Chapter 1
Distributed Optimization Over Networks

Angelia Nedić

Abstract The advances in wired and wireless technology necessitated the development of theory, models and tools to cope with new challenges posed by large-scale optimization problems over networks. The classical optimization methodology works under the premise that all problem data is available to some central entity (computing agent/node). This premise does not apply to large networked systems where typically each agent (node) in the network has access only to its private local information and has a local view of the network structure. This chapter will cover the development of such distributed computational models for time-varying networks, both deterministic and stochastic, which arise due to the use of different synchronous and asynchronous communication protocols in ad-hoc wireless networks. For each of these network dynamics, distributed algorithms for convex constrained minimization will be considered. In order to emphasize the role of the network structure in these approaches, our main focus will be on direct primal (sub)-gradient methods. The development of these methods combines optimization techniques with graph theory and the non-negative matrix theory, which model the network aspect. The lectures will provide some basic background theory on graphs, graph Laplacians and their properties, and the convergence results for related stochastic matrix sequences. Using the graph models and optimization techniques, the convergence and convergence rate analysis of the methods will be presented. The convergence rate results will demonstrate the dependence of the methods' performance on the problem and the network properties, such as the network capability to diffuse the information.

1.1 Introduction

Recent advances in wired and wireless technology have lead to the emergence of large-scale networks such as Internet, mobile ad-hoc networks, and wireless sensor networks. Their emergence gave rise to new network application domains ranging

A. Nedić (✉)
Arizona State University, Tempe, AZ, USA
e-mail: Angelia.Nedich@asu.edu

F. Facchinei, J.-S. Pang (eds.), *Multi-agent Optimization*, Lecture Notes in Mathematics 2224, https://doi.org/10.1007/978-3-319-97142-1_1

from data-base networks, social and economic networks to decentralized in-network operations including resource allocation, coordination, learning, and estimation.

As a result, there is a necessity to develop new models and tools for the design and performance analysis of such large complex networked systems. The problems arising in such networks stem mainly from two aspects, namely, *a lack of central authority or a coordinator* (a master node), and an inherent *dynamic of the network connectivity structure*. The lack of central authority in a network system naturally requires decentralized architecture for operations over the network (such as in the case of Internet). In some applications, the decentralized architecture is often preferred over a centralized architecture due to several reasons: (1) the size of the network (the number of agents) and the resources needed to coordinate (i.e., communicate) with a large number of agents; (2) a centralized network architecture is not desirable since it is not robust to the failure of the central entity; and (3) the privacy of agent information often cannot be preserved in a centralized systems. Furthermore, additional challenges in decentralized operations over such networks are encountered from the network connectivity structure that can vary over time due to unreliable communication links or mobility of the network agents.

The challenge is to control, coordinate, and analyze the performance of such networks. As a particular goal, one would like to develop distributed optimization algorithms that can be deployed in such networks that do not have a central coordinator, but exploit the network connectivity to achieve a global network performance objective. Thus, it is desirable that such algorithms are:

- Locally distributed in the sense that they rely on local information and observations only, i.e., the agents can exchange some limited information with their one-hop neighbors only;
- Robust against changes in the network topology (since the topology is not necessarily static as the communication links may not function perfectly);
- Easily implementable in the sense that the local computations performed by the agents are not expensive.

We next provide some examples of large scale networks and applications that arise within such networks.

Example 1 (Sensor Networks) A new computing concept based on a system of small sensors also referred to as motes or smart dust sensors, see Fig. 1.1. The sensors are of small size and have some computational, sensing and communication capabilities. They can be used in many different ways, such as for example, they may be mixed into concrete in order to monitor the structural health of buildings and bridges (smart structures), or be placed on power grids to monitor the power load (smart grids).

A specific problem of interest that supports a number of applications in sensor networks, such as building a piece-wise approximation of the coverage area, multi-sensor target localization and tracking problems, is the determination of Voronoi cells. A Voronoi cell of a sensor in a network is the locus of points in a sensor field that are the closest to a given sensor among all other sensors [6]. Upon determining

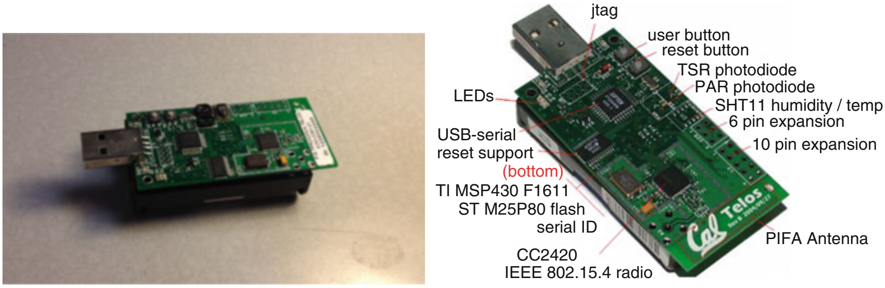


Fig. 1.1 A mote and its functionalities

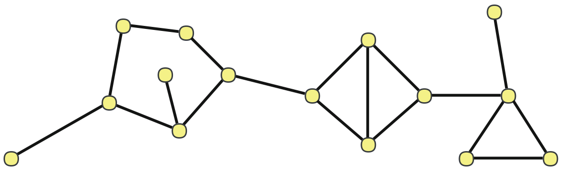


Fig. 1.2 A peer-to-peer network

such a partition in a distributed fashion, each sensor acts as a representative for the points in its cell. □

Example 2 (Computing Aggregates in Peer-to-Peer (P2P) Networks) In a P2P network consisting of m nodes, each node i has its local data/files stored with average size θ_i, which is known to node i only. The nodes are connected over a static undirected network (see Fig. 1.2), and they want to jointly compute the average file size $\frac{1}{m}\sum_{i=1}^{m}\theta_i$ without a central coordinator. In control theory and game theory literature, the problem is known as the *agreement or consensus problem* [15, 34, 60, 67, 171].

The problem has an optimization formulation, as follows:

$$\min_{x\in\mathbb{R}} \sum_{i=1}^{m}(x-\theta_i)^2.$$

It is a convex unconstrained problem with a strongly convex objective. Its unique solution θ^* is the average of the values θ_i, i.e., $\theta^* = \frac{1}{m}\sum_{i=1}^{m}\theta_i$. The solution cannot easily be computed when the agents have to calculate it in a distributed fashion by communicating only locally. In this case, the agents need to agree on the average of the values they hold.

In a more general variant of the consensus problem, the agents want to agree on some common value, which need not be the average of the values they initially have. For example, in a problem of leaderless heading alignment, autonomous agents move in a two-dimensional plane region with the same speed but different headings (they are refracted from the boundary to prevent them from leaving the area) [60, 174]. The objective is to design a local protocol that will ensure the alignment of the agent headings, while the agents communications are constrained by a given maximal distance. □

Another motivating example for distributed optimization over networks is a special machine learning problem, known as Support Vector Machine or maximum margin classifier. We discuss this problem in a centralized setting and then, we will see how it naturally fits in a distributed setting in the situations when the privacy of the data is of concern or when the data is too large to be shared.

Example 3 (Support Vector Machine (SVM)) We are given a data set $\{z_j, y_j\}_{j=1}^d$ consisting of d points, where $z_j \in \mathbb{R}^n$ is a measurement vector and $y_j \in \{+1, -1\}$ is its label. Assuming that the data can be perfectly separated, the problem consists of determining a hyperplane that separates data the best, i.e., solving the following convex problem:

$$\min_{x\in\mathbb{R}^n} F(x), \quad \text{where} \quad F(x) = \frac{\rho}{2}\|x\|^2 + \sum_{j=1}^{d} \max\left\{0, 1 - y_j\langle x, z_j\rangle\right\},$$

where $\rho > 0$ is a regularizing parameter that indicates the importance of having a small-norm solution. Given that the objective function is strongly convex, a solution exists and it is unique (see Fig. 1.3 for an illustration). The problem can be solved by using a subgradient method. If the data is distributed across several data centers, say m centers, then the joint problem can be written as:

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} \left(\frac{\rho}{2m}\|x\|^2 + \sum_{j\in D_i} \max\{0, 1 - y_j\langle x, z_j\rangle\} \right),$$

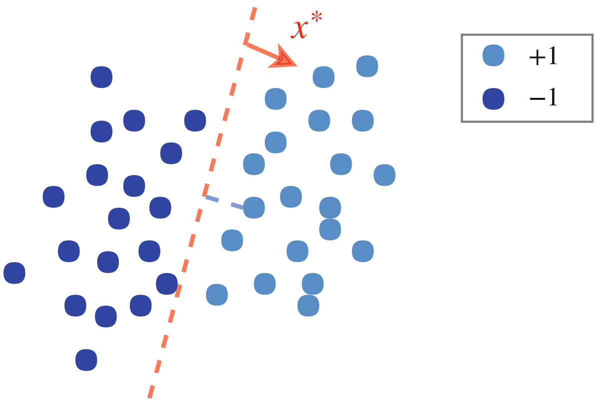


Fig. 1.3 A maximum margin separating hyperplane for a single data center

where D_i is the collection of data points at the center i. Letting

$$f_i(x) = \frac{\rho}{2m}\|x\|^2 + \sum_{j \in D_i} \max\{0, 1 - y_j \langle x, z_j \rangle\},$$

we see that the distributed variant of the problem assumes the following form:

$$\min_{x \in \mathbb{R}^n} F(x) = \sum_{i=1}^{m} f_i(x),$$

where the function f_i is known to center i. In this setting, sharing the function f_i with any other center amounts to sharing the entire data collection D_i available at center i. When the data is private or the data sets are too large, sharing the data is not an option and the problem has to be solved in a distributed manner. □

Many more examples of distributed problems and the use of consensus can be found in the domains of bio-inspired systems, self-organized systems, social networks, opinion dynamics, and autonomous (robotic) systems. For such examples, a reader may refer to some recent books and monographs on robotic networks [18, 94, 96], and social and economic networks [42, 49, 59]. These examples can also be found in related thesis works including [12, 52, 120, 155] dealing with averaging dynamics and [47, 69, 73, 130, 144, 183] dealing with distributed optimization aspects.

In the sequel, we will often refer to networks as graphs, and we will use “agent” and “node” interchangeably. The rest of the chapter is organized as follows: in Sect. 1.2, we formally describe a multi-agent problem in a network and discuss some related aspects of the consensus protocol. Section 1.3 presents a distributed synchronous algorithm for solving the multi-agent problem in time-varying undirected graphs, while Sect. 1.4 deals with asynchronous implementations over a static undirected graph. Section 1.5 concludes this chapter by providing an overview of related literature including the most recent research directions.

1.2 Distributed Multi-Agent Problem

This section provides a formal multi-agent system problem description, introduces our basic notation and gives the underlying assumptions on the multi-agent problem. The agents are embedded in a communication graph which accommodates distributed computations through the use of consensus protocols. A basic consensus protocol for undirected time-varying graphs is presented, and its convergence result is provided for later use.

1.2.1 Problem and Assumptions

Throughout this chapter, we will be focused on solving distributed problems of the generic form

$$\min_{x \in X} f(x) \qquad \text{with} \quad f(x) = \sum_{i=1}^{m} f_i(x), \tag{1.1}$$

in a network of m agents, where each function f_i is known to agent i only while the constraint set $X \subseteq \mathbb{R}^n$ is known to all agents. We will assume that the problem (1.1) is convex.

Assumption 1 *The set $X \subseteq \mathbb{R}^n$ is closed and convex, and each function $f_i : \mathbb{R}^n \to \mathbb{R}$ is convex.*

We will explicitly state when we assume that problem (1.1) has a solution. In such cases, we will let f^* denote the optimal value of the problem and X^* denote the set of its solutions,

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

Throughout the chapter, we will work with the Euclidean norm, denoted by $\|\cdot\|$, unless otherwise explicitly stated. We use $\langle \cdot, \cdot \rangle$ to denote the inner product. We will view all vectors as column vectors, unless stated otherwise. We will use the prime to denote the transpose of a matrix and a vector.

We assume that the agents are embedded in a communication network, which allows the agents to exchange some limited information with their immediate (one-hop) neighbors. Multi-hop communications are not allowed in this setting. The agents' common goal is to solve the problem (1.1) collaboratively.

The communication network structure over time is captured with a sequence of time-varying undirected graphs. More specifically, we assume that the agents exchange their information (and perform some updates) at given discrete time instances, which are indexed by $k = 0, 1, 2, \ldots$. The communication network structure at time k is represented by an undirected graph $G_k = ([m], E_k)$, where $[m]$ is the agent (node) set, i.e., $[m] = \{1, \ldots, m\}$, while E_k is the set of edges. The edge $i \leftrightarrow j \in E_k$ indicates that agents i and j can communicate (send and receive messages) at time k.

Given a graph G_k at a time k, we let $N_i(k)$ denote the set of neighbors of agent i, at time k:

$$N_i(k) = \{j \in [m] \mid i \leftrightarrow j \in E_k\} \cup \{i\}.$$

Note that the neighbor set $N_i(k)$ includes agent i itself, which reflects the fact that agent i has access to some information from its one-hop neighbors and its own information.

The agents' desire to solve the problem (1.1) jointly through local communications translates to the following problem that agents are facing at time k:

$$\begin{aligned} \min \quad & \mathbf{f}(x_1, \ldots, x_m) \quad \text{with} \quad \mathbf{f}(x_1, \ldots, x_m) = \sum_{i=1}^{m} f_i(x_i) \\ \text{subject to} \quad & x_i = x_j \text{ for all } j \in N_i(k) \text{ and all } i \in [m], \\ & x_i \in X \text{ for all } i \in [m]. \end{aligned} \tag{1.2}$$

Thus, the agents are facing a sequence of optimization problems with time-varying constraints, which are capturing the time-varying structure of the underlying communication network. Since this is a nonstandard optimization problem, we need to specify what it means to solve the problem. To do so, we will impose some additional assumptions on the graphs G_k.

Throughout, we will assume that the graphs G_k are connected.

Assumption 2 *Each graph G_k is connected.*

This assumption can be relaxed to the requirement that the union of B consecutive graphs $G_k, \ldots, G_{k+B-1}$ is connected for all $k \geq 0$ and for some positive integer B. However, to keep the exposition simple, we will adopt Assumption 2.

Let C_k be the constraint set of problem (1.2) at time k, i.e.,

$$C_k = \{(x_1, \ldots, x_m) \in X^m \mid x_i = x_j \text{ for all } j \in N_i(k) \text{ and all } i \in [m]\}.$$

Under Assumption 2, the constraint sets C_k are all the same. Their description is given to the agents through a different set of equations at different time instances, as seen from the following lemma.

Lemma 1 *Let Assumption 2 hold. Then, for each k, we have*

$$C_k = \{(x_1, \ldots, x_m) \mid x_i = x \textit{ for some } x \in X \textit{ and all } i \in [m]\}.$$

The proof of Lemma 1 is straightforward and it is omitted. In fact, it can be seen that Lemma 1 also holds when the graphs G_k are directed and each of the graphs contains a directed rooted spanning tree,[1] where the neighbor set $N_i(k)$ is replaced with the in-neighbor set[2] $N_i^{\text{in}}(k)$ of agent i at time k.

[1]There exists a node i such that the graph contains a directed path from node i to any other node in the network.

[2]The set $N_i^{\text{in}}(k)$ of in-neighbors of agent i in a directed graph G_k is the set of all agents j such that the directed edge (j, i) exists in the graph.

In view of Lemma 1, it is now obvious that we can associate a limit problem with the sequence of problems (1.2), where the limit problem is given by:

$$\begin{aligned} &\min && \mathbf{f}(x_1,\ldots,x_m) \quad \text{with} \quad \mathbf{f}(x_1,\ldots,x_m)=\sum_{i=1}^m f_i(x_i) \\ &\text{subject to} && (x_1,\ldots,x_m)\in \cap_{k=1}^{\infty} C_k. \end{aligned} \tag{1.3}$$

As we noted, all sets C_k are the same under Assumption 2. However, we will keep the notation C_k to capture the fact that the agents have a different set of equations that describe the constraint set at different times. Furthermore, the preceding formulation of the limit problem is also suitable for the situations where the graphs G_k are not necessarily connected.

1.2.2 Consensus Problem and Algorithm

The consensus problem is a special case of the limit problem (1.3), where each $f_i \equiv 0$ and $X = \mathbb{R}^n$, i.e., the consensus problem is given by

$$\begin{aligned} &\min && 0 \\ &\text{subject to} && (x_1,\ldots,x_m)\in \cap_{k=1}^{\infty} C_k, \end{aligned} \tag{1.4}$$

with

$$C_k = \{(x_1,\ldots,x_m) \mid x_i = x \text{ for some } x\in\mathbb{R}^n \text{ and all } i\in[m]\} \qquad \text{for all } k\geq 1.$$

As one may observe, the consensus problem is a feasibility problem where the agents need to collectively determine an $\mathbf{x} = (x_1,\ldots,x_m)$ satisfying the constraint in (1.4), while obeying the communication structure imposed by graph G_k at each time k.

A possible way to solve the consensus problem is that each agent considers its own problem, at time k, of the following form:

$$\min_{x\in\mathbb{R}^n} \sum_{j\in N_i(k)} p_{ij}(k)\|x-x_j\|^2,$$

where $p_{ij}(k) > 0$ for all $j \in N_i(k)$ and for all $i \in [m]$. The values x_j are assumed to be communicated to agent i by its neighbors $j \in N_i(k)$. This problem can be viewed as a penalty problem associated with the constraints in the set C_k that involve agent i decision variable. The objective function is strongly convex and it has a unique solution, denoted by $\hat{x}_i$, i.e.,

$$\hat{x}_i(k) = \operatorname*{argmin}_{x\in\mathbb{R}^n} \sum_{j\in N_i(k)} p_{ij}(k)\|x-x_j\|^2.$$

In the following lemma, we provide the closed form of the solution $\hat{x}_i(k)$.

Lemma 2 *Let $X = \mathbb{R}^n$, and consider the feasible set*

$$C_{ik} = \{(x_j)_{j\in N_i(k)} \mid x_j = x \textit{ for all } j \in N_i(k) \textit{ and } x \in \mathbb{R}^n\} \tag{1.5}$$

corresponding to the constraints in C_k that involve agent i at time k. Then, the solution $\hat{x}_i(k)$ of the penalty problem $\min_{x\in\mathbb{R}^n} \sum_{j\in N_i(k)} p_{ij}(k)\|x - x_j\|^2$ *associated with the feasible set C_{ik} is given by*

$$\hat{x}_i(k) = \frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}.$$

Proof We note that

$$\begin{aligned}
&\sum_{j\in N_i(k)} p_{ij}(k)\|x - x_j\|^2 \\
&\quad= \sum_{j\in N_i(k)} p_{ij}(k)\|x\|^2 - 2\left\langle x, \sum_{j\in N_i(k)} p_{ij}(k)x_j\right\rangle + \sum_{j\in N_i(k)} p_{ij}(k)\|x_j\|^2 \\
&\quad= \sum_{j\in N_i(k)} p_{ij}(k)\left[\|x\|^2 - 2\left\langle x, \frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\rangle + \frac{\sum_{j\in N_i(k)} p_{ij}(k)\|x_j\|^2}{\sum_{j\in N_i(k)} p_{ij}(k)}\right] \\
&\quad= \sum_{j\in N_i(k)} p_{ij}(k)\left[\|x\|^2 - 2\left\langle x, \frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\rangle + \left\|\frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\|^2\right. \\
&\qquad\left. - \left\|\frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\|^2 + \frac{\sum_{j\in N_i(k)} p_{ij}(k)\|x_j\|^2}{\sum_{j\in N_i(k)} p_{ij}(k)}\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{j\in N_i(k)} p_{ij}(k)\|x - x_j\|^2 = \sum_{j\in N_i(k)} p_{ij}(k)&\left[\left\|x - \frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\|^2\right. \\
&\left. - \left\|\frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\|^2 + \frac{\sum_{j\in N_i(k)} p_{ij}(k)\|x_j\|^2}{\sum_{j\in N_i(k)} p_{ij}(k)}\right].
\end{aligned}$$

Since the last two terms in the preceding sum do not depend on x, we see that

$$\begin{aligned}\hat{x}_i(k) &= \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\sum_{j\in N_i(k)} p_{ij}(k)\|x-x_j\|^2\right\}\\ &= \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\left(\sum_{j\in N_i(k)} p_{ij}(k)\right)\left\|x-\frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\|^2\right\}.\end{aligned}$$

Furthermore, since $\sum_{j\in N_i(k)} p_{ij}(k) > 0$, we finally have

$$\hat{x}_i(k) = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\left\{\left\|x-\frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}\right\|^2\right\} = \frac{\sum_{j\in N_i(k)} p_{ij}(k)x_j}{\sum_{j\in N_i(k)} p_{ij}(k)}.$$

□

In view of Lemma 2, the penalty problem associated with agent i feasible set C_{ik} at time k can be equivalently be given by

$$\min_{x\in\mathbb{R}^n}\sum_{j\in N_i(k)} w_{ij}(k)\|x-x_j\|^2,$$

where the weights $w_{ij}(k)$, $j\in N_i(k)$, correspond to convex combinations, i.e.,

$$w_{ij}(k) > 0 \quad \text{for all } j\in N_i(k), \qquad \sum_{j\in N_i(k)} w_{ij}(k) = 1. \tag{1.6}$$

Obviously, for the equivalence of the two penalty problems, we need $w_{ij}(k) = p_{ij}(k)/\sum_{j\in N_i(k)} p_{ij}(k)$. In this case, the corresponding solution $\hat{x}_i(k)$ is given by

$$\hat{x}_i(k) = \sum_{j\in N_i(k)} w_{ij}(k)x_j.$$

The preceding discussion motivates the following algorithm, known as a *consensus algorithm* (with projections), for solving the constrained consensus problem (1.4): each agent has a variable $x_i(k)$ at time k. At time $k+1$, every agent i sends $x_i(k)$ to its neighboring agents $j\in N_i(k)$ and receives $x_j(k)$ from them. Then, every agent i updates its variable as follows:

$$x_i(k+1) = \sum_{j\in N_i(k)} w_{ij}(k)x_j(k),$$

where $w_{ij}(k) > 0$ for all $j \in N_i(k)$ and all $i \in [m]$, and $\sum_{j \in N_i(k)} w_{ij}(k) = 1$ for all $i \in [m]$. For a more compact representation, we define $w_{ij}(k) = 0$ for all $j \notin N_i(k)$ and all $i \in [m]$, so we have

$$x_i(k+1) = \sum_{j=1}^{m} w_{ij}(k)x_j(k) \qquad \text{for all } i \in [m] \text{ and all } k \geq 0. \tag{1.7}$$

The initial points $x_i(0) \in \mathbb{R}^n$, $i \in [m]$, are assumed to be arbitrary.

We note here that if a (convex) constraint set $X \subseteq \mathbb{R}^n$ is known to all agents, then the constrained consensus problem (1.4) can also be solved by the consensus algorithm in (1.7) with an adjustment of the initial selections $x_i(0)$ to satisfy $x_i(0) \in X$ for all i. This can be seen by noting that $x_i(k+1)$ is a convex combination of $x_j(k)$ for $j \in N_i(k)$ (see (1.6)), and it will lie in the set X as long as this set is convex and $x_j(k) \in X$ for all $j \in N_i(k)$.

The consensus algorithm in (1.7) has regained interest since the recent work [60], which attracted a significant attention to the consensus problem in various settings (for an overview of the consensus related literature see Sect. 1.5).

For the convergence of the consensus algorithm, some additional assumptions are typically needed for the weights $w_{ij}(k)$ aside from the "convex combination" requirement captured by relation (1.6). To state one such assumption, we will introduce some additional terminology and notation. We let $W(k)$ be the matrix with ijth entry equal to $w_{ij}(k)$. We will say that a matrix W is (row) *stochastic* if its entries are non-negative and the sum of its entries in each row is equal to 1. We will say that W is *doubly stochastic* if both W and its transpose W' are stochastic matrices.

Next, we state an assumption on the matrices $W(k)$ that we will use later on.

Assumption 3 *For every $k \geq 0$, the matrix $W(k)$ has the following properties:*

(a) $W(k)$ is doubly stochastic.
(b) $W(k)$ is compatible with the structure of the graph G_k, i.e.,

$$w_{ij}(k) = 0 \qquad iff \qquad i \leftrightarrow j \notin E_k.$$

(c) $W(k)$ has positive diagonal entries, i.e., $w_{ii}(k) > 0$ for all $i \in [m]$.
(d) There is an $\eta > 0$ such that

$$w_{ij}(k) \geq \eta \qquad iff \qquad i \leftrightarrow j \in E_k.$$

First, let us note that Assumption 3 is much stronger than what is typically assumed to guarantee the convergence of the consensus algorithm. In general, the graph G_k can be directed and the positive weights $w_{ij}(k)$ are assumed for the directed links $(j,i) \in E_k$, while the matrix $W(k)$ is assumed to be just (row) stochastic. We work with a stronger assumption since we want to address the optimization problem (1.3) which has a more general objective function than that of the consensus problem (1.4).

To provide insights into what motivates Assumption 3, consider the consensus algorithm in the case of an unconstrained scalar problem, i.e., $X = \mathbb{R}$. Then, for the consensus algorithm in (1.7), by stacking all the variables $x_i(k)$ into a single vector $x(k)$ at time k, we have

$$x(k+1) = W(k)x(k) = \cdots = W(k)W(k-1)\cdots W(0)x(0).$$

Furthermore, in the case of static graphs G_k, i.e., $G_k = G$ for some graph G, we can use $W(k) = W$ for all k, thus implying that

$$x(k) = W^k x(0).$$

When W is a stochastic matrix which is compatible with a connected graph G, then W is irreducible and, by Perron-Frobenius Theorem (see Theorem 4.2.1, page 101 of [46]), the spectral radius $\rho(W)$ (which is equal to 1 in this case) is a simple positive eigenvalue and the vector $\mathbf{1}$ with all entries equal to 1 is the unique right-eigenvector associated with eigenvalue 1, i.e.,

$$W\mathbf{1} = \mathbf{1}.$$

When, in addition, W has positive diagonal entries, then W is also primitive, so we have

$$\lim_{k\to\infty} W^k = \mathbf{1}v',$$

where v is the normalized (unique positive) left-eigenvector of W associated with eigenvalue 1, i.e., a unique vector satisfying

$$v'W = v' \qquad \text{where } v_i > 0 \text{ for all } i \quad \text{and} \quad \langle v, \mathbf{1}\rangle = 1$$

(see Theorem 4.3.1, page 106, and Theorem 4.4.4, page 119, both in [46]). Thus, when W is stochastic, compatible with a connected graph G, and has a positive diagonal, we obtain

$$\lim_{k\to\infty} x(k) = \left(\lim_{k\to\infty} W^k\right) x(0) = \mathbf{1}v'x(0) = \langle v, x(0)\rangle\, \mathbf{1}.$$

Hence, in this case, the consensus is reached, i.e., the iterates of the consensus algorithm converge to the value $\langle v, x(0)\rangle$, which is a convex combination of the initial agents' values $x_i(0)$. Observe that the behavior of the iterates in the limit, as k increases, is completely determined by the limit behavior of W^k as $k \to \infty$.

In the light of the preceding discussion, Assumption 3 guarantees that a similar behavior is exhibited in the case when the matrices are time-varying and the graphs G_k are connected. Specifically, in this case, we would like to have

$$\lim_{k\to\infty} [W(k)W(k-1)\cdots W(0)] = \mathbf{1}v'$$

for some vector v with all entries v_i positive and $\langle v, \mathbf{1}\rangle = 1$. This relation is guaranteed by Assumptions 2 and 3. In fact, under these assumptions we have a stronger result for the matrix sequence $\{W(k)\}$, as follows:

$$\lim_{k\to\infty} [W(k)W(k-1)\cdots W(s)] = \frac{1}{m}\mathbf{1}\mathbf{1}' \qquad \text{for all } s \geq 0.$$

This result is formalized in the following lemma, which also provides the rate of convergence for the matrix products $W(k)W(k-1)\cdots W(s)$ for all $k \geq s \geq 0$.

Lemma 3 (Lemma 5 in [110]) *Let the graph sequence $\{G_k\}$ satisfy Assumption 2, and let the matrix sequence $\{W(k)\}$ satisfy Assumption 3. Then, we have for all $s \geq 0$ and $k \geq s$,*

$$\sup_{x\in\mathbb{R}^n, \|x\|=1} \left\| \left(W(k)W(k-1)\cdots W(s+1)W(s) - \frac{1}{m}\mathbf{1}\mathbf{1}' \right) x \right\|^2 \leq \left(1 - \frac{\eta}{2m^2}\right)^{k-s}.$$

In particular, for all $s \geq 0$ and $k \geq s$, and for all $i, j \in [m]$,

$$\left([W(k)W(k-1)\cdots W(s+1)W(s)]_{ij} - \frac{1}{m} \right)^2 \leq \left(1 - \frac{\eta}{2m^2}\right)^{k-s}.$$

The first relation in Lemma 3 is a consequence of Lemma 5 in [110]. The second relation follows by letting x be any of the unit-vectors of the standard basis in $\mathbb{R}^n$.

Lemma 3 provides a key insight into the behavior of the products of the matrices $W(k)$, which implies that the consensus method in (1.7) converges geometrically to $\frac{1}{m}\sum_{i=1}^m x_i(0)$. We will use this lemma to show that consensus-based methods for solving a more general optimization problem (1.1) converge to a solution, as discussed in the next section.

1.3 Distributed Synchronous Algorithms for Time-Varying Undirected Graphs

We now consider a distributed algorithm for solving problem (1.3). We assume that the set X is closed and convex, and it has a simple structure so that the projection of a point on the set X is not computationally expensive. The idea is to construct an algorithm to be executed locally by each agent i that at every instant k involves two steps: one step aimed at satisfying agent i feasibility constraint C_{ik} in (1.5), and the other step aimed at minimizing its objective cost f_i over the set X. Thus, the first step is akin to consensus update in (1.7), while the second step is a simple projection-based (sub)gradient update using f_i.

To illustrate the idea, consider agent i and its surrogate objective function at time k:

$$F_{ik}(x) = f_i(x) + \delta_X(x) + \frac{1}{2} \sum_{j \in N_i(k)} w_{ij}(k) \|x - x_j\|^2,$$

where $\delta_X(x)$ is the indicator function of the set X, i.e.,

$$\delta_X(x) = \begin{cases} 0 & \text{if } x \in X, \\ +\infty & \text{otherwise.} \end{cases}$$

The weights $w_{ij}(k)$, $j \in N_i(k)$, are convex combinations (i.e., they are positive and they sum to 1; see (1.6)).

Having the vectors x_j, $j \in N_i(k)$, agent i may take the first step aimed at minimizing $\frac{1}{2} \sum_{j \in N_i(k)} w_{ij}(k) \|x - x_j\|^2$, which would result in setting

$$\hat{x}_i(k) = \sum_{j \in N_i(k)} w_{ij}(k) x_j.$$

In the second step, assuming for the moment that f_i is differentiable, agent i considers solving the problem

$$\min_{x \in \mathbb{R}^n} \left\{ \langle \nabla f_i(\hat{x}_i(k)), x \rangle + \delta_X(x) + \frac{1}{2\alpha_k} \|x - \hat{x}_i(k)\|^2 \right\},$$

which is equivalent to

$$\min_{x \in X} \left\{ \langle \nabla f_i(\hat{x}_i(k)), x \rangle + \frac{1}{2\alpha_k} \|x - \hat{x}_i(k)\|^2 \right\},$$

where $\alpha_k > 0$ is a stepsize. The preceding problem has a closed form solution given by

$$x_i^*(k) = \Pi_X[\hat{x}_i(k) - \alpha_k \nabla f(\hat{x}_i(k))],$$

where $\Pi_X[z]$ is the projection of a point z on the set X, i.e.,

$$\Pi_X[z] = \operatorname*{argmin}_{x \in X} \|x - z\|^2 \qquad \text{for all } z \in \mathbb{R}^n.$$

When the function f_i is not differentiable, we would replace the gradient $\nabla f(\hat{x}_i(k))$ with a subgradient $g_i(\hat{x}_i(k))$. Recall that a subgradient of a convex function $h : \mathbb{R}^n \to \mathbb{R}$ at a given point x is a vector $g(x) \in \mathbb{R}^n$ such that

$$h(x) + \langle g(x), y - x \rangle \leq h(y) \qquad \text{for all } y \in \mathbb{R}^n.$$

In what follows, we will use $g_i(k)$ to abbreviate the notation for a subgradient $g_i(\hat{x}_i(k))$ of the function $f_i(z)$ evaluated at $z = \hat{x}_i(k)$.

Now, based on the preceding discussion, we have the following algorithm: at every time k, each agent $i \in [m]$ maintains two vectors $y_i(k)$ and $x_i(k)$. The agent sends $x_i(k)$ to its neighbors $j \in N_i(k)$ and receives $x_j(k)$ from its neighbors $j \in N_i(k)$. Then, it updates as follows:

$$\begin{aligned} y_i(k+1) &= \sum_{j \in N_i(k)} w_{ij}(k) x_j(k), \\ x_i(k+1) &= \Pi_X[y_i(k+1) - \alpha_{k+1} g_i(k+1)], \end{aligned} \tag{1.8}$$

where $\alpha_{k+1} > 0$ is a stepsize and $g_i(k+1)$ is a subgradient of $f_i(z)$ at point $z = y_i(k+1)$. The process is initialized with arbitrary points $x_i(0) \in X$ for all $i \in [m]$.

Note that the agents use the same stepsize value α_{k+1}. Note further that, due to the projection on the set X, we have $x_i(k) \in X$ for all i and k. Moreover, since $y_i(k+1)$ is a convex combination of points in X and since X is convex, we have $y_i(k+1) \in X$ for all i and k.

By introducing 0-weights for non-existing links in the graph G_k, i.e., by defining

$$w_{ij}(k) = 0 \qquad \text{when } j \notin N_i(k),$$

we can re-write (1.8) as follows: for all $k \geq 0$ and all $i \in [m]$,

$$\begin{aligned} y_i(k+1) &= \sum_{j=1}^{m} w_{ij}(k) x_j(k), \\ x_i(k+1) &= \Pi_X[y_i(k+1) - \alpha_{k+1} g_i(k+1)]. \end{aligned} \tag{1.9}$$

To illustrate the iterations of the algorithm in (1.9), consider a system of three agents in a connected graph, as illustrated in Fig. 1.4. Figure 1.4 shows a typical iteration of the algorithm. Since the graph is fully connected, all weights $w_{ij}(k)$ are positive, so the resulting points $y_i(k+1)$ lie inside the triangle formed by the points $x_i(k)$, $i = 1, 2, 3$. The new points $x_i(k+1)$, $i = 1, 2, 3$, obtained after the subgradient steps do not necessarily lie inside the triangle formed by the points $x_i(k)$, $i = 1, 2, 3$. Under some suitable assumptions on the stepsize and the subgradients, these triangles formed by $x_i(k)$, $i = 1, 2, 3$, as $k \to \infty$, can shrink into a single point, which is solution of the problem. Loosely speaking, while the consensus steps force the agents to agree on some point, the subgradient steps are

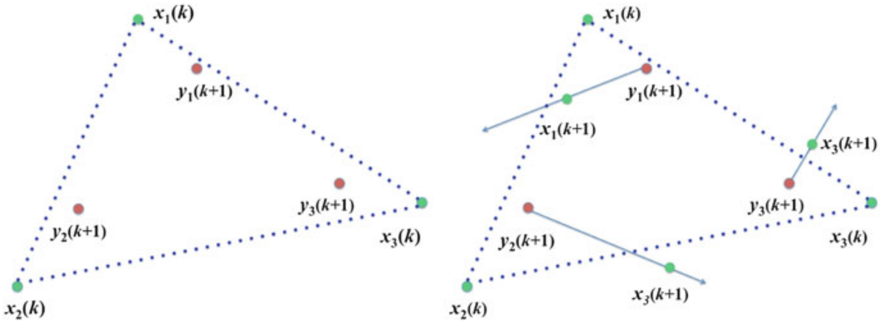


Fig. 1.4 At iteration k, agents hold values $x_i(k)$. The plot to the left illustrates the resulting points $y_i(k+1)$ of the iteration (1.9) which lie inside the triangle formed by the points $x_i(k)$, $i = 1, 2, 3$, (as all weights $w_{ij}(k)$ are positive in this case). The plot to the right depicts the iterates $x_i(k+1)$, $i = 1, 2, 3$, obtained through the subgradient steps of algorithm (1.9). These iterates do not necessarily lie inside the triangle formed by the prior iterates $x_i(k)$, $i = 1, 2, 3$

forcing the agreement point to be a solution of a given problem. Thus, one can think of the algorithm in (1.9) as a process that steers the consensus toward a particular region, in this case the region being the solution set of the agent optimization problem (1.3). To see this, note that from the definition of $x_i(k+1)$ we have

$$x_i(k+1) = y_i(k+1) - \alpha_{k+1} g_i(k+1) + \mathsf{e}_k,$$
$$\mathsf{e}_k = (\Pi_X[y_i(k+1) - \alpha_{k+1} g_i(k+1)] - (y_i(k+1) - \alpha_{k+1} g_i(k+1)).$$

Assuming that the projection error e_k is small, and assuming that the functions are differentiable, we can approximate $x_i(k+1)$ as follows:

$$x_i(k+1) \approx y_i(k+1) - \alpha_{k+1} \nabla f_i(y_i(k+1))$$
$$= \left(\sum_{j=1}^{m} w_{ij}(k) x_j(k) \right) - \alpha_{k+1} \nabla f_i \left(\sum_{j=1}^{m} w_{ij}(k) x_j(k) \right). \quad (1.10)$$

Thus, the algorithm is similar to the consensus process $\sum_{j=1}^{m} w_{ij}(k) x_j(k)$ with an additional force coming from the gradient field, which steers the agreement point toward a solution of the problem $\min_{x \in X} \sum_{i=1}^{m} f_i(x)$. The preceding discussion sketches the approach that we will follow to establish the convergence properties of the method, which is the focus of the next section.

1.3.1 Convergence Analysis of Distributed Subgradient Method

In this section, we provide a main convergence result in Theorem 1 showing that the iterates $x_i(k)$, for all agents $i \in [m]$, converge to a solution of the problem (1.1), as $k \to \infty$. The proof of Theorem 1 relies on a basic relation satisfied by the algorithm in terms of all agents' iterates, as given in Proposition 1. The proof of this proposition is constructed through several auxiliary results that are provided in Lemmas 4–6. Specifically, Lemma 4 provides an elementary relation for the iterates $x_i(k)$ for a single agent, without the use of the network aspect. By viewing the algorithm (1.9) as a perturbation of the consensus algorithm, Lemma 5 establishes a relation for the distances between the iterates $x_i(k)$ and their averages taken across the agents (i.e., $\frac{1}{m}\sum_{j=1}^m x_j(k)$) in terms of the perturbation. The result of Lemma 5 is refined in Lemma 6 by taking into account that the perturbation to the consensus algorithm comes from a subgradient influence controlled by a stepsize choice.

Based on relation (1.10), we see that if $y_i(k+1)$ is close to the average of the points $x_j(k+1)$, $j \in [m]$, then for the iterate $x_i(k+1)$ we have

$$x_i(k+1) \approx x_{av}(k+1) - \alpha_{k+1}\nabla f_i(x_{av}(k+1)) + \epsilon_{k+1},$$

where $x_{av}(k+1) = \frac{1}{m}\sum_{j=1}^m x_j(k+1)$ and ϵ_{k+1} is an error due to using the gradient difference $\nabla f_i(x_{av}(k+1)) - \nabla f_i(y_i(k+1))$. When f_i is not differentiable, the iterates $x_i(k+1)$ would similarly correspond to an approximate subgradient update, where a subgradient $g_i(k+1)$ of $f_i(z)$ at $z = y_i(k+1)$ is used instead of a subgradient of $f_i(z)$ evaluated at $z = x_{av}(k+1)$ (which would have been used if the average $x_{av}(k+1)$ were available to all agents). Thus, the method (1.9) can be interpreted as an approximation of a centralized algorithm, where each agent would have access to the average vector $x_{av}(k+1)$ and could update by computing gradients of its own objective function f_i at the average $x_{av}(k+1)$.

1.3.1.1 Relation for a Single Agent Iterates

To start the analysis, for a single arbitrary agent, we will explore a basic relation for the distances between $x_i(k+1)$ and a point $x \in X$. In doing so, we will use the well-known property of the projection operator, namely

$$\|\Pi_X[z] - x\|^2 \le \|z - x\|^2 - \|\Pi_X[z] - z\|^2 \qquad \text{for all } x \in X \text{ and all } z \in \mathbb{R}^n. \tag{1.11}$$

The preceding projection relation follows from a more general relation which can be found in [44], in Volume II, 12.1.13 Lemma, on page 1120.

Lemma 4 *Let the problem be convex (Assumption 1 holds) and let* $\alpha_{k+1} > 0$. *Then, for the iterate* $x_i(k+1)$ *of the method* (1.9)*, we have for all* $x \in X$ *and all* $i \in [m]$,

$$\|x_i(k+1) - x\|^2 \le \|y_i(k+1) - x\|^2 - 2\alpha_{k+1}\left(f_i(y_i(k+1)) - f_i(x)\right) + \alpha_{k+1}^2\|g_i(k+1)\|^2.$$

Proof From the projection relation in (1.11) and the definition of $x_i(k+1)$ we obtain for any $x \in X$,

$$\|x_i(k+1) - x\|^2 \leq \|y_i(k+1) - \alpha_{k+1} g_i(k+1) - x\|^2 \\ - \|x_i(k+1) - y_i(k+1) + \alpha_{k+1} g_i(k+1)\|^2.$$

By expanding the squared-norm terms, we further have

$$\begin{aligned}\|x_i(k+1) - x\|^2 \leq & \|y_i(k+1) - x\|^2 - 2\alpha_{k+1}\langle y_i(k+1) - x, g_i(k+1)\rangle \\ & + \alpha_{k+1}^2 \|g_i(k+1)\|^2 - \|x_i(k+1) - y_i(k+1)\|^2 \\ & - 2\alpha_{k+1}\langle x_i(k+1) - y_i(k+1), g_i(k+1)\rangle \\ & - \alpha_{k+1}^2 \|g_i(k+1)\|^2 \\ = & \|y_i(k+1) - x\|^2 - 2\alpha_{k+1}\langle y_i(k+1) - x, g_i(k+1)\rangle \\ & - \|x_i(k+1) - y_i(k+1)\|^2 \\ & - 2\alpha_{k+1}\langle x_i(k+1) - y_i(k+1), g_i(k+1)\rangle.\end{aligned}$$

Since $g_i(k+1)$ is a subgradient of f_i at $y_i(k+1)$, by convexity of f_i, we have

$$\langle y_i(k+1) - x, g_i(k+1)\rangle \geq f_i(y_i(k+1)) - f_i(x),$$

implying that

$$\begin{aligned}\|x_i(k+1) - x\|^2 \leq & \|y_i(k+1) - x\|^2 - 2\alpha_{k+1}\left(f_i(y_i(k+1)) - f_i(x)\right) \\ & - \|x_i(k+1) - y_i(k+1)\|^2 \\ & - 2\alpha_{k+1}\langle x_i(k+1) - y_i(k+1), g_i(k+1)\rangle.\end{aligned}$$

The last term in the preceding relation can be estimated by using Cauchy-Schwarz inequality, to obtain

$$\begin{aligned} & - 2\alpha_{k+1}\langle x_i(k+1) - y_i(k+1), g_i(k+1)\rangle \\ & \leq 2\|x_i(k+1) - y_i(k+1)\| \cdot \alpha_{k+1}\|g_i(k+1)\| \\ & \leq \|x_i(k+1) - y_i(k+1)\|^2 + \alpha_{k+1}^2 \|g_i(k+1)\|^2.\end{aligned}$$

By combining the preceding two relations, we find that for any $x \in X$,

$$\|x_i(k+1) - x\|^2 \leq \|y_i(k+1) - x\|^2 - 2\alpha_{k+1}\left(f_i(y_i(k+1)) - f_i(x)\right) \\ + \alpha_{k+1}^2 \|g_i(k+1)\|^2.$$

□

1.3.1.2 Relation for Agents' Iterates and Their Averages Through Perturbed Consensus

We would like to estimate the difference between $x_i(k+1)$ and the average of these vectors, which can be then used in Lemma 4 to get some insights into the behavior of $\|x_i(k) - x^*\|$ for an optimal solution x^*. To do so, we will re-write the iterations of the method (1.9), as follows:

$$\begin{aligned} y_i(k+1) &= \sum_{j=1}^{m} w_{ij}(k) x_j(k), \\ x_i(k+1) &= y_i(k+1) + \underbrace{(\Pi_X[y_i(k+1) - \alpha_{k+1} g_i(k+1)] - y_i(k+1))}_{\epsilon_i(k+1)}. \end{aligned}$$

Thus, we have for all i and $k \geq 0$,

$$\begin{aligned} x_i(k+1) &= \sum_{j=1}^{m} w_{ij}(k) x_j(k) + \epsilon_i(k+1), \\ \epsilon_i(k+1) &= \Pi_X[y_i(k+1) - \alpha_{k+1} g_i(k+1)] - y_i(k+1), \\ y_i(k+1) &= \sum_{j=1}^{m} w_{ij}(k) x_j(k). \end{aligned} \tag{1.12}$$

In this representation, the iterates $x_i(k+1)$ can be viewed as obtained through a perturbed consensus algorithm, where $\epsilon_i(k+1)$ is a perturbation at agent i.

Under suitable conditions (cf. Assumption 3), by Lemma 3, we know that the matrix products $W(k)W(k-1)\cdots W(t)$ are converging as $k \to \infty$, for any t, to the matrix with all entries equal to $1/m$. We will use that result to establish a relation for the behavior of the iterates $x_i(k+1)$.

Lemma 5 *Let the graphs G_k satisfy Assumption 2 and the matrices $W(k)$ satisfy Assumption 3. Then, for the iterate process (1.12), we have for all $k \geq 0$,*

$$\begin{aligned} &\sqrt{\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2} \\ &\quad \leq mp^k \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + m \left(\sum_{t=1}^{k} p^{k-t} \sqrt{\sum_{i=1}^{m} \|\epsilon_i(t)\|^2} \right) \\ &\qquad + \sqrt{m-1} \sqrt{\sum_{i=1}^{m} \|\epsilon_i(k+1)\|^2}, \end{aligned}$$

where $x_{av}(k+1) = \frac{1}{m}\sum_{j=1}^{m} x_j(k+1)$, $p = 1 - \frac{\eta}{4m^2}$ *and* $\eta > 0$ *is a uniform lower bound on the entries of the matrices* $W(k)$ *(see Assumption 3(d)).*

Proof We write the evolution of the iterates $x_i(k+1)$ in (1.12) in a matrix representation. Letting $\ell \in [n]$ be any coordinate index, we can write for the ℓth coordinate (denoted by a superscript)

$$x_i^\ell(k+1) = \sum_{j=1}^{m} w_{ij}(k) x_j^\ell(k) + \epsilon_i^\ell(k+1) \qquad \text{for all } \ell \in [n].$$

Stacking all the ℓth coordinates in a column vector, denoted by $x^\ell(k+1)$, we have

$$x^\ell(k+1) = W(k) x_j^\ell(k) + \epsilon^\ell(k+1) \qquad \text{for all } \ell \in [n].$$

Next, we take the column vectors $x^\ell(k+1)$, $\ell \in [n]$, in a matrix $\mathbf{X}(k+1)$, for all k, and similarly, we construct the matrix $\mathbf{E}(k+1)$ from the perturbation vectors $\epsilon^\ell(k+1)$, $\ell \in [n]$. Thus, we have the following compact form representation for the evolution of the iterates $x_i(k+1)$:

$$\mathbf{X}(k+1) = W(k)\mathbf{X}(k) + \mathbf{E}(k+1) \qquad \text{for all } k \geq 0. \tag{1.13}$$

Using the recursion, from (1.13) we see that for all $k \geq 0$,

$$\begin{aligned}\mathbf{X}(k+1) &= W(k)\mathbf{X}(k) + \mathbf{E}(k+1) \\ &= W(k)W(k-1)\mathbf{X}(k-1) + W(k)\mathbf{E}(k) + \mathbf{E}(k+1) \\ &= \cdots \\ &= W(k:0)\mathbf{X}(0) + \left(\sum_{t=1}^{k} W(k:t)\mathbf{E}(t)\right) + \mathbf{E}(k+1),\end{aligned} \tag{1.14}$$

where

$$W(k:t) = W(k)W(k-1)\cdots W(t+1)W(t) \qquad \text{for all } k \geq t \geq 0.$$

By multiplying both sides of (1.14) with the matrix $\frac{1}{m}\mathbf{1}\mathbf{1}'$, we have

$$\begin{aligned}&\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1) \\ &= \frac{1}{m}\mathbf{1}\mathbf{1}' W(k:0)\mathbf{X}(0) + \left(\sum_{t=1}^{k} \frac{1}{m}\mathbf{1}\mathbf{1}' W(k:t)\mathbf{E}(t)\right) + \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{E}(k+1) \\ &= \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(0) + \left(\sum_{t=1}^{k} \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{E}(t)\right) + \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{E}(k+1),\end{aligned}$$

where the last equality follows from the fact that the matrices $W(k:t)$ are column-stochastic, as inherited from the matrices $W(k)$ being column-stochastic. By subtracting the preceding relation from (1.14), we obtain

$$\begin{aligned}
&\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1) \\
&\quad = \left(W(k:0) - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)\mathbf{X}(0) + \sum_{t=1}^{k}\left(W(k:t) - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)\mathbf{E}(t) \\
&\quad\quad + \left(I - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)\mathbf{E}(k+1), \qquad (1.15)
\end{aligned}$$

where I is the identity matrix. Let $\|A\|_F$ denote the Frobenius norm of an $m \times n$ matrix A, i.e.,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

By taking the Frobenius norm of both sides in (1.15), we further obtain

$$\begin{aligned}
\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F &\le \left\|\left(W(k:0) - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)\mathbf{X}(0)\right\|_F \\
+ \left(\sum_{t=1}^{k}\left\|\left(W(k:t) - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)\mathbf{E}(t)\right\|_F\right) &+ \left\|\left(I - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)\mathbf{E}(k+1)\right\|_F.
\end{aligned}$$

Since the Frobenius norm is sub-multiplicative, i.e., $\|AB\|_F \le \|A\|_F\,\|B\|_F$, it follows that

$$\begin{aligned}
\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F &\le \left\|W(k:0) - \frac{1}{m}\mathbf{1}\mathbf{1}'\right\|_F \|\mathbf{X}(0)\|_F \\
+ \left(\sum_{t=1}^{k}\left\|W(k:t) - \frac{1}{m}\mathbf{1}\mathbf{1}'\right\|_F \|\mathbf{E}(t)\|_F\right) &+ \left\|I - \frac{1}{m}\mathbf{1}\mathbf{1}'\right\|_F \|\mathbf{E}(k+1)\|_F. \quad (1.16)
\end{aligned}$$

By Lemma 3 we have

$$\left([W(k:t)]_{ij} - \frac{1}{m}\right)^2 \le q^{k-t} \qquad \text{for all } k \ge t \ge 0, \text{ with } q = 1 - \frac{\eta}{2m^2}.$$

Hence,

$$\left\|W(k:t) - \frac{1}{m}\mathbf{1}\mathbf{1}'\right\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m}\left([W(k:t)]_{ij} - \frac{1}{m}\right)^2} \le m\sqrt{q^{k-t}}.$$

Since $q = 1 - \frac{\eta}{2m^2}$, by using the fact $\sqrt{1-\mu} \le 1 - \mu/2$ for any $\mu \in (0, 1)$, we see that for all $k \ge t \ge 0$,

$$\left\| W(k:t) - \frac{1}{m}\mathbf{1}\mathbf{1}' \right\|_F \le mp^{k-t} \qquad \text{with } \; p = 1 - \frac{\eta}{4m^2}. \tag{1.17}$$

For the norm $\left\| I - \frac{1}{m}\mathbf{1}\mathbf{1}' \right\|_F$ we have

$$\left\| I - \frac{1}{m}\mathbf{1}\mathbf{1}' \right\|_F = \sqrt{m\left(1 - \frac{1}{m}\right)^2 + (m-1)m\frac{1}{m^2}} = \sqrt{m-1}. \tag{1.18}$$

Using relations (1.17) and (1.18) in inequality (1.16), we obtain

$$\begin{aligned} \left\| \mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1) \right\|_F \le mp^k \|\mathbf{X}(0)\|_F + m\left(\sum_{t=1}^{k} p^{k-t}\|\mathbf{E}(t)\|_F\right) \\ + \sqrt{m-1}\|\mathbf{E}(k+1)\|_F. \end{aligned} \tag{1.19}$$

We next interpret relation (1.19) in terms of the iterates $x_i(k+1)$ and the vectors $\epsilon_i(k+1)$, as given in (1.12). Recalling that the ℓth column of $\mathbf{X}(k)$ consists of the vector $x^\ell(k)$, with the entries $x_i^\ell(k)$, $i \in [m]$, for all $\ell \in [n]$, we can see that

$$\mathbf{1}'\mathbf{X}(k) = \left(\sum_{i=1}^{m} x_i^1(k), \ldots, \sum_{i=1}^{m} x_i^m(k)\right).$$

Thus,

$$\frac{1}{m}\mathbf{1}'\mathbf{X}(k) = x_{av}'(k) \qquad \text{where } \; x_{av}(k) = \frac{1}{m}\sum_{j=1}^{m} x_j(k),$$

and

$$\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k) = \mathbf{1}x_{av}'(k) \qquad \text{for all } k.$$

Hence, $\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)$ is the matrix with all rows consisting of the vector $x_{av}'(k)$. Observing that the matrix $\mathbf{X}(k)$ has rows consisting of $x_1'(k), \ldots, x_m'(k)$, and using the definition of the Frobenius norm, we can see that

$$\left\| \mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k) \right\|_F = \sqrt{\sum_{i=1}^{m} \|x_i(k) - x_{av}(k)\|^2}.$$

Similarly, recalling that $\mathbf{E}(k)$ has rows consisting of $\epsilon_i'(k)$, $i \in [m]$, we also have

$$\|\mathbf{E}(k)\|_F = \sqrt{\sum_{i=1}^{m} \|\epsilon_i(k)\|^2}.$$

Therefore, relation (1.19) is equivalent to

$$\begin{aligned}
&\sqrt{\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2} \\
&\quad \leq mp^k \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + m \left(\sum_{t=1}^{k} p^{k-t} \sqrt{\sum_{i=1}^{m} \|\epsilon_i(t)\|^2} \right) \\
&\quad + \sqrt{m-1} \sqrt{\sum_{i=1}^{m} \|\epsilon_i(k+1)\|^2}.
\end{aligned}$$

□

1.3.1.3 Basic Relation for Agents' Iterates

Recall that each $\epsilon_i(k+1)$ represents the difference between the projection point $\Pi_X[y_i(k+1) - \alpha_{k+1} g_i(k+1)]$ and the point $y_i(k+1)$ (see (1.12)). Thus, there is a structure in $\epsilon_i(k+1)$ that can be further exploited. In particular, we can further refine the result of Lemma 5, under the assumption of bounded subgradients $g_i(k+1)$, as given in the following lemma.

Lemma 6 *Let the problem be convex (i.e., Assumption 1 holds). Also, assume that the subgradients of f_i are bounded over the set X for all i, i.e., there exists a constant C such that*

$$\|s\| \leq C \qquad \text{for every subgradient } s \text{ of } f_i(z) \text{ at any } z \in X.$$

Furthermore, let Assumptions 2 and 3 hold for the graphs G_k and the matrices $W(k)$, respectively. Then, for the iterates $x_i(k)$ of the method (1.9) and their averages $x_{av}(k) = \frac{1}{m}\sum_{j=1}^{m} x_j(k)$, we have for all $i \in [m]$ and $k \geq 0$,

$$\begin{aligned}
\sqrt{\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2} &\leq mp^k \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} \\
&\quad + m\sqrt{m}C \sum_{t=1}^{k} p^{k-t} \alpha_t + mC\alpha_{k+1},
\end{aligned}$$

where $p = 1 - \frac{\eta}{4m^2}$.

Proof By Lemma 5 we have for all $k \geq 0$,

$$\begin{aligned}\sqrt{\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2} &\leq mp^k \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + m \left(\sum_{t=1}^{k} p^{k-t} \sqrt{\sum_{i=1}^{m} \|\epsilon_i(t)\|^2} \right) \\ &\quad + \sqrt{m-1} \sqrt{\sum_{i=1}^{m} \|\epsilon_i(k+1)\|^2}. \end{aligned} \tag{1.20}$$

Since $y_i(k+1)$ is a convex combination of points $x_j(k+1) \in X$, $j \in [m]$, by the convexity of the set X it follows that $y_i(k+1) \in X$ for all i, implying that for all $k \geq 0$,

$$\|\Pi_X[y_i(k+1) - \alpha_{k+1} g_i(k+1)] - y_i(k+1)\| \leq \alpha_{k+1} \|g_i(k+1)\| \leq \alpha_{k+1} C.$$

Therefore, for all i and $k \geq 0$,

$$\|\epsilon_i(k+1)\|^2 \leq \alpha_{k+1}^2 C^2,$$

implying that

$$\sum_{i=1}^{m} \|\epsilon_i(k+1)\|^2 \leq m\alpha_{k+1}^2 C^2 \qquad \text{for all } k \geq 0.$$

By substituting the preceding estimate in (1.19), we obtain

$$\begin{aligned}\sqrt{\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2} &\leq mp^k \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + m \left(\sum_{t=1}^{k} p^{k-t} \sqrt{m\alpha_t^2 C^2} \right) \\ &\quad + \sqrt{m-1}\sqrt{m\alpha_{k+1}^2 C^2} \\ &= mp^k \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + m\sqrt{m}C \left(\sum_{t=1}^{k} p^{k-t} \alpha_t \right) \\ &\quad + \sqrt{m-1}\sqrt{m} C \alpha_{k+1}. \end{aligned}$$

The desired relation follows by using $m - 1 \leq m$. □

We will now put together Lemmas 4 and 6 to provide a key result for establishing the convergence of the method. We assume some conditions on the stepsize, some of which are often used when analyzing the behavior of a subgradient algorithm.

Proposition 1 *Let Assumptions 1–3 hold. Assume that the subgradients of f_i are uniformly bounded over the set X for all i, i.e., there exists a constant C such that*

$$\|s\| \leq C \qquad \text{for every subgradient } s \text{ of } f_i(z) \text{ at any } z \in X.$$

Also, let the stepsize satisfy the following conditions

$$\alpha_{k+1} \leq \alpha_k \qquad \text{for all } k \geq 1, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Then, for the iterates $x_i(k)$ of the method (1.9), we have for all $k \geq 0$ and all $x \in X$,

$$\sum_{i=1}^{m} \|x_i(k+1) - x\|^2 \leq \sum_{j=1}^{m} \|x_j(k) - x\|^2 - 2\alpha_{k+1} \left(f(x_{av}(k)) - f(x)\right) + s_k,$$

where $x_{av}(k) = \frac{1}{m} \sum_{j=1}^{m} x_j(k)$, while s_k is given by

$$s_k = 2\alpha_{k+1} C \sqrt{m} \sqrt{\sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2} + m\alpha_{k+1}^2 C^2,$$

and it satisfies

$$\sum_{k=0}^{\infty} s_k < \infty.$$

Proof By Lemma 4, we have for all i, all $k \geq 0$ and all $x \in X$,

$$\|x_i(k+1) - x\|^2 \leq \|y_i(k+1) - x\|^2 - 2\alpha_{k+1} \left(f_i(y_i(k+1)) - f_i(x)\right) + \alpha_{k+1}^2 \|g_i(k+1)\|^2.$$

Since $y_i(k+1)$ is a convex combination of the points $x_j(k)$, $j \in [m]$, by the convexity of the norm squared, it follows that

$$\|x_i(k+1) - x\|^2 \leq \sum_{j=1}^{m} w_{ij}(k) \|x_j(k) - x\|^2 - 2\alpha_{k+1} \left(f_i(y_i(k+1)) - f_i(x)\right) + \alpha_{k+1}^2 \|g_i(k+1)\|^2.$$

By summing these relations over i and by using the subgradient-boundedness property, we obtain

$$\sum_{i=1}^{m} \|x_i(k+1) - x\|^2 \leq \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}(k)\|x_j(k) - x\|^2 \\ -2\alpha_{k+1} \sum_{i=1}^{m} (f_i(y_i(k+1)) - f_i(x)) \\ +m\alpha_{k+1}^2 C^2.$$

By exchanging the order of summation in the double-sum term, we see that

$$\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}(k)\|x_j(k) - x\|^2 = \sum_{j=1}^{m} \|x_j(k) - x\|^2 \sum_{i=1}^{m} w_{ij}(k) = \sum_{j=1}^{m} \|x_j(k) - x\|^2,$$

where the last equality follows from $\mathbf{1}'W(k) = \mathbf{1}'$. Therefore,

$$\sum_{i=1}^{m} \|x_i(k+1) - x\|^2 \leq \sum_{j=1}^{m} \|x_j(k) - x\|^2 - 2\alpha_{k+1} \\ \sum_{i=1}^{m} (f_i(y_i(k+1)) - f_i(x)) + m\alpha_{k+1}^2 C^2.$$

We next estimate $f_i(y_i(k+1)) - f_i(x)$ by using the average vector $x_{av}(k)$, as follows:

$$f_i(y_i(k+1)) - f_i(x) = f_i(y_i(k+1)) - f_i(x_{av}(k)) + f_i(x_{av}(k)) - f_i(x) \\ \geq -C\|y_i(k+1) - x_{av}(k)\| + f_i(x_{av}(k)) - f_i(x),$$

where the inequality follows by the Lipschitz continuity of f_i (due to the uniform subgradient-boundedness property on the set X and the fact that $y_i(k+1) \in X$ and $x_{av}(k) \in X$). By combining the preceding two relations and using $f = \sum_{i=1}^{m} f_i$, we have for all $k \geq 0$ and all $x \in X$,

$$\sum_{i=1}^{m} \|x_i(k+1) - x\|^2 \leq \sum_{j=1}^{m} \|x_j(k) - x\|^2 - 2\alpha_{k+1} (f(x_{av}(k)) - f(x)) \\ + 2\alpha_{k+1} C \sum_{i=1}^{m} \|y_i(k+1) - x_{av}(k)\| + m\alpha_{k+1}^2 C^2. \quad (1.21)$$

Consider now the vectors $y_i(k+1)$ and note that by the definition of $y_i(k+1)$, we have for any $y \in \mathbb{R}^n$,

$$\sum_{i=1}^{m} \|y_i(k+1) - y\| = \sum_{i=1}^{m} \left\| \sum_{j=1}^{m} w_{ij}(k)(x_j(k) - y) \right\|,$$

where we use $W(k)\mathbf{1} = \mathbf{1}$. By the convexity of the norm, it follows that

$$\sum_{i=1}^{m} \|y_i(k+1) - y\| \leq \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij}(k)\|x_j(k) - y\| = \sum_{j=1}^{m} \|x_j(k) - y\|,$$

where the last equality is obtained by exchanging the order of summation and using $\mathbf{1}'W(k) = \mathbf{1}'$. Hence, for $y = x_{av}(k)$, we obtain

$$\sum_{i=1}^{m} \|y_i(k+1) - x_{av}(k)\| \leq \sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\| \leq \sqrt{m} \sqrt{\sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2},$$

where the last inequality follows by Hölder's inequality. By substituting the preceding estimate in relation (1.21), we have for all $k \geq 0$ and all $x \in X$,

$$\sum_{i=1}^{m} \|x_i(k+1) - x\|^2 \leq \sum_{j=1}^{m} \|x_j(k) - x\|^2 - 2\alpha_{k+1} \left(f(x_{av}(k)) - f(x)\right) + 2\alpha_{k+1} C\sqrt{m} \sqrt{\sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2} + m\alpha_{k+1}^2 C^2.$$

To simplify the notation, we let for all $k \geq 0$,

$$s_k = 2\alpha_{k+1} C\sqrt{m} \sqrt{\sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2} + m\alpha_{k+1}^2 C^2, \tag{1.22}$$

so that we have for all $k \geq 0$ and all $x \in X$,

$$\sum_{i=1}^{m} \|x_i(k+1) - x\|^2 \leq \sum_{j=1}^{m} \|x_j(k) - x\|^2 - 2\alpha_{k+1} \left(f(x_{av}(k)) - f(x)\right) + s_k.$$

We next show that the terms $\alpha_{k+1}\sqrt{\sum_{j=1}^m \|x_j(k) - x_{av}(k)\|^2}$ involved in the definition of s_k are summable over k. According to Lemma 6 we have for all $k \geq 1$,

$$\sqrt{\sum_{i=1}^m \|x_i(k) - x_{av}(k)\|^2} \leq mp^{k-1}\sqrt{\sum_{i=1}^m \|x_i(0)\|^2} + m\sqrt{m}C\sum_{t=1}^{k-1} p^{k-t-1}\alpha_t + mC\alpha_k.$$

Letting

$$r_k = \alpha_{k+1}\sqrt{\sum_{j=1}^m \|x_j(k) - x_{av}(k)\|^2}, \tag{1.23}$$

and using the assumption that the stepsize α_k is non-increasing, we see that

$$r_k \leq mp^{k-1}\alpha_1\sqrt{\sum_{i=1}^m \|x_i(0)\|^2} + m\sqrt{m}C\sum_{t=1}^{k-1} p^{k-t-1}\alpha_t^2 + mC\alpha_k^2.$$

By summing r_k over $k = 2, 3, \ldots, K$, for some $K \geq 2$, we have

$$\begin{aligned}\sum_{k=2}^K r_k &\leq m\left(\sum_{k=1}^K p^{k-1}\right)\alpha_1\sqrt{\sum_{i=1}^m \|x_i(0)\|^2} \\ &\quad + m\sqrt{m}C\sum_{k=2}^K\sum_{t=1}^{k-1} p^{k-t-1}\alpha_t^2 + mC\sum_{k=2}^K \alpha_k^2 \\ &< \frac{m}{1-p}\alpha_1\sqrt{\sum_{i=1}^m \|x_i(0)\|^2} + m\sqrt{m}C\sum_{s=1}^{K-1}\sum_{t=1}^{s} p^{s-t}\alpha_t^2 + mC\sum_{k=2}^K \alpha_k^2,\end{aligned}$$

where we use the fact that $p \in (0, 1)$ and we shift the indices in the double-sum term. Furthermore, by exchanging the order of summation, we see that

$$\sum_{s=1}^{K-1}\sum_{t=1}^{s} p^{s-t}\alpha_t^2 = \sum_{t=1}^{K-1}\alpha_t^2\sum_{s=t}^{K-1} p^{s-t} < \sum_{t=1}^{K-1}\alpha_t^2\frac{1}{1-p}.$$

Therefore,

$$\sum_{k=2}^{K} r_k < \frac{m}{1-p}\alpha_1 \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + \frac{m\sqrt{m}C}{1-p} \sum_{t=1}^{K-1} \alpha_t^2 + mC \sum_{k=2}^{K} \alpha_k^2.$$

In view of the assumption that $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, it follows that $\sum_{k=2}^{\infty} r_k < \infty$. Since

$$s_k = 2C\sqrt{m} r_k + m\alpha_{k+1}^2 C^2$$

(see (1.22) and (1.23)), it follows that

$$\sum_{k=0}^{\infty} s_k < \infty.$$

□

1.3.1.4 Convergence Result for Agents' Iterates

Using Proposition 1, we establish a convergence result for the iterates $x_i(k)$, as given in the following theorem.

Theorem 1 *Let Assumptions 1–3 hold. Assume that there is a constant C such that*

$$\|s\| \leq C \qquad \text{for every subgradient } s \text{ of } f_i(z) \text{ at any } z \in X.$$

Let the stepsize satisfy the following conditions

$$\alpha_{k+1} \leq \alpha_k \qquad \text{for all } k \geq 1, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty,$$

and assume that problem (1.1) has a solution. Then, the iterate sequences $\{x_i(k)\}$, $i \in [m]$, generated by the method (1.9), converge to an optimal solution of problem (1.1), i.e.,

$$\lim_{k\to\infty} \|x_i(k) - x^*\| = 0 \qquad \text{for all } i \in [m] \text{ and some } x^* \in X^*.$$

Proof By letting $x = x^*$ in Proposition 1, for an arbitrary $x^* \in X^*$, we obtain for all $i \in [m]$ and $k \geq 0$,

$$\sum_{i=1}^{m} \|x_i(k+1) - x^*\|^2 \leq \sum_{j=1}^{m} \|x_j(k) - x^*\|^2 - 2\alpha_{k+1}\big(f(x_{av}(k)) - f(x^*)\big) + s_k,$$

with $s_k > 0$ satisfying $\sum_{k=0}^{\infty} s_k < \infty$. By summing these relations over $k = K, K+1, \ldots, T$ for any $T \geq K \geq 0$, after re-arranging the terms, we further obtain for all $x^* \in X^*$ and all $T \geq K \geq 0$,

$$\sum_{i=1}^{m} \|x_i(T+1) - x^*\|^2 + 2 \sum_{k=K}^{T} \alpha_{k+1} \left(f(x_{av}(k)) - f(x^*) \right)$$
$$\leq \sum_{j=1}^{m} \|x_j(K) - x^*\|^2 + \sum_{k=K}^{T} s_k. \tag{1.24}$$

Note that $f(x_{av}(k)) - f(x^*) > 0$ since $x_{av}(k) \in X$. Thus, the preceding relation implies that the sequences $\{x_i(k)\}$, $i \in [m]$, are bounded and, also, that

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

since $\sum_{k=0}^{\infty} s_k < \infty$, where $f^* = f(x^*)$ for any $x^* \in X^*$. Thus, it follows that

$$\liminf_{k \to \infty} \left(f(x_{av}(k)) - f^* \right) = 0.$$

Let $\{k_\ell\}$ be a sequence of indices that attains the above limit inferior, i.e.,

$$\lim_{\ell \to \infty} f(x_{av}(k_\ell)) = f^*. \tag{1.25}$$

Since the sequences $\{x_i(k)\}$, $i \in [m]$, are bounded, so is the average sequence $\{x_{av}(k)\}$. Hence, $\{x_{av}(k_\ell)\}$ contains a converging subsequence. Without loss of generality, we may assume that $\{x_{av}(k_\ell)\}$ converges to some point $\hat{x}$, i.e.,

$$\lim_{\ell \to \infty} x_{av}(k_\ell) = \hat{x}.$$

Note that $\hat{x} \in X$ since $\{x_{av}(k)\} \subset X$ and the set X is assumed to be closed. Note further that f is continuous on $\mathbb{R}^n$ since it is convex on $\mathbb{R}^n$. Hence, we have

$$\lim_{\ell \to \infty} f(x_{av}(k_\ell)) = f(\hat{x}) \qquad \text{with} \qquad \hat{x} \in X,$$

which together with relation (1.25) yields $f(\hat{x}) = f^*$. Therefore, $\hat{x}$ is an optimal point.

Next, we show that $\{x_i(k_\ell)\}$ converges to $\hat{x}$ for all i. By Lemma 6 we have for all $k \geq 0$ and all $i \in [m]$,

$$\sqrt{\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2} \leq mp^k \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + m\sqrt{m}C \sum_{t=1}^{k} p^{k-t}\alpha_t + mC\alpha_{k+1}.$$

Letting $k = k_\ell - 1$ for any $k_\ell \geq 1$, we see that for all $i \in [m]$,

$$\sqrt{\sum_{i=1}^{m} \|x_i(k_\ell) - x_{av}(k_\ell)\|^2} \leq mp^{k_\ell - 1} \sqrt{\sum_{i=1}^{m} \|x_i(0)\|^2} + m\sqrt{m}C \sum_{t=1}^{k_\ell - 1} p^{k_\ell - 1 - t}\alpha_t + mC\alpha_{k_\ell}.$$

Since $p \in (0, 1)$ and $\alpha_k \to 0$ (due to $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$), it follows that

$$\limsup_{\ell \to \infty} \sqrt{\sum_{i=1}^{m} \|x_i(k_\ell) - x_{av}(k_\ell)\|^2} \leq m\sqrt{m}C \limsup_{\ell \to \infty} \sum_{t=1}^{k_\ell - 1} p^{k_\ell - 1 - t}\alpha_t.$$

We note that

$$\begin{aligned}
\limsup_{\ell \to \infty} \sum_{t=1}^{k_\ell - 1} p^{k_\ell - 1 - t}\alpha_t &= \lim_{k \to \infty} \sum_{t=1}^{k-1} p^{k-1-t}\alpha_t \\
&= \lim_{k \to \infty} \left(\left(\sum_{\tau=1}^{k-1} p^{k-1-\tau} \right) \frac{1}{\sum_{\tau=1}^{k-1} p^{k-1-\tau}} \sum_{t=1}^{k-1} p^{k-1-t}\alpha_t \right) \\
&= \lim_{k \to \infty} \left(\sum_{\tau=1}^{k-1} p^{k-1-\tau} \right) \left(\lim_{k \to \infty} \frac{1}{\sum_{\tau=1}^{k-1} p^{k-1-\tau}} \sum_{t=1}^{k-1} p^{k-1-t}\alpha_t \right) \\
&= \frac{1}{1-p} \lim_{t \to \infty} \alpha_t, \qquad (1.26)
\end{aligned}$$

where in the last equality we use that fact that any convex combination of a convergent sequence $\{\alpha_k\}$ converges to the same limit as the sequence itself. Hence, we have

$$\limsup_{\ell \to \infty} \sum_{t=1}^{k_\ell - 1} p^{k_\ell - 1 - t}\alpha_t = 0,$$

implying that

$$\limsup_{\ell\to\infty} \sqrt{\sum_{i=1}^{m} \|x_i(k_\ell) - x_{av}(k_\ell)\|^2} = 0.$$

Therefore, since $\lim_{\ell\to\infty} x_{av}(k_\ell) = \hat{x}$, it follows that

$$\lim_{\ell\to\infty} x_i(k_\ell) = \hat{x} \quad \text{for all } i \in [m], \qquad \text{with} \quad \hat{x} \in X^*. \tag{1.27}$$

Now, since $\hat{x} \in X^*$, we let $x^* = \hat{x}$ in (1.24). Then, we let $K = k_\ell$ in (1.24) and by omitting the term involving the function values, from (1.24) we obtain for all $\ell \geq 1$,

$$\limsup_{T\to\infty} \sum_{i=1}^{m} \|x_i(T+1) - \hat{x}\|^2 \leq \sum_{j=1}^{m} \|x_j(k_\ell) - \hat{x}\|^2 + \sum_{k=k_\ell}^{\infty} s_k.$$

Letting $\ell \to \infty$, and using relation (1.27), we see that

$$\limsup_{T\to\infty} \sum_{i=1}^{m} \|x_i(T+1) - \hat{x}\|^2 \leq \lim_{\ell\to\infty} \sum_{k=k_\ell}^{\infty} s_k = 0,$$

where $\lim_{\ell\to\infty} \sum_{k=k_\ell}^{\infty} s_k = 0$ holds since $\sum_{k=0}^{\infty} s_k < \infty$. Thus, it follows that for $\hat{x} \in X^*$,

$$\lim_{k\to\infty} \|x_i(k) - \hat{x}\| = 0 \qquad \text{for all } i \in [m].$$

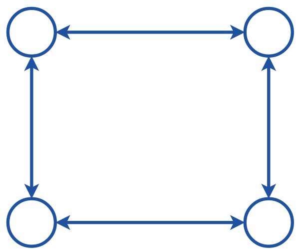
□

1.3.2 Numerical Examples

Here, we show some numerical results obtained for a variant of the algorithm in (1.9) as applied to a data classification problem from Example 3. We will consider an extension of that problem to the case when the data set is not perfectly separable. In this case, there is an additional slack variable u that enters the model, and the distributed version of the problem assumes the following form:

$$\min_{(x,u)\in\mathbb{R}^n\times\mathbb{R}} f(x,u) \qquad f(x,u) = \sum_{i=1}^{m} f_i(x,u),$$

Fig. 1.5 An undirected network with four nodes



where each $f_i : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is given by:

$$f_i(x, u) = \frac{\rho}{2m}\|x\|^2 + \sum_{j \in D_i} \max\{0, 1 - y_j\langle x, z_j\rangle + u\},$$

where D_i is the collection of the data points at the center i (agent i).

Letting $\mathbf{x} = (x, u) \in \mathbb{R}^n \times \mathbb{R}$, we consider the following distributed algorithm[3] for this problem over a static graph $G = ([m], E)$:

$$\mathbf{y}_i(k+1) = \mathbf{x}_i(k) - \eta_{k+1}\sum_{j=1}^{m} r_{ij}\mathbf{x}_j(k) \qquad (r_{ij} = 0 \text{ when } i \leftrightarrow j \notin E),$$

$$\mathbf{x}_i(k+1) = \mathbf{y}_i(k+1) - \alpha_{k+1}g_i(k+1), \tag{1.28}$$

where $g_i(k+1)$ is a subgradient of f_i at $\mathbf{y}_i(k+1)$. We note that the weights used in the update of $\mathbf{y}_i(k+1)$ are different from the weights used in (1.9). The weights here are based on a Laplacian formulation of the consensus problem, which include another parameter $\eta_{k+1} > 0$. This parameter can be viewed as an additional stepsize that is associated with the feasibility step for the consensus constraints. Under some (boundedness) conditions on η_{k+1} and standard conditions on the stepsize (akin to those in Theorem 1), the method converges to a solution of the problem [144].

We illustrate the behavior of the method in (1.28) for the case of a network with four nodes organized in a ring, as depicted in Fig. 1.5. The simulations are generated for the regularization parameter $\rho = 6$. The stepsize values used in the experiment are: $\eta_k = 0.8$ and $\alpha_k = \frac{1}{k}$, for all $k \geq 1$. The behavior of the method is depicted in Fig. 1.6 where the resulting hyperplanes produced by agents are shown after 20 and after 500 iterations. The plots also show the true separating hyperplane that solves the centralized problem.

The algorithm (1.28) assumes perfect communication links, which is not typically the case in wireless networks. To capture the effect of communication noise,

[3] See [144, 146] for more details.

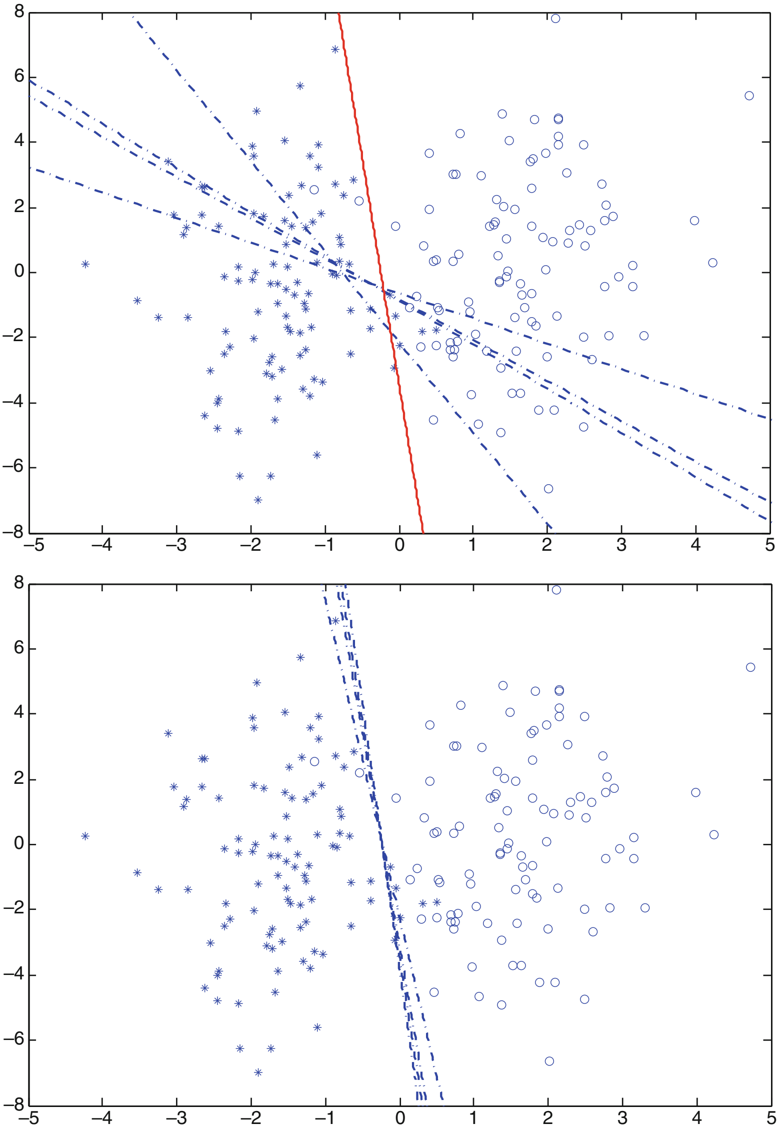


Fig. 1.6 The top plot shows the agents' iterates $\mathbf{x}_i(k)$ after 20 iterations and the true solution (the hyperplane in red color), while the bottom plot shows the agents' iterates after 500 iterations

consider the following variant of the method:

$$\mathbf{y}_i(k+1) = \mathbf{x}_i(k) - \eta_{k+1} \sum_{j=1}^{m} r_{ij}(\mathbf{x}_j(k) + \xi_{ij}(k)) \quad (r_{ij} = 0 \text{ when } i \leftrightarrow j \notin E),$$

$$\mathbf{x}_i(k+1) = \mathbf{y}_i(k+1) - \alpha_{k+1} g_i(k+1), \tag{1.29}$$

where $\xi_{ij}(k)$ is a link dependent noise associated with messages received by agent i from a neighbor j. The parameter $\eta_{k+1} > 0$ can be viewed as a noise-damping stepsize. In this case, for the method to converge to a solution of the problem, the noise-damping stepsize η_k has to be coordinated with the subgradient-related stepsize α_k. In particular, the following conditions are imposed:

$$\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \eta_k = \infty,$$

$$\sum_{k=1}^{\infty} \eta_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k \eta_k < \infty, \quad \sum_{k=1}^{\infty} \frac{\alpha_k^2}{\eta_k} < \infty.$$

The simulations are performed for a different set of data and the ring graph shown in Fig. 1.5, where the links were assumed to be noisy. The noise is modeled by the i.i.d. zero mean Gaussian process with the variance equal to 1. The regularization parameter is $\rho = 6$, while the noise-damping parameter and the stepsize are $\eta_k = \frac{1}{k^{0.55}}$ and $\alpha_k = \frac{1}{k}$ for all $k \geq 1$. The results for one of the typical simulation run are shown in Fig. 1.7. These simulation results are taken from [144], where more simulation results can be found.

1.4 Distributed Asynchronous Algorithms for Static Undirected Graphs

There are several drawbacks of synchronous updates that limit their applications, including

- All agents have to update *at the same time*. Imagine that each agent has its own clock and it updates at each tick of its clock. Then, the requirement that the agents update synchronously, as in method (1.9), means that the agents must have their local clocks perfectly synchronized throughout the computation task. This is hard to ensure in practice for some networks, such as wireless networks where communication interference is an issue.
- The communication links in the connectivity graphs G_k have to be *perfectly activated* to transmit and receive information. Some communication protocols

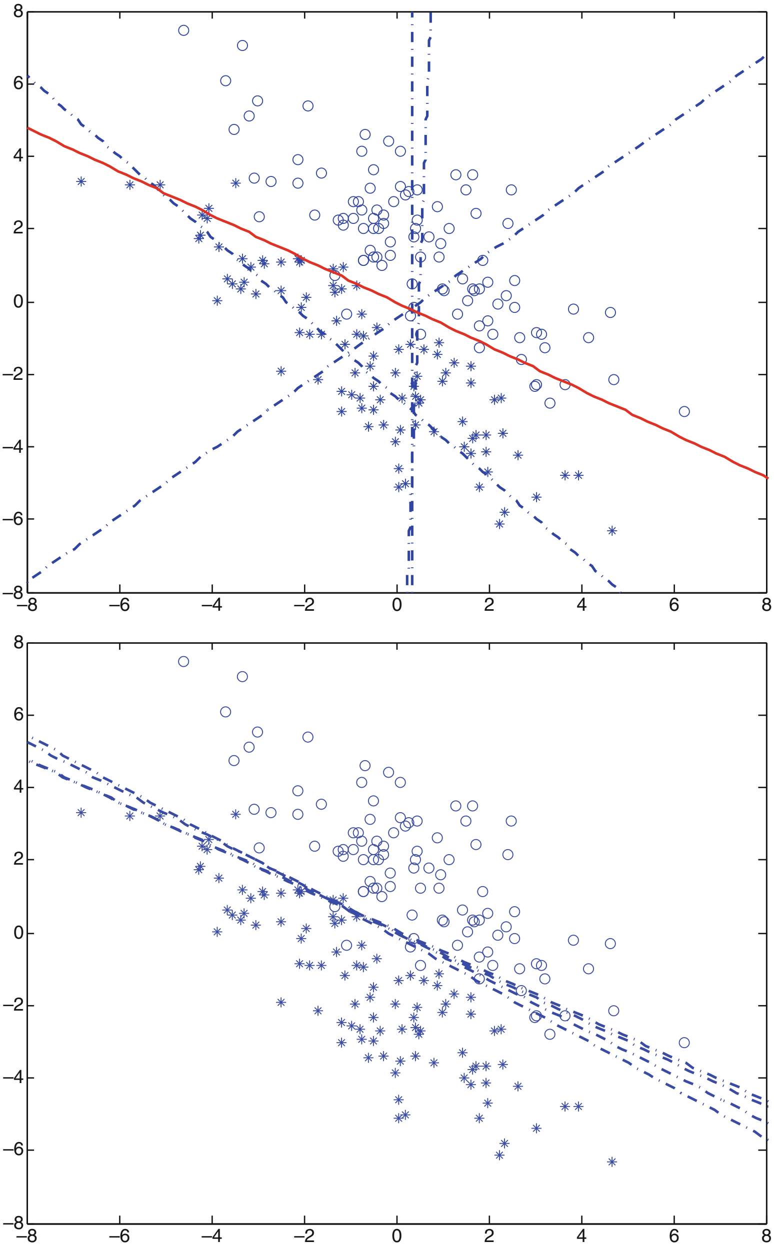


Fig. 1.7 The top plot shows the agents' iterates $\mathbf{x}_i(k)$ after the first iteration and the true solution (the hyperplane in red color), while the bottom plot shows the agents' iterates after 500 iterations

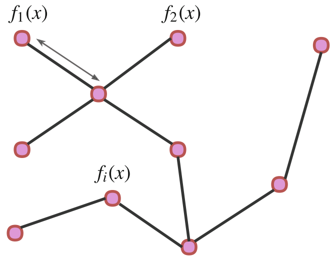


Fig. 1.8 The agent communication graph is static, undirected and connected. However, not all the links will necessarily be active at each time instance. A function f_i is a local objective of agent i

require receiving message acknowledgements (“acks"), which can lead to deadlocks when the links are not perfect.

- Communications can be *costly* (consume power) and it may not be efficient for agents to be activated too frequently to communicate, due to their limited power supply for example.

In order to alleviate these drawbacks of simultaneous updates, one possibility is to randomize the activation of communication links in the network or the activation of the agents. We will discuss two such random activations: gossip and broadcast protocols. Gossip could be viewed as a random link activation process, while broadcast is a random-agent activation process. We will here treat both as a random-agent activation, by assuming that the agents are equipped with local clocks that tick according to the same rate (the same inter-click time), but do not click synchronously.

Throughout this section, we assume that the underlying communication graph is static and undirected, denoted by $G = ([m], E)$, see Fig. 1.8 for an illustration of the graph. The randomization we will use to develop asynchronous algorithms will have some stepsizes that can be easily analyzed for a static graph. While their extensions to time-varying graphs may be possible, we will not consider them here.

To develop asynchronous algorithms for solving problem (1.1), we will use the asynchronous methods for consensus. Thus, we will at first discuss random asynchronous consensus methods in Sect. 1.4.1 and, then, we give the corresponding asynchronous optimization methods in Sect. 1.4.2.

1.4.1 Random Gossip and Random Broadcast for Consensus

Both random gossip and random broadcast algorithms can be used to achieve a consensus. These two approaches share in common the mechanism that triggers the update events, but they differ in the update rule specifications.

More concretely, these two approaches use the same random process to wake up an agent that will initiate a communication event that includes an iterate update. However, in the random gossip algorithm the agent that wakes up contacts one of its neighbors at random, thus, randomly activating a single undirected link for communication, and both the agent and the selected neighbor perform updates. Unlike this, in the random broadcast approach, the agent that wakes up broadcasts its information to all of its neighbors, thus resulting in using directed communication links (even-though the links are undirected). Moreover, upon the broadcast, the agent that triggered the communication event goes to an inactive mode and only its neighbors perform updates.

Let us now describe the random process that triggers the communication events for both gossip and broadcast models. Each agent has its own local Poisson clock that ticks with a rate equal to 1 (the rate can take any other positive value as long as all agents have the same rate). Each agent's clock ticks independently of the other agents' clocks. At a tick of its clock, an agent wakes up and initiates a communication event. The ticks of the local agents' clocks can be modeled as single virtual Poisson clock that ticks with a rate m. Letting $\{Z_k\}$ be the Poisson process of the virtual clock tick-times, we discretize the time according to $\{Z_k\}$ since these are the only times when a change will occur in some of the agent values $x_i(k)$. The inter-tick times $\{Z_{k+1} - Z_k\}$ are i.i.d. exponentially distributed with the rate m.

1.4.1.1 Gossip-Based Consensus Algorithm

In the gossip-model, the randomly activated agent wakes up and selects randomly one of its neighbors, as depicted in Fig. 1.9. The activated agent and its selected neighbor exchange their information and perform an iterate update. Upon the update, both agents go to sleep.

To formalize the process, we let I_k be the index of the agent that is activated at time Z_k, i.e., the agent whose clock ticks at time Z_k. The variables $\{I_k\}$ are i.i.d. with a uniform distribution over $\{1, \ldots, m\}$, i.e.,

$$\text{Prob}\{I_k = i\} = \frac{1}{m} \qquad \text{for all } i \in [m].$$

For agent i, let $p_{ij} > 0$ be the probability of contacting its neighbor $j \in N_i$, $j \neq i$. Let J_k be the index of a neighbor of agent I_k that is selected randomly for communication at time Z_k. Let $P = [p_{ij}]$ be the matrix of contact probabilities, where $p_{ij} = 0$ if $j \notin N_i$, and note that P is row-stochastic.

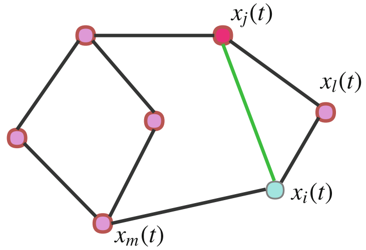


Fig. 1.9 Random gossip communication protocol: an agent that wakes up establishes a connection with a randomly selected neighbor. Thus, a random link is activated (shown in green color)

At time k, the active agents I_k and J_k exchange their current values $x_{I_k}(k)$ and $x_{J_k}(k)$, and then both update as follows:

$$x_{I_k}(k+1) = \frac{1}{2}\left(x_{I_k}(k) + x_{J_k}(k)\right), \qquad x_{J_k}(k+1) = \frac{1}{2}\left(x_{J_k}(k) + x_{I_k}(k)\right), \tag{1.30}$$

while the other agents do nothing (they sleep),

$$x_i(k+1) = x_i(k) \qquad \text{for all } i \notin \{I_k, J_k\}.$$

A value other than $1/2$ can be used in the updates in (1.30); however, we will work with $1/2$.

Assuming that the agent values $x_i(k)$ are scalars, the gossip iteration update can be compactly written, as follows:

$$x(k+1) = W_g(k)x(k) \qquad \text{for all } k \geq 0, \tag{1.31}$$

where $x(k)$ is a vector with components $x_i(k)$, $i \in [m]$, and the matrix $W_g(k)$ is symmetric with the entries given by

$$[W_g(k)]_{I_k,J_k} = [W_g(k)]_{J_k,I_k} = \frac{1}{2}, \qquad [W_g(k)]_{I_k,I_k} = \frac{1}{2}, \qquad [W_g(k)]_{J_k,J_k} = \frac{1}{2},$$

$$[W_g(k)]_{ii} = 1 \quad \text{for all } i \in [m] \setminus \{I_k, J_k\}, \qquad \text{and else} \quad [W_g(k)]_{ij} = 0.$$

Equivalently, the random matrix $W_g(k)$ is given by

$$W_g(k) = W^{(I_k J_k)}, \qquad \text{with} \quad W^{(ij)} = I - \frac{1}{2}(e_i - e_j)(e_i - e_j)' \quad \text{for all } i, j \in [m],$$

where e_i is the unit vector with its ith entry equal to 1 and the other entries equal to 0. Thus, the random matrix $W_g(k)$ takes values $W_g(k) = W^{(ij)}$ with the probability $(p_{ij} + p_{ji})/m$.

Every realization of $W_g(k)$ is a symmetric and stochastic matrix, hence, $W_g(k)$ is doubly stochastic. Furthermore, it can be seen that every realization $W^{(ij)}$ of $W_g(k)$ is a projection matrix[4] on the sub-space

$$S_{ij} = \{x \in \mathbb{R}^m \mid x_i = x_j\}.$$

Therefore, we have

$$W_g'(k)W_g(k) = W_g^2(k) = W_g(k) \qquad \text{for all } k \geq 0.$$

[4]A matrix A is a projection matrix if $A^2 = A$.

The convergence of the gossip algorithm has been shown in [16]. In the following theorem, we provide a statement based on the result in [16] that is relevant to our subsequent discussion on distributed asynchronous methods.

Theorem 2 *([16]) Assume that the graph G is connected. Then, the iterate sequences $\{x_i(k)\}$, $i \in [m]$, produced by the gossip algorithm (1.31) satisfy the following relation*

$$\mathbb{E}\left[\left\| x(k) - \frac{1}{m}\sum_{j=1}^{m} x_j(0)\,\mathbf{1}\right\|^2\right] \le \lambda_g^k \left\| x(0) - \frac{1}{m}\sum_{j=1}^{m} x_j(0)\,\mathbf{1}\right\|^2 \qquad \text{for all } k \ge 0,$$

where $0 < \lambda_g < 1$ is the second largest eigenvalue of $\bar{W}_g = \mathbb{E}\left[W_g(k)\right]$.

Proof Defining

$$z(k) = x(k) - \frac{1}{m}\sum_{j=1}^{m} x_j(0)\,\mathbf{1},$$

and using the gossip updates in (1.31), the following relation has been shown in [16]:

$$\mathbb{E}\left[\|z(k+1)\|^2 \mid z(k)\right] = \langle z(k), \mathbb{E}\left[W_g'(k) W_g(k)\right] z(k)\rangle$$

(see equation (14) in [16]). Since $W_g'(k) W_g(k) = W_g(k)$ and since $\{W_g(k)\}$ is an i.i.d. matrix sequence, by letting $\mathbb{E}\left[W_g(k)\right] = \bar{W}_g$, we obtain

$$\mathbb{E}\left[\|z(k+1)\|^2 \mid z(k)\right] = \langle z(k), \bar{W}_g z(k)\rangle.$$

The matrix $\bar{W}_g$ is symmetric and doubly stochastic, since each realization $W^{(ij)}$ of any $W_g(k)$ is symmetric and doubly stochastic. Furthermore, since each realization $W^{(ij)}$ is a projection matrix, each $W^{(ij)}$ is positive semi-definite. Hence, $\bar{W}_g$ is also positive semi-definite and, consequently, all eigenvalues of $\bar{W}_g$ are non-negative. Moreover, we have

$$[\bar{W}_g]_{ii} > 0 \qquad \text{for all } i \in [m],$$

and for all $i \neq j$,

$$[\bar{W}_g]_{ij} > 0 \quad \Longleftrightarrow \quad i \leftrightarrow j \in E.$$

Since the graph G is connected, $\bar{W}_g$ is irreducible and by Theorem 4.3.1, page 106 in [46], the matrix $\bar{W}_g$ has 1 as the largest eigenvalue of multiplicity 1, with the associated eigenvector $\mathbf{1}$. Since $z(k) \perp \mathbf{1}$, it follows that

$$\langle z(k), \bar{W}_g z(k)\rangle \leq \lambda_g \|z(k)\|^2,$$

where $0 < \lambda_g < 1$ is the second largest eigenvalue of $\bar{W}_g$.

Therefore, we have

$$\mathbb{E}\left[\|z(k+1)\|^2 \mid z(k)\right] \leq \lambda_g \|z(k)\|^2 \qquad \text{for all } k \geq 0,$$

implying that

$$\mathbb{E}\left[\|z(k+1)\|^2\right] \leq \lambda_g \mathbb{E}\left[\|z(k)\|^2\right] \leq \cdots \leq \lambda_g^{k+1} \|z(0)\|^2 \qquad \text{for all } k \geq 0.$$

□

From Theorem 2 it follows that

$$\sum_{k=0}^{\infty} \mathbb{E}\left[\left\| x(k) - \frac{1}{m}\sum_{j=1}^{m} x_j(0)\,\mathbf{1}\right\|^2\right] < \infty,$$

which (by Fatou's Lemma) implies that with probability 1,

$$\lim_{k\to\infty} \left\| x(k) - \left(\frac{1}{m}\sum_{j=1}^{m} x_j(0)\right)\mathbf{1}\right\|^2 = 0,$$

showing that the iterates converge to the average of their initial values with probability 1. We note that the same result is true if the agents variables $x_j(0)$ were vectors due to the linearity of the update rule (1.31).

1.4.1.2 Broadcast-Based Consensus Algorithm

In the broadcast model, at time Z_k, the randomly activated agent I_k broadcasts its value $x_{I_k}(k)$ to all of its neighbors $j \in \breve{N}_{I_k}$ in the graph $G = ([m], E)$. Here, the neighbor set $\breve{N}_i$ of an agent i does not include the agent i itself,

$$\breve{N}_i = \{j \in [m] \mid i \leftrightarrow j \in E\}.$$

Thus, even though the graph G is undirected, the actual links that are used at any instance of time are virtually directed, as shown in Fig. 1.10.

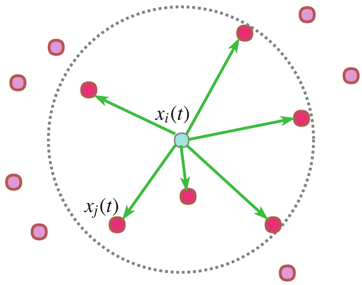


Fig. 1.10 Broadcast communication protocol: an agent that wakes up broadcasts its value to its neighbors, resulting in a random set of agents that are activated for performing an update. In a wireless network, the neighbors of an agent are typically defined as those agents that are within a certain radius of a given agent

Upon receiving the broadcasted value, the agents $j \in \breve{N}_{I_k}$ perform an update of their values, while the other agents do nothing (they sleep), including the agent I_k that broadcasted its information. Formally, the updates are given by

$$x_j(k+1) = (1-\beta)x_j(k) + \beta x_{I_k}(k) \qquad \text{for all } j \in \breve{N}_{I_k},$$
$$x_j(k+1) = x_j(k) \qquad \text{for all } j \notin \breve{N}_{I_k},$$

where $\beta \in (0, 1)$.

We define the matrix $W_b(k)$, as follows:

$$[W_b(k)]_{ii} = 1 - \beta \quad \text{for all } i \in \breve{N}_{I_k}, \qquad [W_b(k)]_{iI_k} = \beta \quad \text{for all } i \in \breve{N}_{I_k},$$
$$[W_b(k)]_{ii} = 1 \quad \text{for all } i \notin N_{I_k} \qquad \text{and else} \quad [W_b(k)]_{ij} = 0.$$

Using this matrix, the broadcast method can be written as:

$$x(k+1) = W_b(k)x(k) \qquad \text{for all } k \geq 0. \tag{1.32}$$

Note that the random matrix $W_b(k)$ is stochastic, but not necessarily doubly stochastic. Also, note that it is not symmetric. The expected matrix $\bar{W}_b = \mathbb{E}[W_b(k)]$ is in fact doubly stochastic. Specifically, as shown in [2, 3], $\bar{W}_b$ is given by

$$\bar{W}_b = I - \frac{\beta}{m} L_G,$$

where L_G is the Laplacian of the graph G, i.e., $L_G = D - A$ where A is the 0-1 adjacency matrix for the graph G and D is the diagonal matrix with entries $d_{ii} = |\breve{N}_i|$, $i \in [m]$. Since G is undirected, its Laplacian L_G is symmetric. Furthermore, since $L_G\mathbf{1} = 0$, it follows that $\bar{W}_b\mathbf{1} = \mathbf{1}$, which due to the symmetry of $\bar{W}_b$ also implies that $\mathbf{1}'\bar{W}_b = \mathbf{1}'$.

In addition, it has been shown in [2, 3] that, when the graph G is connected, the spectral norm of the matrix

$$\bar{W}_b - \frac{1}{m}\mathbf{1}\mathbf{1}'$$

is less than 1 (see Lemma 2 in [3]). This spectral property of the matrix $\bar{W}_b - \frac{1}{m}\mathbf{1}\mathbf{1}'$ is sufficient to guarantee the convergence of the random broadcast algorithm to a consensus in expectation only. Its convergence with probability 1 requires some additional analysis of the properties of the random matrices $W_b(k)$. In particular, the spectral norm of the matrix $\mathbb{E}\left[W_b'(k)(I - \frac{1}{m}\mathbf{1}\mathbf{1}')W_b(k)\right]$ plays a crucial role in establishing such a convergence result. Let Q denote this matrix, i.e.,

$$Q = \mathbb{E}\left[W_b'(k)\left(I - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)W_b(k)\right], \tag{1.33}$$

where I denotes the identity matrix of the appropriate dimension.

It has been shown in Proposition 2 of [3] that, when the graph G is connected, then the matrix Q has a spectral radius less than 1 for any $\beta \in (0, 1)$ (see the role of β in the definition of matrix $W_b(k)$). This property is a key in proving the convergence of the method with probability 1, as given in Theorem 1 in [3]. In the next theorem, we summarize some key relations which have been established in [3].

Theorem 3 (Lemma 3 and Proposition 2 of [3]) *Assume that the graph G is connected. Then, for any $\beta \in (0, 1)$, we have*

(a) The spectral radius of the matrix Q in (1.33) is less than 1.
(b) The iterate sequences $\{x_i(k)\}$, $i \in [m]$, produced by the random broadcast algorithm (1.32) satisfy the following relation

$$\mathbb{E}\left[\left\|x(k) - \frac{1}{m}\sum_{j=1}^{m} x_j(k)\,\mathbf{1}\right\|^2\right] \le \lambda_b^k \left\|x(0) - \frac{1}{m}\sum_{j=1}^{m} x_j(0)\,\mathbf{1}\right\|^2 \quad \text{for all } k \ge 0,$$

where $0 < \lambda_b < 1$ is the spectral norm of the matrix Q given in (1.33).

An extension of Theorem 3 to the case when the links are unreliable can be found in [101]. We note that the random broadcast algorithm does not lead to the consensus on the average of the initial agents' values with probability 1. It guarantees, with probability 1, that the agents will reach a consensus on a random point whose expected value is the average of the initial agents' values. Concretely, as shown in Theorem 1 of [3], there holds

$$\text{Prob}\left\{\lim_{k\to\infty} x(k) = c\mathbf{1}\right\} = 1,$$

where c is a random scalar satisfying

$$\mathbb{E}[c] = \frac{1}{m}\sum_{i=1}^{m} x_i(0).$$

1.4.2 *Distributed Asynchronous Algorithm*

In this section, we consider a general distributed asynchronous algorithm for optimization problem (1.1) based on random matrices. The random matrices are employed for the alignment of the agents iterates. As special cases of this algorithm, one can obtain the algorithms that use the random gossip and the random broadcast communications.

In particular, we will consider an algorithm with random asynchronous updates, as follows. We assume that there is some random i.i.d. process that triggers the update times Z_k (as for the cases of gossip and broadcast). Without going into details of such a process, we can simply keep a virtual index to count the update times (corresponding to the times when at least one agent is active). We also assume that the agents communicate over a network with connectivity structure captured by an undirected graph G.

At the time of the $k+1$st update, a random stochastic matrix $W(k)$ is available that captures the communication pattern among the agents, i.e., $w_{ij}(k) > 0$ if and only if agent i receives $x_j(k)$ from its neighbor $j \in N_i$. We let A_k be the set of agents that are active (perform an update) at time $k+1$. Then, the agents iterates at time $k+1$ are described through the following two steps:

$$\begin{aligned} v_i(k+1) &= \sum_{j=1}^{m} w_{ij}(k)x_j(k), \\ x_i(k+1) &= \Pi_X[v_i(k+1) - \alpha_{i,k+1}g_i(k+1)]\chi_{\{i\in A_k\}} \\ &\quad + v_i(k+1)\chi_{\{i\notin A_k\}}, \end{aligned} \tag{1.34}$$

where $\alpha_{i,k+1} > 0$ is a stepsize of agent i, $g_i(k+1)$ is a subgradient of f_i at $v_i(k+1)$ and χ_{E} is the characteristic function of an event E. We will *assume that the initial points* $\{x_i(0), i \in [m]\} \subset X$ *are deterministic.*

Note that each agent uses its own stepsize $\alpha_{i,k+1}$. It is important to note that, since $W(k)$ is stochastic, the event $\{i \in A_k\}$ is equivalent to $\{w_{ii}(k) \neq 1\}$. Thus, when $i \notin A_k$, which is equivalent to $\{w_{ii}(k) = 1\}$, we have $v_i(k+1) = x_i(k)$

and $x_i(k+1) = x_i(k)$. Hence, the relation $i \notin A_k$ corresponds to agent i not updating at all, so the iterate updates in (1.34) are equivalent to the following update scheme:

$$v_i(k+1) = \sum_{j=1}^{m} w_{ij}(k) x_j(k),$$

$$x_i(k+1) = \Pi_X[v_i(k+1) - \alpha_{i,k+1} g_i(k+1)] \qquad \text{for all } i \in A_k,$$

and otherwise

$$x_i(k+1) = x_i(k).$$

Moreover, when the matrices $W(k)$ are stochastic, we have

$$x_i(k+1) \in X \qquad \text{for all } k \geq 0 \text{ and all } i \in [m].$$

There is an alternative view for the updates of $x_i(k+1)$ in (1.34) that will be useful in our analysis. Specifically, noting that $\chi_{\{i \notin A_k\}} = 1 - \chi_{\{i \in A_k\}}$, from the definition of $x_i(k+1)$ in (1.34) it follows that

$$x_i(k+1) = \chi_{\{i \in A_k\}} \Pi_X[v_i(k+1) - \alpha_{i,k+1} g_i(k+1)] + \left(1 - \chi_{\{i \in A_k\}}\right) v_i(k+1). \tag{1.35}$$

Thus, we can view $x_i(k+1)$ as a convex combination of two points, namely, a convex combination of $\Pi_X[v_i(k+1) - \alpha_{i,k+1} g_i(k+1)] \in X$ and $v_i(k+1)$. When $W(k)$ is a stochastic matrix, the point $v_i(k+1)$ is in the set X.

If the random gossip protocol is used for communications, then $W(k) = W_g(k)$. Similarly, if the agents communicate using the random broadcast protocol, then $W(k) = W_b(k)$. Thus, the random gossip and random broadcast algorithms can be viewed as a special case of a more general random communication model, where the weight matrices $W(k)$ are random, drawn independently in time from the same distribution, and have the properties as specified in the following assumption.

Assumption 4 *Let $\{W(k)\}$ be a sequence of $m \times m$ random i.i.d. matrices such that the following conditions are satisfied:*

(a) Each realization of $W(k)$ is a stochastic matrix compatible with the graph $G = ([m], E)$, i.e., $w_{ij}(k) > 0$ only if $j \in N_i$.

(b) The spectral norm of the matrix $Q = \mathbb{E}\left[W'(k)\left(I - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)W(k)\right]$ is less than 1.

(c) The expected matrix $\mathbb{E}[W(k)] = \bar{W}$ is doubly stochastic.

In view of Assumption 4(a), the entry $\bar{W}_{ij}$ of the expected matrix may be positive only if $j \in N_i$. We do not assume explicitly that the graph G is connected, however, this property of the graph is subsumed within Assumption 4(b).

Note that the random matrices corresponding to the gossip and the broadcast model satisfy Assumption 4 when the graph G is connected. In fact, the random matrices corresponding to the gossip model in (1.31) satisfy a stronger condition than that of Assumption 4(a), since each realization of $W_g(k)$ is a doubly stochastic matrix.

Under Assumption 4(a), the event $\{i \in A_k\}$ that agent i updates (is awaken) at time $k+1$ has a stationary probability, denoted by p_i, i.e.,

$$p_i = \text{Prob}\{i \in A_k\}.$$

We now specify the stepsize rule for the algorithm. We consider the case when every agent i choses its stepsize value $\alpha_{i,k+1}$ based on its own local count of the update times. Letting $\Gamma_i(k+1)$ be the number of times the agent was awaken up to (including) time k, i.e.,

$$\Gamma_i(k+1) = \sum_{t=0}^{k} \chi_{\{i \in A_t\}},$$

we define the stepsize $\alpha_{i,k+1}$, as follows

$$\alpha_{i,k+1} = \frac{1}{\Gamma_i(k+1)} \qquad \text{for all } i \in [m] \text{ and } k \geq 0. \tag{1.36}$$

We note that

$$\Gamma_i(k+1) \geq \Gamma_i(k) \qquad \text{for all } k \geq 0 \text{ and } i \in [m],$$

implying that

$$\alpha_{i,k+1} \leq \alpha_{i,k} \qquad \text{for all } k \geq 0 \text{ and } i \in [m]. \tag{1.37}$$

In what follows, we will work with the conditional expectations with respect to the past iterates of the algorithm. For this, we let F_k denote the history of the algorithm (1.34), i.e.,

$$\mathsf{F}_k = \{W(0), \ldots, W(k-1)\} \qquad \text{for all } k \geq 1,$$

and $\mathsf{F}_0 = \emptyset$.

1.4.3 Convergence Analysis of Asynchronous Algorithm

We investigate the convergence properties of the algorithm assuming that the stepsize $\alpha_{i,k+1}$ is selected by agent i based on its local information. Prior to specifying the stepsize, we provide a result that is valid for any stepsize choice. It is also valid for any matrix sequence $\{W(k)\}$.

Lemma 7 *Assume that the problem is convex (i.e., Assumption 1 holds). Then, for the iterates of the algorithm (1.34) with any stepsize $\alpha_{i,k+1} > 0$ we have for all $x \in X$, for all $i \in [m]$ and all $k \geq 0$,*

$$\sum_{i=1}^{m} \|x_i(k+1) - x\|^2 \leq \sum_{i=1}^{m} \|v_i(k+1) - x\|^2$$
$$- 2\sum_{i=1}^{m} \alpha_{i,k+1}\chi_{\{i\in A_k\}} (f_i(v_i(k+1)) - f_i(x))$$
$$+ \sum_{i=1}^{m} \alpha_{i,k+1}^2 \chi_{\{i\in A_k\}} \|g_i(k+1)\|^2.$$

Proof From the relation in (1.35) by the convexity of the squared norm, it follows that for any $x \in X$, all $k \geq 0$ and all $i \in [m]$,

$$\|x_i(k+1) - x\|^2 \leq \chi_{\{i\in A_k\}} \|\Pi_X[v_i(k+1) - \alpha_{i,k+1} g_i(k+1)] - x\|^2$$
$$+ \left(1 - \chi_{\{i\in A_k\}}\right) \|v_i(k+1) - x\|^2.$$

By Lemma 4, for the point $\Pi_X[v_i(k+1) - \alpha_{i,k+1} g_i(k+1)]$ and any $x \in X$, we have

$$\|\Pi_X[v_i(k+1) - \alpha_{i,k+1} g_i(k+1)] - x\|^2 \leq \|v_i(k+1) - x\|^2$$
$$- 2\alpha_{i,k+1} (f_i(v_i(k+1)) - f_i(x)) + \alpha_{i,k+1}^2 \|g_i(k+1)\|^2.$$

By combining the preceding two relations, we obtain

$$\|x_i(k+1) - x\|^2 \leq \|v_i(k+1) - x\|^2 - 2\alpha_{i,k+1}\chi_{\{i\in A_k\}} (f_i(v_i(k+1)) - f_i(x))$$
$$+ \alpha_{i,k+1}^2 \chi_{\{i\in A_k\}} \|g_i(k+1)\|^2.$$

The desired relation follows by summing the preceding inequalities over $i \in [m]$. □

We have the following refinement of Lemma 7 for the random stepsizes $\alpha_{i,k+1}$ given by (1.36), which are measurable with respect to F_k for all $i \in [m]$. The result is developed under the assumption that the set X is compact, which is used to bound the error induced by the asynchronous updates and, in particular, the error due to a different frequency of agents' updates. The result assumes that the matrix sequence $\{W(k)\}$ is just an i.i.d. sequence.

Proposition 2 *Let the problem be convex (Assumption 1) and, also, assume that the set X is bounded. Let the random matrix sequence $\{W(k)\}$ be i.i.d. Consider the iterates produced by method (1.34) with the random stepsizes $\alpha_{i,k+1}$ as given in (1.36). Then, with probability 1, we have for all $k \geq 0$ and all $x \in X$,*

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \leq \sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] - \frac{2}{k+1}\left(f(x_{av}(k)) - f(x)\right) + r_k,$$

where

$$\begin{aligned} r_k =& 2CD \sum_{i=1}^{m} \left| \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right) \chi_{\{i \in A_k\}} \mid \mathsf{F}_k\right]\right| \\ &+ \frac{2C\sqrt{\sum_{i=1}^{m} \frac{1}{p_i}}}{k+1} \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]} \\ &+ 2\sqrt{m}CD \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)^2 \chi_{\{i \in A_k\}} \mid \mathsf{F}_k\right]} + C^2 \alpha_{i,k}^2, \end{aligned}$$

with p_i denoting the probability of the event $\chi_{\{i \in A_k\}}$, C being the uniform upper bound on the subgradient norms of f_i over the set X, and $D = \max_{x,y \in X} \|x - y\|$.

Proof In view of the compactness of the set X, it follows that the subgradients of f_i are uniformly bounded over the set X for all i, i.e., there exists a constant C such that

$$\|s\| \leq C \qquad \text{for every subgradient } s \text{ of } f_i(z) \text{ at any } z \in X. \tag{1.38}$$

Therefore, each function f_i is Lipschitz continuous on X, so that for all $x \in X$, all $k \geq 0$, and all $i \in [m]$,

$$\begin{aligned} f_i(v_i(k+1)) - f_i(x) &= f_i(v_i(k+1)) - f_i(x_{av}(k)) + f_i(x_{av}(k)) - f_i(x) \\ &\geq -C\|v_i(k+1) - x_{av}(k)\| + f_i(x_{av}(k)) - f_i(x), \end{aligned}$$

where $x_{av}(k) = \frac{1}{m}\sum_{j=1}^m x_j(k)$.

By using the preceding estimate in Lemma 7 and the fact that the subgradients are bounded, we obtain

$$\begin{aligned} &\sum_{i=1}^m \|x_i(k+1) - x\|^2 \\ &\quad \leq \sum_{i=1}^m \|v_i(k+1) - x\|^2 - 2\sum_{i=1}^m \alpha_{i,k+1}\chi_{\{i\in A_k\}}\left(f_i(x_{av}(k)) - f_i(x)\right) \\ &\qquad 2C\sum_{i=1}^m \alpha_{i,k+1}\chi_{\{i\in A_k\}}\|v_i(k+1) - x_{av}(k)\| + C^2\sum_{i=1}^m \alpha_{i,k+1}^2\chi_{\{i\in A_k\}}. \end{aligned}$$

We take the conditional expectation with respect to F_k in both sides of the preceding relation and further obtain, with probability 1, for all $x \in X$ and all $k \geq 0$,

$$\begin{aligned} \sum_{i=1}^m \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] &\leq \sum_{i=1}^m \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \\ &- 2\sum_{i=1}^m \mathbb{E}\left[\alpha_{i,k+1}\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]\left(f_i(x_{av}(k)) - f_i(x)\right) \\ &+ 2C\sum_{i=1}^m \mathbb{E}\left[\alpha_{i,k+1}\chi_{\{i\in A_k\}}\|v_i(k+1) - x_{av}(k)\| \mid \mathsf{F}_k\right] \\ &+ C^2\sum_{i=1}^m \mathbb{E}\left[\alpha_{i,k+1}^2\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]. \end{aligned} \tag{1.39}$$

Since $\alpha_{i,k+1}\chi_{\{i\in A_k\}} \leq \alpha_{i,k+1}$ and the stepsize is non-increasing (see (1.37)), it follows that

$$\alpha_{i,k+1}\chi_{\{i\in A_k\}} \leq \alpha_{i,k} \qquad \text{for all } i \in [m] \text{ and all } k \geq 0.$$

Hence, with probability 1,

$$\mathbb{E}\left[\alpha_{i,k+1}^2\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right] \leq \mathbb{E}\left[\alpha_{i,k}^2 \mid \mathsf{F}_k\right] = \alpha_{i,k}^2, \tag{1.40}$$

where in the last equality we use the fact that $\alpha_{i,k}$ is completely determined given the past F_k. By substituting relation (1.40) in inequality (1.39), we obtain

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \le \sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right]$$
$$- 2\sum_{i=1}^{m} \mathbb{E}\left[\alpha_{i,k+1}\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right](f_i(x_{av}(k)) - f_i(x))$$
$$+ 2C\sum_{i=1}^{m} \mathbb{E}\left[\alpha_{i,k+1}\chi_{\{i\in A_k\}}\|v_i(k+1) - x_{av}(k)\| \mid \mathsf{F}_k\right] + C^2\alpha_{i,k}^2. \tag{1.41}$$

By adding and subtracting $2\sum_{i=1}^{m}\frac{\mathbb{E}[\chi_{\{i\in A_k\}}|\mathsf{F}_k]}{(k+1)p_i}(f_i(x_{av}(k)) - f_i(x))$ to the second term on the right hand side of (1.41), and by doing similarly with a corresponding expression for the third term, we have

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \le \sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right]$$
$$- 2\sum_{i=1}^{m} \frac{\mathbb{E}\left[\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]}{(k+1)p_i}(f_i(x_{av}(k)) - f_i(x))$$
$$+ 2\sum_{i=1}^{m}\left|\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right](f_i(x_{av}(k)) - f_i(x))\right|$$
$$+ \frac{2C}{k+1}\underbrace{\sum_{i=1}^{m}\frac{1}{p_i}\mathbb{E}\left[\chi_{\{i\in A_k\}}\|v_i(k+1) - x_{av}(k)\| \mid \mathsf{F}_k\right]}_{T_1}$$
$$+ 2C\underbrace{\left|\sum_{i=1}^{m}\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)\chi_{\{i\in A_k\}}\|v_i(k+1) - x_{av}(k)\| \mid \mathsf{F}_k\right]\right|}_{T_2}$$
$$+ C^2\alpha_{i,k}^2, \tag{1.42}$$

where p_i is the probability of the event that agent i is updating.

To estimate the term T_1, we use Hölder's inequality

$$\sum_{i=1}^{r}\mathbb{E}\left[|a_i b_i|\right] \le \sqrt{\sum_{i=1}^{r}\mathbb{E}\left[a_i^2\right]}\sqrt{\sum_{i=1}^{r}\mathbb{E}\left[b_i^2\right]},$$

and obtain

$$T_1 \le \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\frac{1}{p_i^2}\chi_{\{i\in A_k\}}\right]} \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]}$$
$$\le \sqrt{\sum_{i=1}^{m} \frac{1}{p_i}} \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]}, \qquad (1.43)$$

where in the first inequality we also use the fact that the event $\{i \in A_k\}$ is independent from the past, while in the last inequality we use the fact that the probability that the event $\{i \in A_k\}$ occurs is p_i.

For the term T_2 in (1.42), by Hölder's inequality, we have

$$T_2 \le \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)^2 \chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]}$$
$$\times \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]}. \qquad (1.44)$$

We now substitute the estimates (1.43) and (1.44) in the inequality (1.42) and obtain that, with probability 1, there holds for all $x \in X$ and $k \ge 0$,

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \le \sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right]$$
$$- 2\sum_{i=1}^{m} \frac{\mathbb{E}\left[\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]}{(k+1)p_i} \left(f_i(x_{av}(k)) - f_i(x)\right)$$
$$+ 2\sum_{i=1}^{m} \left|\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right) \chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right] \left(f_i(x_{av}(k)) - f_i(x)\right)\right|$$
$$+ \frac{2C\sqrt{\sum_{i=1}^{m}\frac{1}{p_i}}}{k+1} \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]}$$
$$+ 2C \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)^2 \chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]}$$
$$\times \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]} + C^2 \alpha_{i,k}^2.$$

In view of compactness of X, and $x_{av}(k) \in X$ and $v_i(k+1) \in X$ for all i and k, it follows that for all $x \in X$,

$$|f_i(x_{av}(k)) - f_i(x)| \le C\|x_{av}(k) - x\| \le CD, \qquad \|v_i(k+1) - x_{av}(k)\| \le D,$$

where $D = \max_{y,z \in X} \|y - z\|$. We also note that

$$\frac{\mathbb{E}\left[\chi_{\{i \in A_k\}} \mid \mathsf{F}_k\right]}{(k+1)p_i} = \frac{1}{k+1}.$$

By using the preceding relations, we have that, with probability 1, there holds for all $x \in X$ and $k \ge 0$,

$$\begin{aligned}
&\sum_{i=1}^m \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \\
&\quad \le \sum_{i=1}^m \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] - \frac{2}{k+1}\left(f(x_{av}(k)) - f(x)\right) \\
&\quad + 2CD \sum_{i=1}^m \left|\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)\chi_{\{i \in A_k\}} \mid \mathsf{F}_k\right]\right| \\
&\quad + \frac{2C\sqrt{\sum_{i=1}^m \frac{1}{p_i}}}{k+1}\sqrt{\sum_{i=1}^m \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]} \\
&\quad + 2\sqrt{m}CD\sqrt{\sum_{i=1}^m \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)^2 \chi_{\{i \in A_k\}} \mid \mathsf{F}_k\right]} + C^2\alpha_{i,k}^2.
\end{aligned}$$

The desired relation follows by introducing the notation for the sum of the last four terms on the right hand side of the preceding relation. □

To establish the convergence of the method, one of the goals is to show that the error terms r_k in Proposition 2 are well behaved in the sense that $\sum_{k=0}^{\infty} r_k < \infty$ with probability 1. We note that the error r_k has two types of terms, one type related to the stepsize and the other related to the distances of iterates $v_i(k+1)$ and the average vector $x_{av}(k)$, which also involves the stepsize implicitly. So we start by investigating some properties of the stepsize.

1.4.3.1 Stepsize Analysis

We consider the random agent based stepsize defined in (1.36), which is the inverse of the number $\Gamma_i(k+1)$ of agent i updates from time $t = 0$ up to time $t = k$, inclusively. We establish some relations for the stepsize that involve expectations and a set of results for the stepsize sums.

We start with the relations involving the expectations of the stepsize in the term r_k of Proposition 2.

Lemma 8 *Let the matrix sequence* $\{W(k)\}$ *be an i.i.d. random sequence. Then, for the stepsize* $\alpha_{i,k}$ *in* (1.36)*, with probability 1, we have for all* $k \geq 0$ *and* $i \in [m]$*,*

$$\left|\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]\right| \leq p_i\left|\alpha_{i,k} - \frac{1}{kp_i}\right| + (1-p_i)\frac{\alpha_{i,k}}{k},$$

$$\sqrt{\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)^2\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]} \leq \sqrt{2p_i}\left|\alpha_{i,k} - \frac{1}{kp_i}\right| + (1-p_i)\sqrt{\frac{2}{p_i}}\frac{\alpha_{i,k}}{k}.$$

Proof Recall that the event $\chi_{\{i\in A_k\}}$ that agent i updates has probability p_i. Thus, using the independence of the event $\chi_{\{i\in A_k\}}$ given the past F_k, we have with probability 1 for all $k \geq 0$ and $i \in [m]$,

$$\mathbb{E}\left[\alpha_{i,k+1}\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right] = \frac{p_i}{\Gamma_i(k)+1}.$$

Using the preceding relation and $\mathbb{E}\left[\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right] = p_i$, we obtain

$$\begin{aligned} M_1 &:= \left|\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]\right| \\ &= \left|p_i\left(\frac{1}{\Gamma_i(k)+1} - \frac{1}{(k+1)p_i}\right)\right| \\ &= p_i\frac{|kp_i - \Gamma_i(k) + p - 1|}{(k+1)p_i(\Gamma_i(k)+1)}, \end{aligned}$$

where the last equality is obtained by re-grouping the terms in the numerator. Thus, it follows that

$$M_1 \leq p_i\frac{|kp_i - \Gamma_i(k)| + (1-p_i)}{(k+1)p_i(\Gamma_i(k)+1)} \leq p_i\frac{|kp_i - \Gamma_i(k)| + (1-p_i)}{kp_i\Gamma_i(k)}.$$

By separating the terms, we have

$$M_1 \leq p_i\frac{|kp_i - \Gamma_i(k)|}{kp_i\Gamma_i(k)} + \frac{(1-p_i)}{k\Gamma_i(k)} = p_i\left|\frac{1}{\Gamma_i(k)} - \frac{1}{kp_i}\right| + \frac{(1-p_i)}{k\Gamma_i(k)}.$$

Recognizing that $\alpha_{i,k} = \frac{1}{\Gamma_i(k)}$, we obtain

$$M_1 \le p_i \left|\alpha_{i,k} - \frac{1}{kp_i}\right| + (1-p_i)\frac{\alpha_{i,k}}{k},$$

thus showing the first relation stated in the lemma.

For the second relation we have

$$\begin{aligned}M_2 :=& \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)^2 \chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right] = p_i\left(\frac{1}{\Gamma_i(k)+1} - \frac{1}{(k+1)p_i}\right)^2 \\ =& p_i\left(\frac{(k+1)p_i - \Gamma_i(k) - 1}{(k+1)p_i(\Gamma_i(k)+1)}\right)^2 = p_i\frac{(kp_i - \Gamma_i(k) + p - 1)^2}{(k+1)^2 p_i^2(\Gamma_i(k)+1)^2}.\end{aligned}$$

Now using the relation $(a+b)^2 \le 2(a^2+b^2)$, which is valid for any scalars a and b, we obtain

$$M_2 \le 2p_i\frac{(kp_i - \Gamma_i(k))^2 + (1-p_i)^2}{(k+1)^2 p_i^2(\Gamma_i(k)+1)^2} \le 2p_i\frac{(kp_i - \Gamma_i(k))^2 + (1-p_i)^2}{k^2 p_i^2 \Gamma_i^2(k)}.$$

By separating the terms we further have

$$M_2 \le 2p_i\frac{(kp_i - \Gamma_i(k))^2}{k^2 p_i^2 \Gamma_i^2(k)} + \frac{2(1-p_i)^2}{k^2 p_i \Gamma_i^2(k)} = 2p_i\left(\frac{1}{\Gamma_i(k)} - \frac{1}{kp_i}\right)^2 + \frac{2(1-p_i)^2}{k^2 p_i \Gamma_i^2(k)}.$$

By substituting $\alpha_{i,k} = \frac{1}{\Gamma_i(k)}$, it follows that

$$M_2 \le 2p_i\left(\alpha_{i,k} - \frac{1}{kp_i}\right)^2 + \frac{2(1-p_i)^2\alpha_{i,k}^2}{k^2 p_i}.$$

Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, which is valid for any $a, b \ge 0$, we have

$$\begin{aligned}\sqrt{M_2} &\le \sqrt{2p_i\left(\alpha_{i,k} - \frac{1}{kp_i}\right)^2} + \sqrt{\frac{2(1-p_i)^2\alpha_{i,k}^2}{k^2 p_i}} \\ &= \sqrt{2p_i}\left|\alpha_{i,k} - \frac{1}{kp_i}\right| + (1-p_i)\sqrt{\frac{2}{p_i}}\frac{\alpha_{i,k}}{k},\end{aligned}$$

which establishes the second relation of the lemma. □

We next investigate some properties of the stepsize sums under the assumption that the random matrix sequence $\{W(k)\}$ is i.i.d. In this case, for each $i \in [m]$, the events $\{i \in A_k\}$ are i.i.d., so that we have

$$\mathbb{E}\left[\Gamma_i(k)\right] = (k+1)p_i.$$

By the law of iterated logarithms [41] (pages 476–479), we have that for any $q > 0$,

$$\text{Prob}\left\{\lim_{k\to\infty} \frac{|\Gamma_i(k) - (k+1)p_i|}{(k+1)^{\frac{1}{2}+q}} = 0\right\} = 1 \qquad \text{for all } i \in [m]. \tag{1.45}$$

We use this relation to establish some results for the sums involving the stepsize, as given is the following lemma.

Lemma 9 *Let the random matrix sequence* $\{W(k)\}$ *be i.i.d., and consider the stepsize* $\alpha_{i,k}$ *as given in* (1.36)*. Then, we have*

$$\text{Prob}\left\{\sum_{k=1}^{\infty} \alpha_{i,k}^2 < \infty, \ \textit{for all } i \in [m]\right\} = 1,$$

$$\text{Prob}\left\{\sum_{k=1}^{\infty} \left|\alpha_{i,k} - \frac{1}{kp_i}\right| < \infty, \ \textit{for all } i \in [m]\right\} = 1,$$

$$\text{Prob}\left\{\sum_{k=1}^{\infty} \frac{\alpha_{i,k}}{k} < \infty, \ \textit{for all } i \in [m]\right\} = 1.$$

Proof The proof is based on considering sample paths, where a sample path corresponds to a sequence of realizations of the matrices, which is denoted by ω. We fix a sample path ω for which the limit in (1.45) is zero. Then, using relation (1.45), we can show that for every $q \in (0, \frac{1}{2})$ there exists an index[5] $\tilde{k}(\omega)$ such that for all $k \geq \tilde{k}(\omega)$ and for all $i \in [m]$, we have[6]

$$\alpha_{i,k}(\omega) \leq \frac{2}{kp_i}, \qquad \left|\alpha_{i,k}(\omega) - \frac{1}{kp_i}\right| \leq \frac{1}{k^{\frac{3}{2}-q}\,p_i^2}. \tag{1.46}$$

Thus, there holds for all $i \in [m]$,

$$\sum_{k\geq\tilde{k}(\omega)} \alpha_{i,k}^2(\omega) < \infty, \qquad \sum_{k\geq\tilde{k}(\omega)} \left|\alpha_{i,k}(\omega) - \frac{1}{kp_i}\right| < \infty,$$

[5] The index $\tilde{k}(\omega)$ also depends on q, but this dependence is suppressed in the notation.

[6] The derivation of the relations in (1.46) can be found in the proof of Lemma 3 in [100], where the analysis is to be performed on a sample path.

where the last relation holds due to $q \in (0, \frac{1}{2})$. Furthermore, we have for all $k \geq \tilde{k}(\omega)$ and for all $i \in [m]$,

$$\frac{\alpha_{i,k}(\omega)}{k} \leq \frac{2}{k^2 p_i},$$

implying that

$$\sum_{k \geq \tilde{k}(\omega)} \frac{\alpha_{i,k}(\omega)}{k} < \infty \qquad \text{for all } i \in [m].$$

Since the preceding relations are true for almost all but zero measure sample paths ω that satisfy relation (1.45), the stated results follow. □

1.4.3.2 Relation for Agents' Iterates and Their Averages

We now turn our attention to the disagreements $\|x_i(k) - x_{av}(k)\|$. We establish a relation for these disagreements which will be combined with Proposition 2 to assert the convergence behavior of the distances $\|x_i(k) - x^*\|$ for an optimal solution x^*.

We start by re-writing the iterations of the method (1.34), as follows:

$$v_i(k+1) = \sum_{j=1}^{m} w_{ij}(k)x_j(k),$$
$$x_i(k+1) = v_i(k+1) + \underbrace{\left(\Pi_X[v_i(k+1) - \alpha_{i,k+1}g_i(k+1)] - v_i(k+1)\right)\chi\{i \in A_k\}}_{\phi_i(k+1)}.$$

Hence, for all $i \in [m]$ and $k \geq 0$,

$$\begin{aligned} x_i(k+1) &= \sum_{j=1}^{m} w_{ij}(k)x_j(k) + \phi_i(k+1), \\ \phi_i(k+1) &= \left(\Pi_X[v_i(k+1) - \alpha_{i,k+1}g_i(k+1)] - v_i(k+1)\right)\chi\{i \in A_k\}, \\ v_i(k+1) &= \sum_{j=1}^{m} w_{ij}(k)x_j(k). \end{aligned} \tag{1.47}$$

Thus, we perceive the iterates $x_i(k+1)$ as obtained through a perturbed random consensus algorithm with random perturbations $\phi_i(k+1)$.

Under Assumption 4(a) and (c), in the following lemma, we establish a relation for the iterates $x_i(k+1)$.

Lemma 10 *Let the matrices $W(k)$ satisfy Assumptions 4(a) and (b). Then, for the iterate process in (1.47), we have for all $k \geq 0$ with probability 1,*

$$\sqrt{\mathbb{E}\left[\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2 \mid \mathsf{F}_k\right]} \leq \sqrt{\sum_{i=1}^{m} \rho \|x_i(k) - x_{av}(k)\|^2} + \sqrt{\sum_{i=1}^{m} \|\phi_i(k+1)\|^2},$$

where $x_{av}(k) = \frac{1}{m}\sum_{j=1}^{m} x_j(k)$ and $\rho \in (0,1)$ is the spectral norm of the matrix $Q = \mathbb{E}\left[W'(k)(I - \frac{1}{m}\mathbf{1}\mathbf{1}')W(k)\right]$ (see Assumption 4(b)).

Proof We first write the iterates $x_i(k+1)$ in (1.47) in a matrix form. We construct a matrix $\mathbf{X}(k)$ by placing the vectors $x_i'(k)$ in its rows, and similarly, we construct the matrix $\Phi(k)$ by placing the vectors $\phi_i'(k)$ in its rows. By doing so, we have the following representation for the evolution of the iterates $x_i(k+1)$:

$$\mathbf{X}(k+1) = W(k)\mathbf{X}(k) + \Phi(k+1) \qquad \text{for all } k \geq 0. \tag{1.48}$$

By multiplying both sides of (1.48) with the matrix $\frac{1}{m}\mathbf{1}\mathbf{1}'$, we have

$$\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1) = \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\mathbf{X}(k) + \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1).$$

By subtracting the preceding relation from (1.48), we obtain for all $k \geq 0$,

$$\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1) = \left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\mathbf{X}(k) + \Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1).$$

Since $W(k)$ is stochastic, we have $W(k)\mathbf{1} = \mathbf{1}$, implying that $\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\mathbf{1} = 0$, so that we have

$$\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\mathbf{X}(k) = \left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right).$$

Therefore, for all $k \geq 0$,

$$\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1) = \left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right) + \Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1). \tag{1.49}$$

By taking the squared Frobenius norms of both sides in (1.49), and using the fact that $\|AB\|_F \le \|A\|_F\|B\|_F$ for any two (compatible) matrices, we obtain for all $k \ge 0$,

$$\begin{aligned}
&\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F^2 \\
&\quad \le \left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \\
&\qquad + \left\|\Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F^2 \\
&\qquad + 2\left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F \\
&\qquad \times \left\|\Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F.
\end{aligned}$$

Next, we take conditional expectation with respect to F_k and obtain for all $k \ge 0$ with probability 1,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F^2 \mid \mathsf{F}_k\right] \\
&\quad \le \mathbb{E}\left[\left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \mid \mathsf{F}_k\right] \\
&\quad + \mathbb{E}\left[\left\|\Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F^2 \mid \mathsf{F}_k\right] \\
&\quad + 2\mathbb{E}\left[\left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F\right. \\
&\quad \left. \times \left\|\Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F \mid \mathsf{F}_k\right].
\end{aligned}$$

By using Hölder's inequality for expectations (see [8], page 242), we can see that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F^2 \mid \mathsf{F}_k\right] \\
&\quad \le \mathbb{E}\left[\left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \mid \mathsf{F}_k\right]
\end{aligned}$$

$$+\,\mathbb{E}\left[\left\|\Phi(k+1)-\frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F^2 \mid \mathsf{F}_k\right]$$
$$+\,2\sqrt{\mathbb{E}\left[\left\|\left(W(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \mid \mathsf{F}_k\right]}$$
$$\times\sqrt{\mathbb{E}\left[\left\|\Phi(k+1)-\frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F^2 \mid \mathsf{F}_k\right]}.$$

Hence, for all $k \geq 0$ with probability 1 we have

$$\mathbb{E}\left[\left\|\mathbf{X}(k+1)-\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F^2 \mid \mathsf{F}_k\right]$$
$$\leq \left(\sqrt{\mathbb{E}\left[\left\|\left(W(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \mid \mathsf{F}_k\right]}\right.$$
$$\left.+\sqrt{\mathbb{E}\left[\left\|\Phi(k+1)-\frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F^2 \mid \mathsf{F}_k\right]}\right)^2. \tag{1.50}$$

Next, using the column vectors $\mathbf{x}^\ell(k)$ of the matrix $X(k)$ and the definition of the Frobenius norm, we can write

$$\left\|\left(W(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2$$
$$=\sum_{\ell=1}^{m}\left\|\left(W(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{x}^\ell(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{x}^\ell(k)\right)\right\|^2$$
$$=\sum_{\ell=1}^{m}\left\langle \mathbf{x}^\ell(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{x}^\ell(k),\, Q(k)\left(\mathbf{x}^\ell(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{x}^\ell(k)\right)\right\rangle,$$

where

$$Q(k)=\left(W(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)'\left(W(k)-\frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right).$$

After some elementary algebra, it can be seen that for $Q(k)$ we have

$$Q(k) = W(k)'W(k) - \frac{1}{m}W'(k)\mathbf{1}\mathbf{1}'W(k) = W'(k)\left(I - \frac{1}{m}\mathbf{1}\mathbf{1}'\right)W(k).$$

Given the history F_k, the column vectors $\mathbf{x}^\ell(k)$ are deterministic, so that by using $Q = \mathbb{E}[Q(k)]$ (see Assumption 4(b)), we have with probability 1 for all $k \geq 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \mid \mathsf{F}_k\right] \\
&\quad = \sum_{\ell=1}^m \left\langle \mathbf{x}^\ell(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{x}^\ell(k), \mathbb{E}[Q(k)]\left(\mathbf{x}^\ell(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{x}^\ell(k)\right)\right\rangle \\
&\quad \leq \sum_{\ell=1}^m \|Q\| \left\|\mathbf{x}^\ell(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{x}^\ell(k)\right\|^2 \\
&\quad \leq \rho \sum_{\ell=1}^m \left\|\mathbf{x}^\ell(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{x}^\ell(k)\right\|^2 \\
&\quad = \rho \left\|\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right\|_F^2,
\end{aligned}$$

where $\rho \in (0, 1)$ is the spectral norm of the matrix Q (note that ρ is smaller than 1 by Assumption 4(b)). Hence, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \mid \mathsf{F}_k\right] \\
&\quad \leq \rho \left\|\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right\|_F^2.
\end{aligned}$$

Finally by noticing that the matrix $\frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)$ has identical rows, where each its row is given by the vector $\frac{1}{m}\mathbf{1}'\mathbf{X}(k) = x_{av}'(k)$ with $x_{av}(k) = \frac{1}{m}\sum_{j=1}^m x_j(k)$, we see that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(W(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'W(k)\right)\left(\mathbf{X}(k) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k)\right)\right\|_F^2 \mid \mathsf{F}_k\right] \\
&\quad \leq \rho \left\|\mathbf{X}(k) - \mathbf{1}x_{av}'(k)\right\|_F^2 \\
&\quad = \rho \sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2. \qquad (1.51)
\end{aligned}$$

Similarly, using the definition of the Frobenius norm and the matrices $\Phi(k)$, we can see that

$$\mathbb{E}\left[\left\|\Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F^2 \mid \mathsf{F}_k\right]$$
$$= \mathbb{E}\left[\sum_{j=1}^{m} \left\|\phi_j(k+1) - \phi_{av}(k+1)\right\|^2 \mid \mathsf{F}_k\right],$$

where $\phi_{av}(k) = \sum_{\ell=1}^{m} \phi_\ell(k)$. Since the distance $\sum_{j=1}^{m} \|\phi_i(k+1) - y\|^2$ is minimized over all $y \in \mathbb{R}^n$ at $y^* = \phi_{av}(k+1)$, we obtain (using $y = 0$)

$$\mathbb{E}\left[\left\|\Phi(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\Phi(k+1)\right\|_F^2 \mid \mathsf{F}_k\right] \leq \mathbb{E}\left[\sum_{j=1}^{m} \left\|\phi_j(k+1)\right\|^2 \mid \mathsf{F}_k\right]. \tag{1.52}$$

Using relations (1.51) and (1.52) in inequality (1.50), we have for all $k \geq 0$ with probability 1,

$$\mathbb{E}\left[\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F^2 \mid \mathsf{F}_k\right]$$
$$\leq \left(\sqrt{\rho\sum_{j=1}^{m} \left\|x_j(k) - x_{av}(k)\right\|^2} + \sqrt{\mathbb{E}\left[\sum_{j=1}^{m} \left\|\phi_j(k+1)\right\|^2 \mid \mathsf{F}_k\right]}\right)^2.$$

Hence, by taking square roots on both sides, we see that

$$\sqrt{\mathbb{E}\left[\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F^2 \mid \mathsf{F}_k\right]}$$
$$\leq \sqrt{\rho\sum_{j=1}^{m} \left\|x_j(k) - x_{av}(k)\right\|^2} + \sqrt{\mathbb{E}\left[\sum_{j=1}^{m} \left\|\phi_j(k+1)\right\|^2 \mid \mathsf{F}_k\right]}. \tag{1.53}$$

The desired relation follows from (1.53) by using the fact that

$$\left\|\mathbf{X}(k+1) - \frac{1}{m}\mathbf{1}\mathbf{1}'\mathbf{X}(k+1)\right\|_F^2 = \sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2.$$

□

Based on Lemma 10, we can prove that the agents' disagreements are well behaved. The proof makes use of an (almost) supermartingale convergence result. The result is due to Robbins and Siegmund [136], and it can also be found in [129] (Chapter 2.2, Lemma 11).

Lemma 11 ([136]) *Let V_k, u_k, β_k and γ_k be non-negative random variables adapted to some σ-algebra σ_k. If with probability 1 we have $\sum_{k=0}^{\infty} u_k < \infty$, $\sum_{k=0}^{\infty} \beta_k < \infty$, and*

$$\mathbb{E}\left[V_{k+1} \mid \sigma_k\right] \leq (1+u_k)V_k - \gamma_k + \beta_k \qquad \textit{for all } k \geq 0,$$

then V_k converges to some non-negative scalar and $\sum_{k=0}^{\infty} \gamma_k < \infty$ with probability 1.

We have the following result for the disagreement sum $\sum_{j=1}^{m} \left\|x_j(k) - x_{av}(k)\right\|^2$.

Lemma 12 *Let the problem be convex (Assumption 1 holds) and let the set X be bounded. Let the matrices $W(k)$ satisfy Assumption 4(a) and (b). Then, with probability 1,*

$$\sum_{k=0}^{\infty} \frac{1}{k+1} \sqrt{\sum_{i=1}^{m} \|x_i(k) - x_{av}(k)\|^2} < \infty,$$

$$\liminf_{k\to\infty} \sum_{j=1}^{m} \left\|x_j(k) - x_{av}(k)\right\|^2 = 0.$$

Proof By Lemma 10 and by using the relation $\mathbb{E}\left[a\right] \leq \sqrt{\mathbb{E}\left[a^2\right]}$, we have with probability 1 for all $k \geq 0$,

$$\mathbb{E}\left[\sqrt{\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2} \mid \mathsf{F}_k\right] \leq \sqrt{\rho \sum_{j=1}^{m} \left\|x_j(k) - x_{av}(k)\right\|^2} + \sqrt{\mathbb{E}\left[\sum_{j=1}^{m} \left\|\phi_j(k+1)\right\|^2 \mid \mathsf{F}_k\right]}, \tag{1.54}$$

with $\rho \in (0, 1)$. From the definition of $\phi_j(k+1)$ in (1.47) and the non-expansiveness property of the projection operator, we can se that

$$\|\phi_i(k+1)\|^2 \le \alpha_{i,k+1}^2 \|g_i(k+1)\|^2 \chi\{i \in A_k\} \le \alpha_{i,k+1}^2 C^2 \chi\{i \in A_k\},$$

where the last inequality follows from the compactness of X, which implies that the subgradients of each f_i are uniformly bounded over the set X. Therefore, we have with probability 1,

$$\begin{aligned}\mathbb{E}\left[\|\phi_i(k+1)\|^2 \mid F_k\right] \le C^2 \mathbb{E}\left[\alpha_{i,k+1}^2 \chi\{i \in A_k\}\right] &= \frac{C^2 p_i}{(\Gamma_i(k)+1)^2} \le \frac{C^2 p_i}{\Gamma_i^2(k)} \\ &= C^2 p_i \alpha_{i,k}^2,\end{aligned}$$

implying that

$$\sqrt{\mathbb{E}\left[\sum_{j=1}^m \left\|\phi_j(k+1)\right\|^2 \mid \mathsf{F}_k\right]} \le C \sqrt{\sum_{i=1}^m p_i \alpha_{i,k}^2} \le C \sum_{i=1}^m \sqrt{p_i} \alpha_{i,k},$$

where the last inequality follows from relation $\sum_{i=1}^r a_i^2 \le (\sum_{i=1}^r a_i)^2$ which holds for any nonnegative scalars a_i. By using the preceding estimate in relation (1.54), we obtain with probability 1 for all $k \ge 0$,

$$\begin{aligned}&\mathbb{E}\left[\sqrt{\sum_{i=1}^m \|x_i(k+1) - x_{av}(k+1)\|^2 \mid \mathsf{F}_k}\right] \\ &\quad \le \sqrt{\rho \sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2} + C \sum_{i=1}^m \sqrt{p_i} \alpha_{i,k}.\end{aligned} \tag{1.55}$$

We divide both sides of (1.55) by $\frac{1}{k+1}$ and obtain with probability 1 for all $k \ge 0$,

$$\begin{aligned}&\frac{1}{k+1}\mathbb{E}\left[\sqrt{\sum_{i=1}^m \|x_i(k+1) - x_{av}(k+1)\|^2 \mid \mathsf{F}_k}\right] \\ &\quad \le \frac{\sqrt{\rho}}{k+1}\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2} + C \sum_{i=1}^m \frac{\sqrt{p_i}\alpha_{i,k}}{k+1}.\end{aligned}$$

We add and subtract $\frac{1}{k}\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2}$ to the right hand side of the preceding relation, so we can write

$$\frac{1}{k+1}\mathbb{E}\left[\sqrt{\sum_{i=1}^m \|x_i(k+1) - x_{av}(k+1)\|^2} \mid \mathsf{F}_k\right]$$
$$\leq \frac{1}{k}\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2} - \left(\frac{1}{k} - \frac{\sqrt{\rho}}{k+1}\right)\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2}$$
$$+ C\sum_{i=1}^m \frac{\sqrt{p_i}\alpha_{i,k}}{k+1}.$$

We note that

$$\frac{1}{k} - \frac{\sqrt{\rho}}{k+1} = \frac{k+1-\sqrt{\rho}k}{k(k+1)} = \frac{k(1-\sqrt{\rho})+1}{k(k+1)} \geq \frac{(1-\sqrt{\rho})}{(k+1)},$$

where we use the fact that $\rho \in (0, 1)$. Therefore, we have with probability 1 for all $k \geq 0$,

$$\frac{1}{k+1}\mathbb{E}\left[\sqrt{\sum_{i=1}^m \|x_i(k+1) - x_{av}(k+1)\|^2} \mid \mathsf{F}_k\right] \leq \frac{1}{k}\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2}$$
$$- \frac{1-\sqrt{\rho}}{k+1}\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2} + C\sum_{i=1}^m \frac{\sqrt{p_i}\alpha_{i,k}}{k+1}. \tag{1.56}$$

We now apply Lemma 11 with the following identification

$$V_k = \frac{1}{k}\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2}, \quad u_k = 0, \quad \beta_k = C\sum_{i=1}^m \frac{\sqrt{p_i}\alpha_{i,k}}{k+1},$$
$$\gamma_k = \frac{1-\sqrt{\rho}}{k+1}\sqrt{\sum_{j=1}^m \left\|x_j(k) - x_{av}(k)\right\|^2}.$$

Note that by Lemma 9 we have

$$\sum_{k=1}^\infty \sum_{i=1}^m \frac{\sqrt{p_i}\alpha_{i,k}}{k+1} < \infty \quad \text{with probability 1,}$$

so that the condition $\sum_{k=0}^{\infty} \beta_k < \infty$ with probability 1 is also satisfied. Thus, by the Robinson-Siegmund (almost) supermartingale convergence result in Lemma 11, from (1.56) we have that

$$\sum_{k=0}^{\infty} \frac{1-\sqrt{\rho}}{k+1} \sqrt{\sum_{j=1}^{m} \left\| x_j(k) - x_{av}(k) \right\|^2} < \infty \quad \text{with probability 1.}$$

Since $1 - \sqrt{\rho} > 0$, the preceding relation implies that with probability 1,

$$\sum_{k=0}^{\infty} \frac{1}{k+1} \sqrt{\sum_{j=1}^{m} \left\| x_j(k) - x_{av}(k) \right\|^2} < \infty, \tag{1.57}$$

$$\liminf_{k\to\infty} \sum_{j=1}^{m} \left\| x_j(k) - x_{av}(k) \right\|^2 = 0,$$

where the last relation can be shown by arguing per sample path ω for which relation (1.57) is valid. □

1.4.3.3 Convergence Result for Asynchronous Algorithm

We are now in position to assert the convergence of the iterates of the algorithm (1.34) with probability 1, as stated in the following theorem. The proof is based on Proposition 2 and the almost supermartingale convergence result as given in Lemma 11. To verify that the conditions of Lemma 11 are satisfied, we rely on Lemmas 8, 9 and 12.

Theorem 4 *Let the problem be convex (Assumption 1) and, also, assume that the set X is bounded. Let the random matrix sequence $\{W(k)\}$ satisfy Assumption 4. Consider the iterates produced by method (1.34) with the random stepsizes $\alpha_{i,k+1}$ as given in (1.36). Then, with probability 1, the sequences $\{x_i(k)\}$, $i \in [m]$, converge to a common (random) optimal solution of the problem.*

Proof The proof has two main parts: (i) proving that $x_i(k) - x_{av}(k)$ converges to 0, as k tends to infinity, with probability 1 for all i, and (ii) proving that the iterate sequences $\{x_i(k)\}$, $i \in [m]$, converge to the same (random) optimal solution[7] with probability 1. We start by deriving a relation that will be used in both parts (i) and (ii) of the proof.

[7]Different sample paths may converge to different solutions.

By Proposition 2 we have with probability 1 for all $k \geq 0$ and all $x \in X$,

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \leq \sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] - \frac{2}{k+1}\,(f(x_{av}(k)) - f(x)) + r_k, \tag{1.58}$$

with

$$\begin{aligned} r_k =& 2CD \sum_{i=1}^{m} \left| \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)\,p_i} \right) \chi_{\{i \in A_k\}} \mid \mathsf{F}_k \right] \right| \\ &+ \frac{2C\sqrt{\sum_{i=1}^{m} \frac{1}{p_i}}}{k+1} \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]} \\ &+ 2\sqrt{m}CD \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)\,p_i} \right)^2 \chi_{\{i \in A_k\}} \mid \mathsf{F}_k \right]} \\ &+ C^2 \alpha_{i,k}^2, \end{aligned} \tag{1.59}$$

where p_i is the probability of the event $\chi_{\{i \in A_k\}}$, C is a uniform upper bound on the subgradient norms of f_i over the set X, and $D = \max_{x,y \in X} \|x - y\|$. By the definition, $v_i(k+1)$ is a convex combination of $x_j(k)$, $j \in [m]$, so that by the convexity of the norm and the assumption that $W(k)$ is a stochastic matrix (see Assumption 4(a)), we have for any $x \in \mathbb{R}^n$,

$$\|v_i(k+1) - x\|^2 \leq \sum_{j=1}^{m} w_{ij}(k)\|x_j(k) - x\|^2,$$

implying that

$$\begin{aligned} \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] &\leq \sum_{j=1}^{m} \mathbb{E}\left[w_{ij}(k)\|x_j(k) - x\|^2 \mid \mathsf{F}_k\right] \\ &= \sum_{j=1}^{m} \mathbb{E}\left[w_{ij}(k)\right]\|x_j(k) - x\|^2 \\ &= \sum_{j=1}^{m} \bar{W}_{ij}\|x_j(k) - x\|^2. \end{aligned}$$

By summing the preceding inequalities over $i \in [m]$ and by using the assumption that $\bar{W}$ doubly stochastic (cf. Assumption 4(c)), we obtain for any $x \in \mathbb{R}^n$,

$$\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x\|^2 \mid \mathsf{F}_k\right] \leq \sum_{j=1}^{m} \|x_j(k) - x\|^2. \tag{1.60}$$

By substituting (1.60) in relation (1.58), we obtain with probability 1 for all $k \geq 0$ and all $x \in X$,

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x\|^2 \mid \mathsf{F}_k\right]$$
$$\leq \sum_{j=1}^{m} \|x_j(k) - x\|^2 - \frac{2}{k+1}\left(f(x_{av}(k)) - f(x)\right) + r_k. \tag{1.61}$$

We will use relation (1.61) with $x = x_{av}(k)$ and $x = x^*$ for an arbitrary $x^* \in X^*$. In either case, we will use the almost supermartingale convergence result of Lemma 11. To use Lemma 11, we need to verify that $\sum_{k=0}^{\infty} r_k < \infty$ with probability 1. For this, we break down r_k in its terms, as follows:

$$r_k = \underbrace{2CD \sum_{i=1}^{m} \left|\mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)\chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]\right|}_{r_{1,k}}$$
$$+ 2\sqrt{m}CD \underbrace{\sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\left(\alpha_{i,k+1} - \frac{1}{(k+1)p_i}\right)^2 \chi_{\{i\in A_k\}} \mid \mathsf{F}_k\right]}}_{r_{2,k}}$$
$$+ \underbrace{\frac{2C\sqrt{\sum_{i=1}^{m}\frac{1}{p_i}}}{k+1} \sqrt{\sum_{i=1}^{m} \mathbb{E}\left[\|v_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right]}}_{r_{3,k}} + C^2 \alpha_{i,k}^2. \tag{1.62}$$

By Lemmas 8 and 9 we can see that

$$\sum_{k=0}^{\infty} r_{1,k} < \infty \quad \text{and} \quad \sum_{k=0}^{\infty} r_{2,k} < \infty \qquad \text{with probability 1.}$$

By Lemma 12 it follows that

$$\sum_{k=0}^{\infty} r_{3,k} < \infty \qquad \text{with probability 1,}$$

while by Lemma 9 we have

$$\sum_{k=0}^{\infty} \alpha_{i,k}^2 < \infty \qquad \text{with probability 1.}$$

Hence, from the preceding relations and relation (1.62) we have

$$\sum_{k=0}^{\infty} r_k < \infty \qquad \text{with probability 1.} \tag{1.63}$$

(i) We now use relation (1.61) with $x = x_{av}(k)$ and, thus, obtain with probability 1 for all $k \geq 0$,

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x_{av}(k)\|^2 \mid \mathsf{F}_k\right] \leq \sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2 + r_k. \tag{1.64}$$

By noting that $\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2 \leq \sum_{i=1}^{m} \|x_i(k+1) - y\|^2$ for any $y \in \mathbb{R}^n$, we have $\sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k+1)\|^2 \leq \sum_{i=1}^{m} \|x_i(k+1) - x_{av}(k)\|^2$. Thus relation (1.64) implies that with probability 1 for all $k \geq 0$,

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x_{av}(k+1)\|^2 \mid \mathsf{F}_k\right] \leq \sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2 + r_k. \tag{1.65}$$

By the almost supermartingale convergence of Lemma 11, we conclude that

$$\left\{\sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2\right\} \quad \text{is convergent with probability 1.}$$

Since by Lemma 12 $\liminf_{k\to\infty} \sum_{j=1}^{m} \left\|x_j(k) - x_{av}(k)\right\|^2 = 0$, we must have that

$$\lim_{k\to\infty} \sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2 = 0 \ \text{ with probability 1.} \tag{1.66}$$

(ii) Since the set X is compact, the optimal set X^* must be nonempty. Thus, by letting $x = x^*$ in relation (1.61) for any $x^* \in X^*$, we have with probability 1 for all $k \geq 0$ and all $x^* \in X^*$,

$$\sum_{i=1}^{m} \mathbb{E}\left[\|x_i(k+1) - x^*\|^2 \mid \mathsf{F}_k\right]$$
$$\leq \sum_{j=1}^{m} \|x_j(k) - x^*\|^2 - \frac{2}{k+1}\left(f(x_{av}(k)) - f^*\right) + r_k.$$

By the almost supermartingale convergence of Lemma 11, it follows that

$$\left\{\sum_{j=1}^{m} \|x_j(k) - x^*\|^2\right\} \text{ is convergent with probability 1 for all } x^* \in X^*, \tag{1.67}$$

$$\sum_{k=0}^{\infty} \frac{2}{k+1}\left(f(x_{av}(k)) - f^*\right) < \infty \quad \text{with probability 1.}$$

Thus, with probability 1 we must have

$$\liminf_{k\to\infty}\left(f(x_{av}(k)) - f^*\right) = 0.$$

Since X is bounded so is the sequence $\{x_{av}(k)\}$, and we can choose a subsequence $\{x_{av}(k_\ell)\}$ such that, with probability 1, $\lim_{\ell\to\infty} f(x_{av}(k_\ell)) = f^*$, $x_{av}(k_\ell) \to \tilde{x}$ where $\tilde{x} \in X$. By continuity of f, it follows that $f(\tilde{x}) = f^*$, i.e., $\tilde{x} \in X^*$ with probability 1.

Since we have shown that $\sum_{j=1}^{m} \|x_j(k) - x_{av}(k)\|^2 \to 0$ with probability 1 (see relation (1.66)), it follows that

$$\lim_{\ell\to\infty} \sum_{j=1}^{m} \|x_j(k_\ell) - \tilde{x}\|^2 = 0 \quad \text{with probability 1.}$$

By using relation (1.67) with $x^* = \tilde{x}$, we find that $\lim_{k\to\infty} \sum_{j=1}^{m} \|x_j(k) - \tilde{x}\|^2 = 0$ with probability 1. □

Theorem 4 shows that the random updates driven by an i.i.d. matrix sequence $\{W(k)\}$, where agents are updating using their own stepsizes (based on their frequency of updates) leads to a convergence to some consensual (random) solution of the problem. Theorem 4 includes as special cases the random gossip and random broadcast methods, by setting $W(k) = W_g(k)$ and $W(k) = W_b(k)$. It can accommodate other asynchronous approaches whose corresponding matrices $W(k)$ satisfy Assumption 4.

Prior work on convex multi-agent problems that explored the use of random gossip method, with frequency-based stepsize, can be found in [131], while the random broadcast method with such a stepsize has been proposed and studied in [100]. Error rate results for these methods with a constant stepsize can be found in [132] (see also [130]). Another related work on distributed convex optimization that considers random matrices is reference [86].

1.5 Literature Overview and New Research Directions

1.5.1 Literature on Consensus Algorithms

The consensus algorithms have drawn a renewed interest in more recent years, which was inspired by the work [60]. There has been a large body of work on the stability property and the convergence behavior of consensus dynamic in continuous-time [98, 118, 119, 134, 135], including the variants with switching network topology and time delays. Also, there are many papers investigating convergence of consensus algorithm in discrete-time, such as [21–24, 97, 99, 135, 180–182], where different aspects are also considered including time-switching network topology and asynchrony. The convergence rate properties have been studied in [122, 124] (for a more detailed overview see [102]).

Consensus has been used in rendezvous problems [78, 79], flocking [11], opinion dynamics [90, 91], and parameter estimation [65]. There are several Ph.D. theses that provide a comprehensive study of the averaging dynamics and its properties for deterministic and random matrices, including [12, 52, 120, 144, 155, 156]. Moreover, various communication effects have been investigated such as the effects of the delay [9, 10, 107], quantization effects [20, 25, 26, 43, 66, 71, 110], and the effects of link failures and noise [56, 63, 64, 125, 127, 128, 158]. A question of (non)existence of quadratic Lyapunov function for consensus dynamic has been investigated in [123, 163], while in [53, 54, 166] some cut-related properties have been explored.

Consensus algorithms implemented in a network using a gossip-based or a broadcast-based communications have been studied in [2, 3, 16, 82], while a different consensus algorithm (the so called push-sum method) has been considered in [7, 67]. Paper [93] explores a connection between consensus problems and potential games. An averaging dynamic for opinion spread has been proposed in [51] (leading to a formation of agents' groups with each group reaching its own consensus value).

The robustness of consensus including robustness to adversarial agents, faulty nodes, resilience and privacy preserving are studied in [29, 72, 149–151, 185]. Fast convergence of consensus algorithms in networks with quantized communications has been recently investigated in [4, 30]. There are also recent works on complex consensus [84], consensus stability [83, 85], and the use of semi-norms [32, 81] to study the stability. The best known convergence rate (in terms of its dependence on the number of agents) has been accomplished in a recent paper [121], which provides a consensus algorithm inspired by the fast Nesterov method [117].

There is a stream of work focused on the consensus problem over random graphs and weighted (random) averaging dynamics [5, 13, 27, 45, 50, 55–58, 126, 152, 153, 156, 157, 159–162, 164–166]. One may refer to survey paper [38] for a detailed account of gossip algorithms and their applications to signal processing in sensor networks. An application of an asynchronous gossip algorithm to the problem of spectral ranking has been explored in [14], while a nonlinear gossip is investigated in [95]. A monograph [138] discusses applications of consensus-based approaches to parameter estimation, learning and adaptation in networks, while monograph [102] discusses weighted-averaging algorithms for solving constrained and unconstrained consensus problems over time-varying (deterministic) graphs.

1.5.2 Literature on Distributed Methods for Optimization in Networks

1.5.2.1 Weighted Averaging-Based Approaches

The approaches that use consensus models with stochastic matrices are often referred to as weighted-averaging methods. Thus, both algorithms (1.9) and (1.34) are based on weighted-averaging methods, where the former uses deterministic weights while the latter employs random weights.

There are various extensions and modifications to the algorithm (1.9) which were developed over the past few years. The early work on consensus-based optimization can be found in [171], where the agents share a common objective function. The first work on distributed optimization in a network with agent based local objective functions can be found in [88, 106, 108]. In [106, 108] a slightly different algorithm has been considered (with a fixed stepsize), namely

$$x_i(k+1) = \sum_{j=1}^{m} w_{ij}(k)x_j(k) - \alpha_k d_i(k), \tag{1.68}$$

where $d_i(k)$ is a subgradient of $f_i(x)$ at $x = x_i(k)$. The convergence rate of this algorithm has been investigated in [105]. An extension of this algorithm to the case of quantized messages has been investigated in [109], while its implementation over random networks has been studied in [86]. In [89, 172] both of these alternative approaches, namely, algorithm (1.9) and (1.68) have been studied for distributed estimation.

A variant of the distributed optimization problem, where agents want to solve the problem of minimizing $\sum_{i=1}^{m} f_i(x)$ over $X = \cap_{i=1}^{m} X_i$, with each agent i handling its own function f_i and its constraint set X_i, has been considered in [111]. To solve the problem, a modification of the algorithm (1.9) has been proposed, where each agent i updates using the projection on its constraint set X_i, instead of the common set X. Thus, agent i updates assume the following form:

$$x_i(k+1) = \Pi_{X_i}\left[\sum_{j=1}^{m} w_{ij}(k)x_j(k) - \alpha_k g_i(k)\right],$$

where $g_i(k)$ is a subgradient of $f_i(x)$ at $x = \sum_{j=1}^{m} w_{ij}(k)x_j(k)$. This algorithm has been studied in [73–75] for synchronous updates over time-varying graphs and for gossip-based asynchronous updates over a static graph. A different variant of this algorithm (using the Laplacian formulation of the consensus problem) for distributed optimization with distributed constraints in noisy networks has been studied in [144–146]. Distributed algorithms for special quadratic convex problems arising in parameter estimation in sensor networks have been developed and studied in [28, 33, 88, 89, 172, 173].

The consensus-based algorithms for other types of network objective functions have been considered in [133]. A Bregman-distance based distributed algorithm has been developed in [147], as well as consensus-based algorithms for solving certain min-max problems. Distributed algorithms, both synchronous and asynchronous, for solving special type of games (aggregative games) have been studied in [68].

A distributed dual Nesterov algorithm has been proposed in [40], while a distributed algorithm using the gradient differences has been recently proposed in [142]. A distributed algorithm that is based on the idea of preserving an optimality condition at every stage of the algorithm has been proposed in [92]. Distributed convex optimization algorithms for weight-balanced directed graphs have been investigated in continuous-time [48].

A different type of a distributed algorithm for convex optimization has been proposed in [77], where each agent keeps an estimate for all agents' decisions. This algorithm solves a problem where the agents have to minimize a global cost function $f(x_1, \ldots, x_m)$ while each agent i can control only its variable x_i. The algorithm of [77] has been recently extended to the online optimization setting in [76, 113].

Distributed algorithms using augmented Lagrangian approach with gossip-type communications have been studied in [61], while accelerated versions of distributed gradient methods have been proposed and studied in [62].

A consensus-based algorithm for solving problems with a separable constraint structure and the use of primal-dual distributed methods has been studied in [147, 189, 190], while a distributed primal-dual approach with perturbations has been explored in [31]. Work [176] provides algorithms for centralized and distributed convex optimization from control perspective, while [175] considers an event-triggered distributed optimization for sensor networks. In [19], a distributed simplex algorithm has been developed for linear programming problems, while a Newton-

Raphson consensus-based method has been proposed in [186] for distributed convex problems.

All of the prior work relies on the use of state-independent weights, i.e., the weights that do not depend on agents' iterates. A consensus-based algorithm employing state-dependent weights have been proposed and analyzed in [87].

The application of distributed methods to hypothesis testing problems in graphs has recently attracted attention, resulting in a stream of papers [70, 112, 139, 140]. While these works deal mainly with finitely many hypothesis, a recent paper [114] extends the framework to the case of infinitely many hypotheses. Paper [116] considers several algorithms that use different types of consensus models, namely, weighted-averaging as well as push-sum models, which are discussed in the next paragraph with some more details.

1.5.2.2 Push-Sum Based Approaches

Another class of distributed algorithms has recently been developed that employs a different type of consensus strategy known as push-sum model. It has also been referred to as doubly-linear iteration or ratio consensus algorithm, due to its form which involves ratio of two variables that evolve according to the same linear dynamics, but differ in the choice of the initial point. This algorithm was originally proposed in [67] for consensus problem over a static network (in a random gossip-based form), and has been recently investigated in [39] in a deterministic setting. It has been extended to time-varying networks in [7].

The first work that has employed the push-sum consensus model to develop distributed optimization methods is [169], which has been further investigated in [167, 168, 170]. This work has been focused on static graphs and it has been proposed as an alternative to the algorithm based on weighted averages in order to eliminate deadlocks and synchronization issues among others. The prior work also offers a push-sum algorithm that can deal with constraints by using Nesterov dual-averaging approach. Recently, this push-sum consensus-based algorithm has been extended to a *subgradient-push* algorithm in [103, 104] that can deal with convex optimization problems over time-varying directed graphs. More recently, the paper [148] has extended the push-sum algorithm to a larger class of distributed algorithms that are applicable to nonconvex objective functions, convex constraint sets, and time-varying graphs (see the subsequent paragraph on new directions for more details).

1.5.2.3 ADMM Based Approaches

Another approach for solving problem (1.1) in a distributed fashion over a static network can be constructed by using the Alternate Direction Multiplier Method (ADMM). This method is based on an equivalent formulation of consensus constraints. Unlike consensus-based (sub)-gradient method (1.9) which operates in the space of the primal-variables, the ADMM solves a corresponding Lagrangian dual problem (obtained by relaxing the equality constraints that are associated with consensus requirement). Just as any dual method, the ADMM is applicable to problems where the structure of the objective functions f_i is simple enough so that the ADMM updates can be executed efficiently. The algorithm has potential of solving the problem with geometric convergence rate, which requires global knowledge of some parameters including eigenvalues of a weight matrix associated with the graph. A recent survey on the ADMM and its various applications is given in [17].

The first works to address the development of distributed ADMM over a network are [80, 177, 178], while a linear convergence rate of the ADMM has been shown in [141]. In [1] the ADMM with linearization has been proposed for special composite optimization problems over graphs.

1.5.2.4 New Directions

Within the area of distributed (multi-agent) optimization over networks, loosely speaking, two main directions of research can be noted, namely toward efficiency improvements (to develop "fast" distributed algorithms whose performance can meet the best known performance guarantees in a centralized setting) and toward addressing non-convex problems over networks.

In the domain of efficiency improvements, there are new approaches that are rooted in the idea of distributing the optimality conditions for multi-agent problems and the approaches that investigate gradient-consensus models. The consensus-based primal algorithms with a constant step-size are not likely to reach geometric convergence rate even when the overall objective function is strongly convex. Papers [142, 143] develop the algorithm EXTRA and its proximal-gradient variant by employing a carefully selected gradient-difference scheme to cancel out the steady-state error that occurs in some distributed methods with a constant stepsize [106, 108]. The EXTRA algorithm converges at an $o(1/k)$ rate when the objective network function is convex, and it has a geometric rate when the objective function is strongly convex. These developments have considered a static and undirected graph.

References [179, 187] combine EXTRA with the push-sum protocol of [67] to produce DEXTRA (Directed Extra-Push) algorithm for optimization over a directed graph. It has been shown that DEXTRA converges at a geometric (R-linear) rate for a strongly convex objective function, but it requires a careful stepsize selection. It has been noted in [179] that the feasible region of stepsizes which guarantees this convergence rate can be empty in some cases.

The work [183, 184] utilizes an adapt-then-combine (ATC) strategy [137, 138] of dynamic weighted-average consensus approach [188] to develop a distributed algorithm, termed Aug-DGM algorithm. This algorithm can be used over static directed or undirected graphs (but requires doubly stochastic matrices). The most interesting aspect of the Aug-DGM algorithm is that it can produce convergent iterates even when different agents use different (constant) stepsizes.

Simultaneously and independently, the idea of tracking the gradient averages through the use of consensus has been proposed in [184] for convex unconstrained problems and in [35] for non-convex problems with convex constraints. The work in [35–37] develops a large class of distributed algorithms, referred to as NEXT, which utilizes various “function-surrogate modules” thus providing a great flexibility in its use and rendering a new class of algorithms that subsumes many of the existing distributed algorithms. The ideas in [36, 37] and in [183] have also been proposed independently, with the former preceding the latter. The algorithm framework of [35–37] is applicable to nonconvex problems with convex constraint sets over time-varying graphs, but requires the use of doubly stochastic matrices. This assumption was recently removed in [148] by using column-stochastic matrices, which are more general than the degree-based column-stochastic matrices of the push-sum method. Simultaneously and independently, the papers [37] and [154] have appeared to treat nonconvex problems over graphs. The work in [154] proposes and analyzes a distributed gradient method based on the push-sum consensus in deterministic and stochastic setting for unconstrained problems.

The idea of using a consensus process to track gradients has also been recently used in [115] to develop a distributed algorithm, referred to as DIGing, with a geometric convergence rate over time-varying graphs. This is the first paper to establish such a rate for a consensus-based algorithms for convex optimization over time-varying graphs. We note that the algorithm uses a fixed stepsize, and the rate result is applicable to the problems with a strongly convex smooth objective function.

Acknowledgment The work presented in this chapter has been partially supported by the NSF grant DMS-1312907.

References

1. N.S. Aybat, Z. Wang, T. Lin, S. Ma, Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. IEEE Trans. Autom. Control **63**(1), 5–20 (2018)
2. T.C. Aysal, M.E. Yildiz, A.D. Sarwate, A. Scaglione, Broadcast gossip algorithms: design and analysis for consensus, in *Proceedings of the 47th IEEE Conference on Decision and Control*, pp. 4843–4848 (2008)
3. T.C. Aysal, M.E. Yildriz, A.D. Sarwate, A. Scaglione, Broadcast gossip algorithms for consensus. IEEE Trans. Signal Process. **57**, 2748–2761 (2009)
4. T. Başar, S.R. Etesami, A. Olshevsky, Fast convergence of quantized consensus using Metropolis chains, in *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp. 1330–1334 (2014)
5. D. Bajović, J. Xavier, J.M.F. Moura, B. Sinopoli, Consensus and products of random stochastic matrices: exact rate for convergence in probability. IEEE Trans. Signal Process. **61**(10), 2557–2571 (2013)

6. B.A. Bash, P.J. Desnoyers, Exact distributed Voronoi cell computation in sensor networks, in *IPSN'07: Proceedings of the 6th International Conference on Information Processing in Sensor Networks* (ACM Press, New York, 2007), pp. 236–243
7. F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis, M. Vetterli, Weighted gossip: distributed averaging using non-doubly stochastic matrices, in *Proceedings of the 2010 IEEE International Symposium on Information Theory*, pp. 1753–1757 (2010)
8. P. Billingsley, *Probability and Measure*, 3rd edn. (Wiley, Hoboken, 1995)
9. P.A. Bliman, G. Ferrari-Trecate, Average consensus problems in networks of agents with delayed communications. Automatica **44**(8), 1985–1995 (2008)
10. P.A. Bliman, A. Nedich, A. Ozdaglar, Rate of convergence for consensus with delays, in *Proceedings of the 47th IEEE Conference on Decision and Control*, pp. 4849–4854 (2008)
11. V.D. Blondel, J.M. Hendrickx, A. Olshevsky, J.N. Tsitsiklis, Convergence in multiagent coordination, consensus, and flocking, in *Proceedings of the IEEE Conference on Decision and Control*, pp. 2996–3000 (2005)
12. S. Bolouki, Linear consensus algorithms: structural properties and connections with Markov chains, Ph.D. thesis, Department of Electrical Engineering, École Polytechnique de Montréal, 2014
13. S. Bolouki, R.P. Malhamé, Theorems about ergodicity and class-ergodicity of chains with applications in known consensus models, in *Proceedings of the 50th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1425–1431 (2012)
14. V.S. Borkar, R. Makhijani, R. Sundaresan, Asynchronous gossip for averaging and spectral ranking. IEEE J. Sel. Top. Sign. Proces. **8**(4), 703–716 (2014)
15. V. Borkar, P.P. Varaiya, Asymptotic agreement in distributed estimation. IEEE Trans. Autom. Control **27**(3), 650–655 (1982)
16. S. Boyd, A. Ghosh, B. Prabhakar, D. Shah, Gossip algorithms: design, analysis, and applications, in *Proceedings of IEEE INFOCOM*, vol. 3, pp. 1653–1664 (2005)
17. S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2010)
18. F. Bullo, J. Cortés, S. Martínez, *Distributed Control of Robotic Networks: A Mathematical Approach to Motion Coordination Algorithms* (Princeton University Press, Princeton, 2009)
19. M. Bürger, G. Notarstefano, F. Bullo, F. Allgöwer, A distributed simplex algorithm for degenerate linear programs and multi-agent assignments. Automatica **48**(9), 2298–2304 (2012)
20. K. Cai, H. Ishii, Quantized consensus and averaging on gossip digraphs. IEEE Trans. Autom. Control **56**(9), 2087–2100 (2011)
21. M. Cao, D.A. Spielman, A.S. Morse, A lower bound on convergence of a distributed network consensus algorithm, in *Proceedings of IEEE CDC*, pp. 2356–2361 (2005)
22. M. Cao, A.S. Morse, B.D.O. Anderson, Agreeing asynchronously. IEEE Trans. Autom. Control **53**(8), 1826–1838 (2008)
23. M. Cao, A.S. Morse, B.D.O. Anderson, Reaching a consensus in a dynamically changing environment: convergence rates, measurement delays, and asynchronous events. SIAM J. Control. Optim. 47(2), 601–623 (2008)
24. M. Cao, A.S. Morse, B.D.O. Anderson, Reaching a consensus in a dynamically changing environment: a graphical approach. SIAM J. Control. Optim. **47**(2), 575–600 (2008)
25. R. Carli, F. Fagnani, A. Speranzon, S. Zampieri, Communication constraints in coordinated consensus problem, in *Proceedings of IEEE American Control Conference*, pp. 4189–4194 (2006)
26. R. Carli, F. Fagnani, P. Frasca, T. Taylor, S. Zampieri, Average consensus on networks with transmission noise or quantization, in *Proceedings of European Control Conference*, pp. 1852–1857 (2007)
27. R. Carli, F. Fagnani, P. Frasca, S. Zampieri, Gossip consensus algorithms via quantized communication. Automatica **46**(1), 70–80 (2010)

28. F. Cattivelli, A.H. Sayed, Diffusion LMS strategies for distributed estimation. IEEE Trans. Signal Process. **58**(3), 1035–1048 (2010)
29. M. El Chamie, T. Başar, Optimal strategies for dynamic weight selection in consensus protocols in the presence of an adversary, in *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp. 735–740 (2014)
30. M. El Chamie, J. Liu, T. Başar, Design and analysis of distributed averaging with quantized communication, in *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp. 3860–3865 (2014)
31. T-H. Chang, A. Nedich, A. Scaglione, Distributed constrained optimization by consensus-based primal-dual perturbation method. IEEE Trans. Autom. Control **59**(6), 1524–1538 (2014)
32. B. Charron-Bost, Orientation and connectivity based criteria for asymptotic consensus (2013). Available at http://arxiv.org/abs/1303.2043
33. J. Chen, A.H. Sayed, Diffusion adaptation strategies for distributed optimization and learning over networks. IEEE Trans. Signal Process. **60**(8), 4289–4305 (2012)
34. M.H. DeGroot, Reaching a consensus. J. Am. Stat. Assoc. **69**(345), 118–121 (1974)
35. P. Di Lorenzo, G. Scutari, Distributed nonconvex optimization over networks, in *Proceedings of IEEE International Conference on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2015), Dec. 13–16, 2015, Cancun, Mexico* (2015)
36. P. Di Lorenzo, G. Scutari, Distributed nonconvex optimization over time-varying networks, in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 16), March 20–25, 2016, Shanghai, China* (2016)
37. P. Di Lorenzo, G. Scutari, NEXT: in-network nonconvex optimization. IEEE Trans. Signal Inf. Process. Netw. **2**(2), 120–136 (2016)
38. A.G. Dimakis, S. Kar, J.M.F. Moura, M.G. Rabbat, A. Scaglione, Gossip algorithms for distributed signal processing. Proc. IEEE **98**(11), 1847–1864 (2010)
39. A.D. Dominguez-Garcia, C. Hadjicostis, Distributed strategies for average consensus in directed graphs, in *Proceedings of the IEEE Conference on Decision and Control*, pp. 2124–2129 (2011)
40. J.C. Duchi, A. Agarwal, M.J. Wainwright, Dual averaging for distributed optimization: convergence analysis and network scaling. IEEE Trans. Autom. Control **57**(3), 592 –606 (2012)
41. R. Dudley, *Real Analysis and Probability* (Cambridge University Press, Cambridge, 2002)
42. D. Easley, J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World* (Cambridge University Press, New York, 2010)
43. S.R. Etesami, T. Başar, Convergence time for unbiased quantized consensus, in *Proceedings of the 52nd IEEE Conference on Decision and Control*, pp. 6190–6195 (2013)
44. F. Facchinei, J-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, vols. I and II (Springer, New York, 2003)
45. F. Fagnani, S. Zampieri, Randomized consensus algorithms over large scale networks. IEEE J. Sel. Areas Commun. **26**(4), 634–649 (2008)
46. M. Fiedler, *Special Matrices and Their Applications in Numerical Mathematics*, 2nd edn. (Dover Publications, New York, 2008)
47. E. Ghadimi, Accelerating convergence of large-scale optimization algorithms, Ph.D. thesis, School of Electrical Engineering, Department of Automatic Control, KTH Royal Institute of Technology, 2015
48. B. Gharesifard, J. Cortés, Distributed continuous-time convex optimization on weight-balanced digraphs. Eur. J. Control. **18**(6), 539–557 (2012)
49. S. Goyal, *Connections: An Introduction to the Economics Networks* (Princeton University Press, Princeton, 2007)
50. Y. Hatano, A.K. Das, M. Mesbahi, Agreement in presence of noise: pseudogradients on random geometric networks, in *Proceedings of the 44th IEEE Conference on Decision and Control, and European Control Conference*, pp. 6382–6387 (2005)

51. R. Hegselmann, U. Krause, Opinion dynamics and bounded confidence models, analysis, and simulation. J. Artif. Soc. Soc. Simul. **5** (2002)
52. J.M. Hendrickx, Graphs and networks for the analysis of autonomous agent systems, Ph.D. thesis, Université Catholique de Louvain, 2008
53. J.M. Hendrickx, J.N. Tsitsiklis, A new condition for convergence in continuous-time consensus seeking systems, in *Proceedings of the 50th IEEE Conference on Decision and Control*, pp. 5070–5075 (2011)
54. J.M. Hendrickx, J.N. Tsitsiklis, Convergence of type-symmetric and cut-balanced consensus seeking systems. IEEE Trans. Autom. Control **58**(1), 214–218 (2013)
55. M. Huang, J.H. Manton, Stochastic approximation for consensus seeking: mean square and almost sure convergence, in *Proceedings of the 46th IEEE Conference on Decision and Control*, pp. 306–311 (2007)
56. M. Huang, J.H. Manton, Stochastic lyapunov analysis for consensus algorithms with noisy measurements, in *Proceedings of the 2007 IEEE American Control Conference*, pp. 1419–1424 (2007)
57. M. Huang, J.H. Manton, Stochastic consensus seeking with measurement noise: convergence and asymptotic normality, in *Proceedings of the 2008 IEEE American Control Conference*, pp. 1337–1342 (2008)
58. M. Huang, J.H. Manton, Coordination and consensus of networked agents with noisy measurements: stochastic algorithms and asymptotic behavior. SIAM J. Control. Optim. **48**(1), 134–161 (2009)
59. M.O. Jackson, *Social and Economic Networks* (Princeton University Press, Princeton, 2008)
60. A. Jadbabaie, J. Lin, A.S. Morse, Coordination of groups of mobile autonomous agents using nearest neighbor rules. IEEE Trans. Autom. Control **48**(6), 988–1001 (2003)
61. D. Jakovetić, J. Xavier, J.M.F. Moura, Cooperative convex optimization in networked systems: augmented Lagrangian algorithms with directed gossip communication. IEEE Trans. Signal Process. **59**(8), 3889–3902 (2011)
62. D. Jakovetić, J. Xavier, J.M.F. Moura, Fast distributed gradient methods. IEEE Trans. Autom. Control **59**(5), 1131–1146 (2014)
63. S. Kar, J.M.F. Moura, Distributed consensus algorithms in sensor networks: link failures and channel noise. IEEE Trans. Signal Process. **57**(1), 355–369 (2009)
64. S. Kar, J.M.F. Moura, Distributed consensus algorithms in sensor networks: quantized data and random link failures. IEEE Trans. Signal Process. **58**(3), 1383 –1400 (2010)
65. S. Kar, J.M.F. Moura, Convergence rate analysis of distributed gossip (linear parameter) estimation: fundamental limits and tradeoffs. IEEE J. Sel. Top. Sign. Proces. **5**(4), 674–690 (2011)
66. A. Kashyap, T. Basar, R. Srikant, Quantized consensus. Automatica **43**(7), 1192–1203 (2007)
67. D. Kempe, A. Dobra, J. Gehrke, Gossip-based computation of aggregate information, in *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pp. 482–491 (2003)
68. J. Koshl, A. Nedich, U.V. Shanbhag, Distributed algorithms for aggregative games on graphs. Oper. Res. **64**, 680–704 (2016)
69. K. Kvaternik, Decentralized coordination control for dynamic multiagent systems, Ph.D. thesis, Department of Electrical and Computer Engineering, University of Toronto, 2014
70. A. Lalitha, T. Javidi, A. Sarwate, Social learning and distributed hypothesis testing, in *Proceedings of 2014 IEEE International Symposium on Information Theory*, pp. 551–555 (2014)
71. J. Lavaei, R.M. Murray, Quantized consensus by means of gossip algorithm. IEEE Trans. Autom. Control **57**(1), 19–32 (2012)
72. H. LeBlanc, H. Zhang, X. Koutsoukos, S. Sundaram, Resilient asymptotic consensus in robust networks. IEEE J. Sel. Areas Commun.: Spec. Issue In-Netw. Comput. **31**(4), 766–781 (2013)
73. S. Lee, Optimization over networks: efficient algorithms and analysis, Ph.D. thesis, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 2013

74. S. Lee, A. Nedich, Distributed random projection algorithm for convex optimization. IEEE J. Sel. Top. Sign. Proces. **48**(6), 988–1001 (2012)
75. S. Lee, A. Nedich, Asynchronous gossip-based random projection algorithms over networks. IEEE Trans. Autom. Control **61**(4), 953–968 (2016)
76. S. Lee, A. Nedich, M. Raginsky, Coordinate dual averaging for decentralized online optimization with non-separable global objectives. IEEE Trans. Control Netw. Syst. **5**(1), 34–44 (2018)
77. N. Li, J.R. Marden, Designing games for distributed optimization. IEEE J. Sel. Top. Sign. Proces. **7**(2), 230–242 (2013)
78. J. Lin, A.S. Morse, B.D.O. Anderson, The multi-agent rendezvous problem, in *Proceedings of the 42nd IEEE Conference on Decision and Control*, pp. 1508–1513 (2003)
79. J. Lin, A.S. Morse, B.D.O. Anderson, The multi-agent rendezvous problem - the asynchronous case, in *Proceedings of the 43rd IEEE Conference on Decision and Control*, pp. 1926–1931 (2004)
80. Q. Ling, A. Ribeiro, Decentralized dynamic optimization through the alternating direction method of multiplier. IEEE Trans. Signal Process. **62**(5), 1185–1197 (2014)
81. J. Liu, A.S. Morse, B.D.O. Anderson, C. Yu, Contractions for consensus processes, in *Proceedings of the 50th IEEE Conference on Decision and Control*, pp. 1974–1979 (2011)
82. J. Liu, S. Mou, A.S. Morse, B.D.O. Anderson, C. Yu, Deterministic gossiping. Proc. IEEE **99**(9), 1505–1524 (2011)
83. J. Liu, A.S. Morse, A. Nedich, T. Başar, Internal stability of linear consensus processes, in *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp. 922–927 (2014)
84. J. Liu, A. Nedich, T. Başar, Complex constrained consensus, in *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp. 1464–1469 (2014)
85. J. Liu, A.S. Morse, A. Nedich, T. Başar, Stability of a distributed algorithm for solving linear algebraic equations, in *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp. 3707–3712 (2014)
86. I. Lobel, A. Ozdaglar, Distributed subgradient methods for convex optimization over random networks. IEEE Trans. Autom. Control **56**(6), 1291–1306 (2011)
87. I. Lobel, A. Ozdaglar, D. Feijer, Distributed multi-agent optimization with state-dependent communication. Math. Program. **129**(2), 255–284 (2011)
88. C. Lopes, A.H. Sayed, Distributed processing over adaptive networks, in *Adaptive Sensor Array Processing Workshop, MIT Lincoln Laboratory, MA*, pp. 1–5 (2006)
89. C.G. Lopes, A.H. Sayed, Diffusion least-mean squares over adaptive networks: formulation and performance analysis. IEEE Trans. Signal Process. **56**(7), 3122–3136 (2008)
90. J. Lorenz, A stabilization theorem for continuous opinion dynamics. Phys. A Stat. Mech. Appl. **355**, 217–223 (2005)
91. J. Lorenz, Repeated averaging and bounded confidence-modeling, analysis and simulation of continuous opinion dynamics, Ph.D. thesis, Universität Bremen, 2007
92. J. Lu, C.Y. Tang, Zero-gradient-sum algorithms for distributed convex optimization: the continuous-time case. IEEE Trans. Autom. Control **57**(9), 2348–2354 (2012)
93. J.R. Marden, G. Arslan, J.S. Shamma, Connections between cooperative control and potential games illustrated on the consensus problem, in *Proceedings of the European Control Conference*, pp. 4604–4611 (2007)
94. A. Martinoli, F. Mondada, G. Mermoud, N. Correll, M. Egerstedt, A. Hsieh, L. Parker, K. Stoy, *Distributed Autonomous Robotic Systems*. Springer Tracts in Advanced Robotics (Springer, Berlin, 2013)
95. A.S. Mathkar, V.S. Borkar, Nonlinear gossip (2014). Preprint
96. M. Mesbahi, M. Egerstedt, *Graph Theoretic Methods for Multiagent Networks* (Princeton University Press, Princeton, 2010)
97. C.C. Moallemi, B. Van Roy, Consensus propagation. IEEE Trans. Inf. Theory **52**(11), 4753–4766 (2006)
98. L. Moreau, Stability of continuous-time distributed consensus algorithms, in *Proceedings of the 43rd IEEE Conference on Decision and Control*, pp. 3998–4003 (2004)

99. L. Moreau, Stability of multiagent systems with time-dependent communication links. IEEE Trans. Autom. Control **50**(2), 169–182 (2005)
100. A. Nedich, Asynchronous broadcast-based convex optimization over a network. IEEE Trans. Autom. Control **56**(6), 1337–1351 (2011)
101. A. Nedich, Random projection algorithms for convex minimization problems. Math. Program. Ser. B **129**, 225–253 (2011)
102. A. Nedich, *Convergence Rate of Distributed Averaging Dynamics and Optimization in Networks*. Found. Trends Syst. Control **2** (2015)
103. A. Nedich, A. Olshevsky, Distributed optimization over time-varying directed graphs. IEEE Trans. Autom. Control **60**(3), 601–615 (2015)
104. A. Nedich, A. Olshevsky, Stochastic gradient-push for strongly convex functions on time-varying directed graphs. IEEE Trans. Autom. Control **61**(12), 3936–3947 (2016)
105. A. Nedich, A. Ozdaglar, On the rate of convergence of distributed subgradient methods for multi-agent optimization, in *Proceedings of the 46th IEEE Conference on Decision and Control*, pp. 4711–4716 (2007)
106. A. Nedich, A. Ozdaglar, Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control **54**(1), 48–61 (2009)
107. A. Nedich, A. Ozdaglar, Convergence rate for consensus with delays. J. Glob. Optim. **47**(3), 437–456 (2010)
108. A. Nedich, A. Ozdaglar, Cooperative distributed multi-agent optimization, in *Convex Optimization in Signal Processing and Communications*, ed. by Y. Eldar, D. Palomar (Cambridge University Press, Cambridge, 2010), pp. 340–386
109. A. Nedich, A. Olshevsky, A. Ozdaglar, J.N. Tsitsiklis, Distributed subgradient methods and quantization effects, in *Proceedings of 47th IEEE Conference on Decision and Control*, pp. 4177–4184 (2008)
110. A. Nedich, A. Olshevsky, A. Ozdaglar, J.N. Tsitsiklis, On distributed averaging algorithms and quantization effects. IEEE Trans. Autom. Control **54**(11), 2506–2517 (2009)
111. A. Nedich, A. Ozdaglar, P.A. Parrilo, Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control **55**(4), 922–938 (2010)
112. A. Nedich, A. Olshevsky, C.A. Uribe, Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs, in *Proceedings of American Control Conference 2015*, pp. 5884–5889 (2015)
113. A. Nedich, S. Lee, M. Raginsky, Decentralized online optimization with global objectives and local communication, in *Proceedings of the 2016 American Control Conference (ACC), Boston, MA, July 6–8, 2016*, pp. 4497–4503 (2016)
114. A. Nedich, A. Olshevsky, C.A. Uribe, Network independent rates in distributed learning, in *Proceedings of the 2016 American Control Conference (ACC), Boston, MA, July 6–8, 2016*, pp. 1072–1077 (2016)
115. A. Nedich, A. Olshevsky, W. Shi, Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM J. Optim. **27**(4), 2597–2633 (2017)
116. A. Nedich, A. Olshevsky, C.A. Uribe, Fast convergence rates for distributed non-bayesian learning. IEEE Trans. Autom. Control **62**(11), 5538–5553 (2017)
117. Yu.E. Nesterov, A method of solving a convex programming problem with convergence rate of $o(1/k^2)$. Soviet Math. Dokl. **27**(2), 372–376 (1983)
118. R. Olfati-Saber, R.M. Murray, Consensus problems in networks of agents with switching topology and time-delays. IEEE Trans. Autom. Control **49**(9), 1520–1533 (2004)
119. R. Olfati-Saber, J.A. Fax, R.M. Murray, Consensus and cooperation in networked multi-agent systems. Proc. IEEE **95**(1), 215–233 (2007)
120. A. Olshevsky, Efficient information aggregation for distributed control and signal processing, Ph.D. thesis, MIT, 2010
121. A. Olshevsky, Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control. SIAM J. Control. Optim. **55**(6), 3990–4014 (2017)
122. A. Olshevsky, J.N. Tsitsiklis, Convergence rates in distributed consensus averaging, in *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 3387–3392 (2006)

123. A. Olshevsky, J.N. Tsitsiklis, On the nonexistence of quadratic lyapunov functions for consensus algorithms. IEEE Trans. Autom. Control **53**(11), 2642–2645 (2008)
124. A. Olshevsky, J.N. Tsitsiklis, Convergence speed in distributed consensus and averaging. SIAM J. Control. Optim. **48**(1), 33–55 (2009)
125. S. Patterson, B. Bamieh, Distributed consensus with link failures as a structured stochastic uncertainty problem, in *Proceedings of the 46th Allerton Conference on Communication, Control, and Computing*, pp. 623–627 (2008)
126. S. Patterson, B. Bamieh, Convergence rates of consensus algorithms in stochastic networks, in *Proceedings of the 49th IEEE Conference on Decision and Control*, pp. 6608–6613 (2010)
127. S. Patterson, B. Bamieh, A. El Abbadi, Distributed average consensus with stochastic communication failures, in *Proceedings of the 46th IEEE Conference on Decision and Control*, pp. 4215–4220 (2007)
128. S. Patterson, B. Bamieh, A. El Abbadi, Distributed average consensus with stochastic communication failures. IEEE Trans. Signal Process. **57**, 2748–2761 (2009)
129. B.T. Polyak, *Introduction to Optimisation* (Optimization Software, Inc., New York, 1987)
130. S.S. Ram, Distributed optimization in multi-agent systems: applications to distributed regression, Ph.D. thesis, University of Illinois at Urbana-Champaign, 2009
131. S.S. Ram, A. Nedich, V.V. Veeravalli, Asynchronous gossip algorithms for stochastic optimization, in *Proceedings of the 48th IEEE Conference on Decision and Control (CDC), Shanghai, China*, pp. 3581–3586 (2009)
132. S.S. Ram, A. Nedich, V.V. Veeravalli, Asynchronous gossip algorithms for stochastic optimization: constant stepsize analysis, in *Recent Advances in Optimization and its Applications in Engineering*. Volume of the 14th Belgian-French-German Conference on Optimization (BFG), pp. 51–60 (2010)
133. S.S. Ram, A. Nedich, V.V. Veeravalli, A new class of distributed optimization algorithms: application to regression of distributed data. Optim. Methods Softw. **27**(1), 71–88 (2012)
134. W. Ren, Consensus seeking in multi-vehicle systems with a time-varying reference state, in *Proceedings of IEEE American Control Conference*, pp. 717–722 (2007)
135. W. Ren, R. Beard, Consensus seeking in multi-agent systems under dynamically changing interaction topologies. IEEE Trans. Autom. Control **50**(5), 655–661 (2005)
136. H. Robbins, D. Siegmund, A convergence theorem for nonnegative almost supermartingales and some applications, in *Optimizing Methods in Statistics* (Academic, New York, 1971), pp. 233–257
137. A.H. Sayed, Diffusion adaptation over networks, in *Academic Press Library in Signal Processing*, vol. 3, pp. 323–454 (2013)
138. A.H. Sayed, Adaptation, learning, and optimization over networks. Found. Trends Mach. Learn. **7** (2014)
139. S. Shahrampour, A. Jadbabaie, Exponentially fast parameter estimation in networks using distributed dual averaging, in *Proceedings of the 52nd IEEE Conference on Decision and Control*, pp. 6196–6201 (2013)
140. S. Shahrampour, A. Rakhlin, A. Jadbabaie, Distributed detection: finite-time analysis and impact of network topology. IEEE Trans. Autom. Control **61**(11), 3256–3268 (2016)
141. W. Shi, Q. Ling, K. Yuan, G. Wu, W. Yin, On the linear convergence of the admm in decentralized consensus optimization. IEEE Trans. Signal Process. **62**, 1750–1761 (2014)
142. W. Shi, Q. Ling, G. Wu, W. Yin, Extra: an exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. **25**, 944–966 (2015)
143. W. Shi, Q. Ling, G. Wu, W. Yin, A proximal gradient algorithm for decentralized composite optimization. IEEE Trans. Signal Process. **63**, 6013–6023 (2015)
144. K. Srivastava, Distributed optimization with applications to sensor networks and machine learning, Ph.D. thesis, University of Illinois at Urbana-Champaign, Industrial and Enterprise Systems Engineering, 2011
145. K. Srivastava, A. Nedich, Distributed asynchronous constrained stochastic optimization. IEEE J. Sel. Top. Sign. Proces. **5**(4), 772–790 (2011)

146. K. Srivastava, A. Nedich, D.M. Stipanović, Distributed constrained optimization over noisy networks, in *Proceedings of the 49th IEEE Conference on Decision and Control (CDC)*, pp. 1945–1950 (2010)
147. K. Srivastava, A. Nedich, D. Stipanovic, Distributed Bregman-distance algorithms for min-max optimization, in *Agent-Based Optimization*. Springer Studies in Computational Intelligence (SCI) (Springer, Berlin, 2013), pp. 143–174
148. Y. Sun, G. Scutari, D. Palomar, Distributed nonconvex multiagent optimization over time-varying networks, in *Proceedings of the 50th Asilomar Conference on Signals, Systems, and Computers 2016* (2016). Available on arXiv at http://arxiv.org/abs/1607.00249
149. S. Sundaram, C.N. Hadjicostis, Distributed function calculation and consensus using linear iterative strategies. IEEE J. Sel. Areas Commun.: Issue Control Commun. **26**(4), 650–660 (2008)
150. S. Sundaram, C.N. Hadjicostis, Distributed function calculation via linear iterative strategies in the presence of malicious agents. IEEE Trans. Autom. Control **56**(7), 1495–1508 (2011)
151. S. Sundaram, S. Revzen, G.J. Pappas, A control-theoretic approach to disseminating values and overcoming malicious links in wireless networks. Automatica **48**(11), 2894–2901 (2012)
152. A. Tahbaz-Salehi, A. Jadbabaie, A necessary and sufficient condition for consensus over random networks. IEEE Trans. Autom. Control **53**(3), 791–795 (2008)
153. A. Tahbaz-Salehi, A. Jadbabaie, Consensus over ergodic stationary graph processes. IEEE Trans. Autom. Control **55**(1), 225–230 (2010)
154. T. Tatarenko, B. Touri, Non-convex distributed optimization. IEEE Trans. Autom. Control **62**(8), 3744–3757 (2017)
155. B. Touri, Product of random stochastic matrices and distributed averaging, Ph.D. thesis, University of Illinois at Urbana-Champaign, Industrial and Enterprise Systems Engineering, 2011
156. B. Touri, *Product of Random Stochastic Matrices and Distributed Averaging* (Springer, Berlin, 2012)
157. B. Touri, C. Langbort, On endogenous random consensus and averaging dynamics. IEEE Trans. Control Netw. Syst. **1**(3), 241–248 (2014)
158. B. Touri, A. Nedich, Distributed consensus over network with noisy links, in *Proceedings of the 12th International Conference on Information Fusion*, pp. 146–154 (2009)
159. B. Touri, A. Nedich, Approximation and limiting behavior of random models, in *Proceedings of 49th IEEE Conference on Decision and Control*, pp. 2656–2663 (2010)
160. B. Touri, A. Nedich, When infinite flow is sufficient for ergodicity, in *Proceedings of 49th IEEE Conference on Decision and Control*, pp. 7479–7486 (2010)
161. B. Touri, A. Nedich, Alternative characterization of ergodicity for doubly stochastic chains, in *Proceedings of the 50th IEEE Conference on Decision and Control*, pp. 5371–5376 (2011)
162. B. Touri, A. Nedich, On ergodicity, infinite flow and consensus in random models. IEEE Trans. Autom. Control **56**(7), 1593–1605 (2011)
163. B. Touri, A. Nedich, On existence of a quadratic comparison function for random weighted averaging dynamics and its implications, in *Proceedings of the 50th IEEE Conference on Decision and Control*, pp. 3806–3811 (2011)
164. B. Touri, A. Nedich, On approximations and ergodicity classes in random chains. IEEE Trans. Autom. Control **57**(11), 2718–2730 (2012)
165. B. Touri, A. Nedich, On backward product of stochastic matrices. Automatica **48**(8), 1477–1488 (2012)
166. B. Touri, A. Nedich, Product of random stochastic matrices and distributed averaging. IEEE Trans. Autom. Control **59**(2), 437–448 (2014)
167. K.I. Tsianos, The role of the network in distributed optimization algorithms: convergence rates, scalability, communication/computation tradeoffs and communication delays, Ph.D. thesis, McGill University, Dept. of Electrical and Computer Engineering, 2013
168. K.I. Tsianos, M.G. Rabbat, Distributed consensus and optimization under communication delays, in *Proceedings of the 49th IEEE Allerton Conference on Communication, Control, and Computing*, pp. 974–982 (2011)

169. K.I. Tsianos, S. Lawlor, M.G. Rabbat, Consensus-based distributed optimization: practical issues and applications in large-scale machine learning, in *Proceedings of the 50th Allerton Conference on Communication, Control, and Computing*, pp. 1543–1550 (2012)
170. K.I. Tsianos, S. Lawlor, M.G. Rabbat, Push-sum distributed dual averaging for convex optimization, in *Proceedings of the 51st IEEE Conference on Decision and Control*, pp. 5453–5458 (2012)
171. J.N. Tsitsiklis, Problems in decentralized decision making and computation, Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1984
172. S.-Y. Tu, A.H. Sayed, Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks. IEEE Trans. Signal Process. **60**(12), 6217–6234 (2012)
173. S.-Y. Tu, A.H. Sayed, On the influence of informed agents on learning and adaptation over networks. IEEE Trans. Signal Process. **61**(6), 1339–1356 (2012)
174. T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, O. Schochet, Novel type of phase transitions in a system of self-driven particles. Phys. Rev. Lett. **75**(6), 1226–1229 (1995)
175. P. Wan, M.D. Lemmon, Event-triggered distributed optimization in sensor networks, in *Symposium on Information Processing of Sensor Networks, (San Francisco, CA)*, pp. 49–60 (2009)
176. J. Wang, N. Elia, A control perspective for centralized and distributed convex optimization, in *Proceedings of the IEEE Conference on Decision and Control, (Florida, USA)*, pp. 3800–3805 (2011)
177. E. Wei, A. Ozdaglar, Distributed alternating direction method of multipliers, in *Proceedings of the 51st IEEE Conference on Decision and Control and European Control Conference*, pp. 5445–5450 (2012)
178. E. Wei, A. Ozdaglar, On the O(1/k) convergence of asynchronous distributed alternating direction method of multipliers, in *Proceedings of IEEE Global Conference on Signal and Information Processing*, pp. 551–554 (2013)
179. C. Xi, U. Khan, DEXTRA: a fast algorithm for optimization over directed graphs. IEEE Trans. Autom. Control **62**(10), 4980–4993 (2017)
180. L. Xiao, S. Boyd, Fast linear iterations for distributed averaging. Syst. Control Lett. **53**, 65–78 (2004)
181. L. Xiao, S. Boyd, S. Lall, A scheme for robust distributed sensor fusion based on average consensus, in *Proceedings of the 4th International Conference on Information Processing in Sensor Networks*, pp. 63–70 (2005)
182. L. Xiao, S. Boyd, S.-J. Kim, Distributed average consensus with least mean square deviation. J. Parallel Distrib. Comput. **67**(1), 33–46 (2007)
183. J. Xu, Augmented distributed optimization for networked systems, Ph.D. Thesis, School of Electrical and Electronic Engineering, Nanyang Technological University, 2016
184. J. Xu, S. Zhu, Y. Soh, L. Xie, Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes, in *Proceedings of the 54th IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060 (2015)
185. F. Yan, S. Sundaram, S.V.N. Vishwanathan, Y. Qi, Distributed autonomous online learning: regrets and intrinsic privacy-preserving properties. IEEE Trans. Data Knowl. Eng. **25**(11), 2483–2493 (2013)
186. F. Zanella, D. Varagnolo, A. Cenedese, G. Pillonetto, L. Schenato, Newton-Raphson consensus for distributed convex optimization, in *Proceedings of the IEEE Conference on Decision and Control, (Florida, USA)*, pp. 5917–5922 (2011)
187. J. Zeng, W. Yin, Extrapush for convex smooth decentralized optimization over directed networks (2015). Available on arXiv at http://arxiv.org/abs/1511.02942
188. M. Zhu, S. Martínez, Discrete-time dynamic average consensus. Automatica **46**, 322–329 (2010)

189. M. Zhu, S. Martínez, On distributed convex optimization under inequality and equality constraints. IEEE Trans. Autom. Control **57**(1), 151–164 (2012)
190. M. Zhu, S. Martínez, An approximate dual subgradient algorithm for distributed non-convex constrained optimization. IEEE Trans. Autom. Control **58**(6), 1534–1539 (2013)