DISTRIBUTED OPTIMIZATION IN MULTI-AGENT SYSTEMS: APPLICATIONS TO DISTRIBUTED REGRESSION

BY

SUNDHAR RAM SRINIVASAN

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2010

Urbana, Illinois

Doctoral Committee:

Professor Venugopal V. Veeravalli, Chair
Assistant Professor Angelia Nedić
Professor Rayadurgam Srikant
Professor Sean P. Meyn
Professor Douglas L. Jones

ABSTRACT

The context for this work is cooperative *multi-agent systems* (MAS). An *agent* is an intelligent entity that can measure some aspect of its environment, process information and possibly influence the environment through its action. A cooperative MAS can be defined as a loosely coupled network of agents that interact and cooperate to solve problems that are beyond the individual capabilities or knowledge of each agent.

The focus of this thesis is *distributed stochastic optimization* in multi-agent systems. In *distributed optimization*, the complete optimization problem is not available at a single location but is distributed among different agents. The distributed optimization problem is additionally stochastic when the information available to each agent is with stochastic errors. Communication constraints, lack of global information about the network topology and the absence of coordinating agents make it infeasible to collect all the information at a single location and then treat it as a centralized optimization problem. Thus, the problem has to be solved using algorithms that are distributed, i.e., different parts of the algorithm are executed at different agents, and local, i.e., each agent uses only information locally available to it and other information it can obtain from its immediate neighbors.

In this thesis, we will primarily focus on the specific problem of minimizing a *sum* of functions over a constraint set, when each component function is known partially (with stochastic errors) to a unique agent. The constraint set is known to all the agents. We propose three distributed and local algorithms, establish asymptotic convergence with diminishing stepsizes and obtain rate of

convergence results. Stochastic errors, as we will see, arise naturally when the objective function known to an agent has a random variable with unknown statistics. Additionally, stochastic errors also model communication and quantization errors. The problem is motivated by distributed regression in sensor networks and power control in cellular systems.

We also discuss an important extension to the above problem. In the extension, the network goal is to minimize a global function of a sum of component functions over a constraint set. Each component function is known to a unique network agent. The global function and the constraint set are known to all the agents. Unlike the previous problem, this problem is not stochastic. However, the objective function in this problem is more general. We propose an algorithm to solve this problem and establish its convergence.

To my father, mother and brother.
Yes, this was fun. No, this was not easy.

And to all those who asked me "So what do you work on?",
If you really cared you would read this.

ACKNOWLEDGMENTS

This thesis would not have been possible without my advisers, Prof. V. V. Veeravalli and Prof. A. Nedić. I am grateful to Prof. Veeravalli for his continuous support and patience. Without his emphasis on the practical aspect of things, I would have gotten lost in the bad world of αs and βs. Prof. Nedić is a great source of inspiration as a researcher and as a person. I could not have wished for a better mentor for my Ph.D thesis and I would like to thank her for the enormous amount of time she spent in helping me develop this thesis. I would also like to thank my masters thesis advisor, Prof. D. Manjunath (IIT Bombay). He taught me my first lessons on research and these lessons will stay with me for life.

Besides my advisers, I would like to thank my committee — Prof. R. Srikant, Prof. D. Jones and Prof. S. Meyn — for their insights and excellent advice. Special thanks are due to Prof. T. Coleman for the wonderful course on random processes and the friendly smile in the corridors. I would also like to thank Terri Hovde (CSL), Barbara Horner (CSL) and Debbie Hilligoss (Industrial Engg.) for smoothing out the numerous administrative kinks for me.

When I dedicated this thesis to my parents and brother, I really meant it. When I finished my master's, I had absolutely no clue what I wanted to do. Their insistence that I study further is the main reason I decided to do this Ph.D. While I am at it, I would like to "unthank" my mother for her constant stream of unsolicited, unwarranted and unappreciated advice on how I should keep the house clean, wear warm clothes in the winter and travel safely. I would like to "unthank" my father for refusing to call me a real scientist as I did not

wear a white lab coat to work. And finally I would like to "unthank" my brother for threatening to call up my advisers and request that they hold me back for a couple of years more so that I would get a "good" Ph.D.

Unlike most, I think I had more fun in graduate school than during my undergraduate studies. I owe this largely to the following people:

- 511 gumbal: Karthik (JK JK JK), Dinesh (Touser pandi), Shankar (CID), Jayanand (Kung Fu Panda), Rajan (Baniyan pota saniyan), Venkat (Mama), Kumaresh (Kums), Sriram (Loose Sriram), Aswin (BLEEP mama), Vishal (Suruuthi) and Anjan (Thatha).
- Kappor gumbal: Nikhil, Jayanta, Aftab, Khan and Country
- Cricket teammates
- CSL labmates: Sivakumar (GSK), Sreekant, Nilesh, Adnan, Jay, Che Lin, Jason and others

And finally a special thanks to the Indian cricket team for finally winning a few games.

TABLE OF CONTENTS

NOTATION

Upper case letters denote matrices and random quantities. Lower case letters denote parameters and constants.

i, j, k, n, m, p: Non-negative integers

$\{x_1, \ldots, x_n\}$: Set consisting of the elements $x_1, \ldots, x_n$

$\forall$: Universal qualifier

$\Re^n$: n-dimensional Euclidean space

$f(x)$: A function f of x

$F(x)$: Random variable F that is parameterized by x or a matrix F that is a function of x

$[x]_i$: i-th component of a vector x in $\Re^n$

$\|x\|$: Euclidean norm of a vector x in $\Re^n$

$|x|$: Asbolute value of a scalar x

e_i: Unit vector in $\Re^n$ with i-th component equal to 1

e: Vector with each entry equal to 1

$P_X[x]$: Euclidean projection of a point x onto a set X

A^T: Transpose of a matrix or vector A

$[A]_{i,j}$: (i, j)-th entry of a matrix A

A^{-1}: Inverse of matrix A

$[A]_i$: i-th row of matrix A

$[A]^j$: j-th column of matrix A

$\nabla f(x)$: Gradient of a function $f(x)$, $x \in \Re^n$

$f'(x)$: Derivative of a function $f(x)$, $x \in \Re$

$\nabla^2 f(x)$: Matrix of second partial derivatives of $f(x)$, $x \in \Re^n$

$\partial f(x)$: Subgradient of a convex function $f(x)$, $x \in \Re^n$

$\operatorname{argmin}_X f(x)$: Any global minimum point of $f(x)$ over the set X

$\operatorname{Argmin}_X f(x)$: Set of all global minimum points of $f(x)$ over the set X

$(\Omega, \mathcal{F}, \mathsf{P})$: Underlying probability space

$\mathsf{E}[X]$: Expected value of a random vector X

$\sigma(X)$: $\sigma-$algebra generated by random vector X

$\mathsf{E}[X \mid Y = y]$: Expectation of a random vector X conditioned on the vector Y taking value y

$\mathsf{E}[X \mid \mathcal{G}]$: Expectation of a random vector X conditioned on the σ-algebra $\mathcal{G}$

CHAPTER 1

INTRODUCTION

1.1 Context for this Work

The context for this work is cooperative *multi-agent systems* (MAS). An *agent* is an intelligent entity that can measure some aspect of its environment, process information and possibly influence the environment through its action. A cooperative MAS can be defined as a loosely coupled network of agents that interact and cooperate to solve problems that are beyond the individual capabilities or knowledge of each individual agent [1]. These systems have the following characteristics:

- *Distributed*: There are no central co-ordinating agents and each agent is autonomous.

- *Limited connectivity:* An agent can interact, i.e., communicate, only with a subset of the agents. Typically, the agent interactions are modeled by a communication graph with the agents as the nodes. Two agents are neighbors on the graph, i.e., connected by a graph edge, if they can communicate with each other. See Fig. 1.1 for an illustration.

- *Local view*: Each agent in the MAS has a local view of the environment, i.e., an agent can only measure and control its local environment. In addition, each agent has a limited view of the MAS. The only information that an agent can have about the MAS is the identity of its immediate neighbors. The agent can have no global information about the MAS

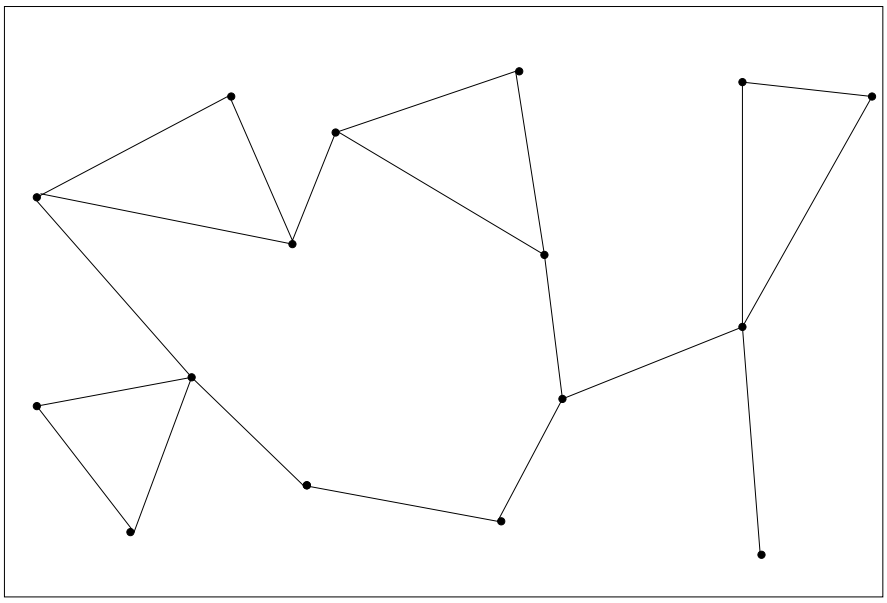

Figure 1.1: A MAS communication graph. The graph nodes are the agents and neighbors are agents that can communicate with each other.

including the number of agents in the MAS and features of the communication graph.

Multi-agent systems encompass a wide range of disciplines including animal behavior [2], social sciences [3], computer animations [4], artificial intelligence [5] and massive sensing applications in engineering. Though the scope of this work is general and can be adapted to different co-operative multi-agent systems, we will study it from an engineering perspective and primarily focus on wireless sensor networks (WSN) [6]. Sensor networks consist of a large number of spatially deployed sensors that sense their local environment across time. Each sensor is equipped with on-board processing units and communication units that allow it to communicate with other sensors over the wireless medium. Over and above the features in a generic MAS, sensors networks have the following additional features:

- *Limited energy:* The sensors in a sensor network are powered by a fixed battery supply and communicating over the wireless medium is a power intensive activity. Thus, to extend the network lifetime each agent must

communicate only limited information.

- *Unreliable communication:* Communication over the wireless medium is unreliable and suffers from quantization and channel errors. In addition, if the sensors move then the network connectivity graph may change with time.

1.2 Focus of this Thesis

The focus of this thesis is on *distributed stochastic optimization* in multi-agent systems. In *distributed optimization*, the complete optimization problem is not available at a single location but is distributed among different agents. Communication constraints, lack of global information about the network topology and the absence of coordinating agents make it infeasible to collect all the information at a single location and then treat it as a centralized optimization problem. Thus, the problem has to be solved using algorithms that are

- *Distributed:* In a distributed algorithm, different parts of the algorithm are executed by different agents, possibly simultaneously, without any coordinating agents.

- *Local:* Each agent uses only information locally available to it and other information it can obtain from its immediate neighbors.

- *Communication efficient:* Each agent must exchange minimal information with its neighbors.

The distributed optimization problem is additionally stochastic when the information available to each agent is with stochastic errors.

In this thesis, we will primarily deal with the specific problem of minimizing a sum of functions over a constraint set, when each component function is known partially (with stochastic errors) to a unique agent. The constraint set is known

to all the agents. We propose three distributed and local algorithms, establish asymptotic convergence with diminishing stepsizes and obtain rate of convergence results. Stochastic errors, as we will see, arise naturally when the objective function known to an agent has a random variable with unknown statistics. Additionally, stochastic errors also model communication and quantization errors.

We also discuss an important extension to the above problem. In the extension, the network goal is to minimize a global function of a sum of component functions over a constraint set. Each component function is known to a unique network agent. The global function and the constraint set are known to all the agents. Unlike the previous problem, this problem is not stochastic. However, the objective function in this problem is more general. We propose an algorithm to solve this problem and establish its convergence.

The optimization algorithms developed are then used to address the problem of distributed regression in sensor networks and distributed power control in cellular systems.

1.3 Organization of Thesis

The thesis is organized as follows. In Chapter 2 we discuss the problem, algorithms and an overview of the complete thesis. This chapter provides a succint summary of the thesis. In Chapters 3, 4 and 5 we discuss the details and proofs of the convergence of the algorithms. In Chapter 6 we discuss the generalization. Chapters 7 and 8 discuss applications of the algorithm to the problem of distributed regression in sensor networks and power control in cellular systems. Some broad research directions are discussed in Chapter 9.

CHAPTER 2

PROBLEM, ALGORITHM AND RESULTS

2.1 System Goal

Consider a multi-agent system of m agents indexed by $1, \ldots, m$. When convenient we will also use $V = \{1, \ldots, m\}$. We make the following assumption.

Assumption 1 *The agents are time synchronized.*

The system goal is to solve the following optimization problem:

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

where $X \subseteq \Re^n$ is a constraint set and $f_i : O \to \Re$ for all i. Here O is an open set containing X. The problem is a distributed stochastic optimization problem because the function f_i is known only partially to agent i. By partially, we mean that the agent can only obtain a noisy estimate of the function gradient. The goal is to solve problem (2.1) using an algorithm that is distributed and local.

Related to the problem, we use the following notation:

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

We are interested in the case when the problem in (2.1) is convex. Specifically, we assume that the following assumption holds.

Assumption 2 *The functions f_i and the set X are such that*

(a) The set X is closed and convex.

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

We make no assumption on the differentiability of the functions f_i. At points where the gradient does not exist, we use the notion of subgradients. A vector ∇f_i is a subgradient of f_i at a point $x \in \mathrm{dom} f$ if the following relation holds:

$$\nabla f_i(x)^T(y - x) \leq f_i(y) - f_i(x) \qquad \text{for all } y \in \mathrm{dom}\, f. \tag{2.2}$$

Since the set X is contained in an open set over which the functions are defined and convex, a subgradient of f_i exists at any point of the set X (see [7] or [8]).

We will 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 stepsize, e.g., [9, 10].

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.,*

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

Assumption 3 holds, for example, when each f_i is a polyhedral function or when the set X is compact. In some of the results, we will directly assume that the set X is compact and thus we will not make Assumption 3.

2.2 Algorithms

We next discuss three distributed and local algorithms to solve the problem in (2.1). To conceptualize, we assume iteration k of these algorithms is performed at time k and that each iteration is performed instantaneously.

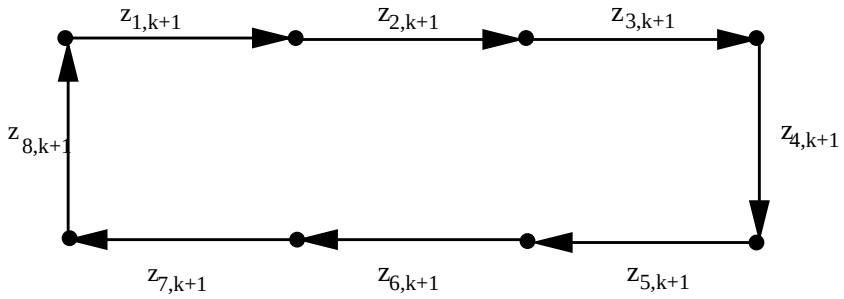

Figure 2.1: A network of 8 agents with cyclical incremental processing. The estimate is cycled through the network. The quantity $z_{i,k}$ is the intermediate value after agent i updates at time $k+1$.

2.2.1 Cyclic incremental algorithm

Each agent designates another agent as an upstream neighbor and another agent as a downstream neighbor so that they form a cycle. See Fig. 2.1 for an example. Without loss of generality, we will index the upstream neighbor of agent i as $i+1$, with the understanding that the upstream neighbor of agent m is agent 1. In iteration $k+1$, agent i receives the current iterate $z_{i-1,k+1}$ from agent $i-1$, updates the iterate using the gradient of its local function, evaluated with errors, and passes it to its upstream neighbor. The update rule is

$$\begin{aligned} z_{0,k+1} &= z_{m,k} = x_k, \\ z_{i,k+1} &= P_X \left[z_{i-1,k+1} - 2\alpha_{k+1} \left(\nabla f_i(z_{i-1,k+1}) + \epsilon_{i,k+1} \right) \right], \end{aligned} \tag{2.3}$$

where the initial iterate x_0 is chosen at random. Here, α_{k+1} is the stepsize, $\epsilon_{i,k+1}$ is the stochastic error at agent i, P_X denotes Euclidean projection onto set X and ∇f_i is the (sub)gradient of function f_i. Note that the algorithm does not have a central coordinating agent, and hence it is distributed. Further, agent i only uses gradient information of its function f_i, and hence the algorithm is local.

This algorithm is not suited for all multi-agent systems. First, the network must be "sufficiently" connected in every time slot for a cycle to exist. Second, even if a cycle exists, the agents would need to identify a suitable upstream and downstream neighbor in a distributed and local manner, e.g., using the

algorithm in [11]. Every time the network graph changes the cycles may have to be recomputed. Therefore, the algorithm requires the following assumption on the communication graph.

Assumption 4 *The communication graph does not change with time and has a Hamiltonian cycle.*

2.2.2 Markov incremental algorithm

In this algorithm, the order in which the agents update the iterate is not fixed and could be random. Suppose in iteration k, agent s_k updates and generates the estimate x_k. Then, agent s_k may either pass this estimate to a neighbor s_{k+1}, with probability

$$[P]_{s_k,s_{k+1}} = a_{s_k,s_{k+1}}(k+1),$$

or choose to keep the iterate with the remaining probability, in which case $s_{k+1} = s_k$. See Fig. 2.2 for an illustration. The update rule for this method is given by

$$x_{k+1} = P_X\left[x_k - \alpha_{k+1}\left(\nabla f_{s_{k+1}}(x_k) + \epsilon_{s_{k+1},k+1}\right)\right], \tag{2.4}$$

where $x_0 \in X$ is some random initial vector. Observe that the value of s_{k+1} depends only on the value of s_k. Thus, the sequence of agents $\{s_k\}$ can be viewed as a Markov chain with states $V = \{1, \ldots, m\}$ and transition matrix P.

Note that the algorithm does not have a central coordinating agent and hence it is distributed. Further, the algorithm requires agent i to use only gradient information related to f_i and requires no coordinating agent. Therefore, it is local and distributed. As we will see, the algorithm converges when the network topology satisfies the following assumption.

Assumption 5 *Let $G_k = (V, E_k)$ be the communication graph at time k. There exists an integer $Q \geq 1$ such that the graph $\left(V, \cup_{l=k}^{k+Q-1} E(\ell)\right)$ is strongly*

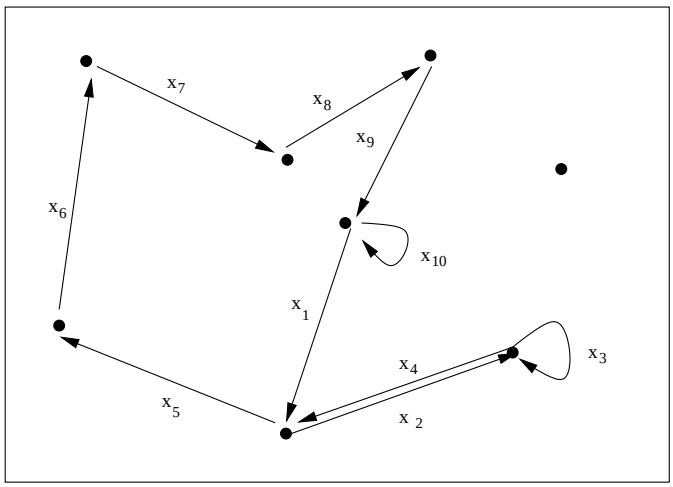

Figure 2.2: A network of 8 agents with incremental processing. The estimate is randomly passed through the network.

connected for all k.

Note that the network can change with time. Further, the network need not be connected in every time slot. It needs to be connected only in blocks of Q.

2.2.3 Parallel algorithm

In this algorithm each agent maintains and updates an iterate sequence. This is fundamentally different from the incremental algorithms in which a single iterate sequence is incrementally updated by the agents. We will use $w_{i,k}$ to denote the iterate with agent i at the end of time slot k. One iteration of the algorithm is performed in each sampling interval. Each agent receives the current iterate of its present neighbors. We denote agent i's neighbors at time $k+1$ by $N_i(k+1)$. See Fig. 2.3 for an illustration. Each agent then calculates the following weighted sum $v_{i,k}$ given by

$$v_{i,k} = \sum_{j \in N_{i,k+1}} a_{i,j}(k+1) w_{j,k} + \left(1 - \sum_{j \in N_{i,k+1}} a_{i,j}(k+1)\right) w_{i,k}.$$

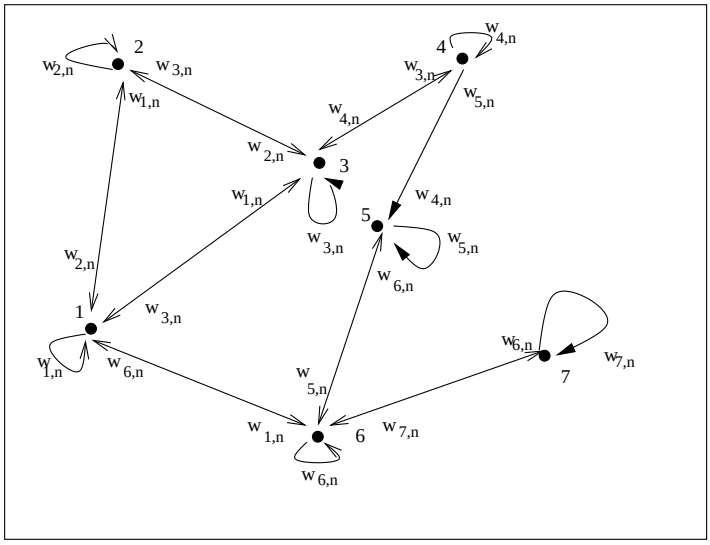

Figure 2.3: A network of 8 agents with parallel processing. Each agent shares its current iterate with its neighbors.

Here $a_{i,j}(k)$ are the weights. Agent i then obtains its new iterate $w_{i,k+1}$ from $v_{i,k}$ according to

$$w_{i,k+1} = P_X \left[v_{i,k} - 2\alpha_{k+1} \left(\nabla f_{s_{k+1}}(v_{i,k}) + \epsilon_{i,k+1} \right) \right]. \tag{2.5}$$

The initial points $\{w_{i,0}\}, i \in V$, are chosen randomly.

As in the Markov incremental algorithm one iteration of the algorithm is performed in the time slot $(k, k+1)$. Note that the algorithm does not have a central coordinating agent and hence it is distributed. Further, agent i only uses gradient information of its function f_i and hence the algorithm is local. The restrictions on the network are similar to Assumption 5.

2.3 Sources of Stochastic Errors

In this section we discuss the main sources of the error $\epsilon_{i,k}$ in the subgradient evaluation in the algorithms.

2.3.1 Random objective function

The primary source of stochastic errors in the subgradient evaluation is when the objective function is not completely known to the agent and has some randomness in it. This is the case in recursive regression (Chapter 7).

Let the function $f_i(x)$ be given by $f_i(x) = \mathsf{E}[\psi_i(x, R_i)]$, where R_i is a random variable with statistics that are independent of x. The statistics of R_i are not available to agent i and hence the function f_i is not known to agent i. Instead, agent i observes samples of R_i in time. Thus, in a subgradient algorithm for minimizing the function, the subgradient must be suitably approximated using the observed samples. In the Robbins-Monro stochastic approximation [12], the subgradient $\nabla f_i(x)$ is approximated by $\nabla \psi_i(x, r_i)$, where r_i denotes a sample of R_i. Thus, the parallel Robbins-Monro stochastic optimization algorithm is

$$w_{i,k+1} = P_X \left[v_{i,k} - \alpha_{k+1} \nabla \psi_i \left(v_{i,k}, r_{i,k+1} \right) \right], \tag{2.6}$$

where $r_{i,k+1}$ is a sample of R_i obtained at time k. The expression for the error is

$$\epsilon_{i,k+1} = \nabla \psi_i(v_{i,k}, r_{i,k+1}) - \mathsf{E}[\nabla \psi_i(v_{i,k}, R_i)] .$$

Let us next consider the case when $f_i(x) = \mathsf{E}[\psi_i(x, R_i(x))]$, where $R_i(x)$ is a random variable that is parametrized by x. To keep the discussion simple, let us assume that $x \in \Re$. As in the preceding case, the statistics of $R_i(x)$ are not known to agent i, but the agent can obtain samples of $R_i(x)$ for any value of x. In the Kiefer-Wolfowitz approximation [12],

$$\nabla f_i(x) \approx \frac{\psi_i \left(x, r_i(x + \beta) \right) - \psi_i \left(x, r_i(x) \right)}{\beta},$$

where $r_i(x)$ is a sample of the random variable $R_i(x)$. The corresponding

distributed optimization algorithm is

$$w_{i,k+1} = P_X \left[v_{i,k} - \alpha_{k+1} \frac{\psi_i \left(v_{i,k}, r_i(v_{i,k} + \beta_{i,k+1}) \right) - \psi_i \left(v_{i,k}, r_i(v_{i,k}) \right)}{\beta_{i,k+1}} \right],$$

where $\beta_{i,k+1}$ is a positive scalar. In this case, the error is

$$\epsilon_{i,k+1} = \frac{\psi_i \left(v_{i,k}, r_i(v_{i,k} + \beta_{i,k+1}) \right) - \psi_i \left(v_{i,k}, r_i(v_{i,k}) \right)}{\beta_{i,k+1}} - \nabla f_i(v_{i,k}).$$

If the function ψ_i is differentiable then $\mathsf{E}[\epsilon_{i,k+1} \mid v_{i,k}]$ is of the order $\beta_{i,k+1}$. Thus, the errors can be controlled through the sequence $\{\beta_{i,k}\}$.

2.3.2 Quantization errors

Typically, the iterates are first quantized before they are communicated in wireless networks. We will discuss the scalar case. The quantization lattice is a discrete set of values in $\Re$. For a vector $x \in \Re$, $Q[x]$ denotes its quantized value.

When there is quantization, the choice of stepsize in iterative algorithms is critical. It seems reasonable to require the algorithms, with quantization, to converge to $Q[x^*]$, i.e., the lattice point that is the closest to x^*. Consider the standard gradient descent algorithm to minimize $f(x)$ over the set $X = \Re$ with a basic quantizer without any dither. The iterates are generated according to

$$x_{k+1} = Q \left[x_k - \alpha_k \nabla f(x_k) \right].$$

Suppose x_0 is a lattice point but not $Q[x^*]$, and $\alpha_k < \frac{\Delta}{2C}$, where C is the bound on the subgradient of f_i; then it can be immediately seen that $x_k = x_0$, for all k. More generally, one can conclude that stepsizes should always be large enough to push the iterate from a non-optimal lattice point, but small enough to ensure that the iterate gets caught in the optimal lattice point. Thus, in the presence of quantization, non-diminishing step-sizes should be used.

In the dither quantizer, the quantized value of a vector $x \in X$ is the Euclidean

projection of x added with a dither signal. Thus, $Q[x] = P_Q[x + D]$, where D is a dither signal whose components are uniformly and independently chosen in $[\frac{-\Delta}{2}, \frac{\Delta}{2}]$. In this case, $Q[x] - x$, is random, statistically independent of x and uniformly distributed in $\left[\frac{-\Delta}{2}, \frac{\Delta}{2}\right]$ [13, 14]. Thus we can write the unconstrained cyclic incremental algorithm with quantization as

$$\begin{aligned} z_{i,k+1} &= Q\left[z_{i-1,k+1} - 2\alpha \nabla f_i(z_{i-1,k+1})\right] \\ &= z_{i-1,k+1} - 2\alpha\left(\nabla f_i(z_{i-1,k+1}) + \epsilon_{i,k+1}\right) \end{aligned}$$

where $\epsilon_{i,k+1}$ is zero mean i.i.d. error that is uniform in $\left[\frac{-\Delta}{2\alpha}, \frac{\Delta}{2\alpha}\right]$.

2.4 Overview of Convergence Results

We will make the following assumption on the errors. Define F_k^i as the σ-algebra generated by all the random variables till agent i computes the gradient of f_i with error $\epsilon_{i,k}$. In essence, F_k^i is the entire history[1] of the algorithm till the exact step when the error $\epsilon_{i,k}$ occurs.

Assumption 6 *There exist deterministic scalar sequences $\{\mu_k\}$ and $\{\nu_k\}$ such that*

$$\|\mathsf{E}\left[\epsilon_{i,k} \mid F_k^{i-1}\right]\| \leq \mu_k \qquad \textit{for all } i \textit{ and } k,$$

$$\mathsf{E}\left[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\right] \leq \nu_k^2 \qquad \textit{for all } i \textit{ and } k.$$

Assumption 6 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

$$\left\|\mathsf{E}\left[\epsilon_{i,k} \mid F_k^{i-1}\right]\right\| \leq \sqrt{\mathsf{E}\left[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\right]} \leq \nu_k. \tag{2.7}$$

We will also, without any loss of generality, assume that $\mu_k < \nu_k$.

[1]Depending on the algorithm the σ-algebra is defined differently.

For the three algorithms we obtain three different forms of convergence result. All the results obtained require Assumptions 1, 2, 3 or boundedness of set X, and Assumption 6.

Theorems 1, 5 and 10 are the basic convergence result and obtain sufficient conditions on the errors and the stepsizes for the iterate sequences to converge to an optimal point in some probabilistic sense. Typically, the stepsize α_k must be chosen so that $\sum_k \alpha_k = \infty$ and $\sum_k \alpha_k^2 < \infty$. In addition, the sequence $\{\mu_k\}$ must diminish fast enough so that $\sum_k \alpha_k \mu_k < \infty$ and the sequence $\{\nu_k\}$ must be uniformly bounded. The value of this result is primarily theoretical and is essentially a statement of correctness of the proposed algorithms. Theorems 2, 6, and 11 provide error bounds on the performance of the algorithm when a constant stepsize is used and the errors do not diminish. The bound obtained is a function of $\limsup \mu_k$ and $\limsup \nu_k$. This result is useful in understanding the effects of quantization as discussed in Section 2.3.2. Theorems 4, 7 and 12 characterize the "expected" performance of the algorithm after a finite number of iterations. These results are expected to be useful in determining stopping times for the algorithm.

2.5 Comparison of the Algorithms

In this section we compare the algorithms along different metrics.

- *Rate of convergence:* When a stepsize of $\frac{1}{k}$ is used, all the algorithms converge as $O\left(\frac{1}{k}\right)$. This can be seen from Theorems 4, 7 and 12. These results only provide an upper bound on the rate of convergence. Therefore, a direct comparison of the constants in the results is not meaningful.

- *Communication requirements:* A quantitative measure of the communication requirements cannot be obtained. However, we can make some qualitative observations about the relative communication requirements of the algorithms. Intuitively, one expects the cyclic

incremental algorithm to perform better than the Markov incremental algorithm. This is because the Markov incremental algorithm is random and therefore the iterate might get "caught" in some part of the network. Between the cyclic and parallel algorithm it is not clear which algorithm will perform better. These observations are validated in the simulation results in Chapter 7.

- *Setup phase:* The cyclic incremental algorithm requires a setup phase where the agents identify a cycle in a distributed and local manner. The other algorithms do not require a setup phase.

- *Network connectivity requirements:* The cyclic incremental algorithm requires the communication graph to not change with time. In addition, the cyclic incremental requires the graph to have a cycle, which is a stronger requirement than connectivity. The Markov incremental and the parallel algorithms allow the network to change with time and require only a weak form of periodic connectivity.

- *Strength of convergence result:* Theorems 1 and 10 prove that the cyclic incremental and the parallel algorithms converge to an optimal point with probability 1. The corresponding result for the Markov incremental guarantees convergence in a weaker sense. In addition, a smaller range of stepsizes is allowed in the Markov incremental algorithm.

2.6 Extension: General Distributed Optimization

We next discuss an important extension to the problem discussed in (2.1). The network goal is to solve the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \tilde{f}(x) := g\left(\sum_{i=1}^{m} h_i(x)\right) \\ \text{subject to} \quad & x \in X, \end{aligned} \tag{2.8}$$

where $g : \Re \rightarrow \Re$, $X \subseteq \Re^p$, $h_i : X \rightarrow \Re$ for all $i \in V$. Different parts of the optimization problem are known to different agents in the network. The function h_i is known only to agent i. The function g and the set X are globally known, i.e., to every agent.

We assume Assumption 5 on the network topology. We make the following assumptions on the functions.

Assumption 7 *The functions g and h_i, $i \in V$, and the set X in (2.8) satisfy the following:*

(a) The set X is convex and closed.

(b) The set X is bounded, i.e., there exists a scalar $D > 0$ such that $\sup_{x \in X} \|x\| \leq D$.

(c) The function $\tilde{f}$ is convex over an open set that contains the set X.

(d) The functions g and h_i are differentiable. Further, g' and ∇h_i are Lipschitz continuous with constant L.

Assumption 7(a) and 7(c) imply that (2.8) is a convex optimization problem. From Assumption 7(b) and 7(d) we can conclude that the norms of the gradients, i.e., $|g'|$ and $\|\nabla h_i\|$, are bounded. We denote the bound on the gradients by C.

Observe that there are three key differences between (2.8) and (2.8). First, observe that the objective function in (2.8) is a generalization of the objective function in (2.1). Second, unlike (2.1), the problem in (2.8), is not a stochastic optimization problem. There are no stochastic errors. Third, note that Assumption 7(d) imposes additional restrictions on the objective function.

We next describe an iterative algorithm to solve (2.8). At the end of iteration k, agent i maintains two statistics: $x_{i,k}$ and $s_{i,k}$. The statistic $x_{i,k}$ is agent i's estimate of an optimal point and $s_{i,k}$is agent i's estimate of $\frac{1}{m} \sum_{i=1}^{m} h_i(x_{i,k})$.

In the next iteration, agent i recursively generates $x_{i,k+1}$ and $s_{i,k+1}$ as follows:

$$\begin{aligned}
\begin{bmatrix} \bar{x}_{i,k} \\ \bar{s}_{i,k} \end{bmatrix} &= \sum_{j \in N_i(k+1)} a_{i,j}(k+1) \begin{bmatrix} x_{i,k} \\ s_{i,k} \end{bmatrix}, \\
x_{i,k+1} &= P_X \left[\bar{x}_{i,k} - \alpha_{k+1} g' \left(m \bar{s}_{i,k} \right) \nabla h_i \left(\bar{x}_{i,k} \right) \right], \\
s_{i,k+1} &= \bar{s}_{i,k} + h_i(x_{i,k+1}) - h_i(x_{i,k}).
\end{aligned}$$

In the $(k+1)$-th iteration, agent i receives $x_{i,k}$ and $s_{i,k}$ from its current immediate neighbors and calculates weighted averages $\bar{x}_{i,k}$ and $\bar{s}_{i,k}$. The weighted average is then updated using locally available information (functions g and h_i, and set X) to generate $x_{i,k+1}$ and $s_{i,k+1}$. The algorithm is initialized with $x_{i,0} \in X$ and $s_{i,0} = h_i(x_{i,0})$ for all $i \in V$.

2.7 Related Literature and Main Contributions

In parallel optimization, the optimization problem is completely known to all the agents and the emphasis is on distributing the processing to reduce the computational burden on a single processor. See [15] for a complete exposition on this topic.

To the best our knowledge this is the first study that deals with distributed stochastic optimization problems. All prior literature deals with deterministic distributed optimization problems. The first paper to formulate the distributed optimization problem was [16], where the incremental subgradient algorithm of [9] was used. A modified cyclic incremental algorithm was proposed in [17]. A version of the Markov incremental algorithm for constant stepsizes was analyzed in [18]. The parallel algorithm for the unconstrained problem was proposed in [19] and then extended in [20, 21]. The parallel algorithm was based on the distributed consensus algorithm that was studied in [15, 20–29]. In addition, since we are interested in the effect of stochastic errors, the thesis is also related to the literature on stochastic gradient methods [30–32].

We next summarize the main contributions made in this thesis.

- This thesis is the first study of the effect of stochastic errors on distributed optimization algorithms. The stochastic errors made by each agent propagate across agents and time, introducing statistical dependence between the agent iterates.

- The convergence analysis for the Markov incremental algorithm, for the error-free case, is available only for the constant stepsize case [18]. Thus, the convergence analysis for diminishing stepsize for the Markov incremental algorithm is an important contribution.

- Chapter 6 identifies a new class of distributed optimization problems that allow distributed and local solution. The problem is very general and we expect it to be useful in a number of applications.

- The thesis formulates regression in vertically and horizontally distributed data as distributed optimization problems and uses the algorithms developed to solve these problems.

2.8 Discussion

We have proposed three algorithms and discussed convergence results for these. We next discuss some relevant issues.

2.8.1 Asynchronous algorithms

The three algorithms proposed require the agents to be time synchronized. A natural extension would be to study algorithms that are asynchronous and do not require the agents to have a common clock. One approach is to base the optimization algorithms on the gossip averaging algorithm of [33]. This is work in progress and we briefly discuss the algorithm. Some initial results are presented in [34].

The algorithm requires the network topology to not change with time. Let $N(i)$ be the set of neighbors of agent i, i.e., $N(i) = \{j \in V : \{i, j\} \in E\}$. Each agent has a local clock that ticks at a Poisson rate of 1. At each tick of its clock, agent i averages its iterate with a randomly selected neighbor $j \in N(i)$, where each neighbor has an equal chance of being selected. Agents i and j then adjust their averages along the negative direction of ∇f_i and ∇f_j, respectively, which are computed with stochastic errors.

As in [33] we will find it easier to study the gossip algorithms in terms of a single virtual clock that ticks whenever any local Poisson clock ticks. Thus, the virtual clock ticks according to a Poisson process with rate m. Let Z_k denote the k-th tick of the virtual clock and let I_k denote the index of the agent whose local clock actually ticked at that instant. The fact that the Poisson clocks at each agent are independent implies that I_k is uniformly distributed in the set V. In addition, the memoryless property of the Poisson arrival process ensures that the process $\{I_k\}$ is i.i.d. Let J_k denote the random index of the agent communicating with agent I_k. Observe that J_k, conditioned on I_k, is uniformly distributed in the set $N(I_k)$. Let $x_{i,k-1}$ denote agent i iterate at time immediately before Z_k. The iterates evolve according to

$$x_{i,k} = \begin{cases} \bar{x}_{I_k,J_k} - \frac{1}{\Gamma_k(i)} \left(\nabla f_i \left(\bar{x}_{I_k,J_k} \right) + \epsilon_{i,k} \right) & \text{if } i \in \{I_k, J_k\} \\ x_{i,k-1} & \text{otherwise,} \end{cases} \tag{2.9}$$

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

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

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

While the algorithm is asynchronous, convergence has been established only for networks that do not change with time. One direction of future research is to develop algorithms that can be shown to converge over time-varying networks.

2.8.2 Non-convex optimization

This thesis primarily deals with convergence when the optimization problem is convex. It is possible to obtain convergence results under the following alternate assumption.

Assumption 8 *The functions $f_i(x)$ are differentiable with gradient that is Lipschitz continuous. Further, the gradient $\nabla f_i(x)$ is bounded and the set X is $\Re^n$.*

A point $x^* \in \Re$ is a stationary point of $f(x)$ if $\nabla f(x^*) = 0$. A global minimum of $f(x)$ is also a stationary point of $f(x)$. Typically, when the objective function is non-convex and iterative methods are employed, the iterates may converge to a stationary point. Observe that, in view of Lipschitz continuity of the gradient, the assumption that the gradients are bounded, i.e., Assumption 3, is equivalent to the following standard assumption.

Assumption 9 *The iterate sequences are bounded with probability 1.*

This assumption is implicit and not very easy to establish. However, this is a standard assumption in stochastic optimization literature and we refer the reader to Chapter 3 of [35] for a discussion of techniques to verify this assumption.

It is possible to obtain results similar to Theorems 1, 5 and 10 with Assumption 2 and convergence to an optimal point replaced with Assumption 8 and convergence to a stationary point, respectively. The proof is along the lines of those in [34] and [36].

CHAPTER 3

CYCLIC INCREMENTAL ALGORITHM

In this chapter we study the properties of the cyclic incremental algorithm discussed in Section 2.2.1. The algorithm is

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

where 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 stepsize and $\mathcal{P}_X$ denotes Euclidean projection onto the set X.

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 sub-iteration is not necessarily a subgradient of the objective function f. For this reason, we cannot directly apply the classic stochastic approximation convergence results of [12, 30, 37] to study the convergence of method in (3.1).

3.1 Basic Iterate Relation

We first derive a key lemma. Define $d_k(y) = x_k - y$ and $d_{i,k+1}(y) = z_{i,k+1} - y$ for all k. We will make Assumption 6 and the σ-algebra $\mathcal{F}_k^i$ is the σ-algebra generated by $\epsilon_{1,1}, \ldots, \epsilon_{1,m}, \epsilon_{2,1}, \ldots, \epsilon_{i-1,k}$.

Lemma 1 *Let Assumptions 1, 2, 3, 4 and 6 hold. Then, the iterates generated*

by algorithm (3.1) are such that for any stepsize rule and for any $y \in X$,

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

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

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

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

$$\begin{aligned}\mathsf{E}\left[\|d_{i,k+1}(y)\|^2 \mid F_{k+1}^{i-1}\right] \leq & \|d_{i-1,k+1}(y)\|^2 - 2\alpha_{k+1}d_{i-1,k+1}(y)^T\nabla f_i(z_{i-1,k+1}) \\ & - 2\alpha_{k+1}d_{i-1,k+1}(y)^T\mathsf{E}\left[\epsilon_{i,k+1} \mid F_{k+1}^{i-1}\right] \\ & + \alpha_{k+1}^2\mathsf{E}\left[\left\|\epsilon_{i,k+1} + \nabla f_i(z_{i-1,k+1})\right\|^2 \mid F_{k+1}^{i-1}\right]. \qquad (3.2)\end{aligned}$$

We now estimate the last two terms in the right-hand side of the preceding equation by using Assumption 6 on the error moments. In particular, we have for all i and k,

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

Next, we estimate the last term in (3.2) by using Assumption 6 on the error

moments in and Assumption 3 on the subgradient norms. We have all i and k,

$$\begin{aligned}\mathsf{E}\big[\|\epsilon_{i,k+1}+\nabla f_i(z_{i-1,k+1})\|^2 \mid F_{k+1}^{i-1}\big] =&\mathsf{E}\big[\|\epsilon_{i,k+1}\|^2 \mid F_{k+1}^{i-1}\big] + \|\nabla f_i(z_{i-1,k+1})\|^2\\ &+ 2\nabla f_i(z_{i-1,k+1})^T\mathsf{E}\big[\epsilon_{i,k+1} \mid F_{k+1}^{i-1}\big]\\ =&\mathsf{E}\big[\|\epsilon_{i,k+1}\|^2 \mid F_{k+1}^{i-1}\big] + \|\nabla f_i(z_{i-1,k+1})\|^2\\ &+ 2\|\nabla f_i(z_{i-1,k+1})\|\|\mathsf{E}\big[\epsilon_{i,k+1} \mid F_{k+1}^{i-1}\big]\|\\ \leq&(\nu_{k+1}+C_i)^2,\end{aligned}$$

where in the last inequality we use $\mathsf{E}\big[\|\epsilon_{i,k}\|^2 \mid F_k^{i-1}\big] \leq \nu_k^2$ and $\big\|\mathsf{E}\big[\epsilon_{i,k} \mid F_k^{i-1}\big]\big\| \leq \nu_k$ for all k [cf. Eqn. (2.7)]. Combining the preceding two relations and the inequality in (3.2), we obtain for all $y \in X$,

$$\begin{aligned}\mathsf{E}\big[\|d_{i,k+1}(y)\|^2 \mid F_{k+1}^{i-1}\big] \leq&\|d_{i-1,k+1}(y)\|^2 - 2\alpha_{k+1}d_{i-1,k+1}(y)^T\nabla f_i(z_{i-1,k+1})\\ &+ 2\alpha_{k+1}\mu_{k+1}\|d_{i-1,k+1}(y)\| + \alpha_{k+1}^2(\nu_{k+1}+C_i)^2. \qquad (3.3)\end{aligned}$$

We now estimate the second term in the right-hand side of the preceding relation. From the subgradient inequality in (2.2) we have

$$\begin{aligned}-d_{i-1,k+1}(y)^T\nabla f_i(z_{i-1,k+1}) &= -(z_{i-1,k+1}-y)^T\nabla f_i(z_{i-1,k+1})\\ &\leq -\left(f_i(z_{i-1,k+1})-f_i(y)\right)\\ &= -\left(f_i(x_k)-f_i(y)\right)-\left(f_i(z_{i-1,k+1})-f_i(x_k)\right)\\ &\leq -\left(f_i(x_k)-f_i(y)\right)\\ &\quad -\left(\nabla f_i(x_k)\right)^T\left(z_{i-1,k+1}-x_k\right) \qquad (3.4)\\ &\leq -\left(f_i(x_k)-f_i(y)\right)+C_i\left\|z_{i-1,k+1}-x_k\right\|. \qquad (3.5)\end{aligned}$$

In (3.4) we have again used the subgradient inequality (2.2) to bound $f_i(z_{i-1,k+1})-f_i(x_k)$, while in (3.5) we have used the subgradient norm bound from Assumption 3. 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 non-expansive property of the projection, we further have

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

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

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

By substituting the preceding estimate in the inequality in (3.3), we obtain for all $y \in X$,

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

Taking the expectation conditional on F_k^m, we obtain

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

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. Eqn. (2.7)]. Summing over $i = 1, \ldots, m$, and noting

that $d_{0,k+1}(y) = x_k - y$, we see that

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

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 for all $y \in X$, and all i and k,

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

□

3.2 Convergence Results

We next use Lemma 1 to obtain convergence results for the algorithm with diminishing and constant stepsizes. Further, we also obtain a rate of convergence result.

3.2.1 Diminishing stepsizes

We first study the convergence of the method in (3.1) for diminishing stepsize rule.

Theorem 1 *Let Assumptions 1, 2, 3, 4 and 6 hold. Assume that the stepsize 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 and in mean square.* □

Proof First note that all the assumptions of Lemma 1 are satisfied. Let x^* be an arbitrary point in X^*. By letting $y = x^*$ in Lemma 1, we obtain for any $x^* \in X^*$,

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

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

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

Taking conditional expectations, we further obtain

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

where we have used Assumption 6 and Jensen's inequality to bound $\mathsf{E}[\|\epsilon_{j,k+1}\| \mid F_k^m]$ by ν_{k+1}. Using the preceding inequality, we have

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

Next, using the inequality

$$2\|d_k(x^*)\| \leq 1 + \|d_k(x^*)\|^2,$$

we obtain

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

By the assumptions on the stepsize, and the sequences $\{\mu_k\}$ and $\{\nu_k\}$, we

further have

$$\sum_{k=0}^{\infty} m\alpha_{k+1}\mu_{k+1} < \infty,$$

$$\sum_{k=0}^{\infty} 2\alpha_{k+1}^2\mu_{k+1}\sum_{i=1}^{m}\sum_{j=1}^{i-1}(C_j+\nu_{k+1}) \le 2\sum_{k=0}^{\infty}\sum_{i=1}^{m}\sum_{j=1}^{i-1}\left(\alpha_{k+1}^2\mu_{k+1}C_j+\alpha_{k+1}^2\nu_{k+1}^2\right) < \infty,$$

$$\sum_{k=0}^{\infty}\alpha_{k+1}^2\left(m\nu_{k+1}+\sum_{i=1}^{m}C_i\right)^2 \le 2\sum_{k=0}^{\infty}\alpha_{k+1}^2\left(m^2\nu_{k+1}^2+\left(\sum_{i=1}^{m}C_i\right)^2\right) < \infty,$$

where in the second relation above, we have used $\mu_{k+1} \le \nu_{k+1}$ [cf. Eqn. (2.7)], while in the last inequality, we have used $(a+b)^2 \le 2(a^2+b^2)$ valid for any scalars a and b. Thus, the conditions of Lemma 16 are satisfied with $u_k = \|d_k(x^*)\|^2$, $\mathcal{F}_k = F_k^m$, $q_k = m\alpha_{k+1}\mu_{k+1}$, $v_k = 2\alpha_{k+1}\left(f(x_k)-f^*\right)$ 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 non-negative random variable for every $x^* \in X^*$. Also with probability 1, we have

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

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 $x^* \in X^*$ in view of continuity of f. Hence, the sequence $\{x_k\}$ converges to some vector in X^* with probability 1.

3.2.2 Constant stepsizes

Here, we study the behavior of the iterates $\{x_k\}$ generated by the method (3.1) with a constant stepsize 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 [38].

Theorem 2 *Let Assumptions 1, 2, 4 and 6 hold. Also, assume that the set X is bounded and the sequence $\{x_k\}$ is generated by the method (3.1) with a constant stepsize rule, i.e., $\alpha_k = \alpha$ for all $k \geq 1$. Let*

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

We then have

$$\liminf_{k\to\infty} \mathsf{E}[f(x_k)] \leq f^* + m\mu \max_{x,y\in X} \|x-y\| + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2, \tag{3.8}$$

and with probability 1,

$$\inf_{k\geq 0} f(x_k) \leq f^* + m\mu \max_{x,y\in X} \|x-y\| + \frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2. \tag{3.9}$$

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 1 are satisfied. Furthermore, the optimal set X^* is non-empty. 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 1, we have for $y = x^* \in X^*$,

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2 \mid F_k^m\right] \leq & \|d_k(x^*)\|^2 - 2\alpha\left(f(x_k) - f^*\right) + 2m\alpha\mu \max_{x,y} \|x-y\| \\ & + \alpha^2\left(\sum_{i=1}^m C_i + m\nu\right)^2. \end{aligned} \tag{3.10}$$

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

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

Now, assume that the relation (3.8) does not hold. Then there will exist a $\gamma > 0$ and an index k_γ such that for all $k > k_\gamma$,

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

Therefore, for $k > k_\gamma$, we have

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

Hence, for $k \geq k_\gamma$,

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

For sufficiently large k, the right-hand side of the preceding relation is negative, yielding a contradiction. Thus the relation (3.8) must hold.

We now prove the relation in (3.9). 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

$$\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid F_k^m\Big] = \hat{d}_k(x^*). \tag{3.11}$$

When $\hat{x}_k \notin L_N$, $\hat{x}_k = x_k$ and $\hat{x}_{k+1} = x_{k+1}$. Using relation (3.10), 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 $\hat{x}_k \notin L_N$,

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

Therefore, from (3.11) and (3.12), we can write

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

where

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

Observe that (3.13) satisfies the conditions of Lemma 16 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 with probability 1,

$$\sum_{k=0}^{\infty} \Delta_{k+1} < \infty.$$

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 (3.9).

As seen from relation (3.9) of Theorem 2, the error bound on the "best function" value $\inf_k f(x_k)$ depends on the stepsize 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 2 hold with $\mu = 0$. The resulting error bound is $\frac{\alpha}{2}\left(\sum_{i=1}^m C_i + m\nu\right)^2$, which can be controlled with the stepsize α. However, this result also holds when the boundedness of X is relaxed by requiring subgradient boundedness instead, as seen in the following theorem. The proof of this theorem is similar to that of Theorem 2, with some extra details to account for the possibility that the optimal set X^* may be empty.

Theorem 3 *Let Assumptions 1, 2,3, 4 and 6 hold. Let the sequence $\{x_k\}$ be generated by the method (3.1) with a constant stepsize rule, i.e., $\alpha_k = \alpha$ for all $k \in \mathbb{N}$. Also, assume that the subgradient errors $\epsilon_{i,k}$ have zero mean and bounded second moments, i.e.,*

$$\mu_k = 0 \quad \textit{for all } k \geq 1, \qquad \nu = \sup_{k \geq 1} \nu_k < \infty.$$

We then have

$$\liminf_{k\to\infty} \mathsf{E}[f(x_k)] \leq f^* + \frac{\alpha}{2}\left(\sum_{i=1}^{m} C_i + m\nu\right)^2, \tag{3.14}$$

and with probability 1,

$$\inf_{k\geq 0} f(x_k) \leq f^* + \frac{\alpha}{2}\left(\sum_{i=1}^{m} C_i + m\nu\right)^2. \tag{3.15}$$

The proof is discussed in [39]. In the absence of errors ($\nu = 0$), the error bound of Theorem 3 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 [9], Proposition 2.1.

3.2.3 Rate of convergence

We next obtain a result that captures the rate of convergence in the expected sense. Define

$$\bar{x}_t = \frac{\sum_{k=0}^{t} \alpha_{k+1} x_k}{\sum_{k=0}^{t} \alpha_{k+1}} \qquad \text{and} \qquad \bar{z}_{i,t} = \frac{\sum_{k=0}^{t} \alpha_{k+1} z_{i,k}}{\sum_{k=0}^{t} \alpha_{k+1}}.$$

We next obtain bounds on the optimality of $\{\bar{x}_k\}$ after a finite number of iterations. The result is applicable for both diminishing and non-diminishing stepsizes.

Theorem 4 *Let Assumptions 1, 2, 4 and 6 hold. Additionally, let the set X be*

bounded. Then,

$$\begin{aligned}\mathsf{E}[f(\bar{z}_{i,t})] \leq f^* + \frac{1}{2\sum_{k=0}^{t}\alpha_{k+1}}\mathsf{E}\left[\|d_0(x^*)\|^2\right] + m\max_{x,y\in X}\|x-y\|\frac{\sum_{k=0}^{t}\alpha_{k+1}\mu_{k+1}}{\sum_{k=0}^{t}\alpha_{k+1}} \\ + \frac{\sum_{k=0}^{t}\alpha_{k+1}^2\left(m\nu_{k+1}+\sum_{j=1}^{m}C_j\right)^2}{\sum_{k=0}^{t}\alpha_{k+1}} \\ + \frac{\sum_{k=0}^{t}\alpha_k\alpha_{k+1}\left(\sum_{j=1}^{m}C_j\right)\left(\sum_{j=1}^{i}C_j+\nu_k\right)}{\sum_{k=0}^{t}\alpha_{k+1}}.\end{aligned}$$

Proof If the set X is bounded then the subgradients are bounded. Thus Assumption 3 holds. Therefore the conditions of Lemma 1 are satisfied. Also note that since the set X is compact X^* is non-empty. Fixing $x^* \in X^*$, taking expectation in Lemma 1 and using the boundedness of the set X we obtain

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

Next note that

$$\begin{aligned}\mathsf{E}[f(x_k)] &\geq \mathsf{E}[f(z_{i,k})] - \left(\sum_{j=1}^{m}C_j\right)\mathsf{E}[\|z_{i,k}-x_k\|] \\ &\geq \mathsf{E}[f(z_{i,k})] - \alpha_k\left(\sum_{j=1}^{m}C_j\right)\left(\sum_{j=1}^{i}C_j+\nu_k\right).\end{aligned}$$

Using the preceding relation we obtain

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2\right] \leq &\mathsf{E}\left[\|d_k(x^*)\|^2\right] - 2\alpha_{k+1}\left(\mathsf{E}[f(z_{i,k})]-f^*\right) \\ &+ 2m\alpha_{k+1}\mu_{k+1}\max_{x,y\in X}\|x-y\| + \alpha_{k+1}^2\left(\sum_{i=1}^{m}C_i+m\nu_{k+1}\right)^2 \\ &+ 2\alpha_k\alpha_{k+1}\left(\sum_{j=1}^{m}C_j\right)\left(\sum_{j=1}^{i}C_j+\nu_k\right).\end{aligned}$$

Summing over k from 0 to t, rearranging terms and dividing by $\sum_{k=0}^{t} \alpha_{k+1}$ we obtain

$$\begin{aligned}\frac{\sum_{k=0}^{t} \alpha_{k+1} f(z_{i,k})}{\sum_{k=0}^{t} \alpha_{k+1}} &\leq f^* + \frac{1}{2\sum_{k=0}^{t} \alpha_{k+1}} \mathsf{E}\left[\|d_0(x^*)\|^2\right] \\ &+ m \max_{x,y \in X} \|x - y\| \frac{\sum_{k=0}^{t} \alpha_{k+1} \mu_{k+1}}{\sum_{k=0}^{t} \alpha_{k+1}} \\ &+ \frac{\sum_{k=0}^{t} \alpha_{k+1}^2 \left(m\nu_{k+1} + \sum_{j=1}^{m} C_j\right)^2}{\sum_{k=0}^{t} \alpha_{k+1}} \\ &+ \frac{\sum_{k=0}^{t} \alpha_k \alpha_{k+1} \left(\sum_{j=1}^{m} C_j\right) \left(\sum_{j=1}^{i} C_j + \nu_k\right)}{\sum_{k=0}^{t} \alpha_{k+1}}.\end{aligned}$$

Next note from the convexity of the function f that

$$f(\bar{z}_{i,t}) = f\left(\frac{\sum_{k=0}^{t} \alpha_{k+1} z_{i,k}}{\sum_{k=0}^{t} \alpha_{k+1}}\right) \leq \frac{\sum_{k=0}^{t} \alpha_{k+1} f(z_{i,k})}{\sum_{k=0}^{t} \alpha_{k+1}}.$$

The result now follows from the preceding two relations. □

CHAPTER 4

MARKOV INCREMENTAL ALGORITHM

We consider the Markov 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:

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

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 that performs the update at time $k+1$, and the sequence $\{s(k)\}$ is modeled as a time non-homogeneous Markov chain with state space $\{1, \ldots, m\}$. 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 $[A(k)]_{i,j}$. Formally, we have

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

When there are no errors ($\epsilon_{s(k+1),k+1} = 0$) and the probabilities $[A(k)]_{i,j}$ are all equal to $\frac{1}{m}$, the method in (4.1) coincides with the incremental method with randomization that was proposed and studied in [9].

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 $A(\ell)\cdots A(k)$ 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 $[A(k)]$. We will make Assumption 5. In addition, we will assume that the probability matrix $A(k)$ satisfies Assumption 10.

Assumption 10 *For $i \in V$ and all k,*

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

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

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

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

Assumptions 10(a) and (b) ensure that the information from each and every agent is persistent in time. Assumption 10(c) ensures that the limiting Markov chain probability distribution (if one exists) is uniform. Assumptions 5 and 10 together guarantee the existence of the uniform limiting distribution, as shown in [20].

Note that the cyclic incremental algorithm (3.1) does not satisfy Assumption 10. 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 10(c) is violated.

We now provide some examples of transition matrices $[A(k)]$ satisfying Assumption 10. The second and third examples are variations of the Metropolis-Hasting weights [26, 29], 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:

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

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

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

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

- *Weighted Metropolis-Hastings scheme.* The probabilities that agent i uses

at time k are given by

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

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 Basic Iterate Relation

We use the estimate of Lemma 8 to establish a key relation in Lemma 2, which is repeatedly invoked in our subsequent analysis. The idea behind Lemma 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 $\frac{1}{m}\sum_{i=1}^m \nabla f_i(x)$ 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 the 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 Assumption 6, we take $F_k^{s_k}$ to be the σ-field G_k generated by the initial vector x_0 and $\left\{ s(n), \epsilon_{s(n),n}; 0 \leq n < k \right\}$.

Lemma 2 *Let Assumptions 1, 2, 3, 5, 6, and 10 hold. Then, the iterates generated by algorithm (4.1) are such that for any step-size rule, for any $y \in X$, and any non-negative integer sequence $\{n(k)\}$, $n(k) \leq k$, we have*

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

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

Proof Using the iterate update rule in (4.1) and the non-expansive property of the Euclidean projection, we obtain for any $y \in X$ and $k \geq 0$,

$$\begin{aligned}
\|d_{k+1}(y)\|^2 = & \left\| \mathcal{P}_X \left[x_k - \alpha_{k+1} \nabla f_{s(k+1)}(x_k) - \alpha_{k+1} \epsilon_{s(k+1),k+1} \right] - y \right\|^2 \\
\leq & \left\| x_k - \alpha_{k+1} \nabla f_{s(k+1)}(x_k) - \alpha_{k+1} \epsilon_{s(k+1),k+1} - y \right\|^2 \\
= & \|d_k(y)\|^2 - 2\alpha_{k+1} d_k(y)^T \nabla f_{s(k+1)}(x_k) \\
& - 2\alpha_{k+1} d_k(y)^T \epsilon_{s(k+1),k+1} + \alpha_{k+1}^2 \left\| \epsilon_{s(k+1),k+1} + \nabla f_{s(k+1)}(x_k) \right\|^2.
\end{aligned}$$

Using the subgradient inequality in (2.2) to bound $d_k(y)^T \nabla f_{s(k+1)}(x_k)$, we get

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

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

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

We next use the subgradient inequality in (2.2) to estimate the second term in the preceding relation.

$$\begin{aligned}\mathsf{E}\left[f_{s(k+1)}(x_k) - f_{s(k+1)}\left(x_{n(k)}\right) \mid G_{n(k)}\right] \geq & \mathsf{E}\left[\nabla f_{s(k+1)}\left(x_{n(k)}\right)^T \left(x_{n(k)} - x_k\right) \mid G_{n(k)}\right] \\ \geq & -\mathsf{E}\left[\left\|\nabla f_{s(k+1)}\left(x_{n(k)}\right)\right\| \left\|x_{n(k)} - x_k\right\| \mid G_{n(k)}\right] \\ \geq & -C\mathsf{E}\left[\left\|x_{n(k)} - x_k\right\| \mid G_{n(k)}.\right] \quad (4.3)\end{aligned}$$

In the last step we have used the subgradient boundedness from Assumption 3 to bound the subgradient norms $\left\|\nabla f_{s(k+1)}\left(x_{n(k)}\right)\right\|$ by $C = \max_{1\leq i\leq m} C_i$. We 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

non-expansive property of the Euclidean projection as follows:

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

where in the last step we have used the boundedness of subgradients and of the second moments of $\epsilon_{i,k}$ [cf. Eqn. (2.7)]. From the preceding relation and Eqn. (4.3), we obtain

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

By substituting the preceding estimate in (4.2), we further obtain

$$\begin{aligned}
\mathsf{E}\big[\|d_{k+1}(y)\|^2 \mid G_{n(k)}\big] \leq & \mathsf{E}\big[\|d_k(y)\|^2 \mid G_{n(k)}\big] + 2C\alpha_{k+1} \sum_{l=n(k)}^{k-1} \alpha_{l+1} \left(C + \nu_{l+1}\right) \\
& - 2\alpha_{k+1} \left(\mathsf{E}\big[f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}\left(y\right) \mid G_{n(k)}\big]\right) \\
& - 2\alpha_{k+1} \mathsf{E}\big[d_k(y)^T \epsilon_{s(k+1),k+1} \mid G_{n(k)}\big] \\
& + \alpha_{k+1}^2 \mathsf{E}\Big[\big\|\epsilon_{s(k+1),k+1} + \nabla f_{s(k+1)}\left(x_k\right)\big\|^2 \mid G_{n(k)}\Big]. \qquad (4.4)
\end{aligned}$$

We estimate the last term in (4.4) by using the subgradient boundedness of Assumption 3 and the boundedness of the second moments of $\epsilon_{i,k}$ [cf. Eqn.

(2.7)], as follows:

$$\begin{aligned}
&\mathsf{E}\Big[\big\|\epsilon_{s(k+1),k+1}+\nabla f_{s(k+1)}\left(x_k\right)\big\|^2 \mid G_{n(k)}\Big] \\
&\quad\leq \mathsf{E}\Big[\big\|\epsilon_{s(k+1),k+1}\big\|^2+\big\|\nabla f_{s(k+1)}\left(x_k\right)\big\|^2+2\big\|\epsilon_{s(k+1),k+1}\big\|\,\big\|\nabla f_{s(k+1)}\left(x_k\right)\big\| \mid G_{n(k)}\Big] \\
&\quad\leq \nu_k^2+C^2+2\nu_k C \\
&\quad= (\nu_k+C)^2.
\end{aligned}$$

Substituting the preceding estimate in Eqn. (4.4), we have

$$\begin{aligned}
\mathsf{E}\big[\|d_{k+1}(y)\|^2 \mid G_{n(k)}\big] \leq & \mathsf{E}\big[\|d_k(y)\|^2 \mid G_{n(k)}\big]+2C\alpha_{k+1}\sum_{l=n(k)}^{k-1}\alpha_{l+1}\left(C+\nu_{l+1}\right) \\
&-2\alpha_{k+1}\left(\mathsf{E}\big[f_{s(k+1)}\left(x_{n(k)}\right)-f_{s(k+1)}\left(y\right) \mid G_{n(k)}\big]\right) \\
&-2\alpha_{k+1}\mathsf{E}\big[d_k(y)^T\epsilon_{s(k+1),k+1} \mid G_{n(k)}\big] \\
&+\alpha_{k+1}^2(\nu_k+C)^2. \qquad (4.5)
\end{aligned}$$

We next estimate the term $\mathsf{E}\big[d_k(y)^T\epsilon_{s(k+1),k+1} \mid G_{n(k)}\big]$. Since $G_{n(k)}\subset G_k$ and $d_k(y)$ is G_k-measurable

$$\begin{aligned}
\mathsf{E}\big[d_k(y)^T\epsilon_{s(k+1),k+1} \mid G_{n(k)}\big] =& \mathsf{E}\big[\mathsf{E}\big[d_k(y)^T\epsilon_{s(k+1),k+1} \mid G_k\big] \mid G_{n(k)}\big] \\
=& \mathsf{E}\big[d_k(y)^T\mathsf{E}\big[\epsilon_{s(k+1),k+1} \mid G_k\big] \mid G_{n(k)}\big] \\
\geq& -\mathsf{E}\big[\|d_k(y)\|\,\big\|\mathsf{E}\big[\epsilon_{s(k+1),k+1} \mid G_k\big]\big\| \mid G_{n(k)}\big] \\
\geq& -\mu_{k+1}\mathsf{E}\big[\|d_k(y)\| \mid G_{n(k)}\big],
\end{aligned}$$

where the first equality follows from the law of iterated conditioning. Using the

preceding estimate in (4.5), we obtain

$$\begin{aligned}\mathsf{E}\big[\|d_{k+1}(y)\|^2 \mid G_{n(k)}\big] \leq & \mathsf{E}\big[\|d_k(y)\|^2 \mid G_{n(k)}\big] + 2C\alpha_{k+1}\sum_{l=n(k)}^{k-1}\alpha_{l+1}\left(C+\nu_{l+1}\right) \\ & - 2\alpha_{k+1}\left(\mathsf{E}\big[f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}\left(y\right) \mid G_{n(k)}\big]\right) \\ & + 2\alpha_{k+1}\mu_{k+1}\mathsf{E}\big[\|d_k(y)\| \mid G_{n(k)}\big] + \alpha_{k+1}^2(\nu_k+C)^2. \quad (4.6)\end{aligned}$$

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 $\Phi(k+1,n(k)) = A(n(k))\cdots A(k)$. We have

$$\begin{aligned}&\mathsf{E}\big[f_{s(k+1)}\left(x_{n(k)}\right) - f_{s(k+1)}\left(y\right) \mid G_{n(k)}\big] \\ &= \sum_{i=1}^m \left[\Phi(k+1,n(k))\right]_{s(n(k)),i}\left(f_i\left(x_{n(k)}\right) - f_i(y)\right) \\ &\geq \sum_{i=1}^m \frac{1}{m}\left(f_i\left(x_{n(k)}\right) - f_i(y)\right) - \sum_{i=1}^m \left|\left[\Phi(k+1,n(k))\right]_{s(n(k)),i} - \frac{1}{m}\right|\left|f_i\left(x_{n(k)}\right) - f_i(y)\right| \\ &\geq \frac{1}{m}\left(f\left(x_{n(k)}\right) - f(y)\right) - b\beta^{k+1-n(k)}\sum_{i=1}^m \left|f_i\left(x_{n(k)}\right) - f_i(y)\right|, \quad (4.7)\end{aligned}$$

where at the last step we have used Lemma 8. Using the subgradient inequality (2.2), we further have

$$\left|f_i\left(x_{n(k)}\right) - f_i(y)\right| \leq C_i\left\|x_{n(k)} - y\right\| = C_i\|d_{n(k)}(y)\|. \quad (4.8)$$

The result now follows by combining the relations in Eqs. (4.6), (4.7) and (4.8). □

4.2 Convergence Results

In this section, we establish the convergence of the Markov randomized method in (4.1) for a diminishing step-size and constant step-size. In addition, we also

obtain a rate of convergence result

4.2.1 Diminishing stepsizes

In this section, we establish the convergence of the Markov randomized method in (4.1) for a diminishing stepsize. Recall that in Theorem 1 for the cyclic incremental method, we showed an almost sure convergence result for a diminishing stepsize α_k subject to some conditions that coordinate the choice of the stepsize, and the bounds $\min_k$ and ν_k on the moments of the errors $\epsilon_{i,k}$. To obtain an analogous result for the Markov randomized method, we use boundedness of the set X and more restricted stepsize. In particular, we consider a stepsize of the form $\alpha_k = \frac{a}{k^p}$ for a range of values of p, as seen in the following.

Theorem 5 *Let Assumptions 1, 2, 5, 6, and 10 hold. Assume that the stepsize is $\alpha_k = \frac{a}{k^p}$, where a and p are positive scalars with $\frac{2}{3} < p \leq 1$. In addition, assume that the bounds $\min_k$ and ν_k on the error moments satisfy*

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

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 2.

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 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 $\mathsf{dist}(x_{k+1}) \leq \|x_{k+1} - x_k^*\|$, from Lemma 2, we obtain

for all $k > 1$[1],

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

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

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

where

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

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

$$\sum_{k=2}^{\infty} 2b\left(\sum_{i=1}^{m} C_i\right)\alpha_{k+1}\beta^{\lceil k^\gamma\rceil}\left\|d_{n(k)}(x_k^*)\right\| < \infty. \tag{4.9}$$

[1]The equivalent expression for the case when $k = 1$ is obtained by setting the fourth term to 0.

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 $k \geq 2$,

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

By combining the preceding two relations, we have

$$\sum_{k=2}^{\infty} 2C(C+\nu)\alpha_{k+1}\alpha_{n(k)+1}(\lceil k^\gamma \rceil - 2) < 2C(C+\nu)\frac{a^2}{(1-2^{\gamma-1})^p}\sum_{k=2}^{\infty}\frac{1}{k^{2p-\gamma}} < \infty, \tag{4.10}$$

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

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

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

Thus, from Eqs. (4.9) and (4.10), and the preceding two relations, we see that $\sum_{k=2}^{\infty} \tau_{k+1} < \infty$.

From the deterministic analog of Lemma 16 we conclude that $\mathsf{E}\big[\mathsf{dist}(x_k)^2\big]$ converges to a non-negative scalar and

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

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

$$\liminf_{k\to\infty} \mathsf{E}\left[f\left(x_{n(k)}\right)\right] = f^*. \tag{4.11}$$

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

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

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

As seen in the proof of Theorem 5, $\mathsf{E}\big[\mathsf{dist}(x_k)^2\big]$ converges to a non-negative scalar, but we have no guarantee that its limit is zero. However, this can be shown, for example, for a function with a sharp set of minima, i.e., f satisfying

$$f(x) - f^* \geq \zeta\, \mathsf{dist}(x)^\xi \qquad \text{for all } x \in X,$$

for some positive scalars ζ and ξ. Under the assumptions of Theorem 5, we have that $\liminf_{k\to\infty} \mathsf{E}[f\left(x_k\right)] = f^*$ [cf. Eqn. (4.11)] and therefore,

$$0 = \liminf_{k\to\infty} \mathsf{E}[f\left(x_k\right) - f^*] \geq \zeta \liminf_{k\to\infty} \mathsf{E}\Big[\mathsf{dist}(x_k)^\xi\Big] \geq 0.$$

Hence, $\liminf_{k\to\infty} \mathsf{E}\Big[\mathsf{dist}(x_k)^\xi\Big] = 0$, and since $\mathsf{E}\big[\mathsf{dist}(x_k)^2\big]$ converges, it has to converge to 0.

4.2.2 Constant stepsizes

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

Theorem 6 *Let Assumptions 1, 2, 5, 6, and 10 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*

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

Then for any integer $T \geq 0$,

$$\begin{aligned}\liminf_k \mathsf{E}[f(x_k)] \leq & f^* + \mu \max_{x,y\in X} \|x-y\| + \frac{1}{2}\alpha(\nu + C)^2 + \alpha TC\,(C+\nu) \\ & + b\left(\sum_{i=1}^m C_i\right)\beta^{T+1} \max_{x,y\in X}\|x-y\|, \end{aligned} \tag{4.12}$$

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 2 are satisfied. Let T be a non-negative 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 2, we have for $y = x^* \in X^*$ and $k \geq T$,

$$\begin{aligned}\mathsf{E}\left[\|d_{k+1}(x^*)\|^2 \mid G_{n(k)}\right] \leq & \mathsf{E}\left[\|d_k(x^*)\|^2 \mid G_{n(k)}\right] - \frac{2\alpha}{m}\left(f\left(x_{k-T}\right) - f^*\right) \\ & + 2b\left(\sum_{i=1}^m C_i\right)\alpha\beta^{T+1}\max_{x,y\in X}\|x-y\| \\ & + 2\alpha^2 TC\,(C+\nu) \\ & + 2\alpha\mu \max_{x,y\in X}\|x-y\| + \alpha^2(\nu+C)^2. \end{aligned} \tag{4.13}$$

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

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

Now assume that the relation (4.12) does not hold. Then, there will exist a $\gamma > 0$ and an index $k_\gamma \geq T$ such that for all $k \geq k_\gamma$,

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

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 (4.12) must hold for all $T \geq 0$.

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

$$\begin{aligned}\inf_k f(x_k) \leq & f^* + \mu\max_{x,y\in X}\|x-y\| + \frac{1}{2}\alpha(\nu+C)^2 + \alpha TC\left(C+\nu\right) \\ & + b\left(\sum_{i=1}^{m} C_i\right)\beta^{T+1}\max_{x,y\in X}\|x-y\|, \qquad (4.14)\end{aligned}$$

with probability 1. Define the set

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

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

$$\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid G_{n(k)} \Big] = \mathsf{E}\Big[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)} \Big]. \tag{4.15}$$

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

$$\begin{aligned} f(x_{k-T}) - f^* \geq & \frac{1}{N} + \mu \max_{x,y \in X} \|x - y\| + \frac{1}{2}\alpha(\nu + C)^2 + \alpha TC\,(C + \nu) \\ & + b \left(\sum_{i=1}^{m} C_i \right) \beta^{T+1} \max_{x,y \in X} \|x - y\|. \end{aligned} \tag{4.16}$$

By using relations (4.13) and (4.16), we conclude that for $\hat{x}_k \notin L_N$,

$$\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid G_{n(k)} \Big] \leq \mathsf{E}\Big[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)} \Big] - \frac{2\alpha}{N}. \tag{4.17}$$

Therefore, from (4.15) and (4.17), we can write

$$\mathsf{E}\Big[\|\hat{d}_{k+1}(x^*)\|^2 \mid G_{n(k)}\Big] \leq \mathsf{E}\Big[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)}\Big] - \Delta_{k+1}, \tag{4.18}$$

where

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

Observe that (4.18) satisfies the conditions of Lemma 16 with $u_k = \mathsf{E}\Big[\|\hat{d}_k(x^*)\|^2 \mid G_{n(k)}\Big]$, $\mathcal{F}_k = G_{n(k)}$, $q_k = 0$, $w_k = 2\Delta_{k+1}$ 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 (4.14). □

Under Assumptions of Theorem 6, the function f is bounded over the set X, and by Fatou's lemma, we have

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

It follows that the estimate of Theorem 6 also holds for $\mathsf{E}[\liminf_{k\to\infty} f(x_k)]$.

In the absence of errors ($\min_k = 0$ and $\nu_k = 0$), the error bound in Theorem 6 reduces to

$$f^* + \frac{1}{2}\,\alpha C^2 + \alpha T C^2 + b\left(\sum_{i=1}^{m} C_i\right)\beta^{T+1} \max_{x,y\in X} \|x - y\|. \tag{4.19}$$

With respect to the parameter β, the error bound is obviously smallest when $\beta = 0$. This corresponds to uniform transition matrices $A(k)$, i.e., $A(k) = \frac{1}{m} e e^T$

for all k (see Lemma 8). As mentioned, the Markov randomized method with uniform transition probability matrices $A(k)$ reduces to the incremental method with randomization in [9]. In this case, choosing $T = 0$ in (4.19) 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 [9].

When transition matrices are non-uniform ($\beta > 0$), 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 (4.19) with respect to integer T for $T \geq 0$. In particular, one may optimize the term $\alpha TC^2 + b\left(\sum_{i=1}^m C_i\right)\beta^{T+1}\max_{x,y\in X}\|x-y\|$ over integers $T \geq 0$. It can be seen that the optimal integer T^* is given by

$$T^* = \begin{cases} 0 & \text{when } \frac{\alpha C^2}{C_0(-\ln\beta)} \geq 1, \\ \left\lceil (\ln\beta)^{-1}\ln\left(\frac{\alpha C^2}{C_0(-\ln\beta)}\right)\right\rceil - 1 & \text{when } \frac{\alpha C^2}{C_0(-\ln\beta)} < 1, \end{cases} \tag{4.20}$$

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

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 min 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.

Corollary 1 *Let the conditions of Theorem 6 hold. Then,*

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

where

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

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 6) we get

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

When $\alpha<\beta$ we can choose $T=\left\lceil\frac{\ln(\alpha)}{\ln(\beta)}\right\rceil-1$. Then, from (Theorem 6),

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

□

It can be seen that the error bounds in (4.20) and Corollary 1 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 [18] is obtained assuming that there are no errors in subgradient evaluations and that the sequence of computing agents forms 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 1 and the results in [18] is not possible. However, some qualitative comparisons on the nature of the bounds can be made. The bound in [18] is obtained for each individual agent's sequence of iterates (by sampling the iterates). This is a stronger result than our results in (4.20) and Corollary 1, which provide guarantees only on the entire iterate sequence (and not on the sequence of iterates at an individual agent). However, the bound in [18] 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 (4.20) and Corollary 1 can be evaluated without knowing the network topology. We require that the topology satisfies a connectivity assumption, as specified by Assumption 5, but we do not assume the knowledge of the exact network topology.

4.2.3 Rate of convergence

We next obtain a result that is similar to the rate of convergence result in Section 4.2.3. Define

$$\bar{x}_t = \frac{\sum_{k=0}^{t} \alpha_{k+1} x_k}{\sum_{k=0}^{t} \alpha_{k+1}}$$

We only state the result. The proof is very similar to the proof in Section 4.2.3.

Theorem 7 *Let Assumptions 1, 2, 5 6, and 10 hold. Further, let the set X be bounded. Then, the iterates generated by algorithm (4.1) are such that for any stepsize rule,*

$$\begin{aligned}
\mathsf{E}[f(\bar{x}_t)] \leq & f^* \frac{1}{2\sum_{k=0}^{t} \alpha_{k+1}} \mathsf{E}\left[\|d_0(x^*)\|^2\right] \\
& + b\left(\sum_{i=1}^{m} C_i\right) \max_{x,y \in X} \|x-y\| \frac{\sum_{k=0}^{t} \alpha_{k+1} \beta^{k+1-n(k)}}{\sum_{k=0}^{t} \alpha_{k+1}} \\
& + C \frac{\sum_{k=0}^{t} \alpha_{k+1} \sum_{l=n(k)}^{k-1} \alpha_{l+1} \left(C + \nu_{l+1}\right)}{\sum_{k=0}^{t} \alpha_{k+1}} \\
& + \max_{x,y \in X} \|x-y\| \frac{\sum_{k=0}^{t} \alpha_{k+1} \mu_{k+1}}{\sum_{k=0}^{t} \alpha_{k+1}} + \alpha_{k+1}^2 (\nu_k + C)^2,
\end{aligned}$$

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

CHAPTER 5

PARALLEL ALGORITHM

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

$$w_{i,k+1} = P_X \left[v_{i,k} - \alpha_{k+1} \left(\nabla f_i \left(v_{i,k} \right) + \epsilon_{i,k+1} \right) \right], \tag{5.1}$$

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

$$v_{i,k} = \sum_{j \in N_i(k+1)} a_{i,j}(k+1) w_{j,k}, \tag{5.2}$$

where $N_i(k+1)$ denotes the set of agents whose current iterates are available to agent i in the $(k+1)$-st iteration. We assume that $i \in N_i(k+1)$ for all agents and at all times k. The scalars $a_{i,j}(k+1)$ are the non-negative weights that agent i assigns to agent j's iterate. We will find it convenient to define

$a_{i,j}(k+1)$ as 0 for $j \notin N_i(k+1)$ and rewrite (5.2) as

$$v_{i,k} = \sum_{j=1}^{m} a_{i,j}(k+1) w_{j,k}. \tag{5.3}$$

This is a "consensus"-based step ensuring that, in a long run, the information of each f_i reaches every agent with the same frequency, directly or through a sequence of local communications. This is similar to the distributed averaging algorithm in Appendix A.5. Due to this, the iterates $w_{j,k}$ become eventually "the same" for all j and for large enough k. The update step in (5.1) is just a subgradient iteration for minimizing f_i over X taken after the "consensus"-based step.

In addition to Assumptions 2 and 5, we make some assumptions on the weights to ensure that the influence of the functions f_i is "equal" in the long run so that the sum, rather than a weighted sum, of the component functions is minimized. The influence of a component f_j on the iterates of agent i depends on the weights that agent i uses. To ensure equal influence, we assume that the weights satisfy Assumption 10.

5.1 Basic Iterate Relation

In this section, we derive two basic relations that form the basis for the analysis in this chapter. The first deals with the disagreements among the agents, and the second deals with the agent iterate sequences.

5.1.1 Disagreement estimate

The agent disagreements are typically thought of as the norms $\|w_{i,k} - w_{j,k}\|$ of the differences between the iterates $w_{i,k}$ and $w_{j,k}$ generated by different agents according to (5.1)–(5.2). Alternatively, the agent disagreements can be measured with respect to a reference sequence, which we adopt here. In particular, we

study the behavior of $\|y_k - w_{i,k}\|$, where $\{y_k\}$ is the auxiliary vector sequence defined by

$$y_k = \frac{1}{m}\sum_{i=1}^{m} w_{i,k} \qquad \text{for all } k. \tag{5.4}$$

In the next lemma, we provide a basic estimate for $\|y_k - w_{j,k}\|$. The rate of convergence result from Lemma 8 plays a crucial role in obtaining this estimate. Equivalently, the result also charachterizes the effect of errors in distributed averaging.

Lemma 3 *Let Assumptions 1, 2a, 5 and 10 hold. Assume that the subgradients of f_i are uniformly bounded over the set X, i.e., there are scalars C_i such that*

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

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

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

Proof Define for all $i \in V$ and all k,

$$p_{i,k+1} = w_{i,k+1} - \sum_{j=1}^{m} a_{i,j}(k+1)w_{j,k}. \tag{5.5}$$

Using the matrices $\Phi(k,s)$ defined in (A.5) we can write

$$w_{j,k+1} = \sum_{i=1}^{m}[\Phi(k+1,0)]_{j,i}w_{i,0} + p_{j,k+1} + \sum_{\ell=1}^{k}\left(\sum_{i=1}^{m}[\Phi(k+1,\ell)]_{j,i}p_{i,\ell}\right). \tag{5.6}$$

Using (5.5), we can also rewrite y_k, defined in (5.4) as

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

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

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

Therefore

$$y_{k+1} = y_0 + \frac{1}{m}\sum_{\ell=1}^{k+1}\sum_{i=1}^{m} p_{i,\ell} = \frac{1}{m}\sum_{i=1}^{m} w_{i,0} + \frac{1}{m}\sum_{\ell=1}^{k+1}\sum_{i=1}^{m} p_{i,\ell}. \tag{5.7}$$

Substituting for y_{k+1} from (5.7) and for $w_{j,k+1}$ from (5.6), we obtain

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

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

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

We can bound $\|w_{i,0}\| \leq \max_{i \in V} \|w_{i,0}\|$. Further, we can use the rate of convergence result from Lemma 8 to bound $\left|\frac{1}{m} - [\Phi(k,\ell)]_{j,i}\right|$. We obtain

$$\begin{aligned}\|y_{k+1} - w_{j,k+1}\| \leq & m\theta\beta^{k+1} \max_{i \in V} \|w_{i,0}\| + \theta \sum_{\ell=1}^{k} \beta^{k+1-\ell} \sum_{i=1}^{m} \|p_{i,\ell}\| \\ & + \frac{1}{m} \sum_{i=1}^{m} \|p_{i,k+1}\| + \|p_{j,k+1}\| . \end{aligned} \tag{5.8}$$

We next estimate the norms of the vectors $\|p_{i,k}\|$ for any k. From the definition of $p_{i,k+1}$ in (5.5) and the definition of the vector $v_{i,k}$ in (5.2), we have $p_{i,k+1} = w_{i,k+1} - v_{i,k}$. Note that, being a convex combination of vectors $w_{j,k}$ in the convex set X, the vector $v_{i,k}$ is in the set X. By the definition of the iterate $w_{i,k+1}$ in (5.1) and the non-expansive property of the Euclidean projection in (A.4), we have

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

In the last step we have used the subgradient boundedness. By substituting the preceding relation in (5.8), we obtain the desired relation. □

5.1.2 Iterate relation

Here, we derive a relation for the distances $\|v_{i,k+1} - z\|$ and the function value differences $f(y_k) - f(z)$ for an arbitrary $z \in X$. This relation together with Lemma 3 provides the basis for our subsequent convergence analysis. In what follows, recall that $f = \sum_{i=1}^{m} f_i$.

Lemma 4 *Let Assumptions 1, 5, 2, 3, 5 and 10 hold. Then, for any $z \in X$ and*

all k,

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

Proof Using the Euclidean projection property in (A.4), from the definition of the iterate $w_{i,k+1}$ in (5.1), we have for any $z\in X$ and all k,

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

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

$$\begin{aligned}\|w_{i,k+1}-z\|^2 \leq& \|v_{i,k}-z\|^2-2\alpha_{k+1}\left(f_i(v_{i,k})-f_i(z)\right)\\ &-2\alpha_{k+1}\epsilon_{i,k+1}^T(v_{i,k}-z)+\alpha_{k+1}^2\left\|\nabla f_i(v_{i,k})+\epsilon_{i,k+1}\right\|^2. \quad (5.9)\end{aligned}$$

Note that by the convexity of the squared norm [cf. Eq. (A.3)], we have

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

In view of Assumption 10, we have $\sum_{i=1}^{m}a_{i,j}(k+2)=1$ for all j and k, implying that

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

By summing the relations in (5.9) over all $i\in V$ and by using the preceding

relation, we obtain

$$\begin{aligned}\sum_{i=1}^{m}\|v_{i,k+1}-z\|^2 \;\leq\; & \sum_{i=1}^{m}\|v_{i,k}-z\|^2 - 2\alpha_{k+1}\sum_{i=1}^{m}\left(f_i(v_{i,k})-f_i(z)\right)\\ & -2\alpha_{k+1}\sum_{i=1}^{m}\epsilon_{i,k+1}^T(v_{i,k}-z)\\ & +\alpha_{k+1}^2\sum_{i=1}^{m}\left\|\nabla f_i(v_{i,k})+\epsilon_{i,k+1}\right\|^2. \end{aligned} \tag{5.10}$$

From (2.2) we have

$$\begin{aligned} f_i(v_{i,k})-f_i(z) &\geq \left(f_i(v_{i,k})-f_i(y_k)\right)+\left(f_i(y_k)-f_i(z)\right)\\ &\geq -\left\|\nabla f_i(v_{i,k})\right\|\|y_k-v_{i,k}\|+\left(f_i(y_k)-f_i(z)\right). \end{aligned} \tag{5.11}$$

Recall that $v_{i,k}=\sum_{j=1}^{m}a_{i,j}(k+1)w_{j,k}$ [cf. (5.3)]. Substituting for $v_{i,k}$ and using the convexity of the norm [cf. (A.2)], from (5.11) we obtain

$$\begin{aligned}\sum_{i=1}^{m}f_i(v_{i,k})-f_i(z) &\geq -\sum_{i=1}^{m}\left\|\nabla f_i(v_{i,k})\right\|\|y_k-v_{i,k}\|+\left(f(y_k)-f(z)\right)\\ &\geq -\sum_{i=1}^{m}\left\|\nabla f_i(v_{i,k})\right\|\left\|y_k-\sum_{j=1}^{m}a_{i,j}(k+1)w_{j,k}\right\|+\left(f(y_k)-f(z)\right)\\ &\geq -\sum_{i=1}^{m}\left\|\nabla f_i(v_{i,k})\right\|\sum_{j=1}^{m}a_{i,j}(k+1)\|y_k-w_{j,k}\|+\left(f(y_k)-f(z)\right)\\ &\geq -\left(\max_{i\in V}\left\|\nabla f_i(v_{i,k})\right\|\right)\sum_{j=1}^{m}\left(\sum_{i=1}^{m}a_{i,j}(k+1)\right)\|y_k-w_{j,k}\|\\ &\quad +\left(f(y_k)-f(z)\right)\\ &= -\left(\max_{i\in V}\left\|\nabla f_i(v_{i,k})\right\|\right)\sum_{j=1}^{m}\|y_k-w_{j,k}\|+\left(f(y_k)-f(z)\right). \end{aligned}$$

By using the preceding estimate in relation (5.10), we have

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

The result follows by using the subgradient norm boundedness, $\|\nabla f_i(v_{i,k})\| \leq C_i$ for all k and i. □

5.2 Convergence Results

In this section we study the convergence of the algorithm with diminishing and constant stepsizes. We also obtain a rate of convergence result.

5.2.1 Diminishing stepsizes

We first obtain bounds on the expected disagreement between the agents. we provide a bound on the expected disagreement $\mathsf{E}[\|w_{i,k} - y_k\|]$ for non-diminishing stepsize. We later use this bound to provide an estimate for the algorithm's performance in mean. The bound is provided in the following theorem.

Theorem 8 *Let Assumptions 1, 2, 3, 5, 6, and 10 hold. If $\nu_k \leq \nu$ for sufficiently large k and the stepsize $\{\alpha_k\}$ is such that $\lim_{k\to\infty} \alpha_k = \alpha$ for some $\alpha \geq 0$, then for all $j \in V$,*

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

Proof The conditions of Lemma 3 are satisfied. Taking the expectation in the relation of Lemma 3 and using the inequality $\mathsf{E}[\|\epsilon_{i,k}\|] \leq \sqrt{\mathsf{E}[\|\epsilon_{i,k}\|^2]} = \nu$, we

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

$$\begin{aligned}\mathsf{E}[\|y_{k+1} - w_{j,k+1}\|] \leq & m\theta\beta^{k+1} \max_{i\in V} \|w_{i,0}\| + m\theta\beta \max_{i\in V} \{C_i + \nu\} \sum_{\ell=1}^{k} \beta^{k-\ell}\alpha_\ell \\ & + 2\alpha_{k+1} \max_{i\in V}\{C_i + \nu\}. \end{aligned} \tag{5.12}$$

Since $\lim_{k\to\infty} \alpha_k = \alpha$, by Lemma 7(a) we have $\lim_{k\to\infty} \sum_{\ell=1}^{k} \beta^{k-\ell}\alpha_\ell = \frac{\alpha}{1-\beta}$. Using this relation and $\lim_{k\to\infty} \alpha_k = \alpha$, we obtain the result by taking the limit superior in (5.12) as $k \to \infty$. □

When the stepsize is diminishing (i.e., $\alpha = 0$), the result of Theorem 8 implies that the expected disagreements $\mathsf{E}[\|y_{k+1} - w_{j,k+1}\|]$ converge to 0 for all j. Thus, there is an asymptotic consensus in mean. We formally state this as a corollary.

Corollary 2 *Let the conditions of Theorem 8 hold with $\alpha = 0$. Then $\lim_{k\to\infty} \mathsf{E}[\|w_{j,k} - y_k\|] = 0$ for all $j \in V$.*

We next use this to establish convergence with probability 1. We have the following result.

Theorem 9 *Let Assumptions 1, 2, 3,5, 10 and 6 hold. If $\nu_k \leq \nu$ for sufficiently large k and if $\sum_{k=0}^{\infty} \alpha_{k+1}^2 < \infty$, then with probability 1,*

$$\sum_{k=1}^{\infty} \alpha_{k+2}\|y_{k+1} - w_{j,k+1}\| < \infty \qquad \text{for all } j \in V.$$

Furthermore, for all $j \in V$, we have $\lim_{k\to\infty} \|y_{k+1} - w_{j,k+1}\| = 0$ with probability 1 and in mean square.

Proof By Lemma 3 and the subgradient boundedness, we have for all $j \in V$,

$$\begin{aligned}\|y_{k+1} - w_{j,k+1}\| \leq & m\theta\beta^{k+1} \max_{i\in V} \|w_{i,0}\| + \theta \sum_{\ell=1}^{k} \beta^{k+1-\ell} \sum_{i=1}^{m} \alpha_\ell \left(C_i + \|\epsilon_{i,\ell}\|\right) \\ & + \frac{1}{m} \sum_{i=1}^{m} \alpha_{k+1} \left(C_i + \|\epsilon_{i,k+1}\|\right) + \alpha_{k+1} \left(C_j + \|\epsilon_{j,k+1}\|\right). \end{aligned}$$

Using the inequalities

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

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

$$\begin{aligned}
\alpha_{k+2}\|y_{k+1} - w_{j,k+1}\| \leq & \alpha_{k+2}\, m\theta\beta^{k+1} \max_{i\in V} \|w_{i,0}\| \\
& + \theta \sum_{\ell=1}^{k} \beta^{k+1-\ell} \sum_{i=1}^{m} \left(\frac{1}{2}\alpha_{k+2}^2 + \alpha_\ell^2 \left(C_i^2 + \|\epsilon_{i,\ell}\|^2\right)\right) \\
& + \frac{1}{m}\sum_{i=1}^{m} \left(\frac{1}{2}\alpha_{k+2}^2 + \alpha_{k+1}^2 \left(C_i^2 + \|\epsilon_{i,k+1}\|^2\right)\right) \\
& + \frac{1}{2}\alpha_{k+2}^2 + \alpha_{k+1}^2 \left(C_j^2 + \|\epsilon_{j,k+1}\|^2\right).
\end{aligned}$$

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

$$\begin{aligned}
\alpha_{k+2}\|y_{k+1} - w_{j,k+1}\| \leq & \alpha_{k+2}\, m\theta\beta^{k+1} \max_{i\in V} \|w_{i,0}\| + \left(1 + \frac{m\theta\beta}{2(1-\beta)}\right)\alpha_{k+2}^2 \\
& + \theta \sum_{\ell=1}^{k} \alpha_\ell^2 \beta^{k+1-\ell} \sum_{i=1}^{m} \left(C_i^2 + \|\epsilon_{i,\ell}\|^2\right) \\
& + \frac{1}{m}\alpha_{k+1}^2 \sum_{i=1}^{m} \left(C_i^2 + \|\epsilon_{i,k+1}\|^2\right) + \alpha_{k+1}^2 \left(C_j^2 + \|\epsilon_{j,k+1}\|^2\right).
\end{aligned}$$

Taking the conditional expectation and using $\mathsf{E}[\|\epsilon_{i,\ell}\|^2 \mid F_{\ell-1}] \leq \nu^2$, and then taking the expectation again, we obtain

$$\begin{aligned}
\mathsf{E}[\alpha_{k+2}\|y_{k+1} - w_{j,k+1}\|] \leq & \alpha_{k+2}\, m\theta\beta^{k+1} \max_{i\in V} \|w_{i,0}\| + \left(1 + \frac{m\theta\beta}{2(1-\beta)}\right)\alpha_{k+2}^2 \\
& + \theta \left(\sum_{i=1}^{m} \left(C_i^2 + \nu^2\right)\right) \sum_{\ell=1}^{k} \alpha_\ell^2\, \beta^{k+1-\ell} \\
& + \frac{1}{m}\alpha_{k+1}^2 \sum_{i=1}^{m} \left(C_i^2 + \nu^2\right) + \alpha_{k+1}^2 \left(C_j^2 + \nu_j^2\right).
\end{aligned}$$

Since $\sum_k \alpha_k^2 < \infty$ (and hence $\{\alpha_k\}$ bounded), the first two terms and the last

two terms are summable. Furthermore, in view of Lemma 7 [part (b)], we have

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

Thus, the third term is also summable. Hence

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

From the monotone convergence theorem [40], it follows that

$$\mathsf{E}\left[\sum_{k=1}^{\infty}\alpha_{k+2}\|y_{k+1}-w_{j,k+1}\|\right]=\sum_{k=1}^{\infty}\mathsf{E}[\alpha_{k+2}\|y_{k+1}-w_{j,k+1}\|]\,,$$

and it is hence finite for all j. If the expected value of a random variable is finite, then the variable has to be finite with probability 1; thus, with probability 1,

$$\sum_{k=1}^{\infty}\alpha_{k+2}\|y_{k+1}-w_{j,k+1}\|<\infty\qquad\text{for all } j\in V. \tag{5.13}$$

We now show that $\lim_{k\to\infty}\|y_k-w_{j,k}\|=0$ with probability 1 for all $j\in V$. Note that the conditions of Theorem 8 are satisfied with $\alpha=0$. Therefore, $\|y_k-w_{j,k}\|$ converges to 0 in the mean and from (Fatou's) Lemma 9 it follows that

$$0\le\mathsf{E}\Big[\liminf_{k\to\infty}\|y_k-w_{j,k}\|\Big]\le\liminf_{k\to\infty}\mathsf{E}[\|y_k-w_{j,k}\|]=0,$$

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

$$\liminf_{k\to\infty}\|y_k-w_{j,k}\|=0. \tag{5.14}$$

To complete the proof, in view of (5.14) it suffices to show that $\|y_k-w_{j,k}\|$

converges with probability 1. To show this, we define

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

and note that $P_X[r_{i,k+1}] = w_{i,k+1}$ [see (5.1) and (5.2)]. Since $y_k = \frac{1}{m} \sum_{i=1}^{m} w_{i,k}$ and the set X is convex, it follows that $y_k \in X$ for all k. Therefore, by the non-expansive property of the Euclidean projection in (A.4), we have $\|w_{i,k+1} - y_k\|^2 \leq \|r_{i,k+1} - y_k\|^2$ for all $i \in V$ and all k. Summing these relations over all i, we obtain

$$\sum_{i=1}^{m} \|w_{i,k+1} - y_k\|^2 \leq \sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2 \qquad \text{for all } k.$$

From $y_{k+1} = \frac{1}{m} \sum_{i=1}^{m} w_{i,k+1}$ and the fact that the average of vectors minimizes the sum of distances between each vector and arbitrary vector in $\Re^n$ [cf. Eqn (A.1)], we further obtain

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

Therefore, for all k,

$$\sum_{i=1}^{m} \|w_{i,k+1} - y_{k+1}\|^2 \leq \sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2. \tag{5.15}$$

We next relate $\sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2$ to $\sum_{i=1}^{m} \|w_{i,k} - y_k\|^2$. From the definition of $r_{i,k+1}$ and the equality $\sum_{j=1}^{m} a_{i,j}(k+1) = 1$ [cf. Assumption 10b], we have

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

By Assumption 10a and 10b, we have that the weights $a_{i,j}(k+1), j \in V$ yield a convex combination. Thus, by the convexity of the norm [(A.2) and (A.3)] and

by the subgradient boundedness, we have

$$\begin{aligned}\|r_{i,k+1} - y_k\|^2 \leq & \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\|^2 + \alpha_{k+1}^2 \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\|^2 \\ & + 2\alpha_{k+1} \|\nabla f_i(v_{i,k}) + \epsilon_{i,k+1}\| \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\| \\ & \leq \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\|^2 + 2\alpha_{k+1}^2 \left(C_i^2 + \|\epsilon_{i,k+1}\|^2\right) \\ & + 2\alpha_{k+1} \left(C_i + \|\epsilon_{i,k+1}\|\right) \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\| .\end{aligned}$$

Summing over all i and using $\sum_{i=1}^{m} a_{i,j}(k+1) = 1$ [cf. Assumption 10d], we obtain

$$\begin{aligned}\sum_{i=1}^{m} \|r_{i,k+1} - y_k\|^2 \leq & \sum_{j=1}^{m} \|w_{j,k} - y_k\|^2 + 2\alpha_{k+1}^2 \sum_{i=1}^{m} \left(C_i^2 + \|\epsilon_{i,k+1}\|^2\right) \\ & + 2\alpha_{k+1} \sum_{i=1}^{m} \left(C_i + \|\epsilon_{i,k+1}\|\right) \sum_{j=1}^{m} a_{i,j}(k+1) \|w_{j,k} - y_k\| .\end{aligned}$$

Using this in (5.15) and taking the conditional expectation, we see that for all k, we have with probability 1,

$$\begin{aligned}\sum_{i=1}^{m} \mathsf{E}\left[\|w_{i,k+1} - y_{k+1}\|^2 \mid F_k\right] \leq & \sum_{i=1}^{m} \|w_{i,k} - y_k\|^2 + 2\alpha_{k+1}^2 \sum_{i=1}^{m} \left(C_i^2 + \nu^2\right) \\ & + 2\alpha_{k+1} \sum_{i=1}^{m} \left(C_i + \nu\right) \sum_{j=1}^{m} \|w_{j,k} - y_k\| , \qquad (5.16)\end{aligned}$$

where we use $a_{i,j}(k+1) \leq 1$ for all i, j and k, and the relations $\mathsf{E}[\|\epsilon_{i,k+1}\|^2 \mid F_k] \leq \nu^2$, $\mathsf{E}[\|\epsilon_{i,k+1}\| \mid F_k] \leq \nu$ hold with probability 1.

We now apply Theorem 16 to the relation in (5.16). To verify that the conditions of Theorem 16 are satisfied, note that the stepsize satisfies $\sum_{k=1}^{\infty} \alpha_{k+1}^2 < \infty$ for all $i \in V$. We also have $\sum_{k=1}^{\infty} \alpha_{k+1} \|w_{j,k} - y_k\| < \infty$ with probability 1 [cf. (5.13)]. Therefore, the relation in (5.16) satisfies the conditions of Theorem 16 with $\zeta_k = D_k = 0$, thus implying that $\|w_{j,k} - y_k\|$ converges with

probability 1 for every $j \in V$. $\square$

Let us compare Theorem 9 and Corollary 2. Corollary 2 provided sufficient conditions for the different agents to have consensus in the mean. Theorem 9 strengthens this to consensus with probability 1 and in mean square sense, for a smaller class of stepsize sequences under a stricter assumption.

We next show that the consensus vector is actually in the optimal set, provided that the optimal set is nonempty and the conditional expectations $\|\mathsf{E}[\epsilon_{i,k+1} \mid F_k]\|$ are diminishing.

Theorem 10 *Let Assumptions 1, 2, 3, 5, 10 and 6 hold. Also, assume that* $\sum_{k=0}^{\infty} \nu_k^2 < \infty$ *for all* $i \in V$. *Further, let the stepsize sequence* $\{\alpha_k\}$ *be such that* $\sum_{k=1}^{\infty} \alpha_k = \infty$ *and* $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. *Then, if the optimal set* X^* *is nonempty, the iterate sequence* $\{w_{i,k}\}$ *of each agent* $i \in V$ *converges to the same optimal point with probability 1 and in mean square.*

Proof Since $\sum_{k=0}^{\infty} \nu_k^2 < \infty$ for all $i \in V$ it implies that we can find a ν such that $\nu_k < \nu$ for sufficiently large k. Thus the conditions of Lemma 4 are satisfied. Letting $z = x^*$ for some $x^* \in X^*$, taking conditional expectations and using the bounds on the error moments, we obtain for any $x^* \in X^*$ and any k, with probability 1,

$$\begin{aligned}\sum_{i=1}^{m} \mathsf{E}\big[\|v_{i,k+1} - x^*\|^2 \mid F_k\big] \leq & \sum_{i=1}^{m} \|v_{i,k} - x^*\|^2 - 2\alpha_{k+1}\left(f(y_k) - f^*\right) \\ & + 2\alpha_{k+1}\left(\max_{i \in V} C_i\right) \sum_{j=1}^{m} \|y_k - w_{j,k}\| \\ & + 2\alpha_{k+1} \sum_{i=1}^{m} \mu_{i,k+1} \|v_{i,k} - x^*\| + \alpha_{k+1}^2 \sum_{i=1}^{m} \left(C_i + \nu\right)^2,\end{aligned}$$

where $f^* = f(x^*)$, and we use the notation $\mu_{i,k+1} = \|\mathsf{E}[\epsilon_{i,k+1} \mid F_k]\|$. Using the inequality

$$2\alpha_{k+1}\mu_{i,k+1}\|v_{i,k} - x^*\| \leq \alpha_{k+1}^2 \|v_{i,k} - x^*\|^2 + \mu_{i,k+1}^2,$$

we obtain with probability 1,

$$\sum_{i=1}^{m} \mathsf{E}\big[\|v_{i,k+1} - x^*\|^2 \mid F_k\big] \leq \sum_{i=1}^{m} \big(1 + \alpha_{k+1}^2\big) \|v_{i,k} - x^*\|^2 - 2\alpha_{k+1} \left((f(y_k) - f^*) - \left(\max_{i \in V} C_i \right) \sum_{j=1}^{m} \|y_k - w_{j,k}\| + \sum_{i=1}^{m} \mu_{i,k+1}^2 - \frac{1}{2}\, \alpha_{k+1} \sum_{i=1}^{m} (C_i + \nu)^2 \right). \tag{5.17}$$

By Theorem 9, we have with probability 1,

$$\sum_k \alpha_{k+1} \|w_{j,k} - y_k\| < \infty.$$

Further, since $\sum_k \mu_{i,k}^2 < \infty$ and $\sum_k \alpha_k^2 < \infty$ with probability 1, the relation in (5.17) satisfies the conditions of Theorem 16. We therefore have

$$\sum_k \alpha_k (f(y_k) - f^*) < \infty, \tag{5.18}$$

and $\|v_{i,k} - x^*\|$ converges with probability 1 and in mean square. In addition, by Theorem 9, we have $\lim_{k\to\infty} \|w_{i,k} - y_k\| = 0$ for all i, with probability 1. Hence, $\lim_{k\to\infty} \|v_{i,k} - y_k\| \to 0$ for all i, with probability 1. Therefore, $\|y_k - x^*\|$ converges with probability 1 for any $x^* \in X^*$. Moreover, from (5.18) and the fact that $\sum_k \alpha_k = \infty$, by continuity of f, it follows that y_k, and hence $w_{i,k}$, must converge to a vector in X^* with probability 1 and in mean square. □

Note that the result of Theorem 10 holds without assuming compactness of the constraint set X. This was possible due to the assumption that both the stepsize α_k and the norms $\|\mathsf{E}[\epsilon_{i,k+1} \mid F_k]\|$ of the conditional errors are square summable. In addition, note that the result of Theorem 10 remains valid when the condition $\sum_{k=0}^{\infty} \|\mathsf{E}[\epsilon_{i,k+1} \mid F_k]\|^2 < \infty$ for all i is replaced with $\sum_{k=0}^{\infty} \alpha_{k+1} \|\mathsf{E}[\epsilon_{i,k+1} \mid F_k]\| < \infty$ for all i.

5.2.2 Constant stepsizes

Theorem 11 *Let Assumptions 1, 2, 5, 10 and 6 hold. Assume that the set X is bounded. Let $\lim_{k\to\infty}\alpha_k = \alpha$ with $\alpha \geq 0$. If $\alpha = 0$, also assume that $\sum_k \alpha_k = \infty$. Then, for all $j \in V$,*

$$\liminf_{k\to\infty} \mathsf{E}[f(w_{j,k})] \leq f^* + \max_{x,y\in X} \|x-y\| \sum_{i=1}^{m} \bar{\mu} + m\alpha \left(\max_{i\in V}\{C_i + \bar{\nu}\}\right)^2 \left(\frac{9}{2} + \frac{2m\theta\beta}{1-\beta}\right),$$

where $\bar{\mu} = \operatorname{limsup}_{k\to\infty} \mu_k$.

Proof Under Assumption 6, the limit superiors $\bar{\mu} = \operatorname{limsup}_{k\to\infty} \|\mathsf{E}[\epsilon_{i,k+1}]\|$ are finite. Since the set X is bounded the subgradients of f_i over the set X are also bounded for each $i \in V$; hence, the bounds $C_i,\ i \in V$ on subgradient norms exist. Thus, the conditions of Lemma 4 are satisfied. Further, by Assumption 2, the set X is contained in the interior of the domain of f, over which the function is continuous (by convexity; see [8]). Thus, the set X is compact and f is continuous over X, implying that the optimal set X^* is nonempty. Let $x^* \in X^*$, and let $y = x^*$ in Lemma 4. We have, for all k,

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

Since X is bounded, by using $\|v_{i,k} - x^*\| \leq \max_{x,y\in X} \|x-y\|$, taking the expectation and using the error bounds $\mathsf{E}[\|\epsilon_{i,k+1}\|^2] \leq \bar{\nu}^2$ we obtain

$$\begin{aligned}\sum_{i=1}^{m} \mathsf{E}\left[\|v_{i,k+1} - x^*\|^2\right] \ \leq\ & \sum_{i=1}^{m} \mathsf{E}\left[\|v_{i,k} - x^*\|^2\right] - 2\alpha_{k+1}\ (\mathsf{E}[f(y_k)] - f^*) \\ & +2\alpha_{k+1}\left(\max_{i\in V} C_i\right) \sum_{j=1}^{m} \mathsf{E}[\|y_k - w_{j,k}\|]\end{aligned}$$

$$+2\alpha_{k+1} \max_{x,y\in X} \|x-y\| \sum_{i=1}^{m} \|\mathsf{E}[\epsilon_{i,k+1}]\| \tag{5.19}$$

$$+\alpha_{k+1}^2 \sum_{i=1}^{m} (C_i+\bar{\nu})^2 . \tag{5.20}$$

By rearranging the terms and summing over $k = 1, \ldots, K$, for an arbitrary K, we obtain

$$\begin{aligned}
&2\sum_{k=1}^{K} \alpha_{k+1} \left((\mathsf{E}[f(y_k)] - f^*) - \left(\max_{i\in V} C_i\right) \sum_{j=1}^{m} \mathsf{E}[\|y_k - w_{j,k}\|] \right. \\
&\left. - \max_{x,y\in X} \|x-y\| \sum_{i=1}^{m} \|\mathsf{E}[\epsilon_{i,k+1}]\| - \frac{m\alpha_{k+1}}{2} \left(\max_{i\in V}\{C_i+\bar{\nu}\}\right)^2 \right) \\
&\leq \sum_{i=1}^{m} \mathsf{E}\left[\|v_{i,1} - x^*\|^2\right] - \sum_{i=1}^{m} \mathsf{E}\left[\|v_{i,K+1} - x^*\|^2\right] \leq m \max_{x,y\in X} \|x-y\|^2.
\end{aligned}$$

Note that when $\alpha_{k+1} \to \alpha$ and $\alpha > 0$, we have $\sum_k \alpha_k = \infty$. When $\alpha = 0$, we have assumed that $\sum_k \alpha_k = \infty$. Therefore, by letting $K \to \infty$, we have

$$\begin{aligned}
&\liminf_{k\to\infty} \left(\mathsf{E}[f(y_k)] - \left(\max_{i\in V} C_i\right) \sum_{j=1}^{m} \mathsf{E}[\|y_k - w_{j,k}\|] \right. \\
&\left. - \max_{x,y\in X} \|x-y\| \sum_{i=1}^{m} \|\mathsf{E}[\epsilon_{i,k+1}]\| - \frac{m\alpha_{k+1}}{2} \left(\max_{i\in V}\{C_i+\bar{\nu}\}\right)^2 \right) \leq f^*.
\end{aligned}$$

Using $\operatorname{limsup}_{k\to\infty} \|\mathsf{E}[\epsilon_{i,k+1}]\| = \bar{\mu}$ and $\lim_{k\to\infty} \alpha_k = \alpha$, we obtain

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

Next from the convexity inequality in (2.2) and the boundedness of the subgradients it follows that for all k and $j \in V$,

$$\mathsf{E}[f(w_{j,k}) - f(y_k)] \leq \left(\sum_{i=1}^{m} C_i\right) \mathsf{E}[\|y_k - w_{j,k}\|] ,$$

implying

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

By Theorem 8, we have for all $j \in V$,

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

By using the preceding relation, we see that

$$\begin{aligned}\liminf_{k\to\infty}\mathsf{E}[f(w_{j,k})] \leq & f^* + \frac{m\alpha}{2}\left(\max_{i\in V}\{C_i+\bar{\nu}\}\right)^2 + \max_{x,y\in X}\|x-y\|\sum_{i=1}^{m}\bar{\mu}_i \\ & + m\alpha\left(\max_{i\in V} C_i\right)\max_{i\in V}\{C_i+\bar{\nu}\}\left(2+\frac{m\theta\beta}{1-\beta}\right) \\ & + \alpha\left(\sum_{i=1}^{m} C_i\right)\max_{j\in V}\{C_j+\bar{\nu}_j\}\left(2+\frac{m\theta\beta}{1-\beta}\right) \\ \leq & f^* + \max_{x,y\in X}\|x-y\|\sum_{i=1}^{m}\bar{\mu} + m\alpha\left(\max_{i\in V}\{C_i+\bar{\nu}\}\right)^2\left(\frac{9}{2}+\frac{2m\theta\beta}{1-\beta}\right).\end{aligned}$$

□

The network topology influences the error only through the term $\frac{\theta\beta}{1-\beta}$ and can hence be used as a figure of merit for comparing different topologies. For a network that is strongly connected at every time [i.e., $Q = 1$ in Assumption 5], and when η in Assumption 10 does not depend on the number m of agents, the term $\frac{\theta\beta}{1-\beta}$ is of the order m^2 and the error bound scales as m^4.

5.2.3 Rate of convergence

We next obtain a rate of convergence result. The result for the general case, though straightforward, is computationally involved. We therefore consider only

the case when a constant stepsize α is used. The vector $z_{j,t}$ is the running average of all the iterates of agent j until time t, i.e., $z_{j,t} = \frac{1}{t}\sum_{k=1}^{t} w_{j,k}$. The proof is similar to the proof in Chapter 3.2.3.

Theorem 12 *Let Assumptions 1, 2, 5, 10 and 6 hold. Assume that the set X is bounded. Then*

$$\begin{aligned}\mathsf{E}[f(z_{j,t})] \;\leq\; & f^* + \frac{1}{2t\alpha}\sum_{i=1}^{m}\mathsf{E}\big[\|v_{i,1}-x^*\|^2\big] \\ & + \left(\max_{i\in V} C_i\right)\frac{1}{t}\sum_{k=1}^{t}\left(m\mathsf{E}[\|y_k - w_{j,k}\|] + \sum_{i=1}^{m}\mathsf{E}[\|y_k - w_{i,k}\|]\right) \\ & + \frac{\alpha}{2}\sum_{i=1}^{m}(C_i+\bar{\nu})^2. \end{aligned} \tag{5.21}$$

CHAPTER 6

EXTENSION: A GENERAL DISTRIBUTED OPTIMIZATION PROBLEM

In this chapter, we solve the following problem. Consider a network of m agents indexed by $V = \{1, \ldots, m\}$. The goal is to solve the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \tilde{f}(x) := g\left(\sum_{i=1}^{m} h_i(x)\right) \\ \text{subject to} \quad & x \in X, \end{aligned} \tag{6.1}$$

where $g : \Re \to \Re$, $X \subseteq \Re^p$, $h_i : X \to \Re$ for all $i \in V$. The function h_i is known only to agent i. The function g and the set X are globally known, i.e., to every agent. Further, the network size m is also known to all the agents. Associated with the optimization problem we use the notation

$$\tilde{f}^* = \min_{x \in X} \tilde{f}(x), \qquad \tilde{X}^* = \{x \in X : \tilde{f}(x) = \tilde{f}^*\}.$$

We propose the following iterative algorithm to solve problem (6.1). At the end of iteration k, agent i maintains two statistics: $x_{i,k}$ and $s_{i,k}$. The statistic $x_{i,k}$ is agent i's estimate of an optimal point and $s_{i,k}$ is agent i's estimate of $\frac{1}{m}\sum_{i=1}^{m} h_i(x_{i,k})$. These are updated as follows:

$$\begin{aligned} \begin{bmatrix} \bar{x}_{i,k} \\ \bar{s}_{i,k} \end{bmatrix} &= \sum_{j \in N_i(k+1)} a_{i,j}(k+1) \begin{bmatrix} x_{i,k} \\ s_{i,k} \end{bmatrix}, \\ x_{i,k+1} &= P_X\left[\bar{x}_{i,k} - \alpha_{k+1} g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right)\right], \\ s_{i,k+1} &= \bar{s}_{i,k} + h_i(x_{i,k+1}) - h_i(x_{i,k}). \end{aligned} \tag{6.2}$$

Here $N_i(k+1)$ is the set of agents that agent i can communicate with at the time of the k-th iteration and also includes agent i. Futher, $a_{i,j}(k+1)$ are positive weights, α_{k+1} is the stepsize, g' is the derivative of g, ∇h_i is the gradient of h_i and P_X denotes Euclidean projection onto the set X.

The algorithm is distributed and local. Agent i receives $x_{i,k}$ and $s_{i,k}$ from its current immediate neighbors and calculates the weighted averages $\bar{x}_{i,k}$ and $\bar{s}_{i,k}$ using the weights $a_{i,j}(k+1)$. The weighted average is then updated using locally available information (functions g and h_i, number of agents m and set X) to generate $x_{i,k+1}$ and $s_{i,k+1}$. The algorithm is initialized with

$$x_{i,0} \in X, \qquad s_{i,0} = h_i(x_{i,0}) \qquad \text{for all } i \in V. \tag{6.3}$$

When convenient we will use the notation $z_{i,k} = [x_{i,k} \;\; s_{i,k}]^T$ and $\bar{z}_{i,k} = [\bar{x}_{i,k} \;\; \bar{s}_{i,k}]^T$.

We state a result that captures the effect of deterministic errors in the standard distributed averaging algorithm. The result guarantees that the agents achieve consensus when the errors diminish, but the consensus point may not necessarily be the target value. The proof is very similar to the proof of Lemma 3 in Chapter 5.

Theorem 13 *Let Assumptions 5 and 10 hold. Consider the iterates generated by*

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

Suppose there exists a non-negative non-increasing scalar sequence $\{\alpha_k\}$ such that

$$\sum_k \alpha_k \|\epsilon_{i,k}\| < \infty$$

for all $i \in V$; then for all $i, j \in V$,

$$\sum_k \alpha_k \left\|\theta_{i,k} - \theta_{j,k}\right\| < \infty.$$

Since $a_{i,j}(k+1) = 0$ when $j \notin N_i(k+1)$ we can rewrite (6.2) as

$$\begin{aligned}
\begin{bmatrix} \bar{x}_{i,k} \\ \bar{s}_{i,k} \end{bmatrix} &= \sum_{j=1}^{m} a_{i,j}(k+1) \begin{bmatrix} x_{i,k} \\ s_{i,k} \end{bmatrix}, \\
x_{i,k+1} &= P_X \left[\bar{x}_{i,k} - \alpha_{k+1} g' \left(m \bar{s}_{i,k} \right) \nabla h_i \left(\bar{x}_{i,k} \right) \right], \\
s_{i,k+1} &= \bar{s}_{i,k} + h_i(x_{i,k+1}) - h_i(x_{i,k}).
\end{aligned} \tag{6.4}$$

We next provide some intuition for the algorithm. Consider the standard gradient projection algorithm to solve (6.1). The iterates are generated according to

$$x_{k+1} = P_X \left[x_k - \alpha_{k+1} \; g' \left(\sum_{j=1}^{m} h_j \left(x_k \right) \right) \sum_{j=1}^{m} \nabla h_j \left(x_k \right) \right].$$

To replicate the standard gradient projection algorithm in our distributed setting, the computations of $\sum_{j=1}^{m} \nabla h_j \left(x_k \right)$ and $\sum_{j=1}^{m} h_j \left(x_k \right)$ have to be distributed and compliant with the local connectivity structure of each agent. When the function g is the identity function then (6.2) is identical to the distributed subgradient algorithm in [41]. As in [41], by using a weighted average $\bar{x}_{i,k}$ of its own and its neighbors estimates, the agent quantity $\nabla h_i \left(\bar{x}_{i,k} \right)$ approximates $\sum_{j=1}^{m} \nabla h_j \left(\bar{x}_{i,k} \right)$ with decreasing error as time k increases.

The term $\bar{s}_{i,k}$ is essentially an approximation for $\frac{1}{m} \sum_{j=1}^{m} h_j \left(x_{i,k} \right)$, and therefore $g' \left(m \bar{s}_{i,k} \right)$ approximates $g' \left(\sum_{j=1}^{m} h_j \left(x_k \right) \right)$. In iteration $k+1$, each agent in the network is interested in determining $\frac{1}{m} \sum_{j=1}^{m} h_j \left(x_{j,k+1} \right)$ while each term $h_i \left(x_{i,k+1} \right)$ is available only to agent i. If agent i were to use $h_i \left(x_{i,k+1} \right)$ as the start value, then a large number of consensus steps would be required for agent i to obtain a good approximation to $\frac{1}{m} \sum_{j=1}^{m} h_j \left(x_{j,k+1} \right)$. As an alternative, we consider a more efficient procedure that uses $\bar{s}_{i,k} + h_i(x_{i,k+1}) - h_i(x_{i,k})$ as the start value to estimate $\frac{1}{m} \sum_{j=1}^{m} h_j \left(x_{j,k+1} \right)$. In this way, a single consensus step is enough to obtain a sufficiently good approximation. To see this, suppose that

$\bar{s}_{i,k}$ is a good approximation for $\frac{1}{m}\sum_{j=1}^{m} h_j\left(x_{j,k}\right)$. Then

$$\frac{1}{m}\sum_{i=1}^{m} \bar{s}_{i,k} + h_i(x_{i,k+1}) - h_i(x_{i,k}) \approx \frac{1}{m}\sum_{j=1}^{m} h_j\left(x_{j,k+1}\right).$$

When the difference between $x_{i,k+1}$ and $x_{i,k}$ is small, the difference between $\sum_{j=1}^{m} h_j\left(x_{j,k+1}\right)$ and $\sum_{j=1}^{m} h_j\left(x_{j,k}\right)$ is also small. Thus, the value $\bar{s}_{i,k} + h_i(x_{i,k+1}) - h_i(x_{i,k})$ is closer to the target value than just $h_i(x_{i,k+1})$. This approach to tracking the network-wide average of a changing statistic is reminiscent of the consensus filters than have been proposed in the literature [42].

We now formally establish the convergence of the algorithm. We first characterize the rate of consensus.

Lemma 5 *Let Assumptions 7, 5, and 10 hold. If $\{\alpha_k\}$ is a non-negative non-increasing sequence such that $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, then for all $i, j \in V$,*

$$\sum_{k=1}^{\infty} \alpha_k \left\|x_{i,k} - x_{j,k}\right\| < \infty.$$

Proof Note from (6.4) that

$$x_{i,k+1} \;=\; \sum_{j=1}^{m} a_{i,j}(k+1)x_{j,k} + p_{i,k+1}, \tag{6.5}$$

where $p_{i,k+1} = x_{i,k+1} - \bar{x}_{i,k}$. From (6.4), the Euclidean projection property in (A.4) and the boundedness of the gradients (consequence of Assumption 7(b) and Assumption 7(d)) we obtain

$$\left\|p_{i,k+1}\right\| \leq \alpha_{k+1} \left\|g'\left(m\bar{s}_{i,k}\right) \nabla h_i(x_{i,k})\right\| \leq \alpha_{k+1} C^2. \tag{6.6}$$

Since $\sum_k \alpha_k^2 < \infty$ we conclude that $\sum_k \alpha_k \|p_{i,k}\| < \infty$. Therefore (6.5) satisfies the conditions of Theorem. 13 and the result follows.

An immediate consequence is that the agents achieve consensus, i.e.,

asymptotically they agree. Define

$$\hat{x}_k = \frac{1}{m}\sum_{j=1}^{m} x_{j,k}, \qquad \hat{s}_k = \frac{1}{m}\sum_{j=1}^{m} h_j\left(\hat{x}_k\right).$$

We next characterize the rate of consensus of $\{s_{i,k}\}$.

Lemma 6 *Let Assumptions 7, 5 and 10 hold. If $\{\alpha_k\}$ is a non-negative non-increasing sequence such that $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$ then $\sum_{k=1}^{\infty}\|s_{i,k} - \hat{s}_k\| < \infty$ for all $i \in V$.*

Proof Using the triangle inequality, we obtain

$$\|s_{i,k} - \hat{s}_k\| \leq \left\| s_{i,k} - \frac{1}{m}\sum_{j=1}^{m} s_{j,k} \right\| + \left\| \frac{1}{m}\sum_{j=1}^{m} s_{j,k} - \hat{s}_k \right\|. \tag{6.7}$$

We first consider the last term and show that

$$\sum_{i=1}^{m} s_{i,k} = \sum_{i=1}^{m} h_i(x_{i,k}). \tag{6.8}$$

We will use the induction on k. For $k = 0$, from (6.3) we have $s_{i,0} = h_i(x_{i,0})$ and, hence, the hypothesis is true for $k = 0$. Now, assume that the hypothesis is true for $k-1$ and consider $\sum_{i=1}^{m} s_{i,k}$. Observe that from Assumption 10(d) we have

$$\sum_{i=1}^{m} s_{i,k} = \sum_{i=1}^{m}\sum_{j=1}^{m} a_{i,j}(k+1)\bar{s}_{j,k} = \sum_{j=1}^{m} \bar{s}_{j,k}.$$

By the definition of $s_{i,k}$ in (6.4) and the induction hypothesis for $k-1$, we conclude that

$$\sum_{j=1}^{m} \bar{s}_{j,k} \;=\; \sum_{j=1}^{m} s_{j,k-1} + \sum_{j=1}^{m} h_j(x_{j,k}) - \sum_{j=1}^{m} h_j(x_{j,k-1}) = \sum_{j=1}^{m} h_j(x_{j,k}).$$

This proves the induction hypothesis for k and hence (6.8) follows. Using (6.8)

in (6.7) and substituting for $\hat{s}_k$ we obtain

$$\|s_{i,k} - \hat{s}_k\| \leq \left\| s_{i,k} - \frac{1}{m}\sum_{j=1}^{m} s_{j,k} \right\| + \left\| \frac{1}{m}\sum_{j=1}^{m} h_j(x_{j,k}) - \frac{1}{m}\sum_{j=1}^{m} h_j(\hat{x}_k) \right\|. \tag{6.9}$$

We now deal with the first term on the right-hand side of (6.9). Note that we can rewrite

$$s_{i,k+1} = \sum_{j=1}^{m} a_{i,j}(k+1)s_{i,k} + w_{i,k+1}, \tag{6.10}$$

where $w_{i,k+1} = h_i(x_{i,k+1}) - h_i(x_{i,k})$. Since the gradient of h_i is bounded by C, the function h_i is Lipschitz continuous with C. Using this and the definition of $x_{i,k+1}$ in (6.4), we get

$$\|w_{i,k+1}\| \leq C\|x_{i,k+1} - x_{i,k}\| \leq C\|P_X\left[\bar{x}_{i,k} - \alpha_{k+1}g'\left(m\bar{s}_{i,k}\right)\nabla h_i\left(\bar{x}_{i,k}\right)\right] - x_{i,k}\|.$$

Further, by using the Euclidean projection property in (A.4), we have

$$\begin{aligned}
\|w_{i,k+1}\| &\leq C\left\|\bar{x}_{i,k} - \alpha_{k+1}g'\left(m\bar{s}_{i,k}\right)\nabla h_i\left(\bar{x}_{i,k}\right) - x_{i,k}\right\| \\
&\leq C\left(\|\bar{x}_{i,k} - x_{i,k}\| + \alpha_{k+1}\|g'\left(m\bar{s}_{i,k}\right)\nabla h_i\left(\bar{x}_{i,k}\right)\|\right) \\
&\leq C\left(\|\bar{x}_{i,k} - x_{i,k}\| + \alpha_{k+1}C^2\right),
\end{aligned} \tag{6.11}$$

where the last inequality follows from the boundedness of the gradients (a consequence of Assumption 7(b) and Assumption 7(d)). Next note that the conditions of Lemma 5 are satisfied. Therefore for all $i, j \in V$

$$\sum_k \alpha_k \|x_{j,k} - x_{i,k}\| < \infty,$$

and hence for all $i \in V$

$$\sum_k \alpha_k \|x_{j,k} - \bar{x}_{i,k}\| < \infty.$$

Since $\{\alpha_k\}$ is a non-increasing sequence this implies

$$\sum_k \alpha_{k+1} \|x_{j,k} - \bar{x}_{i,k}\| < \infty.$$

Using the preceding inequality in (6.11) and the fact that $\sum_k \alpha_k^2 < \infty$, we can conclude that

$$\sum_k \alpha_{k+1} \|w_{i,k+1}\| < \infty.$$

Thus (6.5) satisfies the conditions of Theorem 13 and we can conclude that for all $i, j \in V$

$$\sum_k \alpha_k \|s_{i,k} - s_{j,k}\| < \infty,$$

and hence

$$\sum_k \alpha_k \left\| s_{i,k} - \frac{1}{m} \sum_{j=1}^m s_{j,k} \right\| < \infty. \tag{6.12}$$

We now consider the term $\left\| \frac{1}{m} \sum_{j=1}^m h_j(x_{j,k}) - \frac{1}{m} \sum_{j=1}^m h_j(\hat{x}_k) \right\|$ on the right-hand side of (6.9). From the Lipschitz continuity of h_j, we have

$$\|h_j(x_{j,k}) - h_j(\hat{x}_k)\| \leq L \|x_{j,k} - \hat{x}_k\|.$$

According to our notation, we have $\hat{x}_k = \frac{1}{m} \sum_{j=1}^m x_{j,k}$, so that

$$\|h_j(x_{j,k}) - h_j(\hat{x}_k)\| \leq L \left\| x_{j,k} - \frac{1}{m} \sum_{i=1}^m x_{i,k} \right\| \leq \frac{L}{m} \sum_{i=1}^m \|x_{j,k} - x_{i,k}\|.$$

Since the conditions of Lemma 5 are satisfied we can conclude that for all $i \in V$

$$\sum_k \alpha_k \|h_i(\hat{x}_k) - h_i(x_{i,k})\| \leq \frac{L}{m} \sum_k \alpha_k \sum_{i=1}^m \|x_{j,k} - x_{i,k}\| < \infty.$$

Therefore, from (6.9) and (6.12) we get for all $i \in V$

$$\sum_k \alpha_k \|s_{i,k} - \hat{s}_k\| < \infty.$$

We next use Lemmas 5 and 6 to prove convergence to an optimal point.

Theorem 14 *Let Assumptions 7, 5 and 10 hold. If the stepsize sequence* $\{\alpha_k\}$ *is non-increasing, and such that* $\sum_k \alpha_k = \infty$ *and* $\sum_k \alpha_k^2 < \infty$, *then there exists a vector* $x^* \in \tilde{X}^*$ *such that* $\lim_{k\to\infty} \|x_{i,k} - x^*\| = 0$ *for all* $i \in V$.

Proof Note that the solution set $\tilde{X}^*$ is nonempty since h is continuous and X is compact. Fix an arbitrary $x^* \in \tilde{X}^*$. By the Euclidean projection property in (A.4) we have

$$\begin{aligned} \|x_{i,k+1} - x^*\|^2 &= \|P_X\left[\bar{x}_{i,k} - \alpha_{k+1} g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right)\right] - x^*\|^2 \\ &\leq \|\bar{x}_{i,k} - x^*\|^2 + \alpha_{k+1}^2 \left\|g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right)\right\|^2 \\ &\quad -2\alpha_{k+1}(\bar{x}_{i,k} - x^*)^T g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right). \end{aligned} \tag{6.13}$$

Using the boundedness of the gradients (consequence of Assumption 7(b) and Assumption 7(d)), we obtain

$$\left\|g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right)\right\| \leq C^2.$$

By Assumption 10 on the weights, we have

$$\begin{aligned} \sum_{i=1}^m \|\bar{x}_{i,k} - x^*\|^2 &= \sum_{i=1}^m \left\| \sum_{j=1}^m a_{i,j}(k+1) x_{j,k} - x^* \right\|^2 \\ &\leq \sum_{i=1}^m \sum_{j=1}^m a_{i,j}(k+1) \|x_{j,k} - x^*\|^2 \\ &\leq \sum_{i=1}^m \|x_{i,k} - x^*\|^2. \end{aligned}$$

Summing (6.13) over all i and using the preceding two relations, we obtain

$$\sum_{i=1}^{m} \|x_{i,k+1} - x^*\|^2 \leq \sum_{i=1}^{m} \|x_{i,k} - x^*\|^2 + m\alpha_{k+1}^2 C^4$$
$$-2\alpha_{k+1} \sum_{i=1}^{m} (\bar{x}_{i,k} - x^*)^T g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right). \tag{6.14}$$

Recall that we defined $\hat{x}_k$ as $\frac{1}{m}\sum_{j=1}^{m} x_{j,k}$. Note that $\nabla \tilde{f}(\hat{x}_k) = \sum_{i=1}^{m} g'\left(m\hat{s}_k\right) \nabla h_i\left(\hat{x}_k\right)$. Using this we can write

$$\sum_{i=1}^{m} (\bar{x}_{i,k} - x^*)^T \left(g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right)\right) = (\hat{x}_k - x^*)^T \nabla \tilde{f}(\hat{x}_k)$$
$$+ \sum_{i=1}^{m} (\bar{x}_{i,k} - \hat{x}_k)^T \left(g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right)\right)$$
$$+ \sum_{i=1}^{m} (\hat{x}_k - x^*)^T \left(g'\left(m\bar{s}_{i,k}\right) - g'\left(m\hat{s}_k\right)\right) \nabla h_i\left(\bar{x}_{i,k}\right)$$
$$+ \sum_{i=1}^{m} (\hat{x}_k - x^*)^T g'\left(m\hat{s}_k\right) \left(\nabla h_i\left(\bar{x}_{i,k}\right) - \nabla h_i\left(\hat{x}_k\right)\right).$$

We next use Assumption 7(b), (c) and the boundedness of the gradients (a consequence of Assumption 7(b) and 7(d)) to bound the last two terms. We obtain

$$\begin{aligned}
\sum_{i=1}^{m} (\bar{x}_{i,k} - x^*)^T \left(g'\left(m\bar{s}_{i,k}\right) \nabla h_i\left(\bar{x}_{i,k}\right)\right) &\geq (\hat{x}_k - x^*)^T \nabla \tilde{f}(\hat{x}_k) - C^2 \sum_{i=1}^{m} \|\bar{x}_{i,k} - \hat{x}_k\| \\
&\quad - DCL \sum_{i=1}^{m} (m\|\bar{s}_{i,k} - \hat{s}_k\| + \|\bar{x}_{i,k} - \hat{x}_k\|) \\
&\geq \left(\tilde{f}(\hat{x}_k) - \tilde{f}(x^*)\right) - C^2 \sum_{i=1}^{m} \|\bar{x}_{i,k} - \hat{x}_k\| \\
&\quad - DCL \sum_{i=1}^{m} (m\|\bar{s}_{i,k} - \hat{s}_k\| + \|\bar{x}_{i,k} - \hat{x}_k\|).
\end{aligned}$$

Here D is the diameter of the set, which is finite from Assumption 7(b). The last step follows from the convexity of the function f in Assumption 7(c). Using the

preceding relation in (6.14) we obtain

$$\sum_{i=1}^{m}\|x_{i,k+1}-x^*\|^2 \;\leq\; \sum_{i=1}^{m}\|x_{i,k}-x^*\|^2 - 2\alpha_{k+1}\left(\tilde{f}(\hat{x}_k)-f^*\right) + 2C^2\sum_{i=1}^{m}\alpha_{k+1}\|\bar{x}_{i,k}-\hat{x}_k\| \\ +2DCL\sum_{i=1}^{m}\alpha_{k+1}(m\|\bar{s}_{i,k}-\hat{s}_k\|+\|\bar{x}_{i,k}-\hat{x}_k\|)+m\alpha_{k+1}^2C^4.$$

From Lemma 5 we have

$$\sum_k \alpha_{k+1}\|\bar{x}_{i,k}-\hat{x}_k\|<\infty \qquad \text{for all } i\in V,$$

and from Lemma 6 we have

$$\sum_k \alpha_{k+1}\|\bar{s}_{i,k}-\hat{s}_k\|<\infty.$$

Thus the conditions of Lemma 16 are satisfied and we can conclude that $\|x_{i,k+1}-x^*\|$ converges for every $x^*\in\tilde{X}^*$ and every $i\in V$, and

$$\sum_k \alpha_{k+1}\left(\tilde{f}(\hat{x}_k)-f^*\right)<\infty.$$

This implies that the sequences $\{x_{i,k}\},\ i\in V$, must converge to a common point in the set $\tilde{X}^*$

6.1 Extensions

We next discuss two extensions of the problem in (6.1) and generalize the algorithm in (6.4) to solve these extensions.

6.1.1 Extension I

Consider the following general distributed optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{m} g_j \left(\sum_{i=1}^{m} h_{i,j}(x) \right) \\ \text{subject to} \quad & x \in X, \end{aligned} \tag{6.15}$$

where $g_j : \Re \to \Re$, $X \subseteq \Re^p$, $h_{i,j} : X \to \Re$ for all $i, j \in V$. The functions $h_{i,j}$, $j \in V$, are known only to agent i. The functions g_j and the set X are globally known. Note here that the index set for j can be other than V. We prefer to keep $j \in V$ for simplicity of notation.

We can modify (6.4) to solve (6.15) as follows. Let $s_{i,k}$ now denote agent i's estimate of the vector $\left[\frac{1}{m} \sum_{i=1}^{m} h_{i,1}(x_{i,k}) \quad \dots \quad \frac{1}{m} \sum_{i=1}^{m} h_{i,m}(x_{i,k}) \right]^T$. Agent i recursively generates $x_{i,k+1}$ and $s_{i,k+1}$ according to the following rules:

$$\begin{aligned} \begin{bmatrix} \bar{x}_{i,k} \\ \bar{s}_{i,k} \end{bmatrix} &= \sum_{j=1}^{m} a_{i,j}(k+1) \begin{bmatrix} x_{j,k} \\ s_{j,k} \end{bmatrix}, \\ x_{i,k+1} &= P_X \left[\bar{x}_{i,k} - \alpha_{k+1} \sum_{j=1}^{m} g_j' \left(m[\bar{s}_{i,k}]_j \right) \nabla h_{i,j} \left(\bar{x}_{i,k} \right) \right], \\ s_{i,k+1} &= \bar{s}_{i,k} + \begin{bmatrix} h_{i,1} \left(\bar{x}_{i,k+1} \right) \\ \vdots \\ h_{i,m} \left(\bar{x}_{i,k+1} \right) \end{bmatrix} - \begin{bmatrix} h_{i,1} \left(\bar{x}_{i,k} \right) \\ \vdots \\ h_{i,m} \left(\bar{x}_{i,k} \right) \end{bmatrix}. \end{aligned} \tag{6.16}$$

Here $[\bar{s}_{i,k}]_j$ denotes the j-th component of the vector $\bar{s}_{i,k}$. In the $(k+1)$-th iteration, agent i receives $x_{i,k}$ and $s_{i,k}$ from its current immediate neighbors and calculates weighted averages $\bar{x}_{i,k}$ and $\bar{s}_{i,k}$. The weighted average is then updated using locally available information (functions g_j and $h_{i,j}$, and the set X) to

generate $x_{i,k+1}$ and $s_{i,k+1}$. The algorithm is initialized with $x_{i,0} \in X$ and

$$s_{i,0} = \begin{bmatrix} h_{i,1}(x_{i,0}) \\ \vdots \\ h_{i,m}(x_{i,0}) \end{bmatrix}.$$

6.1.2 Extension II

We next consider a generalization of problem (6.15), where the objective function has the same form but the knowledge about the functions $h_{i,j}$ is distributed differently. Consider a network of m^2 agents indexed by ℓ, where $\ell \in W = \{1, \ldots, m^2\}$. The objective function has the same form as (6.15).

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{m} g_j \left(\sum_{i=1}^{m} h_{m(j-1)+i}(x) \right) \\ \text{subject to} \quad & x \in X, \end{aligned} \tag{6.17}$$

where $g_j : \Re \to \Re$ for all $j \in V$, $X \subset \Re^p$ and $h_\ell : X \to \Re$ for all $\ell \in W$. The function h_ℓ, $\ell \in W$, is known only to agent ℓ. Let $j(\ell) = \{j : m(j-1) < \ell \leq mj\}$. The function g_j is known to agent ℓ if $j(\ell) = j$. Essentially, function g_j is known to agents $m(j-1), \ldots, mj$. The set X is known to all the agents.

We can modify (6.4) to solve (6.17) as follows. Let $x_{\ell,k}$ denote agent ℓ's estimate of the optimal point at time k. Let $s_{\ell,k}$ denote agent ℓ's estimate of the vector

$$\left[\frac{1}{m} \sum_{r=1}^{m} h_r(x_{r,k}) \quad \frac{1}{m} \sum_{r=m+1}^{2m} h_r(x_{r,k}) \quad \ldots \quad \frac{1}{m} \sum_{r=m(m-1)+1}^{m^2} h_r(x_{r,k}) \right]^T.$$

For each $\ell \in W$, agent ℓ recursively generates $x_{\ell,k+1}$ and $s_{\ell,k+1}$ as follows:

$$\begin{aligned}\begin{bmatrix} \bar{x}_{\ell,k} \\ \bar{s}_{\ell,k} \end{bmatrix} &= \sum_{r=1}^{m} a_{\ell,r}(k+1) \begin{bmatrix} x_{r,k} \\ s_{r,k} \end{bmatrix}, \\ x_{\ell,k+1} &= P_X \left[\bar{x}_{\ell,k} - \alpha_{k+1} g'_{j(\ell)} \left(m[\bar{s}_{\ell,k}]_{j(\ell)} \right) \nabla h_\ell \left(\bar{x}_{\ell,k} \right) \right] \end{aligned} \tag{6.18}$$

$$[s_{\ell,k+1}]_j = \begin{cases} [\bar{s}_{\ell,k}]_j & \text{for } j \neq j(\ell) \\ [\bar{s}_{\ell,k}]_j - h_\ell(x_{\ell,k}) + h_\ell(x_{\ell,k+1}) & \text{for } j = j(\ell). \end{cases}$$

In the $(k+1)$-th iteration, agent ℓ receives $x_{j,k}$ and $s_{j,k}$ from its current immediate neighbors and calculates weighted averages $\bar{x}_{\ell,k}$ and $\bar{s}_{\ell,k}$. The weighted average is then updated using locally available information (functions $g_{j(l)}$ and h_ℓ, and the set X) to generate $x_{\ell,k+1}$ and $s_{\ell,k+1}$. The algorithm is initialized with $x_{i,0} \in X$ and $s_{\ell,0} = h_{j(\ell)}\left(x_{\ell,0}\right) e_{j(\ell)}$, where $e_{j(\ell)}$ is the unit vector with $j(\ell)$-th component equal to 1.

CHAPTER 7

APPLICATION: DISTRIBUTED REGRESSION

An important application of distributed stochastic optimization is distributed regression in sensor networks. Sensor networks are deployed to learn something about the underlying phenomenon that they sense. A canonical sensor network learning problem is regression, which involves modeling a response variable as a function of one or more predictor variables using samples of the response variable at different levels of the predictor variables. The response variable is viewed as a random variable and its mean is modeled as a known function of the predictor variables and an unknown regression parameter. The goal in regression is to determine the regression parameter value that best explains the observed data. This is usually done by defining a "goodness of fit" cost criterion and choosing the parameter value that minimizes the criterion. *Thus regression involves solving an optimization problem in which the objective function is decided by the observed data and the parameter of optimization is the regression parameter.*

In some systems, no sensible causality relationship exists between the response and the predictor variables. While the model function is physically meaningless, it might nevertheless provide a good prediction for the response variable. In other systems, there is a natural causality relationship between the response and the predictor variable. The model function can be obtained through suitable approximations and physical laws. The parameter in the model function may have a physical significance. Thus in these systems the end goal of regression could be to estimate the parameter of significance and not predict future values of the response variable at unobserved predictor values.

In regression in sensor networks, different parts of the regression data are

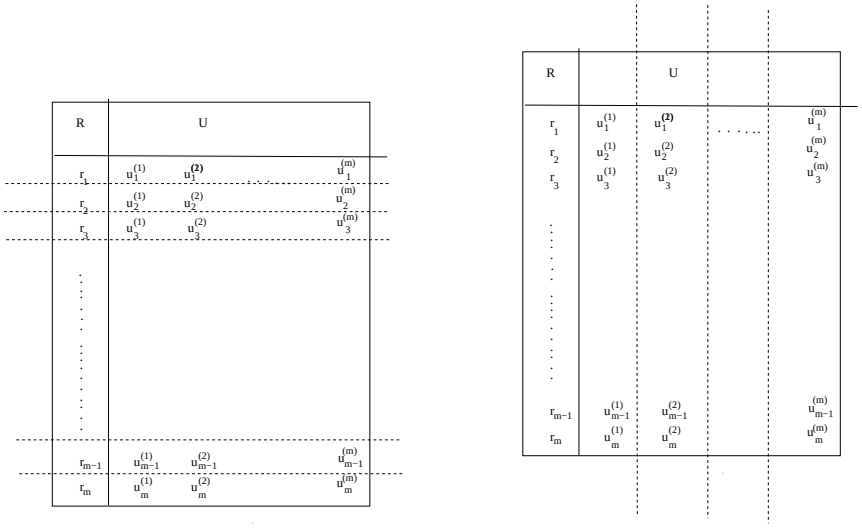

Figure 7.1: Horizontal/vertical distributed regression data.

collected by different agents in the network. Data in the network is distributed *horizontally* when each agent observes complete samples but no agent has access to all the samples [43]. In contrast, in *vertically* distributed data each agent has access only to a subset of the predictor variables' values in each sample. See Fig 7.1 for an illustration. Data could be horizontally and vertically distributed. Based on the manner in which data is distributed, the optimization problem that specifies the regression parameter is a specific distributed optimization problem.

We next discuss some sensor network applications that can be cast as regression problems.

- *Feature estimation:* In feature estimation, the network goal is to understand some feature of interest about the underlying field. Apriori information is used to completely describe the feature, except for some unknown parameters. Physical laws can then be used to map the parametrized model for the feature to a function for the sensor measurement. In the regression framework, the response variable is the sensor measurement and there are no predictor variables. This is an instance of regression where the end goal is to estimate the parameter.

  For example, consider sensors deployed to determine the source of a

spatio-temporal temperature field. A model for the heat source could be the set of all point sources with constant intensity. The model is parametrized by the source location and the constant intensity. The diffusion equation that governs the propagation of heat can then be used to map the model for the heat source to a model function for the sensor measurements. Note that the actual source may have an arbitrary shape and have an exponentially decreasing intensity. In effect, the goal is to only determine the best approximation for the source among all point sources with constant intensity.

- *Spatial field reconstruction:* Field reconstruction essentially involves extrapolating the field values measured at the sensor locations to other locations [44]. In this, the response variable is the sensor measurement and the predictor variable is the sensor location.

Numerous other sensor network problems like sensor localization and calibration can be recast as feature estimation problems.

We mathematically describe the regression problem below. Let R be the response variable and $U^{(i)}$, $i \in \{1, \ldots, q\}$, denote the i-th predictor variable. In regression the goal is to model

$$R = f^* \left(U^{(1)}, \ldots, U^{(q)}\right) + \text{noise}.$$

We emphasize that there need not be a causal relation between R and $U^{(i)}$.

Suppose a total of p observation samples are available. The j-th response sample is denoted by r_j and is measured at the value $u_j^{(i)}$ of the i-th predictor variable. A model function is first decided and the data points are related as follows:

$$r_j = g(u_j^{(1)}, \ldots, u_j^{(q)}; x) + \epsilon_j(x).$$

Here g is the regression function and x is the unknown regression parameter that is to be decided from observed data. The statistics of $\{\epsilon_j(x)\}_{1 \le j \le p}$ are known

functions of the regression parameter x. We will consider only the following class of model functions.

Definition 1 IID models: *In these models, the errors* $\{\epsilon_j(x)\}_{1\leq j\leq p}$ *are samples of independent identically distributed (i.i.d.) random variable* $E(x)$.

We emphasize that this is not an assumption on the actual data generating process.

In maximum likelihood estimation, the optimal parameter value is chosen as the value that minimizes the negative likelihood function.

$$x^* = \arg\min_{x\in X} -\frac{1}{p}\sum_{j=1}^{p} \mathcal{L}_E\left(r_j - g\left(u_j^{(1)},\ldots,u_j^{(q)};x\right)\right). \tag{7.1}$$

Here $\mathcal{L}_E$ is the log likelihood of $E(x)$ in Definition 1. A special case is when $\epsilon_j(x)$ does not depend on x and is Gaussian with zero mean. This corresponds to least square regression.

7.1 Horizontal Regression

Consider a network of m agents indexed by j, $j \in V = \{1,\ldots,m\}$. For the sake of clarity, let the number of data points p equal the number of agents m. When the data is horizontally distributed, only agent i has access to the i-th sample, i.e, $r_i, u_i^{(1)},\ldots,u_i^{(q)}$. The model definition, i.e., the model function and the statistics of $E(x)$, is known to all the agents. Observe that problem (7.1) now is a special case of the problem (2.1) with

$$f_i(x) = -\mathcal{L}_E\left(r_i - g\left(u_i^{(1)},\ldots,u_i^{(q)};x\right)\right).$$

Thus agent i knows function f_i completely and the network goal is to minimize the sum of the functions f_i.

7.2 Sequential Horizontal Regression

There are two regression paradigms: *batch regression* and *sequential regression*. In batch regression, all the data is collected first and the optimization problem that specifies the optimal parameter values is then solved. In contrast, in sequential regression the data is collected sequentially. As and when a data point is collected, the data point is used to update an estimate of the optimal parameter value and then discarded. There are multiple benefits to sequential regression. First, data need not be stored and this reduces the memory requirements. Second, at any time the network has an estimate, possibly coarse, of the optimal parameter. Since information is discarded continuously, the trade-off is that we can only obtain an approximation to the optimal parameter estimate. In a good sequential algorithm, the approximations get increasingly accurate as more measurements are available and asymptotically the approximations become exact.

One approach to designing sequential algorithms is to start from the standard gradient descent algorithm that asymptotically solves the associated optimization problem. Note that the gradient depends on all the data points. To make the algorithm sequential, each iteration of the algorithm is synchronized with one new measurement. At the time of each iteration of the gradient descent, only the new measurement and possibly a summary of the past measurements are available. These are now used to approximate the gradient. Thus, the sequential algorithm can be viewed as the gradient algorithm with errors in the gradient. Typically, the measurements are modeled as a sample path of a random process, making the errors stochastic and the algorithm a stochastic optimization algorithm.

We next mathematically cast the sequential regression problem in a stochastic optimization framework. Since the properties of sequential algorithms that are analyzed are asymptotic, consider the case when each agent sequentially collects infinite data points, i.e., $p \to \infty$. Consider a network of m agents and let each

agent sequentially collect data in every time slot. We denote the k-th response variable collected by agent i as $r_{i,k}$ at the levels $u_{i,k}^{(1)}, \ldots, u_{i,k}^{(q)}$. The associate optimization problem is now

$$x^* = \arg\min_{x \in X} \lim_{p \to \infty} -\frac{1}{p} \sum_{k=1}^{p} \sum_{i=1}^{m} \mathcal{L}_E \left(r_{i,k} - g \left(u_{i,k}^{(1)}, \ldots, u_{i,k}^{(q)}; x \right) \right). \tag{7.2}$$

Suppose we make the following assumption on the actual data generation process.

Assumption 11 *The sequence* $\{(r_{i,k}, u_{i,k}^{(1)}, \ldots, u_{i,k}^{(q)})\}_{k \in \mathbb{N}}$ *are independent samples of* $(R_i, U_i^{(1)}, \ldots, U_i^{(q)})$.

Under Assumption 11 the problem in (7.2) reduces to

$$x^* = \arg\min_{x \in X} - \sum_{i=1}^{m} \mathsf{E}\Big[\mathcal{L}_E \left(R_i - g \left(U_i^{(1)}, \ldots, U_i^{(q)}; x \right) \right) \Big].$$

Observe that the problem now can fit in the distributed stochastic optimization framework discussed in Section 2.3.1. Agent i sequentially collects samples of $R_i, U_i^{(1)}, \ldots, U_i^{(q)}$ and these can be used to obtain the Robbins-Monro approximation for the gradient.

7.2.1 Simulation study

A diffusion field is a spatio-temporal field that follows the diffusion partial differential equation. The diffusion equation models diverse physical phenomena like the conduction of heat, dispersion of plumes in air, dispersion of odors through a medium and the migration of living organisms. We study the problem of determining the source of a diffusion field (specifically, temperature field) generated in a room.

We briefly review the literature on source identification in the diffusion equation. The point source model is a common model for diffusing sources and has been extensively used [45–48]. Usually the intensity is assumed to be a

constant [45, 47, 48] or instantaneous. Localization of sources with time-varying intensity have been studied in a centralized and non-recursive setting in [49, 50]. These studies consider a deterministic evolution of the leak intensity and use a continuous observation model. Most papers study the case in which the medium is infinite or semi-infinite since the diffusion equation has a closed form solution in that case [45, 47]. An exception is [51] where two-dimensional shaped medium is considered.

While centralized recursive source localization has received much interest [45, 47, 50, 52] there are very few papers that discuss a distributed solution. A recursive and distributed solution to the problem in a Bayesian setting is discussed in [46]. A related paper is [53] that deals with the problem of estimating the diffusion coefficient in a distributed and recursive manner. We are not aware of any prior work that solves the source localization problem using a distributed and recursive approach in a non-Bayesian setting. We refer the reader to [46] for a more detailed discussion of the literature.

We consider the case when based on a priori information the model set for the source can be chosen to be the set of all point sources with constant intensity. The model is parametrized by the source location $x = (x_1, x_2)$ and intensity I. Thus the problem of learning the source simplifies to estimating x and I. Additionally, it is also known that the warehouse is large, the initial temperature is a constant throughout the warehouse, and the thermal conductivity is large and uniform throughout the medium, and known.

To map the model for the source to a model function for the sensor measurements we will use the diffusion equation. Let $C(s, t; x, I)$ denote the temperature at a point s at time t when the source is at x and intensity is I. For the source model set and medium, $C(s, t; x, I)$ can be approximated using the steady-state value given by [45]

$$C(s; x, I) = \frac{I}{2\nu\pi\|s - x\|},$$

where ν is the diffusion coefficient. If s_i is the location of the i-th sensor, then the model set for its k-th measurement is

$$\hat{R}_{i,k}(x, I) = C(s_i; x, I) + N_{i,k},$$

where $N_{i,k}$ is a zero mean i.i.d. measurement noise. Thus the estimation problem is a simple nonlinear regression with no measurable inputs.

In the simulation experiments the diffusion coefficient $\nu = 1$. The actual location of the source is $x^* = (37, 48)$ and the actual intensity value is 10. A network of 27 sensors is randomly deployed as shown in Fig 7.2. The initial iterate value is fixed at $(50, 50)$ for the source location and 5 for the intensity. The results are plotted in Figs. 7.3 and 7.4. Observe that about 200 iterations are sufficient to obtain a good estimate of the source location and 1000 iterations to obtain a good estimate of the intensity using the cyclic incremental, Markov incremental and diffusion algorithms. In addition, we observed that the convergence speed of the algorithm is affected by the initial point. If the initial points are taken very far from the actual value then there is convergence to other stationary points in the problem.

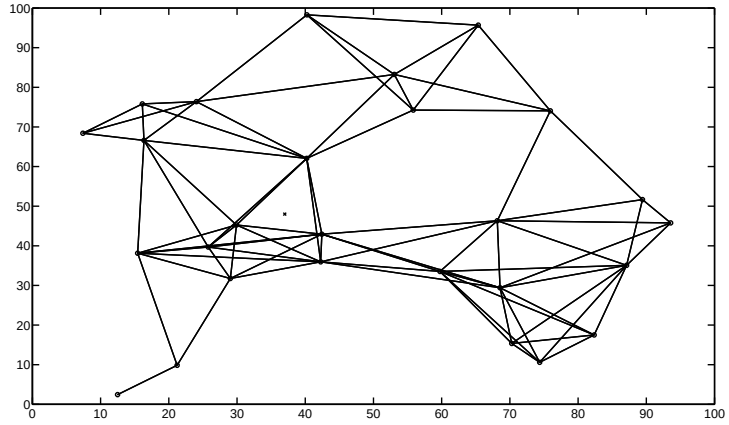

Figure 7.2: A network of 27 sensors that is deployed to determine the source of the temperature field.

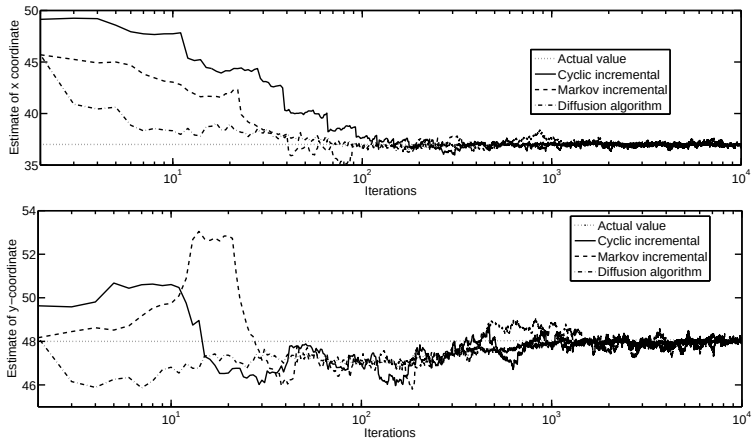

Figure 7.3: Estimates of the x and y coordinates generated by the cyclic incremental, Markov incremental and diffusion gradient algorithms.

7.3 Vertically Distributed Data

Consider a network of m agents indexed by i, $i \in V$. Let the number of predictor variables q equal the number of agents m. Further, let the regression function have the following additive form:

$$g\left(U_i^{(1)}, \ldots, U_i^{(m)}; x\right) = \sum_{i=1}^{m} \Phi_i\left(x, U^{(i)}\right), \qquad x \in X. \tag{7.3}$$

In this case the optimization problem in (7.1) reduces to

$$x^* = \arg\min_{x \in X} -\frac{1}{p} \sum_{j=1}^{p} \mathcal{L}_E\left(r_j - \sum_{i=1}^{m} \Phi_i\left(x, U^{(i)}\right)\right). \tag{7.4}$$

The data is vertically distributed. Therefore, only agent i has access to the samples of the i-th predictor random variable. Thus $\{u_j^{(i)}\}_{j \in V}$ is known only to agent i. The response variable samples are available to all the agents. In distributed regression, the goal is to solve (7.1) in a distributed and local manner. Observe that problem (7.4) can be seen to be a special case of problem

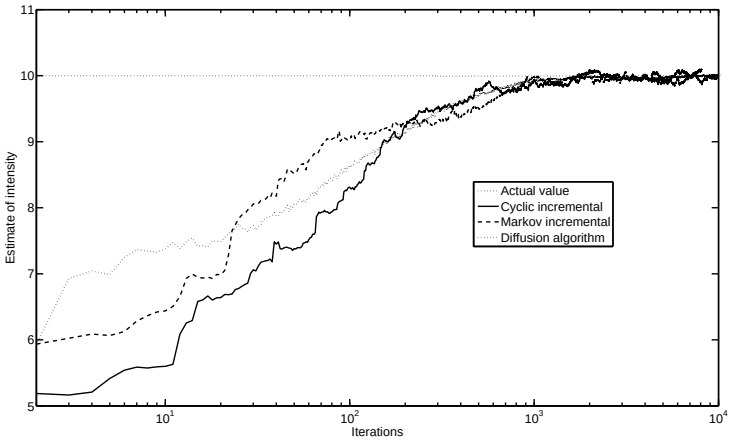

Figure 7.4: Estimate of the intensity generated by the cyclic incremental, Markov incremental and diffusion gradient algorithms.

(6.15) in Chapter 6 by fixing

$$g_j = -\mathcal{L}_E, \qquad h_{i,j}(x) = r_j - \Phi_i\left(x, u_j^{(i)}\right).$$

This formulation can be relaxed to the case when there is an $(m+1)$-th agent that collects the samples of the response variable.

7.4 Horizontally and Vertically Distributed Data

We consider a network of m^2 agents. Each agent is indexed by ℓ where $\ell \in S = \{1, \ldots, m^2\}$. For the sake of clarity, we take the number of predictor variables q and the number of samples p to equal m. Further, the regression function has an additive form as in (7.3). When the data is vertically and horizontally distributed, agent $i + m(j-1)$ has access to only $u_j^{(i)}$ and r_j. Observe that problem (7.4) now is a special case of (6.17) in Chapter 6 with

$$g_j = -\mathcal{L}_E, \qquad h_{m(j-1)+i} = r_j - \Phi_i\left(x, u_j^{(i)}\right).$$

7.5 Extension: Regression with Non i.i.d. Models

In Chapter 7 we have only studied models which are independent and identically distributed. A natural extension is to consider models which assume a specific model for the dependence across time. In [54] we have studied the problem of regression using state space models.

A network of sensors are deployed to sense a spatio-temporal field and infer parameters of interest about the field. Each agent observes the same underlying state-space process through an observation matrix that is parametrized by the same parameter vector. Mathematically,

$$\begin{aligned} \Phi(k+1,x) =& D(x)\Phi(k,x) + W(k+1,x) \\ R_i(k+1,x) =& H_i(x)\Phi(k+1,x) + V_i(k+1,x). \end{aligned} \tag{7.5}$$

Here x is the parameter and is constrained to the set X. The random process $\{\Phi(k+1,x)\}$ is the hidden process that is observed through $R_i(k+1,x)$, which is the model for the i-th sensor's k-th observation Further, D and H_i are matrix functions of appropriate dimensions, and $\{W_i(k+1,x)\}$ and $\{V_i(k+1,x)\}$ are sequences of i.i.d. zero mean random sequences that are parametrized by x.

Let $r_{i,k}$ denote the i-th sensor's k-th measurements. Under the conditional least-square criterion, the best model is the model that best predicts the future based on the past values. Let $\hat{g}_{k+1}(r^k, x)$ be the optimal predictor. Here r^k denotes all the past measurements up to and including time k. For the system in (7.5) the optimal predictor is the Kalman predictor. According to the prediction error criterion, the estimate is

$$x^* = \underset{x\in X}{\operatorname{argmin}} \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{m} \sum_{k=1}^{N} \left\| r_{i,k+1} - \hat{g}_{i,k+1}(x, r^k) \right\|^2 .$$

Solving this in a centralized and recursive manner has been studied in detail in [55].

We are interested in identifying the best model in a distributed manner, i.e., the agents should not share their raw data points with each other. Agent i knows only a part of r^k, i.e., r_i^k, and hence cannot calculate $\hat{g}_{i,k+1}(x, r^k)$. Therefore, it is quite obvious that the agents cannot solve the above optimization problem in a distributed manner. Therefore, we modify the criterion slightly and choose the best model which predicts the best value of $r_{i,k+1}$ from only r_i^k, i.e., past measurement of only sensor i. Thus we choose

$$x^* = \underset{x \in X}{\operatorname{argmin}} \sum_{i=1}^{m} \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \left\| r_{i,k+1} - \hat{g}_{i,k+1}(x, r_i^k) \right\|^2 . \tag{7.6}$$

In [54] a distributed and recursive estimation algorithm, called the incremental recursive prediction error (IRPE) algorithm, is proposed to solve the problem in (7.6). This algorithm has the distributed property of incremental gradient algorithms and the on-line property of recursive prediction error algorithms. The convergence properties of the algorithms are investigated and sufficient conditions are obtained for convergence.

CHAPTER 8

APPLICATION: POWER CONTROL IN CELLULAR SYSTEMS

The multi-access nature of the wireless channel enables multiple transmitters to communicate with their receivers simultaneously using the same channel. This presents a classic resource allocation problem. Each transmitter would prefer to increase its transmission power to improve the signal at its receiver. The higher the signal power, the higher the interference caused by a transmission at the other receiver locations. The power control problem is to assign power levels to the transmissions so that the signal strengths at *all* the receivers are satisfactory.

We will study the inter-cell uplink power control problem in cellular networks. In this setting, a mobile user (MU) from each cell communicates with its base station (BS) on a common channel. The optimal power allocation will depend on the channel between each MU and each BS. In general, the channel is estimated at the base station using pilot signals transmitted by the MUs. Therefore, a BS has information about the channel between the MUs and itself, but has no information about the channel between the MUs and other BSs. Thus, distributed power control algorithms in which each BS uses only locally available information to determine the optimal powers are of interest.

There are different mathematical formulations for the power control problem. We refer the reader to [56] for an extensive survey and we only summarize some points of interest. The first formulation that was studied was to minimize the total power subject to a constraint on the signal to interference and noise ratio (SINR) at each base station [57]. The power control problem was viewed as a closed loop control problem and distributed control algorithms were proposed. In these, each MU iteratively refines its power. In each iteration, all the MUs

transmit at the current iterate value. The SINR is measured at the BS and is communicated to the MU. Based on the SINR value the old iterate is refined to obtain the new iterate. A general framework was later defined in [58] to study such control theoretic solutions. The second approach is to view the power control as a game between non-cooperative users [59, 60]. In this, a utility function is defined for each MU-BS pair that captures their satisfaction as a function of the SINR at the BS and the power used by the MS. The third approach is to view the power control problem as an open loop global optimization problem [61, 62].

All the studies mentioned above do not require any message passing between the MUs and only require communication between each BS and MU. Thus, they are not just distributed but also *autonomous*. In this paper, we view the power control problem as a deterministic convex optimization problem that is to be solved by the base stations through information sharing with neighbors. As we will see, the optimization problem has a special structure where the objective function can be written as the sum of convex functions, each of which is completely known to a BS. Thus the power control problem is a special case of the distributed optimization problem studied in [9, 39, 41]. These algorithms will require the BSs to iteratively exchange information with their immediate neighbors to solve this optimization problem. Since the base stations are connected through a wired backbone this communication overhead may not be an issue. Once the optimal power value (or, a sufficiently good approximation) is determined each BS communicates the optimal power value to the MU in its cell. Thus these algorithms are non-autonomous as the base stations collaborate in solving the problem.

8.1 Problem Formulation

We will consider a finite cellular network. There are m mobile users (MU) in neighboring cells communicating with their respective base station using a

common wireless channel. Any other interfering source is treated as noise. Some of the MUs may be communicating with the same physical base station (BS). However, we will find it convenient to assign a unique index to each MU's receiver. Specifically, we use BS i to denote the base station with which MU i communicates.

We assume that the channel is static and the number of users do not change over the time scales studied in this paper. We denote the channel coefficient between MU j and BS i by $h_{i,j}$. It includes both the effects of large scale and small scale variations. We will assume that BS i has a good estimate of $h_{i,j}$, for all j. MUs that are in cells that are not immediate neighbors cause negligible interference. Therefore, they can be taken to be 0. For MUs in cell i and in the immediate neighborhood, BS i can estimate the channel coefficient from pilot signals that are sent by the mobiles.

Let p_i denote the power used by MU i and σ_i^2 be the receiver noise variance. Define $\bar{p}$ to be the vector with the j-th component equal to p_j, and $\bar{h}_i$ to be the vector with the j-th component equal to $\bar{h}_{i,j}$. The total received SINR at BS i is given by

$$\gamma_i\left(\bar{p}, \bar{h}_i\right) = \frac{p_i h_{i,i}^2}{\sigma_i^2 + \sum_{j \neq i} p_j h_{i,j}^2}.$$

Let U_i denote the utility function that captures the satisfaction of BS i as a function of its received SINR.[1] Depending on the nature of traffic (voice, multimedia or data) between MU i and BS i, the form of the function U_i could be different. The power control problem is to operate at the optimal point on the utility versus power curve. Formally, we have the following optimization

[1]Typically, U_i is assumed to be an increasing function of the SINR.

problem:

$$\max_{\bar{p}} \sum_i U_i\left(\gamma_i\left(\bar{p}, \bar{h}_i\right)\right) - \sum_i V\left(p_i\right)$$
$$\text{subject to} \quad 0 \leq p_i \leq p_t, \quad \forall i. \tag{8.1}$$

Here V is convex and increasing, and captures the cost of the power and p_t is a threshold on the maximum power that a MU can use.

In general there are no closed form solutions for the optimization problem in (8.1) and iterative algorithms have been used. When the problem is non-convex, the iterative algorithms may converge to a local maximum, rather than a global maximum. To avoid this, we impose additional restrictions on the utility functions U_i resulting in convex problem (8.1). These functions have the following property:

$$-\frac{xU_i''(x)}{U_i'(x)} \geq 1, \qquad \forall x \in X_i,$$

where X_i is some convex constraint set (see page 53 of [56]). We will focus on the case when the function $U_i(x) = \log(x)$, although nothing prevents the algorithms developed in this paper from being used for other utility functions such as the α-fair utility [56]. There is a natural motivation for the log utility function. First, it is the standard proportional fairness function used in literature [61, 63]. Second, a common choice for a mobile user utility function is the channel throughput achieved by the user. For each mobile user i, the throughput of the user is modeled as (see [64, 65])

$$T_i(\bar{p}) = \log(1 + \eta\gamma_i(\bar{p})),$$

where η is a constant determined by the modulation scheme that is used. The expression for $T_i(\bar{p})$ is a very good approximation for both additive white Gaussian channels and Rayleigh fading environments. When $\eta = 1$ this is also Shanon's capacity. As such the throughput is a non-convex function of the

powers. A commonly used technique to deal with the non-convexity [64–66] is to approximate

$$T_i(\bar{p}) \approx \log(\eta\gamma_i(\bar{p})).$$

This approximation is valid in the high SINR regime. In summary, the problem that is of interest is

$$\begin{aligned} &\max_{\bar{p}} \sum_i \left[\log\left(\frac{p_i h_{i,i}^2}{\sigma_i^2 + \sum_{j\neq i} p_j h_{i,j}^2} \right) - V\left(p_i\right) \right] \\ &\text{subject to} \quad 0 \leq p_i \leq p_t, \quad \forall i. \end{aligned} \tag{8.2}$$

Using the substitution $p_i = e^{x_i}$ in (8.2) we can rewrite the optimization problem as

$$\begin{aligned} &\min_{x} \sum_{i=1}^{m} \left[\log\left(\sigma_i^2 h_{i,i}^{-2} e^{-x_i} + \sum_{j\neq i} h_{i,i}^{-2} h_{j,i}^2 e^{x_j - x_i} \right) + V(e^{x_i}) \right] \\ &\text{subject to} \quad x \in X. \end{aligned} \tag{8.3}$$

Here x is the vector with the i-th component equal to x_i and X is the set $\{x : x_i \leq \log(p_t) \ \ \forall i\}$. The constraint set X is convex. Furthermore, since the log of a sum of exponentials is convex, the objective function is convex. Thus the problem in (8.3) is a convex optimization problem. Define

$$f_i(x; \bar{h}_i) = \log\left(\sigma_i^2 h_{i,i}^{-1} e^{-x_i} + \sum_{j\neq i} h_{i,i}^{-1} h_{j,i} e^{x_j - x_i} \right) + V(e^{x_i}).$$

Then, the problem in (8.3) can be written as

$$\begin{aligned} &\min_{x} \sum_{i=1}^{m} f_i(x; \bar{h}_i) \\ &\text{subject to} \quad x \in X. \end{aligned} \tag{8.4}$$

Our assumption that the channel coefficients $h_{i,j}$ between users j and base station i are known only to base station i translates to the function $f_i(x; \bar{h}_i)$ being known only to base station i. Therefore, the problem in (8.3) has to be solved in a distributed manner by the base stations, which communicate locally over the wired backbone that connects them.

The problem in (8.4) is a special case of a general distributed optimization problem studied in [9, 18, 39], which has the form

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{m} f_i(x) \\ &\text{subject to} \quad x \in X. \end{aligned} \tag{8.5}$$

Since the channel conditions may change, the algorithm that takes the least time to converge should be used. One can expect the parallel distributed algorithm to converge the fastest since parts of the algorithm run simultaneously at every agent.

8.2 Simulation Results

We consider a cellular network of 25 square cells. Each cell is of dimensions 10×10. Within each cell, the MU is randomly located and the base station is at the center of the cell. The network is shown in Fig. 8.1. The channel coefficient is assumed to decay as the fourth power of the distance between the transmitter and receiver. The shadow fading is assumed to be lognormal with variance 0.1. The receiver noise variance is taken to be 0.01. The cost of the power is modeled as $V(p_i) = 10^{-3} p_i$. The stepsize is taken to be $\alpha_k = \frac{10}{n_{s(k)}}$, where n_i is the number of times agent i receives the iterates in the incremental gradient projection algorithms. For the consensus gradient projection algorithm the stepsize is taken to be $\alpha_k = \frac{7}{k^{0.7}}$. First, observe from Fig. 8.2 that the algorithm chooses power values that are close to the optimal powers. Next, observe from Figs. 8.3, 8.4

and 8.5 that a good estimate is obtained after about 30 iterations of the cyclical incremental algorithm, 1500 iterations of the Markov algorithm and 500 iterations of the parallel algorithm. One can expect a similar behavior when the cells are hexagonal and the the channel coefficient decays differently with distance.

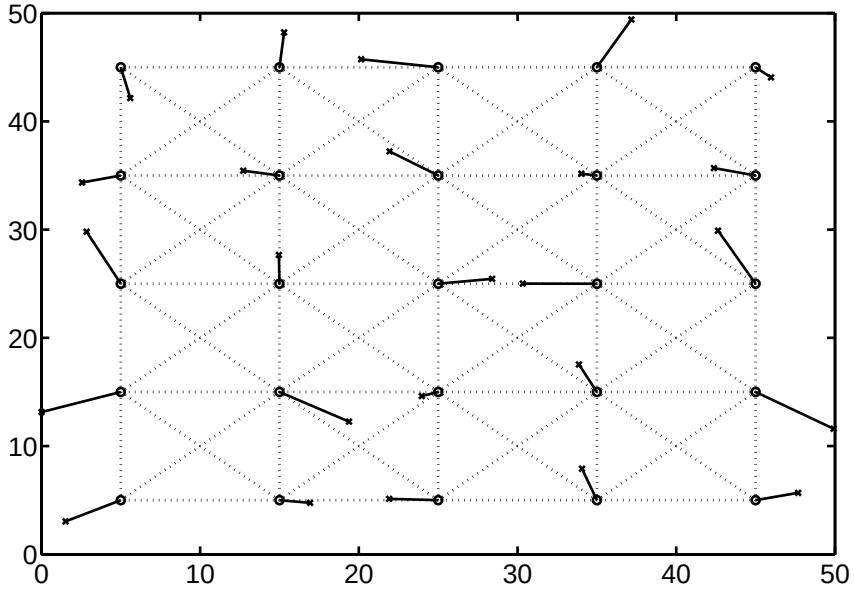

Figure 8.1: The circles denote the base stations. The dotted lines denote the noiseless communication links between adjacent BSs. The cross denotes the MUs. The thick lines connect each MU to its respective base station.

8.3 Discussion

An important criterion that decides the effectiveness of distributed power control algorithms is the time to convergence. If the time to convergence is slow, the channel and the number of users may change by the time the optimal power is determined. The autonomous algorithms discussed earlier require each MU in each iteration to transmit a signal with power equal to the current iterate value. In addition, there is feedback from the BS to the MU. In contrast, the non-autonomous approach proposed here only requires communication between

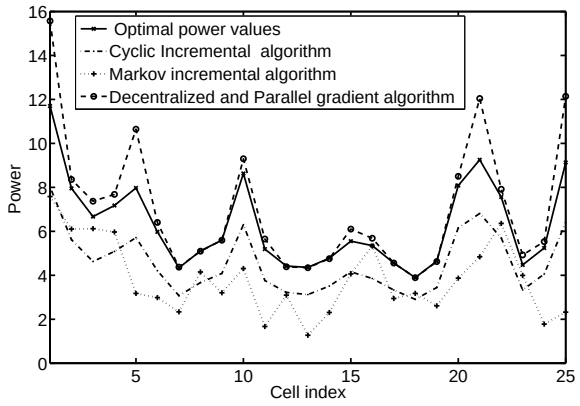

Figure 8.2: The final iterate values after after 2500 iterations of the parallel distributed algorithm.

neighboring base stations over wired links. Thus, the non-autonomous algorithms may potentially converge faster.

While we consider only the uplink of the wireless cellular networks, the algorithm and the discussion can be extended to the downlink channel. More generally, the non-autonomous algorithms proposed here can be used whenever communication between the receivers is cheaper than communication between the receiver and the transmitter. This would be the case in ad hoc wireless networks when the receivers are physically close to each other.

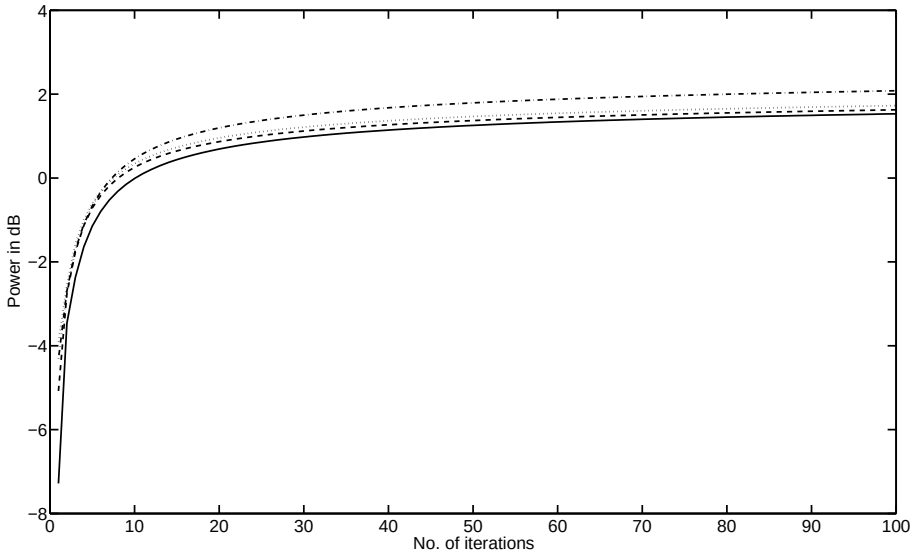

Figure 8.3: The iterate sequence generated at a randomly chosen agent in the incremental algorithm. Only a few components of the parameter vector x are plotted.

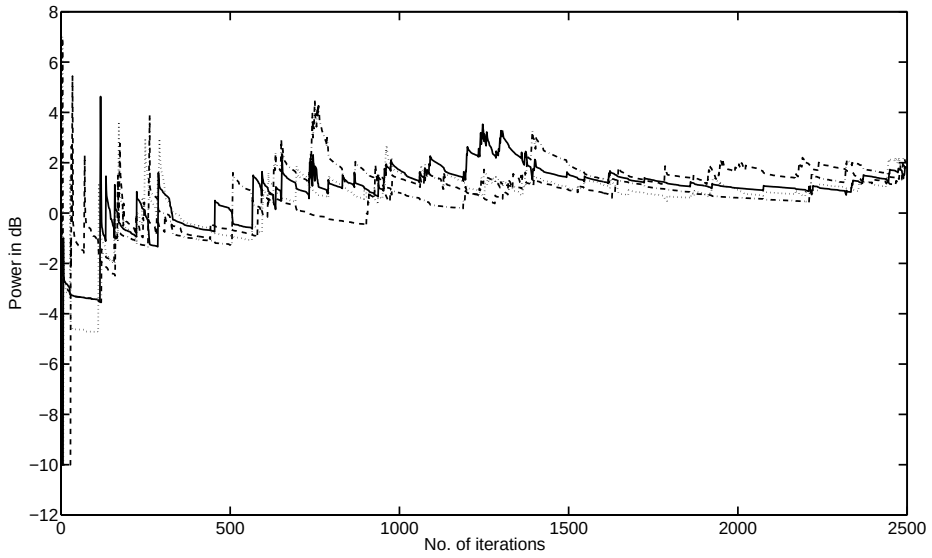

Figure 8.4: The iterate sequence generated at a randomly chosen agent in the Markov incremental algorithm. Only a few components of the parameter vector x are plotted.

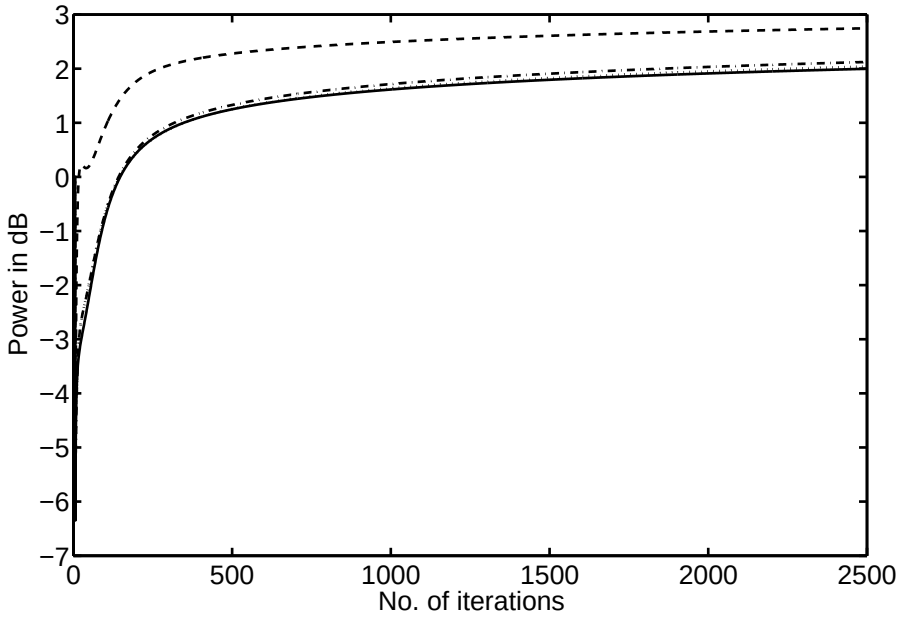

Figure 8.5: The iterate sequence generated at a randomly chosen agent in the parallel distributed algorithm. Only a few components of the parameter vector x are plotted.

CHAPTER 9

FUTURE WORK

In this chapter we outline some broad directions for future research.

9.1 Higher-Order Optimization Algorithms

For the purpose of this section we assume that there are no stochastic errors and the set X is $\Re^n$. All the algorithms that have been proposed in this thesis are first order algorithms. They only use gradient information and try to approximate the standard gradient descent algorithm. Thus they essentially are different distributed and local approximations to

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

where α_k is a fixed scalar sequence. First order optimization algorithms, while conceptually simple, are not efficient in terms of convergence speed. Therefore, one way to develop faster algorithms is to use higher order information.

The key ingredient in the development of higher order algorithms will be distributed consensus average tracking. Implicit in the development of (6.2) is an algorithm to track the network-wide average of a statistic that changes with time. This problem has been studied in detail in [42] as consensus filters. We briefly describe it. Mathematically the problem is formulated as follows. Let agent i observe $\phi_{i,k}$ at time k. The network goal at time k is to evaluate

$$\bar{\phi}_k = \frac{1}{m} \sum_{i=1}^{m} \phi_{i,k}, \tag{9.1}$$

or at least a good approximation to $\bar{\phi}_k$. We propose the following algorithm:

$$\hat{\chi}_{i,k+1} = \sum_{j \in N_i(k+1)} a_{i,j}(k+1)\chi_{i,k}, \tag{9.2}$$

$$\chi_{i,k+1} = \hat{\chi}_{i,k+1} + \phi_{i,k+1} - \phi_{i,k}. \tag{9.3}$$

At time k agent i exchanges $\chi_{i,k}$ with its immediate neighbor and then obtains a weighted average. This step is identical to the the standard distributed averaging step. Since the target value changes across time, agent i compensates for the change by adding an additional innovations term $\phi_{i,k+1} - \phi_{i,k}$ in (9.3). The following result proves that the agents asymptotically track the average.

Theorem 15 *Let Assumptions 5 and 10 hold. Let the sequence* $\{\phi_{i,k}\}_{k\in\mathbb{N}}$ *satisfy*

$$\|\phi_{i,k} - \phi_{j,k}\| \leq D\alpha_k,$$

where $\{\alpha_k\}$ *is a non-negative sequence and* D *is a positive scalar. If* $\alpha_k \to 0$ *then for all* $i \in V$, $\left\|\chi_{i,k} - \bar{\phi}_k\right\|$ *converges to* 0. *Further, if* $\sum_k \alpha_k^2 < \infty$ *then* $\sum_k \alpha_k \|\bar{\chi}_{i,k} - \hat{\bar{\phi}}_k\| < \infty$ *for all* $i \in V$.

An immediate application for this is in change detection in sensor networks. In this context, $\theta_{i,k}$ is the log likelihood of all the observations made at agent i till time k and $\bar{\theta}_k$ is the log likelihood of all the observations made at all the agents till time k. Thus $\|\theta_{k+1} - \theta_k\|$ diminishes with time.

We next use the idea of distributed average tracking to develop higher order optimization algorithms. We illustrate the idea with the steepest descent but the discussion can be extended to other higher order algorithms. The centralized steepest descent algorithm would be

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

where

$$\alpha_k = \arg\min_{\gamma>0} f(x_k - \gamma \nabla f(x_k)).$$

To make this distributed and local, the computation of $\nabla f(x_k)$ and the step-size α_k have to be distributed and local. The three algorithms discussed in this thesis provide an approach to approximate $\nabla f(x_k)$ in a distributed and local manner. The challenge is now in evaluating the stepsize α_k in a distributed manner. Close to the optimal value the stepsize α_k will change "slowly." Thus the network will essentially need to track the sequence $\{\alpha_k\}$ that is slowly changing. This step can possibly be done using the distributed average tracking algorithm that we discussed earlier.

9.2 Saddle Point Problems

An interesting research direction is to extend the distributed optimization framework to solve saddle point problems. Let X and Y be closed and convex sets in $\Re^n$ and $\Re^s$, and let $\mathcal{Q} : X \times M \to R$ be a convex-concave function. The goal in a saddle point problem is to determine a $(x^*, y^*) \in X \times Y$ such that for every $x \in X$ and $y \in Y$

$$\mathcal{Q}(x^*, y) \leq \mathcal{Q}(x^*, y^*) \leq \mathcal{Q}(x, y^*).$$

Equivalently the goal is

$$\max_{y \in Y} \min_{x \in X} \mathcal{Q}(x, y). \tag{9.4}$$

Next let us consider a distributed version of the saddle point problem. The goal is

$$\max_{y \in Y} \min_{x \in X} \sum_{i=1}^{m} \mathcal{Q}_i(x, y). \tag{9.5}$$

The problem is distributed because the function $\mathcal{Q}_i(x, y)$ is known only to agent i. We propose the following distributed, local and parallel algorithm to solve (9.5):

$$\begin{aligned}
\begin{bmatrix} \bar{x}_{i,k} \\ \bar{y}_{i,k} \end{bmatrix} &= \sum_{j=1}^{m} a_{i,j}(k+1) \begin{bmatrix} x_{j,k} \\ y_{j,k} \end{bmatrix} \\
x_{i,k+1} &= P_X \left[\bar{x}_{i,k} - \alpha_{k+1} \nabla_x \mathcal{Q}_i(\bar{x}_k, \bar{y}_k) \right] \\
y_{i,k+1} &= P_X \left[\bar{y}_{i,k} + \alpha_{k+1} \nabla_y \mathcal{Q}(\bar{x}_k, \bar{y}_k) \right].
\end{aligned}$$

Here ∇_x and ∇_y denote gradient with respect to x and y, respectively. This algorithm is essentially a distributed version of the subgradient saddle point algorithm studied in [67]. Other distributed versions like the cyclic and Markov incremental algorithms can also be proposed.

An application of saddle point problems is in robust regression. Let us consider the regression problem in (7.1).

$$x^* = \arg\min_{x \in X} -\frac{1}{p} \sum_{j=1}^{p} \mathcal{L}_E \left(r_j - g \left(u_j^{(1)}, \ldots, u_j^{(q)}; x \right) \right).$$

The log likelihood $\mathcal{L}_E$ is the log likelihood of E, which is essentially the modeling error term defined in Definition 1. In most cases the statistics of E are not accurately known and there are some uncertainties. We can model this uncertainty with an additional parameter y in the log likelihood which can take values in a set Y and minimize the worst case criterion as follows:

$$x^* = \arg\min_{x \in X} \max_{y \in Y} -\frac{1}{p} \sum_{j=1}^{p} \mathcal{L}_E \left(r_j - g \left(u_j^{(1)}, \ldots, u_j^{(q)}; x \right); y \right). \tag{9.6}$$

Another important problem that can be cast as the saddle point problem in

(9.5) is the following constrained optimization problem:

$$\text{minimize} \sum_{i=1}^{m} f_i(x) \tag{9.7}$$

$$\text{subject to} \sum_{i=1}^{m} g_i(x) \leq 0 \tag{9.8}$$

$$x \in X. \tag{9.9}$$

The details of the connection between the above problem and the problem in (9.5) are available in [67]. We expect this problem to have numerous applications in rate control in wireless networks.

9.3 Distributed Kalman Filtering

Consider the following distributed filtering problem:

$$\begin{aligned} \Phi_{k+1} =& D\Phi_k + W_{k+1} \\ R_{i,k+1} =& H_i\Phi_{k+1} + V_{i,k+1}. \end{aligned}$$

Here $\{\Phi_k\}$ is an underlying random process that each agent in the network is interested in tracking. Agent i observes Φ_{k+1} through an observation matrix H_i and the observation is further corrupted by measurement noise $V_{i,k+1}$. The processes $\{V_{i,k+1}\}_{k\in\mathbb{N}}$ are i.i.d. random processes and are also independent across agents and with $\{W_{k+1}\}_{k\in\mathbb{N}}$. The matrix H_i and the statistics of $V_{i,k+1}$ are known only to agent i. The matrix D and the statistics of $\{W_{k+1}\}$ are globally known.

Agent i can track $\{\Phi_{k+1}\}$ only using its observation sequence $\{r_{i,k+1}\}_{k\in\mathbb{N}}$ through a standard Kalman filter. However, one would expect that by collaborating with other agents, agent i can potentially track $\{\Phi_{k+1}\}$ better.

Consider the following algorithm:

$$\hat{\Phi}_{i,k+1}(0) = D\hat{\Phi}_{i,k}(q) + L_i\left(r_{i,k} - H_i\hat{\Phi}_{i,k}(q)\right) \tag{9.10}$$

$$\hat{\Phi}_{i,k+1}(\ell+1) = \sum_{j=1}^{m} a_{i,j}(k+1)\hat{\Phi}_{i,k+1}(\ell) \qquad \text{for } 0 \le \ell \le q-1. \tag{9.11}$$

The algorithm can be understood as follows. The first step, i.e., (9.10), is an innovation step where new information $r_{i,k}$ is used to generate an estimate of Φ_{k+1} by agent i. In the second step, the agents perform q rounds of consensus to obtain the final estimate $\hat{\Phi}_{i,k+1}(q)$.

In [68] it is shown that if $q \to \infty$ then there is a choice of L_i (as a specific function of H_i, D, and statistics of $V_{i,k+1}$ and W_{k+1}) such that the estimate $\hat{\Phi}_{j,k+1}(q)$ converges to the optimal centralized Kalman filter estimate when all the information is available at a single location. In this regime as $q \to \infty$ the rate of convergence is determined by the consensus step. The best choices for the weights in this case are essentially the ones that speed up the consensus [69].

Typically communication is very expensive and therefore performing a large number of consensus steps between every iteration of the filter is not feasible. Therefore, q is usually a small number. In this case, it is not clear what the choice of L_i should be or how the weights should be chosen. Different formulations of this problem are studied in [70–72]. While [71, 72] propose heuristics, [70] provide some interesting optimality results when there is small process noise or small observation noise. However, the results are asymptotic and only provide partial answers to the question of choosing the optimal L_i and weights. A simpler non-trivial problem would be to determine an L_i and weights that guarantee tracking performance, which is better than when the agents do not collaborate. Thus there are some fundamental questions to be answered.

One possible approach to determine the optimal weights for fixed gains could be a learning approach, in which the agents start with uniform weights. At time k^- the latest estimate of the state Φ_k at agent j that is with a neighbor i is

$\hat{\Phi}_{j,k}(q-1)$. The idea is for agent i to essentially use $r_{i,k} - H_i\hat{\Phi}_{j,k}(q-1)$ as a measure of the goodness for the information it receives from its neighbor j. Intuitively, suppose $r_{i,k} - H_i\hat{\Phi}_{j,k}(q-1)$ is significantly less than $r_{i,k} - H_i\hat{\Phi}_{\ell,k}(r)$; then agent i could increase $a_{i,j}(k+1)$ and decrease $a_{i,\ell}(k+1)$. The goal is to develop a trust-update rule that the agents follow, which improves the final estimate of x_k at agent i with k. Effectively the algorithm in (9.11) would have an additional equation

$$a_{i,j}(k+1) = P_A\left[a_{i,j}(k) + K_i\left(r_{i,k} - H_i\hat{\Phi}_{j,k}(q-1)\right)\right]$$

where K_i are gains and A is the set

$$A = \left\{a_{i,j} \geq 0 : \sum_{j=1}^{m} a_{i,j} = 1\right\}.$$

APPENDIX A

BASIC RESULTS

A.1 Euclidean Norm Inequalities

For any vectors $v_1, \ldots, v_M \in \Re^n$, we have

$$\sum_{i=1}^{M} \left\| v_i - \frac{1}{M} \sum_{j=1}^{M} v_j \right\|^2 \leq \sum_{i=1}^{M} \|v_i - x\|^2 \qquad \text{for any } x \in \Re^n. \tag{A.1}$$

The preceding relation states that the average of a finite set of vectors minimizes the sum of distances between each vector and any vector in $\Re^n$, which can be verified using the first-order optimality conditions.

Both the Euclidean norm and its square are convex functions, i.e., for any vectors $v_1, \ldots, v_M \in \Re^n$ and non-negative scalars $\beta_1, \ldots, \beta_M$ such that $\sum_{i=1}^{M} \beta_i = 1$, we have

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

$$\left\| \sum_{i=1}^{M} \beta_i v_i \right\|^2 \leq \sum_{i=1}^{M} \beta_i \|v_i\|^2. \tag{A.3}$$

The following inequality is the well-known[1] *non-expansive* property of the Euclidean projection onto a nonempty, closed and convex set X,

$$\|P_X[x] - P_X[y]\| \leq \|x - y\| \qquad \text{for all } x, y \in \Re^n. \tag{A.4}$$

[1]See for example [7], Proposition 2.2.1.

A.2 Scalar Sequences

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

Lemma 7 *Let $\{\gamma_k\}$ be a scalar sequence.*

(a) If $\lim_{k\to\infty} \gamma_k = \gamma$ and $0 < \beta < 1$, then $\lim_{k\to\infty} \sum_{\ell=0}^{k} \beta^{k-\ell}\gamma_\ell = \frac{\gamma}{1-\beta}$.

(b) If $\gamma_k \geq 0$ for all k, $\sum_k \gamma_k < \infty$ and $0 < \beta < 1$, then $\sum_{k=0}^{\infty} \left(\sum_{\ell=0}^{k} \beta^{k-\ell}\gamma_\ell\right) < \infty$.

(c) If $\operatorname{limsup}_{k\to\infty} \gamma_k = \gamma$ and $\{\zeta_k\}$ is a positive scalar sequence with $\sum_{k=1}^{\infty} \zeta_k = \infty$, then $\operatorname{limsup}_{K\to\infty} \frac{\sum_{k=0}^{K} \gamma_k\zeta_k}{\sum_{k=0}^{K} \zeta_k} \leq \gamma$. In addition, if $\operatorname{liminf}_{k\to\infty} \gamma_k = \gamma$, then $\lim_{K\to\infty} \frac{\sum_{k=0}^{K} \gamma_k\zeta_k}{\sum_{k=0}^{K} \zeta_k} = \gamma$.

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

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

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

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

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

$$\sum_{\ell=0}^{k} \beta^{k-\ell}\gamma_\ell \leq \left(\max_{0\leq t\leq K} \gamma_t\right) \frac{\beta^{k-K}}{1-\beta} + \frac{\gamma+\epsilon}{1-\beta}.$$

Therefore,

$$\operatorname*{limsup}_{k\to\infty} \sum_{\ell=0}^{k} \beta^{k-\ell}\gamma_\ell \leq \frac{\gamma+\epsilon}{1-\beta}.$$

Since ϵ is arbitrary, we conclude that $\limsup_{k\to\infty}\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell \le \frac{\gamma}{1-\beta}$.

Similarly, we have

$$\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell \ge \min_{0\le t\le K}\gamma_t\sum_{\ell=0}^{K}\beta^{k-\ell}+(\gamma-\epsilon)\sum_{\ell=K+1}^{k}\beta^{k-\ell}.$$

Thus,

$$\liminf_{k\to\infty}\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell \ge \liminf_{k\to\infty}\left(\min_{0\le t\le K}\gamma_t\sum_{\ell=0}^{K}\beta^{k-\ell}+(\gamma-\epsilon)\sum_{\ell=K+1}^{k}\beta^{k-\ell}\right).$$

Since $\sum_{\ell=0}^{K}\beta^{k-\ell}\ge\beta^{k-K}$ and $\sum_{\ell=K+1}^{k}\beta^{k-\ell}=\sum_{s=0}^{k-(K+1)}\beta^s$, which tends to $1/(1-\beta)$ as $k\to\infty$, it follows that

$$\liminf_{k\to\infty}\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell \ge \left(\min_{0\le t\le K}\gamma_t\right)\lim_{k\to\infty}\beta^{k-K}+(\gamma-\epsilon)\lim_{k\to\infty}\sum_{s=0}^{k-(K+1)}\beta^s=\frac{\gamma-\epsilon}{1-\beta}.$$

Since ϵ is arbitrary, we have $\liminf_{k\to\infty}\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell\ge\frac{\gamma}{1-\beta}$. This and the relation $\limsup_{k\to\infty}\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell\le\frac{\gamma}{1-\beta}$, imply

$$\lim_{k\to\infty}\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell=\frac{\gamma}{1-\beta}.$$

(b) Let $\sum_{k=0}^{\infty}\gamma_k<\infty$. For any integer $M\ge1$, we have

$$\sum_{k=0}^{M}\left(\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell\right)=\sum_{\ell=0}^{M}\gamma_\ell\sum_{t=0}^{M-\ell}\beta^t\le\sum_{\ell=0}^{M}\gamma_\ell\frac{1}{1-\beta},$$

implying that

$$\sum_{k=0}^{\infty}\left(\sum_{\ell=0}^{k}\beta^{k-\ell}\gamma_\ell\right)\le\frac{1}{1-\beta}\sum_{\ell=0}^{\infty}\gamma_\ell<\infty.$$

(c) Since $\limsup_{k\to\infty}\gamma_k=\gamma$, for every $\epsilon>0$ there is a large enough K such that

$\gamma_k \leq \gamma + \epsilon$ for all $k \geq K$. Thus, for any $M > K$,

$$\frac{\sum_{k=0}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} = \frac{\sum_{k=0}^{K} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} + \frac{\sum_{k=K+1}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} \leq \frac{\sum_{k=0}^{K} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} + (\gamma + \epsilon) \frac{\sum_{k=K+1}^{M} \zeta_k}{\sum_{k=0}^{M} \zeta_k}.$$

By letting $M \to \infty$ and using $\sum_k \zeta_k = \infty$, we see that $\limsup_{M\to\infty} \frac{\sum_{k=0}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} \leq \gamma + \epsilon$, and since ϵ is arbitrary, the result for the limit superior follows.

Analogously, if $\liminf_{k\to\infty} \gamma_k = \gamma$, then for every $\epsilon > 0$ there is a large enough K such that $\gamma_k \geq \gamma - \epsilon$ for all $k \geq K$. Thus, for any $M > K$,

$$\frac{\sum_{k=0}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} = \frac{\sum_{k=0}^{K} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} + \frac{\sum_{k=K+1}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} \geq \frac{\sum_{k=0}^{K} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} + (\gamma - \epsilon) \frac{\sum_{k=K+1}^{M} \zeta_k}{\sum_{k=0}^{M} \zeta_k}.$$

Letting $M \to \infty$ and using $\sum_k \zeta_k = \infty$, we obtain $\liminf_{M\to\infty} \frac{\sum_{k=0}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} \geq \gamma - \epsilon$. Since $\epsilon > 0$ is arbitrary, we have $\liminf_{M\to\infty} \frac{\sum_{k=0}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} \geq \gamma$. This relation and the relation for the limit superior yield $\lim_{M\to\infty} \frac{\sum_{k=0}^{M} \gamma_k \zeta_k}{\sum_{k=0}^{M} \zeta_k} = \gamma$ when $\gamma_k \to \gamma$.

A.3 Matrix Convergence

Let $A(k)$ be the matrix with (i,j)-th entry equal to $a_{i,j}(k)$. As a consequence of Assumptions 10a, 10b and 10d, the matrix $A(k)$ is doubly stochastic[2]. Define, for all k, s with $k \geq s$,

$$\Phi(k,s) = A(k)A(k-1)\cdots A(s+1). \tag{A.5}$$

We next state a result from [27] (Corollary 1) on the convergence properties of the matrix $\Phi(k,s)$. Let $[\Phi(k,s)]_{i,j}$ denote the (i,j)-th entry of the matrix $\Phi(k,s)$, and let $e \in \Re^m$ be the column vector with all entries equal to 1.

Lemma 8 *Let Assumptions 5 and 10 hold. Then*

1. $\lim_{k\to\infty} \Phi(k,s) = \frac{1}{m} ee^T$ *for all* s.

[2]The sum of its entries in every row and in every column is equal to 1.

2. *Further, the convergence is geometric and the rate of convergence is given by*

$$\left| [\Phi(k,s)]_{i,j} - \frac{1}{m} \right| \leq \theta \beta^{k-s},$$

where

$$\theta = \left(1 - \frac{\eta}{4m^2}\right)^{-2} \qquad \beta = \left(1 - \frac{\eta}{4m^2}\right)^{\frac{1}{Q}}.$$

A.4 Stochastic Convergence

We next state some results that deal with the convergence of a sequence of random vectors. The first result is the well known Fatou's lemma [40].

Lemma 9 *Let $\{X_i\}$ be a sequence of non-negative random variables. Then*

$$\mathsf{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathsf{E}[X_n].$$

The next result is due to Robbins and Siegmund (Lemma 11, Chapter 2.2, [12]).

Theorem 16 *Let $\{B_k\}, \{D_k\}$, and $\{H_k\}$ be non-negative random sequences and let $\{\zeta_k\}$ be a deterministic non-negative scalar sequence. Let G_k be the $\sigma-$algebra generated by $B_1, \ldots, B_k, D_1, \ldots, D_k, H_1, \ldots, H_k$. Suppose that $\sum_k \zeta_k < \infty$,*

$$\mathsf{E}[B_{k+1} \mid G_k] \leq (1+\zeta_k)B_k - D_k + H_k \qquad \text{for all } k, \tag{A.6}$$

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

A.5 Distributed Consensus and Averaging

We briefly review the distributed averaging algorithm. See [73] for a recent survey. In the distributed averaging problem, agent i has the value θ_i. The goal in distributed averaging is for the agents to learn $\hat{\theta} = \frac{1}{m}\sum_{i=0}^{m}\theta_i$ in a distributed and local manner. We will refer to $\hat{\theta}$ as the *target* and θ_i as agent i's *start value*.

Distributed averaging is usually achieved iteratively through a sequence of *consensus steps*. In each step, each agent evaluates the new iterate as a weighted average of its current iterate and the current iterates of its neighbors. The initial value of the iterate at agent i is its start value θ_i. Formally, if $\{\theta_{i,k}\}$ denotes the sequence of estimates for the target generated by agent i, then

$$\theta_{i,k+1} = \sum_{j\in N_i(k+1)} a_{i,j}(k+1)\theta_{j,k}, \qquad \theta_i(0) = \theta_i. \tag{A.7}$$

Under Assumptions 5 and 10 it can be shown that $\lim_{k\to\infty}\theta_{i,k} = \hat{\theta}$ for all $i \in V$.

REFERENCES

[1] E. Durfee and V. Lesser, "Negotiating task decomposition and allocation using partial global planning," in *Distributed Artificial Intelligence*, M. Huns, Ed. San Fransisco, CA: Morgan Kaufmann Publishers Inc., 1989, pp. 229–244.

[2] B. Patridge, "The structure and function of fish schools," *Scientific American*, vol. 246, no. 6, pp. 114–123, 1982.

[3] M. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[4] C. Reynolds, "Flocks, herds, and schools: A distributed behavioral model," *Computer Graphics*, vol. 21, no. 4, pp. 25–34, 1987.

[5] J. Ferber, *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*, 1st ed. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[6] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.

[7] D. P. Bertsekas, A. Nedić, and A. Ozdalgar, *Convex Analysis and Optimization*. Nashua, NH: Athena Scientific, 2003.

[8] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton University Press, 1970.

[9] A. Nedić and D. P. Bertsekas, "Incremental subgradient method for nondifferentiable optimization," *SIAM Journal of Optimization*, vol. 12, no. 1, pp. 109–138, 2001.

[10] A. A. Gaivoronski, "Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods. Part 1," *Optimization Methods and Software*, vol. 4, no. 2, pp. 117–134, 1994.

[11] E. Levy, G. Louchard, and J. Petit, "A distributed algorithm to find Hamiltonian cycles in G(n,p) random graphs," in *Combinatorial and*

Algorithmic Aspects of Networking, ser. Lecture Notes in Computer Science, A. López-Ortiz and A. Hamel, Eds., vol. 3405. Berlin, Germany: Springer, 2005, pp. 63–74.

[12] B. T. Polyak, *Introduction to Optimization*. New York, NY: Optimization Software Inc., 1987.

[13] S. Kar, J. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," 2008. [Online]. Available: http://arxiv.org/abs/0809.0009

[14] T. Aysal, M. Coates, and M. Rabbat, "Distributed average consenus with dithered quantization," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4905–4918, 2008.

[15] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Nashua, NH: Athena Scientific, 1997.

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

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

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

[19] A. Nedić and A. Ozdaglar, "Distributed sub-gradient methods for multi-agent optimization," *Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[20] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus," MIT, Boston, MA, Tech. Rep. LIDS 2779, 2008.

[21] I. Lobel and A. Ozdaglar, "Distributed subgradient methods over random networks," MIT, Boston, MA, Tech. Rep. LIDS 2800, 2008.

[22] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, Massachusetts Institute of Technology, 1984.

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

[24] S. Kar and J. Moura, "Distributed consensus algorithms in sensor networks: Link and channel noise," 2007. [Online]. Available: http://arxiv.org/abs/0711.3915

[25] 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.

[26] L. Xiao, S. Boyd, and S.-J. Kim, "Distributed average consensus with least mean square deviation," *Journal of Parallel and Distributed Computing*, vol. 67, no. 1, pp. 33–46, 2007.

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

[28] A. Olshevsky and J. N. Tsitsiklis, "Convergence speed in distributed consensus and control," *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 33–55, 2009.

[29] A. Jadbabaie, J. Lin, and S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 998–1001, 2003.

[30] Y. Ermoliev, *Stochastic Programming Methods*. Moscow, Russia: Nauka, 1976.

[31] Y. Ermoliev, "Stochastic quasi-gradient methods and their application to system optimization," *Stochastics*, vol. 9, no. 1, pp. 1–36, 1983.

[32] Y. Ermoliev, "Stochastic quazigradient methods," in *Numerical Techniques for Stochastic Optimization,*, E. Y. Ermoliev and R. Wets, Eds. New York, NY: Springer-Verlag, 1988, pp. 141–186.

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

[34] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Asynchronous gossip algorithms for stochastic optimization," in *Proceedings of IEEE Conference on Decision and Control*, 2009, to be published.

[35] V. Borkar, *Stochastic Approximation: A Dynamical Viewpoint*. Cambridge, UK: Cambridge University Press, 2008.

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

[37] H. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York, NY: Springer-Verlag, 1978.

[38] A. Nedić and D. P. Bertsekas, "Convergence rate of incremental algorithms," *Stochastic Optimization: Algorithms and Applications*, vol. 54, pp. 223–264, 2001.

[39] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Incremental stochastic sub-gradient algorithms for convex optimization," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, 2009.

[40] P. Billingsley, *Probability and Measure*. London, UK: John Wiley and Sons, 1979.

[41] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient algorithm for convex optimization," 2008. [Online]. Available: http://arxiv.org/abs/0811.2595

[42] R. Olfati-Saber and J. Shamma, "Consensus filters for sensor networks and distributed sensor fusion," in *Proceedings of the IEEE Conference on Decision and Control*, vol. 7, no. 44, 2005, pp. 6698–6703.

[43] D. Hershberger and H. Kargupta, "Distributed multivariate regression using wavelet-based collective data mining," *Journal of Parallel and Distributed Computing*, vol. 61, no. 3, pp. 372–400, 2001.

[44] R. Nowak, U. Mitra, and R. Willet, "Estimating inhomogenous fields using wireless sensor networks," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 6, pp. 999–1009, 2004.

[45] J. Matthes, L. Gröll, and H. B. Keller, "Source localization for spatially distributed electronic noses for advection and diffusion," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1711–1719, 2005.

[46] T. Zhao and A. Nehorai, "Distributed sequential Bayesian estimation of a diffusive source in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 55, pp. 1511–1524, 2007.

[47] S. Vijaykumaran, Y. Levinbook, and T. F. Wong, "Maximum likelihood localization of a diffusive point source using binary observations," *IEEE Transactions on Signal Processing*, vol. 55, no. 2, pp. 665–676, 2007.

[48] M. E. Alpay and M. H. Shor, "Model-based solution techniques for the source localization problem," *IEEE Transactions on Control Systems Technology*, vol. 8, no. 6, pp. 803–810, 2000.

[49] J. R. Cannon and P. DuChateau, "Structural identification of an unknown source term in a heat equation," *Inverse Problems*, vol. 14, no. 3, pp. 535–552, 1998.

[50] C. L. Niliot and F. Lefévre, "Multiple transient point heat sources identification in heat diffusion: Application to numerical two- and three-dimensional problems," *Numerical Heat Transfer Part B Fundamentals*, vol. 39, no. 3, pp. 277–302, 2001.

[51] A. R. Khachfe and Y. Jarny, "Estimation of heat sources within two dimensional shaped bodies," in *Proceedings of 3rd International Conference on Inverse Problems in Engineering*, 1999, pp. 1309–1322.

[52] L. I. Piterbarg and B. L. Rozovskii, "On asymptotic problems of parameter estimation in stochastic PDE's: The case of discrete time sampling," *Mathematical Methods of Statistics*, vol. 6, no. 2, pp. 200–243, 1997.

[53] L. Rossi, B. Krishnamachari, and C. C. J. Kuo, "Distributed parameter estimation for monitoring diffusion phenomena using physical models," in *First IEEE International Conference on Sensor and Ad Hoc Communications and Networks*, 2004, pp. 460–469.

[54] S. S. Ram, V. V. Veeravalli, and A. Nedić, "Distributed and non-autonomous power control through distributed convex optimization," in *Proceedings of the IEEE INFOCOM*, 2009, pp. 973–978.

[55] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Boston, MA: The MIT Press, 1983.

[56] M. Chiang, P. Hande, T. Lan, and C. Tan, "Power control in wireless cellular networks," *Foundations and Trends in Networking*, vol. 2, no. 4, pp. 381–533, 2008.

[57] G. Foschini and Z. Milanjic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Transactions on Vehicular Technology*, vol. 42, no. 4, pp. 641–646, 1993.

[58] R. Yates, "A framework for power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, 1995.

[59] T. Alpcan, T. Basar, R. Srikant, and E. Altman, "CDMA uplink power control as a noncooperative game," *Wireless Networks*, vol. 8, no. 6, pp. 659–670, 2004.

[60] D. Falomari, N. Mandayam, and D. Goodman, "A new framework for power control in wireless data networks: Games utility and pricing," in *Proceedings of Allerton Conference on Communication, Control and Computing*, 1998, pp. 546–555.

[61] P. Hande, S. Rangan, and M. Chiang, "Distributed uplink power control for optimal SIR assignment in cellular data networks," in *Proceedings of the IEEE INFOCOM*, 2006, pp. 1–13.

[62] J. Huang, R. Berry, and M. Honig, "Distributed interference compensation for wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 5, pp. 1074–1085, 2006.

[63] J. Mo and J. Walrand, "Fair end-to-end window based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, 2000.

[64] X. Qui and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Transactions on Communications*, vol. 47, no. 6, pp. 884–895, 1999.

[65] M. Chiang, "Geometric programming for communication systems," *Foundations and Trends in Communications and Information Theory*, vol. 2, no. 1-2, pp. 1–156, 2005.

[66] D. Julian, M. Chiang, D. O'Neill, and S. Boyd, "Qos and fairness constrained convex optimization of resource allocation for wireless cellular and ad hoc networks," in *Proceedings of the IEEE INFOCOM*, 2002, pp. 477–486.

[67] A. Nedić and A. Ozdaglar, "Subgradient methods for saddle-point problems," *Journal of Optimization Theory and Applications*, vol. 142, no. 1, pp. 205–228, 2009.

[68] R. Olfati-Saber, "Distributed kalman filter with embedded consensus filters," in *Proceedings of the IEEE Conference on Decision and Control*, vol. 8, no. 44, 2005, pp. 8179–8184.

[69] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, vol. 53, no. 1, pp. 65–78, 2004.

[70] R. Carli, A. Chuiso, L. Schenato, and S. Zampieri, "Distributed kalman filtering based on consensus strategies," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 622–633, 2008.

[71] P. Alriksson and A. Rantzer, "Experimental evaluation of a distributed kalman filter algorithm," in *Proceedings of the IEEE Conference on Decision and Control*, 2006, pp. 5499–5504.

[72] A. Speranzon, C. Fischione, and K. Johansson, "Distributed and collaborative estimation over wireless sensor networks," in *Proceedings of the IEEE Conference on Decision and Control*, 2006, pp. 1025–1030.

[73] U. Khan, S. Kar, and M. Moura, "Distributed algorithms in sensor networks," in *Handbook on Sensor and Array Processing*, S. Haykin and K. Liu, Eds. New York, NY: Wiley-Interscience, 2009, pp. 716–749.

AUTHOR'S BIOGRAPHY

Sundhar Ram Srinivasan received his B.Tech and M.Tech degrees in Electrical Engineering, both in 2006, from the Indian Institute of Technology, Bombay, India. He worked as a research assistant at the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, from 2006 to 2009. He was a student-researcher at the Mathematics Department, Indian Institute of Science, Bangalore, India, in the summer of 2005 and a summer intern at Adchemy in the summer of 2009. He is the recipient of the Vodafone fellowship for the year 2007-2008.