l) \leq f^* + \frac{\alpha C^2 K}{2}, & \text{if } f^* > -\infty. \end{cases}$$

b) If the set of all optimal x , $\mathcal{X}^* = \{x \in \mathcal{X} | f(x) = f^*\}$, is non-empty, then the sequence $\{x_l\}_{l=0}^\infty$ fulfills

$$\min_{0 \leq l \leq \tau} f(x_l) \leq f^* + \frac{\alpha C^2 K}{2} + \delta,$$

where τ is a stopping time with bounded expected value

$$\mathbb{E}[\tau] \leq \frac{N}{2\alpha\delta} \left(\text{dist}_{\mathcal{X}^*}(x_0) \right)^2,$$

where $\text{dist}_{\mathcal{X}^*}(x_0) = \inf\{\|x_0 - y\|_2 | y \in \mathcal{X}^*\}$.

Proof. We begin with showing a). Denote the starting state of MC by i . With probability 1, all states are visited equally often, and thus we can form the subsequence $\{\bar{x}_k\}_{k=0}^\infty$ of $\{x_l\}_{l=0}^\infty$ by sampling it whenever MC visits the state i ; see (3.1) for an illustration of this sampling.

We attack the problem using a similar approach as in the proof of Proposition 3.1 in [10]. The idea of the proof is to show that the iterates eventually will enter a special level set. For this purpose, let M and J be positive integers. We will now consider the sequences $\{x_l\}_{l=J}^\infty$ and $\{\bar{x}_k\}_{k=\bar{J}}^\infty$, where \bar{J} is chosen such that the first element of $\{\bar{x}_k\}_{k=\bar{J}}^\infty$ corresponds to the first element of $\{x_l\}_{l=J}^\infty$. This implies that $\bar{x}_{\bar{J}} = x_J$ and $\bar{J} \leq J$.

If the function $f(\cdot)$ is such that $\sup_{x \in \mathcal{X}} f(x) < f^* + \frac{1}{M}$, then the iterates trivially fulfill

$$x_l \in \left\{ x \in \mathcal{X} \mid f(x) < f^* + \frac{1}{M} \right\}, \forall l \geq 0.$$

Otherwise, if $\sup_{x \in \mathcal{X}} f(x) \geq f^* + \frac{1}{M}$, we can pick $y_M \in \mathcal{X}$ such that

$$f(y_M) = \begin{cases} -F, & \text{if } f^* = -\infty \\ f^* + \frac{1}{M}, & \text{if } f^* > -\infty, \end{cases}$$

for some $F \geq M$. We now define the special level set, L_M , which the iterates eventually will enter,

$$L_M = \left\{ x \in \mathcal{X} \mid f(x) \leq f(y_M) + \frac{1}{M} + \frac{\alpha C^2 K}{2} \right\}.$$

Note that this set includes y_M . To simplify the analysis, we derive a stopped sequence from $\{\bar{x}_k\}_{k=\bar{J}}^\infty$ by defining the sequence $\{\tilde{x}_k\}_{k=\bar{J}}^\infty$ as follows

$$\tilde{x}_k = \begin{cases} \bar{x}_k & \text{if } \bar{x}_j \notin L_M \forall j \in \{\bar{J}, \dots, k\} \\ y_M & \text{otherwise.} \end{cases}$$

When $\tilde{x}_k \notin L_M$, by setting $y = y_M$ in (3.3) in Lemma 4, we get

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - y_M\|_2^2 \mid \tilde{x}_k, w_k \right] \leq \|\tilde{x}_k - y_M\|_2^2 + \alpha^2 C^2 K - 2\alpha(f(\tilde{x}_k) - f(y_M)).$$

On the other hand, whenever $\tilde{x}_k \in L_M$, the sequence is forced to stay at y_M , and we have the trivial inequality

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - y_M\|_2^2 \mid \tilde{x}_k, w_k \right] \leq \|\tilde{x}_k - y_M\|_2^2 + 0.$$

If we define z_k through

$$z_k = \begin{cases} 2\alpha(f(\tilde{x}_k) - f(y_M)) - \alpha^2 C^2 K & \text{if } \tilde{x}_k \notin L_M \\ 0 & \text{if } \tilde{x}_k \in L_M, \end{cases}$$

we can write

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - y_M\|_2^2 \middle| \tilde{x}_k, w_k \right] \leq \|\tilde{x}_k - y_M\|_2^2 - z_k, \quad \forall k \geq \bar{J}. \quad (3.8)$$

When $\tilde{x}_k \notin L_M$, we have

$$(f(\tilde{x}_k) - f(y_M)) \geq \frac{1}{M} + \frac{\alpha C^2 K}{2},$$

which is equivalent to

$$z_k = 2\alpha(f(\tilde{x}_k) - f(y_M)) - \alpha^2 C^2 K \geq \frac{2\alpha}{M}. \quad (3.9)$$

If we take the expectation of (3.8), the result is

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - y_M\|_2^2 \right] \leq \mathbb{E} \left[\|\tilde{x}_k - y_M\|_2^2 \right] - \mathbb{E} [z_k], \quad \forall k \geq \bar{J},$$

and starting from $\tilde{x}_{\bar{J}}$, we recursively get

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - y_M\|_2^2 \right] \leq \|\tilde{x}_{\bar{J}} - y_M\|_2^2 - \mathbb{E} \left[\sum_{n=\bar{J}}^k z_n \right], \quad \forall k \geq \bar{J}. \quad (3.10)$$

Furthermore, from the iterations (2.1) and the bounded subgradients assumption, we have that $\|\tilde{x}_{\bar{J}} - y_M\|_2^2 \leq \|\tilde{x}_0 - y_M\|_2^2 + \alpha J C$. Let $\tilde{\tau}$ be the stopping time defined as

$$\tilde{\tau} = \inf \{ t \in \mathbb{N} \mid \tilde{x}_t \in L_M, t \geq \bar{J} \},$$

then $\tilde{\tau}$ is the number of non-zero elements in the non-negative sequence $\{z_k\}_{k=\bar{J}}^\infty$. Since z_k is non-negative, the series $\sum_{k=\bar{J}}^\infty z_k$ either converges to a finite real value or diverges to infinity. Thus, from (3.9) it follows that $\sum_{k=\bar{J}}^\infty z_k \geq \frac{2\alpha}{M} \tilde{\tau}$, where the left-hand side always is defined. By letting k go to infinity in (3.10) and using the non-negativity of a norm, we have

$$0 \leq \|\tilde{x}_0 - y_M\|_2^2 + \alpha J C - \mathbb{E} \left[\sum_{n=\bar{J}}^\infty z_n \right] \leq \|\tilde{x}_0 - y_M\|_2^2 + \alpha J C - \frac{2\alpha}{M} \mathbb{E} [\tilde{\tau}] \quad (3.11)$$

and the bound

$$\mathbb{E} [\tilde{\tau}] \leq \frac{M}{2\alpha} \left(\|\tilde{x}_0 - y_M\|_2^2 + \alpha J C \right) = \frac{M}{2\alpha} \left(\|x_0 - y_M\|_2^2 + \alpha J C \right). \quad (3.12)$$

Thus, the stopping time $\tilde{\tau}$ is almost surely finite and at least one element in the sequence $\{\tilde{x}_k\}_{k=\bar{J}}^\infty$ will be in the set L_M . Since $\{\tilde{x}_k\}_{k=\bar{J}}^\infty$ is a subsequence of $\{x_l\}_{l=\bar{J}}^\infty$, it follows that at least one element of $\{x_l\}_{l=\bar{J}}^\infty$ will be in the set L_M . Therefore, we have that

$$\inf_{l \geq \bar{J}} f(x_l) \leq f(y_M) + \frac{1}{M} + \frac{\alpha C^2 K}{2},$$

and since the choice of J is arbitrary and the right-hand side is independent of J , we have that

$$\liminf_{l \rightarrow \infty} f(x_l) \leq f(y_M) + \frac{1}{M} + \frac{\alpha C^2 K}{2}.$$

By letting M go to infinity and noting that Lemma 4 holds for all i , we have shown part a).

Now we proceed with part b), and the proof idea is the same as in part a); we show that the iterates eventually will enter a special level set. If the function $f(\cdot)$ is such that $\sup_{x \in \mathcal{X}} f(x) \leq f^* + \frac{\alpha C^2 K}{2} + \delta$ or $f(x_0) \leq f^* + \frac{\alpha C^2 K}{2} + \delta$, then the claim in b) is trivially fulfilled. Otherwise, let L_δ be the level set defined by

$$L_\delta = \left\{ x \in \mathcal{X} \mid f(x) \leq f^* + \frac{\alpha C^2 K}{2} + \delta \right\}.$$

Define the sequence $\{\tilde{x}_k\}_{k=0}^\infty$ as follows

$$\tilde{x}_k = \begin{cases} \bar{x}_k & \text{if } \bar{x}_j \notin L_\delta \forall j \leq k \\ \tilde{x} \in \mathcal{X}^* & \text{otherwise,} \end{cases}$$

where \tilde{x} is an arbitrary point in \mathcal{X}^* . When $\tilde{x}_k \notin L_\delta$, Lemma 4 gives us

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - \tilde{x}\|_2^2 \mid \tilde{x}_k, w_k \right] \leq \|\tilde{x}_k - \tilde{x}\|_2^2 + \alpha^2 C^2 K - 2\alpha (f(\tilde{x}_k) - f(\tilde{x})).$$

Otherwise, $\tilde{x}_k \in L_\delta$, the sequence will stay at \tilde{x} , and we have the trivial inequality

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - \tilde{x}\|_2^2 \mid \tilde{x}_k, w_k \right] \leq \|\tilde{x}_k - \tilde{x}\|_2^2 + 0.$$

By defining z_k through

$$z_k = \begin{cases} 2\alpha (f(\tilde{x}_k) - f(\tilde{x})) - \alpha^2 C^2 K & \text{if } \tilde{x}_k \notin L_\delta \\ 0 & \text{if } \tilde{x}_k \in L_\delta, \end{cases}$$

we have $z_k \geq 2\alpha\delta$ if $\tilde{x}_k \notin L_\delta$, and we can write

$$\mathbb{E} \left[\|\tilde{x}_{k+1} - \tilde{x}\|_2^2 \mid \tilde{x}_k, w_k \right] \leq \|\tilde{x}_k - \tilde{x}\|_2^2 - z_k, \quad \forall k. \quad (3.13)$$

Let $\tilde{\tau}$ be the stopping time defined as

$$\tilde{\tau} = \inf\{t \mid \tilde{x}_t \in L_\delta, t \geq 0, t \in \mathbb{N}\},$$

then $\tilde{\tau}$ is the random number of non-zero elements in the non-negative sequence $\{z_k\}_{k=0}^\infty$ and $\sum_{k=0}^\infty z_k \geq 2\alpha\delta\tilde{\tau}$, where the series $\sum_{k=0}^\infty z_k$ either converges to a finite real number or diverges to infinity. By letting k go to infinity in (3.13) and using the non-negativity of a norm, we have

$$0 \leq \|\tilde{x}_0 - \tilde{x}\|_2^2 - \mathbb{E} \left[\sum_{n=0}^{\infty} z_n \right] \leq \|\tilde{x}_0 - \tilde{x}\|_2^2 - 2\alpha\delta\mathbb{E}[\tilde{\tau}] \quad (3.14)$$

and the bound

$$\mathbb{E}[\tilde{\tau}] \leq \frac{1}{2\alpha\delta} \|\tilde{x}_0 - \tilde{x}\|_2^2 = \frac{1}{2\alpha\delta} \|x_0 - \tilde{x}\|_2^2. \quad (3.15)$$

Now let τ be the stopping time defined as

$$\tau = \inf\{t | x_t \in L_\delta, w_t = i, t \geq 0, t \in \mathbb{N}\}.$$

This means that the stopping conditions will be fulfilled when x_t is in the set L_δ and the Markov chain is in state i ; note that $f(x_\tau) \leq f^* + \frac{\alpha C^2 K}{2} + \delta$. By using the recurrence time R_k^i , which counts the number of elements in the original sequence $\{x_l\}_{l=0}^\infty$ between the elements in the sampled sequence $\{\bar{x}_k\}_{k=0}^\infty$, we can write

$$\tau = \sum_{k=1}^{\tilde{\tau}} R_{k-1}^i,$$

where $\tilde{\tau} \geq 1$ since $x_0 \notin L_\delta$ by assumption. Since $\tilde{\tau}$ is a stopping time for the sequence $\bar{x}_0, \bar{x}_1, \dots$, occurrence of the event $\{\tilde{\tau} \geq j\}$ is decided by the sequence $\bar{x}_0, \dots, \bar{x}_{j-1}$. In particular, $I\{\tilde{\tau} \geq j\} = \prod_{m=0}^{j-1} I\{\bar{x}_m \notin L_\delta\}$, where $I\{\cdot\}$ is the indicator function of the event $\{\cdot\}$. Furthermore, due to the construction of $\{\bar{x}_k\}_{k=0}^\infty$ and $\{R_k^i\}_{k=0}^\infty$, and the Markov property of the sequence $\{w_k\}_{k=0}^\infty$, the recurrence times $R_{j-1}^i, R_j^i, R_{j+1}^i, \dots$ are independent of $\bar{x}_{j-1}, \bar{x}_{j-2}, \bar{x}_{j-3}, \dots$. More specifically, we have that $\mathbb{E}[I\{\tilde{\tau} \geq j\} R_{j-1}^i] = \mathbb{E}\left[\prod_{m=0}^{j-1} I\{\bar{x}_m \notin L_\delta\} R_{j-1}^i\right] = \mathbb{P}\{\tilde{\tau} \geq j\} \mathbb{E}[R_{j-1}^i]$, where $\mathbb{P}\{\cdot\}$ denotes the probability of the event $\{\cdot\}$. Using the previous properties and a Wald's identity type of argument (see, e.g., [3, Theorem 5.5.3]), we have

$$\mathbb{E}[\tau] = \sum_{l=1}^{\infty} \mathbb{E}\left[I\{\tilde{\tau} = l\} \sum_{k=1}^{\tilde{\tau}} R_{k-1}^i\right] = \quad (3.16)$$

$$= \sum_{l=1}^{\infty} \sum_{k=1}^l \mathbb{E}[I\{\tilde{\tau} = l\} R_{k-1}^i] = \sum_{k=1}^{\infty} \sum_{l=k}^{\infty} \mathbb{E}[I\{\tilde{\tau} = l\} R_{k-1}^i] = \quad (3.17)$$

$$= \sum_{k=1}^{\infty} \mathbb{E}[I\{\tilde{\tau} \geq k\} R_{k-1}^i] = \sum_{k=1}^{\infty} \mathbb{P}\{\tilde{\tau} \geq k\} \mathbb{E}[R_{k-1}^i] = \mathbb{E}[\tilde{\tau}] \mathbb{E}[R_0^i] \leq \quad (3.18)$$

$$\leq \frac{N}{2\alpha\delta} \|x_0 - \tilde{x}\|_2^2. \quad (3.19)$$

The change of summation order in (3.17) holds since the series converges absolutely: $\sum_{k=1}^{\infty} \sum_{l=k}^{\infty} |\mathbb{E}[I\{\tilde{\tau} = l\} R_{k-1}^i]| = \sum_{k=1}^{\infty} \sum_{l=k}^{\infty} \mathbb{E}[I\{\tilde{\tau} = l\} R_{k-1}^i] = \mathbb{E}[\tilde{\tau}] \mathbb{E}[R_0^i] < \infty$, where we used the non-negativity of $\tilde{\tau}$ and R_k^i . The relation $\sum_{k=1}^{\infty} \mathbb{P}\{\tilde{\tau} \geq k\} = \mathbb{E}[\tilde{\tau}]$, used in (3.18), follows from [3, Theorem 3.2.1]. Since (3.19) holds for arbitrary \tilde{x} in \mathcal{X}^* , we can replace $\|x_0 - \tilde{x}\|_2^2$ with $(\text{dist}_{\mathcal{X}^*}(x_0))^2$. \square

Remark: The results in this section show that our proposed algorithm (2.1) can solve the optimization problem (1.1) in a distributed fashion relying only on neighbor-to-neighbor communication. Lemma 1 demonstrates how the forwarding probabilities (and hence the complete Markov chain) can be constructed by each node using only information from neighboring nodes. Theorem 1 establishes that the algorithm becomes increasingly accurate as α decreases, while the convergence rate becomes slower.

TABLE 4.1

Upper bounds of the expected number of iterations, $\mathbb{E}[\tau]$, needed to reach the accuracy $\min_{0 \leq l \leq \tau} f(x_l) \leq f^* + \gamma$. For brevity¹, let $D = NC^2\gamma^{-2} (\text{dist}_{\mathcal{X}^*}(x_0))^2$.

Algorithm	$\mathbb{E}\{\tilde{\tau}\}$
DISM	N^2D
RISM	ND
MISM	KD

4. Comparison with Existing Incremental Subgradient Algorithms. For the DISM and the RISM, there exist results of the same type as Theorem 1, i.e.,

$$\min_{0 \leq l \leq \tau} f(x_l) = f^* + \alpha\beta + \nu \text{ with } \mathbb{E}[\tilde{\tau}] \leq \frac{\rho}{\alpha\nu}, \quad (4.1)$$

where β and ρ are positive constants that depend on the algorithm. To compare the algorithms, we need to compute for each algorithm the minimum expected number of iterations needed for a given accuracy ($\min_{0 \leq l \leq \tau} f(x_l) = f^* + \gamma$). For the general case (4.1), we get the following optimization problem

$$\left\{ \begin{array}{l} \underset{\alpha, \nu}{\text{minimize}} \quad \frac{\rho}{\alpha\nu} \\ \text{subject to} \quad \alpha\beta + \nu \leq \gamma \\ \alpha \geq 0, \nu \geq 0 \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} \underset{\alpha, \nu}{\text{maximize}} \quad \alpha\nu \\ \text{subject to} \quad \alpha\beta + \nu = \gamma \\ \alpha \geq 0, \nu \geq 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} \alpha^* = \frac{\gamma}{2\beta} \\ \nu^* = \frac{\gamma}{2} \end{array} \right.$$

Using these optimal values, we compute an upper bound of the expected number of iterations, $\mathbb{E}[\tau]$, needed to reach the accuracy $\min_{0 \leq l \leq \tau} f(x_l) \leq f^* + \gamma$ for the DISM, RISM, and MISM. The results are presented in Table 4.1. Since

$$K \geq \mathbb{E} \left[(R_k^i)^2 \right] = \mathbb{E} \left[\left(\sum_{n=1}^N v_k(n) \right)^2 \right] \geq \mathbb{E} \left[\sum_{n=1}^N v_k(n)^2 \right] \geq \mathbb{E} \left[\sum_{n=1}^N v_k(n) \right] = N,$$

where we used the non-negativity and integrality of $v_k(n)$, the results in Table 4.1 indicate that the RISM is the best algorithm, and that the ranking between the DISM and the MISM will depend on the topology of the network as well as the transition probability matrix of the Markov chain. However, it is not only the expected number of iterations that are of interest; in applications, the ease of implementation and energy consumption are crucial. Experiments show that the MISM has favorable properties in these two respects, as reported in [5, 4], but this topic will not be further pursued in this paper.

It is interesting to note that we can recover the DISM and RISM from the MISM by choosing the transition probability matrix in the following way:

$$P_{\text{DISM}} = \begin{pmatrix} 0 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \vdots & & & \vdots \\ 1 & 0 & 0 & \dots \end{pmatrix} \text{ and } P_{\text{RISM}} = \begin{pmatrix} \frac{1}{N} & \dots & \frac{1}{N} \\ \vdots & \ddots & \vdots \\ \frac{1}{N} & \dots & \frac{1}{N} \end{pmatrix}$$

¹The constant C is defined in a slightly different way for DISM and RISM in [10]. There it is assumed that the norm of the subgradients for the actual trajectory of the algorithms are upper bounded by C . This is more general and less conservative than our definition of C , but it is very hard to check if it is fulfilled and therefore not practical. Our analysis also holds for the less conservative definition of C .

with

$$\mathbb{E}[R_{\text{DISM}}] = N^2 \text{ and } \mathbb{E}[R_{\text{RISM}}] = 2N^2 - N.$$

The transition matrix P_{DISM} will make the Markov chain deterministically explore the topology in a logical ring and $R_k^i = N$. Note that the Markov chain corresponding to P_{DISM} does not satisfy Assumption 1, since it does not have a stationary distribution and is periodic, but the analysis in Theorem 1 still applies. The transition matrix P_{RISM} will make the Markov chain jump to any node in the topology with equal probability at each time step, precisely as the RISM, and $\mathbb{E}[R_{\text{RISM}}] = 2N^2 - N$ by Lemma 3. The convergence bound given by the MISM analysis for P_{DISM} is identical with the convergence bound given by the DISM analysis. On the other hand, the convergence bound given by the MISM analysis for P_{RISM} is much worse than the original RISM result. This is due to the fact that in the original RISM analysis all iterates are analyzed, while in the MISM analysis, only those iterates at the arbitrary starting state are analyzed.

5. Conclusions. We have proposed a novel randomized incremental subgradient method that is well suited for decentralized implementation in distributed systems. The algorithm is a generalization of the RISM and DISM due to Nedić and Bertsekas. These algorithms can be recovered by choosing the transition probability matrix in a special way. The algorithm has been analyzed in detail with a convergence proof as well as a bound on the expected number of iterations needed to reach an *a priori* specified accuracy.

REFERENCES

- [1] D. BLATT, A. HERO, AND H. GAUCHMAN, *A convergent incremental gradient method with a constant step size*, SIAM J. Optim., 18 (2007), pp. 29–51.
- [2] S. BOYD, P. DIACONIS, AND L. XIAO, *Fastest mixing markov chain on a graph*, SIAM Review, 46 (2004), pp. 667–689.
- [3] K. L. CHUNG, *A Course in Probability Theory*, Academic Press, 1974.
- [4] B. JOHANSSON, C. CARRETTI, AND M. JOHANSSON, *On distributed optimization using peer-to-peer communications in wireless sensor networks*, in Proceedings of IEEE SECON, June 2008.
- [5] B. JOHANSSON, M. RABI, AND M. JOHANSSON, *A simple peer-to-peer algorithm for distributed optimization in sensor networks*, in Proceedings of IEEE CDC, Dec. 2007.
- [6] B. JOHANSSON, A. SPERANZON, M. JOHANSSON, AND K. H. JOHANSSON, *On decentralized negotiation of optimal consensus*, Automatica, 44 (2008), pp. 1175–1179.
- [7] J. G. KEMENY AND J. L. SNELL, *Finite Markov Chains*, Van Nostrand, 1960.
- [8] K. C. KIWIEL, *Convergence of approximate and incremental subgradient methods for convex optimization*, SIAM J. Optim., 14 (2004), pp. 807–840.
- [9] H. J. KUSHNER AND G. G. YIN, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer, 2003.
- [10] A. NEDIĆ AND D. P. BERTSEKAS, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optim., 12 (2001), pp. 109–138.
- [11] J. R. NORRIS, *Markov Chains*, Cambridge University Press, 1998.
- [12] M. RABBAT AND R. NOWAK, *Distributed optimization in sensor networks*, in Proceedings of ACM/IEEE IPSN, 2004.
- [13] N. Z. SHOR, *Minimization methods for non-differentiable functions*, Springer-Verlag, 1985.