Automatica 120 (2020) 109108

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Network flows that solve least squares for linear equations ☆

Yang Liu [a], Youcheng Lou [b,*], Brian D.O. Anderson [a,c,d], Guodong Shi [e]

[a] Research School of Electrical, Energy and Materials Engineering, The Australian National University, Canberra 0200, Australia
[b] MDIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[c] Hangzhou Dianzi University, Hangzhou 310018, China
[d] Data61-CSIRO, Canberra 0200, Australia
[e] Australian Centre for Field Robotics, The University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 13 August 2018
Received in revised form 9 March 2020
Accepted 2 June 2020
Available online 1 July 2020

Keywords:
Distributed algorithms
Linear equation
Least-squares solutions

ABSTRACT

This paper presents a first-order distributed continuous-time algorithm for computing the least-squares solution to a linear equation over networks. Given the uniqueness of the solution, with nonintegrable and diminishing step size, convergence results are provided for fixed graphs. The exact rate of convergence is also established for various types of step size choices falling into that category. For the case where non-unique solutions exist, convergence to one such solution is proved for constantly connected switching graphs with square integrable step size. Validation of the results and illustration of the impact of step size on the convergence speed are made using a few numerical examples.

1. Introduction

In modern engineering systems, there is a great demand for large-scale computing capabilities for solving real-world mathematical problems. Centralized algorithms are effective tools if the computing center possesses information of the entire problem. In some cases, however, due to the comparatively weak computing power of an agent or its limited access to the parameters and measurement data relevant to the whole problem, the notion of distributed computation over networks has been developed (Jadbabaie, Lin, & Morse, 2003; Lynch, 1996; Mesbahi & Egerstedt, 2010; Rabbat, Nowak, & Bucklew, 2005; Tsitsiklis, 1984; Tsitsiklis & Bertsekas, 1986). Nowadays it is widely applied in the areas of analyzing the consensus of complex systems (Olfati-Saber & Murray, 2004), solving various optimization problems (Nedić & Ozdaglar, 2009), carrying out distributed estimation (Cattivelli, Lopes, & Sayed, 2008) and filtering (Kar & Moura, 2011).

Solving systems of linear equations using distributed algorithms over networks emerges as one of the basic tasks in distributed computation. In these scenarios, it is often assumed that each agent of the network only has access to one or a few of the individual linear equations making up the full system due to security issues or memory limitation, and is only permitted to interact with a subset of the other agents. A number of contributions have been made to the development of distributed solvable linear equation solvers, including simple first-order distributed algorithms in discrete-time that incorporate the nature of modern computer (Liu, Morse, Nedic, & Basar, 2014; Liu, Mou, & Morse, 2013; Lu & Tang, 2009; Mou, Liu, & Morse, 2015; Mou, Morse, & S, 2013; Wang & Elia, 2009, 2014). Using differential equations as a tool for the study of discrete optimization and stochastic approximation algorithms has been a long-standing research topic lying at the interface of the areas dynamical systems and optimization (Gharesifard & Cortés, 2014; Helmke & Moore, 2012; Ljung, 2017; Su, Boyd, & Candes, 2016; Wang & Elia, 2014). Recently, the investigation of distributed computation and machine learning algorithms from the perspectives of ODEs has also attracted great interest (Anderson, Mou, Morse, & Helmke, 2015; Orvieto & Lucchi, 2019; Shi, Anderson, & Helmke, 2017). The development of continuous-time linear-equation solvers has also drawn much attention (Anderson et al., 2015; Shi et al., 2017; Wang & Elia, 2014), and the discretization of such algorithms can be easily achieved using existing methods such as Euler approximation. Moreover, the continuous-time perspective itself is also useful in analog circuits and the developing quantum computation (Childs, 2009). The distributed algorithms manage to deliver satisfactory solutions even for switching network structures. As is known to all, however, another frequent case arising in practical problems is concerned with non-solvable linear equations, in which we

☆ This work is supported by the Australian Research Council (ARC) under grants DP-130103610, DP-160104500 and DP190103615. The material in this paper was presented at the 56th IEEE Conference on Decision and Control, December 12–15, 2017, Melbourne, Australia. This paper was recommended for publication in revised form by Associate Editor Claudio De Persis under the direction of Editor Christos G. Cassandras.

* Corresponding author.
E-mail addresses: yang.liu@anu.edu.au (Y. Liu), louyoucheng@amss.ac.cn (Y. Lou), brian.anderson@anu.edu.au (B.D.O. Anderson), guodong.shi@sydney.edu.au (G. Shi).

https://doi.org/10.1016/j.automatica.2020.109108

often seek a least-squares solution by minimizing the associated objective function.

However, it seems a rather challenging problem to develop distributed least-squares solvers for network linear equations, due to the mismatch between individual linear equations at each node and the least-squares solution. Despite the difficulties, there exist a few distributed algorithms developed for the least-squares problem using different approaches, such as second-order algorithms (Gharesifard & Cortés, 2014; Liu, Lageman, Anderson, & Shi, 2019; Wang & Elia, 2010, 2012), state expansion (Mou et al., 2015) and the high gain consensus gain method (Shi et al., 2017). Second-order distributed least-squares solvers (Gharesifard & Cortés, 2014; Liu et al., 2019; Wang & Elia, 2010, 2012) generally can produce good convergence performance. In particular, Nedić, Olshevsky, and Shi (2017) present distributed discrete-time inexact gradient algorithms for undirected and directed graphs with constant step size and exponential convergence rate. However, the algorithms rely on restricted network structures and demand higher communication and storage capacities. Besides, Sun, Scutari, and Palomar (2016) propose discrete-time distributed algorithms for nonconvex constrained optimization over directed graphs. Similar to our algorithms, their algorithms have time-varying step size. Local estimation of gradients is introduced for convexification, and thus the dimension of agent states is necessarily expanded. In addition, the algorithms that work well on fixed networks are not guaranteed to yield desirable results on switching networks (Liu et al., 2019). The state expansion method (Mou et al., 2015) is based on enlarging the state dimension and then applying the existing methods for linear equations with an exact solution, a negative feature is that the nodes must have access to more knowledge than their own linear equations. It was shown in Shi et al. (2017) that first-order algorithms for exact solutions can be adapted to the least-squares case by a high consensus gain, but only in an approximate sense.

In this paper, we propose a first-order continuous-time flow for the least-squares problems of network linear equations, in which each agent keeps averaging the state with those of its neighbors and at the same time descends along the negative gradient of its local cost function. This flow is inspired by the work of Nedić and Olshevsky (2015) on distributed subgradient optimization. If the network linear equation has one unique least-squares solution, we prove that all node states asymptotically converge to that solution along our flow, with constant and connected graphs and a step size tending to zero, but not too fast. We also give analytical results on how the choice of step size, the attributes of linear equations and network size affect the convergence speed. For a switching network structure that is at all times connected, we show that the node states always converge to one of the least-squares solutions with square integrable step size. We also provide a few numerical examples that validate the usefulness of the proposed algorithms and demonstrate the convergence rate.

A preliminary version of this work (Liu, Lou, Anderson, & Shi, 2017) was presented at the 56th IEEE Conference on Decision and Control. Compared to the conference version, we make additional contributions as follows:

(i) Theoretical studies on the rate of convergence of the proposed algorithm are provided.
(ii) Convergence results are clearly stated under a common structure for all network and linear equation scenarios, in addition to the detailed proofs.
(iii) More numerical validations are provided.

The remainder of this paper is organized as follows. In Section 2, a brief introduction to the definition of the problem studied is given. We present the main results in Section 3. We also provide validations and further discussions using numerical examples in Section 4. In Section 5, the main work of this paper is summarized and potential future work directions are provided. The detailed proofs are provided in the Appendices.

2. Problem definition

In this section, a few mathematical preliminaries are provided, regarding linear equations over networks. Also, we establish a distributed network flow that can asymptotically compute the least-squares solution to network linear equations and discuss its relation to existing work.

2.1. Linear equations and networks

Consider the following linear algebraic equation with respect to $\mathbf{y} \in \mathbb{R}^m$

$$\mathbf{z} = \mathbf{H}\mathbf{y}, \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^N$ and $\mathbf{H} \in \mathbb{R}^{N\times m}$ are known and satisfy $N \geq m$. Let $\mathbf{h}_i \in \mathbb{R}^m$ denote the ith row of $\mathbf{H}$ and z_i denote the ith component of $\mathbf{z}$. We can rewrite (1) as $\mathbf{h}_i^\top \mathbf{y} = z_i,\ i = 1,\ldots,N$. Denote the column space of a matrix $\mathbf{M}$ by colsp$\{\mathbf{M}\}$. If $\mathbf{z} \notin$ colsp$\{\mathbf{H}\}$, the least-squares solution is defined by the minimization of $\|\mathbf{z}-\mathbf{H}\mathbf{y}\|^2$ over $\mathbf{y} \in \mathbb{R}^m$. It is well known that if rank$(\mathbf{H}) = m$, then it yields a unique solution $\mathbf{y}^* = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{z}$, while it has a set of non-unique least-squares solutions if rank$(\mathbf{H}) < m$. Define $f(\mathbf{y}) = \|\mathbf{z}-\mathbf{H}\mathbf{y}\|^2 = \sum_{i=1}^N f_i(\mathbf{y})$, where $f_i(\mathbf{y}) = |\mathbf{h}_i^\top\mathbf{y} - z_i|^2$.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a constant, undirected and simple graph with the finite set of nodes $\mathcal{V} = \{1, 2, \ldots, N\}$ and the set of edges $\mathcal{E} = \big\{\{i,j\} : i,j \in \mathcal{V} \text{ are connected}\big\}$. Define a weight function $w : \mathcal{E} \to \mathbb{R}^+$ over the edge set with the weight of edge $\{i,j\}$ being $w(\{i,j\})$. It is worth noting that the weight w for each edge is assumed to be fixed in this paper for ease of the presentation. Generalizations to time-varying weights can be made similar to the analysis of Shi et al. (2017). Based on constant graphs, we next introduce time-varying graphs. Let $\mathcal{Q}$ be the set containing all possible constant and undirected graphs induced by the node set $\mathcal{V}$ and let $\mathcal{Q}^* \subset \mathcal{Q}$ be a subset of $\mathcal{Q}$. Define a piecewise constant mapping $\mathcal{G}_\sigma = (\mathcal{V}, \mathcal{E}_\sigma) : \mathbb{R}^{\geq t_0} \to \mathcal{Q}^*$ for some $t_0 > 0$. Throughout this paper, we assume the set of times corresponding to discontinuities of $\mathcal{G}_{\sigma(t)}$ has measure zero. Note that the time-varying graph $\mathcal{G}_{\sigma(t)} = (\mathcal{V}, \mathcal{E}_{\sigma(t)})$ represents the network topology at time t. Let $\mathcal{N}_i(t)$ be the set of neighbor nodes that are connected to node i at time t, i.e., $\mathcal{N}_i(t) = \big\{j : \{i,j\} \in \mathcal{E}_{\sigma(t)}\big\}$. Define the adjacency matrix $\mathbf{A}(t)$ of the graph $\mathcal{G}_{\sigma(t)}$ by $[\mathbf{A}(t)]_{ij} = w(\{i,j\})$ if $\{i,j\} \in \mathcal{E}_{\sigma(t)}$, and $[\mathbf{A}(t)]_{ij} = 0$ otherwise, and $\mathbf{D}(t) = \mathrm{diag}(\sum_{j=1}^N[\mathbf{A}(t)]_{1j}, \ldots, \sum_{j=1}^N[\mathbf{A}(t)]_{Nj})$. Then $\mathbf{L}(t) = \mathbf{D}(t) - \mathbf{A}(t)$ is the Laplacian of graph $\mathcal{G}_{\sigma(t)}$ at time t.

2.2. Distributed flows

Assume that node i of the network $\mathcal{G}_{\sigma(t)}$ only knows the information of $\mathbf{h}_i, z_i$, i.e., node i is associated with the linear equation $\mathbf{h}_i^\top\mathbf{y} = z_i$. We associate with each node i a state $\mathbf{x}_i(t) \in \mathbb{R}^m$, which, as the notation implies, in general varies with time. Then we propose the following continuous-time network flow starting from time t_0

$$\dot{\mathbf{x}}_i(t) = K \sum_{j\in\mathcal{N}_i(t)} [\mathbf{A}(t)]_{ij}(\mathbf{x}_j(t) - \mathbf{x}_i(t)) - \frac{\alpha(t)}{2}\nabla f_i(\mathbf{x}_i(t)), \tag{2}$$

where $K \in \mathbb{R}^+$ is a positive constant and the step size $\alpha : \mathbb{R}^{\geq t_0} \to \mathbb{R}^+$ is a continuous function which assures the continuity of all

$\mathbf{x}_i(t)$ and their derivatives, except for the time points when the network switches. In vector form, we have

$$\dot{\mathbf{x}}(t) = -\mathbf{M}(t)\mathbf{x}(t) + \alpha(t)\mathbf{z}_H, \tag{3}$$

where $\mathbf{x}(t) = [\mathbf{x}_1(t)^\top \ \ldots \ \mathbf{x}_N(t)^\top]^\top$, $\mathbf{M}(t) = K(\mathbf{L}(t) \otimes \mathbf{I}_m) + \alpha(t)\tilde{\mathbf{H}}$, $\tilde{\mathbf{H}} = \mathrm{diag}\big(\mathbf{h}_1\mathbf{h}_1^\top, \ldots, \mathbf{h}_N\mathbf{h}_N^\top\big)$, $\mathbf{z}_H = [z_1\mathbf{h}_1^\top \ \cdots \ z_N\mathbf{h}_N^\top]^\top$. Now we make several assumptions of $\alpha(t)$ that will be used in our main results.

Assumption 1. (i) $\int_{t_0}^{\infty} \alpha(t)dt = \infty$; (ii) $\lim_{t\to\infty} \alpha(t) = 0$; (iii) $\int_{t_0}^{\infty} \alpha^2(t)dt < \infty$.

Remark 1. Similar to the existing literature on distributed gradient techniques (Nedić, Ozdaglar, & Parrilo, 2010), the step size $\alpha(t)$ is a network-wise synchronized signal, and therefore the flow (2) has to be synchronously executed over the network in its present form. This setup, consistent with Nedić et al. (2010), facilitates a concise analysis and provides a benchmark. Indeed, this means that the algorithm assumes a network-wise synchronized signal $\alpha(t)$. To enforce this, the entire time function $\alpha(t)$ can be specified at the preparation stage of the execution of the algorithm, under the assignment of synchronized clocks at all nodes. Alternatively, an additional step-size-averaging mechanism can be introduced to the network flow so that the step sizes at different nodes are synchronized online. In this case, a step size mismatch function $\delta_i : R^{\geq t_0} \to \mathbb{R}^+$ at each node i would be introduced so that the ideal step size $\alpha(t)$ at node i is replaced with $\alpha(t)+\delta_i(t)$. For two functions $g, h : \mathbb{R}^{\geq t_0} \to \mathbb{R}^+$, it is written as $g(t) = o(h(t))$ if for every $c > 0$, there exists $\tau > t_0$ such that $g(t) \leq c \cdot h(t)$ for all $t \geq \tau$. Then

(i) If $\delta_i(t)$ is integrable, the convergence result still holds, because the effect of the integrable mismatch $\delta_i(t)$ is dominated by the non-integrable $\alpha(t)$.

(ii) We conjecture that if each $\delta_i(t) = o(\alpha(t))$ holds, then the convergence result holds. An example will be provided to validate this conjecture.

2.3. Discussions

Now we clarify the relation between the previous work on distributed least-squares and optimization algorithms, and our algorithm (2) by briefly discussing their structure and applicability. It is clear that (2) has exactly the same structure as the flow in Nedić and Olshevsky (2015) and Nedić et al. (2010) in the sense that they are both in the form of "local averaging consensus"+ "diminishing local objective", with the difference that the flow in Nedić and Olshevsky (2015) and Nedić et al. (2010) is discrete-time but (2) is continuous-time. However, we cannot use the algorithm and the analysis directly because the gradient boundedness of (2) is not directly verifiable. It can be noted that the first-order flow in Shi et al. (2017) is a special case of (2) obtained by letting $\alpha(t)$ be some constant. Due to the existence of the diminishing step size, (2) is a linear time-varying system, while the flow in Shi et al. (2017) is linear time-invariant and can only produce the solution in approximate sense. Hence the approach to analyzing the flow in Shi et al. (2017) is not applicable for (2). Also (2) can be formulated by properly specializing the optimization problem in Touri and Gharesifard (2015) and letting each agent's output scale be constant one. However, because of the specificity of the least-squares cost function, relaxed convergence conditions become possible as will be shown later. Shi and Johansson (2013) investigate a standard distributed averaging algorithm with general disturbance that may be state-dependent, and provides robust convergence results. However, only upper bounds for convergence time are established, but no results on exact convergence rates are presented. Finally, there are also second-order least-squares solvers (Gharesifard & Cortés, 2014; Liu et al., 2019; Wang & Elia, 2010, 2012), which admit quite different nature in terms of dynamical behaviors. Therefore, as arguably one of the most basic distributed flows for distributed computation, it is of interest to understand its convergence conditions and speed limit.

3. Main results

In this section, we investigate the flow (3) over fixed and switching networks, respectively, and establish the convergence conditions regarding $\alpha(t)$ and the graphs.

Proofs of the results appear in the Appendices.

3.1. Convergence over fixed networks

First we consider the case where the linear equation (1) has one unique least-squares solution and the network is a constant graph for all t.

Theorem 2. *Suppose* $\mathrm{rank}(\mathbf{H}) = m$, *and let* $\mathbf{y}^* = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{z}$ *denote the unique least-squares solution of (1). Let Assumption 1 (i) and (ii) hold. If* $\mathcal{G}_{\sigma(t)} = \mathcal{G}$ *is constant and connected for all* $t \geq t_0$, *then along any solution of (2) there holds* $\lim_{t\to\infty} \mathbf{x}_i(t) = \mathbf{y}^*$ *for all* $i \in \mathcal{V}$.

Let $\sigma_{\mathrm{m}}(\cdot)$ and $\sigma_2(\cdot)$ denote the smallest and the second smallest eigenvalue of a real symmetric matrix, respectively. For two functions $g, h : \mathbb{R}^{\geq t_0} \to \mathbb{R}^+$, we say $g(t) = \mathcal{O}(h(t))$ if there exist $c > 0$ and $\tau > t_0$ such that $g(t) \leq c \cdot h(t)$ for all $t \geq \tau$. The following theorem characterizes the convergence speed of the algorithm (2) for different choices of step size known to decay with a t's inverse power that is not greater than one.

Theorem 3. *Suppose the conditions of Theorem 2 hold. Define* $\mathbf{y}^* = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{z}$.

(i) If $\alpha(t) = \mathcal{O}(\frac{1}{t})$, *then along (2) there holds for any* $0 < \epsilon < 1$

$$\left\| \sum_{i=1}^{N} \mathbf{x}_i(t)/N - \mathbf{y}^* \right\| = \mathcal{O}\left(\frac{1}{t^{\min(1-\epsilon, \frac{\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N})}} \right).$$

(ii) If $\alpha(t) = \mathcal{O}(\frac{1}{t^\lambda})$ *for* $\lambda \in (0, 1)$, *then along (2) there holds for any* $0 < \epsilon < \lambda$

$$\left\| \sum_{i=1}^{N} \mathbf{x}_i(t)/N - \mathbf{y}^* \right\| = \mathcal{O}\left(\frac{1}{t^{\lambda-\epsilon}} \right).$$

Theorem 3 is proved based on a novel recursive analysis of the error dynamics, where (2) is written into

$$\dot{\mathbf{x}}_i(t) = K \sum_{j\in\mathcal{N}_i(t)} [\mathbf{A}(t)]_{ij}(\mathbf{x}_j(t) - \mathbf{x}_i(t)) + \alpha(t)\mathbf{h}_i(z_i - \mathbf{h}_i^\top\mathbf{y}^* - \mathbf{h}_i^\top\mathbf{x}_i(t)).$$

By recursively establishing tight bound for $\alpha(t)\mathbf{h}_i(z_i - \mathbf{h}_i^\top\mathbf{y}^* - \mathbf{h}_i^\top\mathbf{x}_i(t))$, we manage to derive the tight bounds on convergence speed in Theorem 3. Clearly, Theorem 3 provides some guidance on the choice of the step size $\alpha(t)$ to guarantee fast convergence speed as follows:

(i) For linear equations and networks with $\frac{\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N} \geq 1$, $\alpha(t) = \mathcal{O}(\frac{1}{t})$ yields the fastest convergence speed.

(ii) For linear equations and networks with $\frac{\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N} < 1$, $\alpha(t) = \mathcal{O}(\frac{1}{t^\lambda})$ with $\frac{\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N} < \lambda < 1$ admits the fastest convergence speed. In this case, the rate of convergence will increase as λ becomes larger. Interestingly, however, when λ reaches one, the rate of convergence suddenly drops to that of the case $\lambda = \frac{\sigma_m(\mathbf{H}^\top\mathbf{H})}{N}$.

These results, especially the discontinuity around the inverse power one of t, would have been difficult to predict. As will be shown later, numerical results demonstrate that the convergence upper bounds established in Theorem 3 are also the asymptotic lower bounds.

3.2. Convergence over switching networks

Now we consider a more general case where the least-squares solutions of (1) can be unique or non-unique, and the network $\mathcal{G}_{\sigma(t)}$ switches among a collection of graphs. Evidently, the Caratheodory solutions of (3) exist for all initial conditions because the set of times corresponding to discontinuities of $\mathcal{G}_{\sigma(t)}$ is assumed to have measure zero.

Theorem 4. *Suppose* $\operatorname{rank}(\mathbf{H}) \le m$ *and denote the set of least-squares solutions of (1) by* $\mathcal{Y}_{\rm LS} = \operatorname{argmin} f(\mathbf{y})$*. In particular,* $|\mathcal{Y}_{\rm LS}| = 1$ *if* $\operatorname{rank}(\mathbf{H}) = m$*. Suppose Assumption 1 (i), (ii) and (iii) hold. If all* $\mathcal{G} \in \mathcal{Q}^*$ *are connected, then along any solution of (2) over the switching graph* $\mathcal{G}_{\sigma(t)}$ *there exists* $\hat{\mathbf{y}} \in \mathcal{Y}_{\rm LS}$ *such that* $\lim_{t\to\infty} \mathbf{x}_i(t) = \hat{\mathbf{y}}$ *for all* $i \in \mathcal{V}$.

Based on Proposition 4.10 in Shi and Johansson (2013), we have that (2) solves the least-squares problem over uniformly jointly connected networks whose definition is given in Shi and Johansson (2013), if an additional assumption that $\mathbf{x}(t)$ is bounded is imposed. We must mention that it is hard to remove the state boundedness assumption. However, numerical examples can show that the state boundedness condition can be satisfied in many circumstances.

3.3. Extensions to discrete time

In this subsection, we investigate the discrete counterpart of the flow (2) and show how the convergence analysis established in Theorems 2 and 3 continues to be useful. Let discrete time be indexed by $k = 0, 1, \ldots$. For a fixed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we introduce $w_{ij} \ge 0$ for all $i, j \in \mathcal{V}$ satisfying (i) $w_{ij} > 0$ if and only if $j \in \mathcal{N}_i$ or $j = i$; (ii) $w_{ij} = w_{ji}$; (iii) $\sum_{j=1}^N w_{ij} = 1$. Similarly, each node $i \in \mathcal{V}$ is associated with a dynamic state $\mathbf{x}_i(k) \in \mathbb{R}^m$. By Euler approximation, one can write the discrete-time algorithm over fixed graphs that corresponds to (3) as

$$\mathbf{x}_i(k+1) = \sum_{j=1}^N w_{ij}\mathbf{x}_j(k) - \frac{\eta(k)}{2}\nabla f_i(\mathbf{x}_i(k)) \tag{4}$$

with $\eta : \mathbb{Z}^{\ge 0} \to \mathbb{R}^+$. This is precisely a distributed multi-agent gradient algorithm of Nedić and Ozdaglar (2009) and Nedić et al. (2010) with a special quadratic cost function.

Theorem 5. *Let the conditions of Theorem 2 hold. Suppose* $\eta(k) = \mathcal{O}(\frac{1}{k^\lambda})$ *for some* $0 < \lambda < 1$*. Then for any* $0 < \epsilon < \lambda$*, along (4) there holds*

$$\left\| \sum_{i=1}^N \mathbf{x}_i(k)/N - \mathbf{y}^* \right\| = \mathcal{O}\Big(\frac{1}{k^{\lambda-\epsilon}}\Big).$$

Theorem 5 clearly advances the convergence rate analysis in Nedić and Ozdaglar (2009) and Nedić et al. (2010). Although the analysis does not shed light on the discontinuity at $\eta(k) = \mathcal{O}(k^{-1})$ as Theorem 3 does, it reveals the potential of applying our novel recursive analysis in the discrete-time domain.

Fig. 1. The trajectories of $e_j(t) := \left\| \sum_{i=1}^4 \mathbf{x}_i(t)/4 - \mathbf{y}_j^* \right\|^2$, $j = 1, 2$ and $\bar{e}_3(t) := \left\| \sum_{i=1}^4 \mathbf{x}_i(t)/4 - \mathbf{y}_3^* \right\| \cdot (\log(t+1))^{-1}$.

4. Numerical examples

In this section, several numerical examples are provided to validate the results of Theorems 2, 4.

4.1. Fixed graphs

Example 1. Consider a 4-node path graph $\mathcal{G}_{\rm ring}$, over which we study two linear algebraic equations with respect to $\mathbf{y} \in \mathbb{R}^2$,

$$\text{(LE. 1)} \begin{bmatrix} 1 & 1 & -0.5 & 0.8 \\ 1 & 2.3 & 0.8 & 0.2 \end{bmatrix}^\top \mathbf{y} = \begin{bmatrix} 1 & 3 & 2 & -1 \end{bmatrix}^\top,$$

$$\text{(LE. 2)} \begin{bmatrix} 2 & 6 & -11 & 1 \\ 7 & 5 & 1 & 0 \end{bmatrix}^\top \mathbf{y} = \begin{bmatrix} 1 & 3 & 2 & -1 \end{bmatrix}^\top.$$

Both (LE. 1) and (LE. 2) yield unique least-squares solutions $\mathbf{y}_1^* = [-1.218\ 1.869]^\top$, $\mathbf{y}_2^* = [-0.092\ 0.361]^\top$, respectively. The resulting $\frac{\sigma_m(\mathbf{H}^\top\mathbf{H})}{N}$ values for (LE. 1) and (LE. 2) are $\big(\frac{\sigma_m(\mathbf{H}^\top\mathbf{H})}{N}\big)_1 = 0.313$, $\big(\frac{\sigma_m(\mathbf{H}^\top\mathbf{H})}{N}\big)_2 = 15.975$, respectively. We also introduce another equation (LE. 3) by multiplying the left-hand side of (LE. 1) with 1.7872 so that $\big(\frac{\sigma_m(\mathbf{H}^\top\mathbf{H})}{N}\big)_3 = 1$, leading to a unique least-squares solution $\mathbf{y}_3^* = \mathbf{y}_1^*/1.7872$. With $K = 100$ and some randomly chosen $\mathbf{x}(t_0)$ with $t_0 = 10^{-6}$, we run the algorithm (3) with $\alpha(t) = \frac{1}{t+1}$ and then plot the trajectories of $e_j(t) := \left\| \sum_{i=1}^4 \mathbf{x}_i(t)/4 - \mathbf{y}_j^* \right\|$, $j = 1, 2$ and $\bar{e}_3(t) := \left\| \sum_{i=1}^4 \mathbf{x}_i(t)/4 - \mathbf{y}_3^* \right\| \cdot \big(\log(t+1)\big)^{-1}$ in logarithmic scales in Fig. 1. As can be seen, each $\mathbf{x}_i(t)$ converges to $\mathbf{y}^*$, which is consistent with the claim of Theorem 2. Further, according to the trajectories in Fig. 1, we directly calculate the slopes $\kappa_1 = -0.313$, $\kappa_2 = -0.997$, $\kappa_3 = -1.040$ for (LE. 1), (LE. 2) and (LE. 3), which implies $e_1(t) = \mathcal{O}(t^{-0.313})$, $e_2(t) = \mathcal{O}(t^{-0.997})$, $\bar{e}_3(t) = \mathcal{O}(t^{-1.040})$. This validates the statement of Theorem 3 when $\alpha(t) = \mathcal{O}(\frac{1}{t})$, where the bounds of $e_1(t)$ and $e_2(t)$ are as predicted as Theorem 3 (i), and that of $\bar{e}_3(t)$ indicates the existence of a tighter bound $\log t/t$.

Example 2. Consider the linear equation (LE. 2) with the same $\mathbf{x}(t_0)$ and K as in Example 1. We run the algorithm (3) on $\mathcal{G}_{\rm ring}$ for $\alpha(t) = \frac{1}{(t+1)^{0.75}}$, $\alpha(t) = \frac{1}{(t+1)^{0.5}}$ and $\alpha(t) = \frac{1}{(t+1)^{0.25}}$, under

Fig. 2. The trajectories of $e(t) := \left\| \sum_{i=1}^{4} \mathbf{x}_i(t)/4 - \mathbf{y}_2^* \right\|$ for $\alpha(t) = t^{-e}$ with $e = 0.75, 0.5, 0.25$, respectively.

which we plot in Fig. 2 the trajectories of $e(t) := \left\| \sum_{i=1}^{4} \mathbf{x}_i(t)/4 - \mathbf{y}_2^* \right\|$. By direct calculation, we find $e(t) = \mathcal{O}(t^{-0.750})$, $e(t) = \mathcal{O}(t^{-0.492})$, $e(t) = \mathcal{O}(t^{-0.249})$ for $\alpha(t) = \frac{1}{(t+1)^{0.75}}, \frac{1}{(t+1)^{0.5}}, \frac{1}{(t+1)^{0.25}}$, respectively. These results validate the statement in Theorem 3 for the step size $\alpha(t) = \mathcal{O}(\frac{1}{t^{\lambda}}),\ \lambda \in (0, 1)$.

Example 3. Consider the same setup as in Example 2 for solving (LE. 1) except that the universal step size $\alpha(t)$ is replaced with local step size $\alpha(t) + \delta_i(t)$ at each node i, where $\delta_i(t) = \frac{i}{(t+1)^2}$. The trajectories of $e(t)$ for different $\alpha(t)$, as depicted in Fig. 3, asymptotically go to zero. This validates the conjecture in Remark 1.

4.2. Comparison with existing least-squares flows

Example 4. Let us now demonstrate the difference between the "consensus + diminishing projection" flow (2), the "consensus + projection" flow in Shi et al. (2017), the AHU flow in Liu et al. (2019), and the damped AHU flow in Wang and Elia (2010). We reconsider (LE. 1) in Example 1 over a 4–node ring graph with uniform edge weights. For (2), we select $K = 10$ and $\alpha(t) = (t+1)^{-0.8}$. For the "consensus + projection" flow, we select $K = 10$ and $\alpha(t) \equiv 0.1$. The flows in Liu et al. (2019) and Wang and Elia (2010) are adopting a uniform gain for the Laplacian. In Fig. 4, we plot $e(t) = \| \sum_{i=1}^{4} \mathbf{x}_i(t)/4 - \mathbf{y}^* \|^2$, respectively for the four flows with $t_0 = 1$. On the one hand, the flow (2) achieves zero error asymptotically with a slower convergence rate compared to the flows in Liu et al. (2019) and Wang and Elia (2010), but under lower communication and computation complexities. The flow in Shi et al. (2017), on the other hand, only produces approximate results.

4.3. A large-scale example

Example 5. Least-squares problems arise in various domains of applications for the purpose of prediction, control, and learning, e.g., Chen, Billings, and Luo (1989), Golub and Van Loan (1980), Lines and Treitel (1984) and Qi et al. (2013). We now demonstrate the feasibility of the algorithm (4) as a distributed least-squares solver for relatively large networks. We randomly

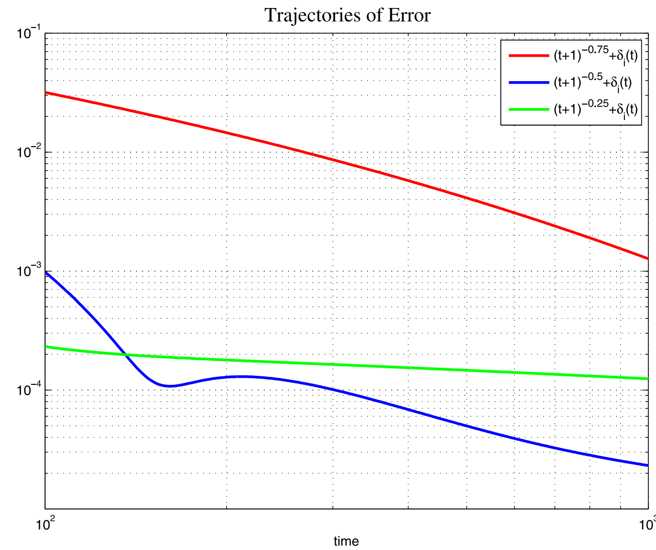


Fig. 3. The trajectories of $e(t) := \left\| \sum_{i=1}^{4} \mathbf{x}_i(t)/4 - \mathbf{y}_1^* \right\|$ with an approximate synchronous step size $\alpha(t) + \frac{i}{(t+1)^2}$ at node i for $\alpha(t) = t^{-e}$, $e = 0.75, 0.5, 0.25$, respectively.

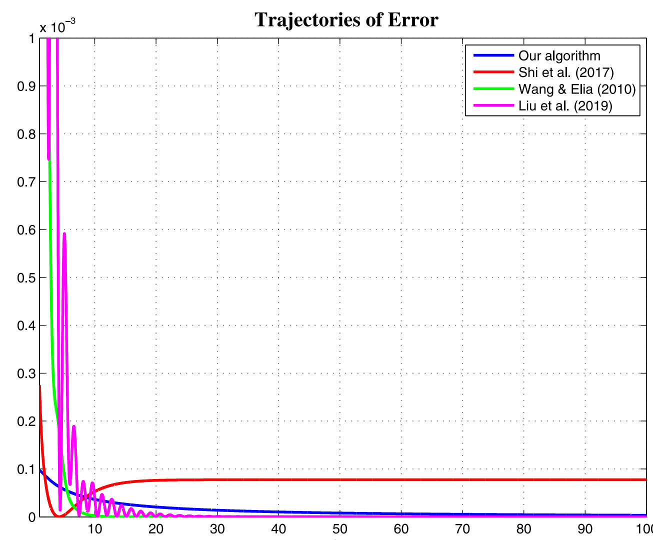


Fig. 4. The trajectories of error illustrating the comparison between the flow (2) and existing continuous least-squares flows.

generate a linear equation in the form of (1), in which $\mathbf{H} \in \mathbb{R}^{100\times 10}$, $\mathbf{z} \in \mathbb{R}^{100}$ with each entry chosen from $[0, 1]$ under a uniform distribution. Then we randomly generate a connected 10-regular graph over 100 nodes. For $\eta(k) = \frac{1}{(k+1)^{0.75}}, \frac{1}{(k+1)^{0.50}}, \frac{1}{(k+1)^{0.25}}$, we plot the error trajectories in logarithmic scales in Fig. 5. The error clearly converges to zero, and moreover, the slopes are $-0.764, -0.577, -0.257$, respectively, which are in strong agreement with Theorem 5.

5. Conclusions

In this paper, a first-order distributed continuous-time least-squares solver over networks was proposed. When the least-squares solution is unique, we proved the convergence results for fixed and connected graphs with an assumption of non-integrable step size. We also carefully analyzed the bound of

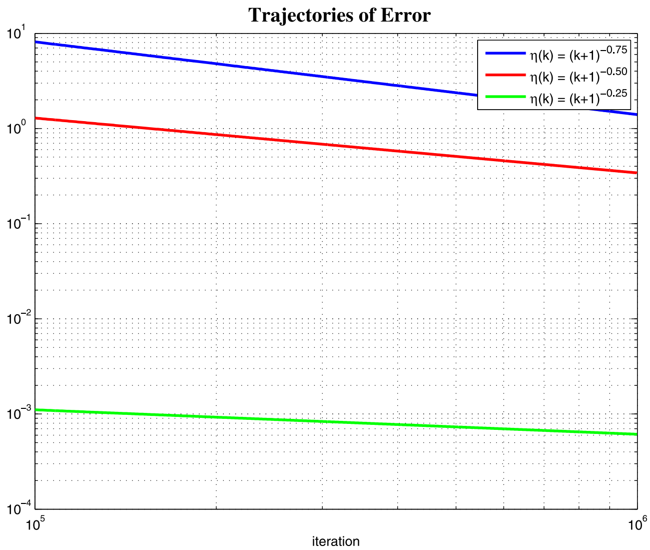


Fig. 5. The trajectories of error in logarithmic scales in the large-scale example.

convergence speed for two classes of step size choices, which provides guidance on the selection of step size to secure the fastest convergence speed. By loosening the requirement for the uniqueness of least-squares solutions and assuming square integrability on step size, we obtained convergence results for a constantly connected switching graph. We also provided some numerical examples, in order to verify the results and illustrate the convergence speed. Potential future work includes proving the convergence over networks without instantaneous connectivity, studying the exact convergence rate, and finding out the convergence limit.

Appendix A. Lemmas

Several lemmas that assist with the proofs of Theorems 2–5 are provided. Let $\langle\cdot,\cdot\rangle$ denote the inner product of two vectors of the same dimension. We say a differentiable function $g:\mathbb{R}^N\to\mathbb{R}$ is θ-strongly convex if $g(\mathbf{y}_1)-g(\mathbf{y}_2)\ge\nabla g(\mathbf{y}_2)^\top(\mathbf{y}_1-\mathbf{y}_2)+\frac{\theta}{2}\|\mathbf{y}_1-\mathbf{y}_2\|^2$ for all $\mathbf{y}_1,\mathbf{y}_2\in\mathbb{R}^N$.

Lemma 6. *Consider a matrix $\mathbf{H}\in\mathbb{R}^{N\times m}$ with $N\ge m$ and a vector $\mathbf{z}\in\mathbb{R}^N$. Define $f(\mathbf{y})=\|\mathbf{H}\mathbf{y}-\mathbf{z}\|^2$. If $\operatorname{rank}(\mathbf{H})=m$, then f is $2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})$-strongly convex.*

The proof of Lemma 6 can be achieved based on Taylor series of f and is omitted here.

Lemma 7. *Let $\mu,\lambda,\theta,t_0>0$. Then*

$$\int_{t_0^\theta}^{t^\theta}\mathcal{O}\left(\frac{e^{\mu s}}{s^\lambda}\right)\mathrm{d}s=\mathcal{O}\left(\frac{e^{\mu t^\theta}}{t^{\lambda\theta}}\right)\tag{5}$$

for $t\ge t_0$. Furthermore, for any $\mu^*>0$ and $0<\lambda_m\le\lambda_M$, there exist $q>0$, $T>t_0$ such that for $\mu=\mu^*$ and all $\lambda_m\le\lambda\le\lambda_M$, the inequality

$$\int_{t_0}^{t}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s<\frac{qe^{\mu t}}{t^\lambda}$$

holds for all $t\ge T$.

Proof. Introduce $\phi\in(0,\mu)$ and define $\tau=\frac{\lambda}{\mu-\phi}$. Then it can be easily shown for $t^\theta>\tau$, there holds

$$\begin{aligned}\int_{t_0^\theta}^{t^\theta}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s&=\int_{t_0^\theta}^{\tau}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s+\int_{\tau}^{t^\theta}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s\\&\le\int_{t_0^\theta}^{\tau}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s+\int_{\tau}^{t^\theta}\left(\frac{1}{\phi}\left(\mu-\frac{\lambda}{s}\right)\right)\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s\\&\le\int_{t_0^\theta}^{\tau}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s+\int_{\tau}^{t^\theta}\frac{\mathrm{d}}{\mathrm{d}s}\frac{e^{\mu s}}{\phi s^\lambda}\\&=\frac{e^{\mu t^\theta}}{\phi t^{\lambda\theta}}+\int_{t_0^\theta}^{\tau}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s-\frac{e^{\mu\tau}}{\phi\tau^\lambda}.\end{aligned}\tag{6}$$

Clearly, $C=\int_{t_0^\theta}^{\tau}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s-\frac{e^{\mu\tau}}{\phi\tau^\lambda}$ is a constant. This completes the proof of (5).

Further, we impose an additional assumption that $\theta=1$ and $\phi\in(0,\mu)$ is fixed so that $\tau=\frac{\lambda}{\mu-\phi}\ge\frac{\lambda_m}{\mu^*-\phi}>1$. Note that such a choice of ϕ only depends on μ^* and λ_m. We first suppose $t_0<1$. Then it is evident that

$$C=\int_{t_0}^{\tau}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s-\frac{e^{\mu\tau}}{\phi\tau^\lambda}\le\int_{t_0}^{1}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s+\int_{1}^{\tau}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s.\tag{7}$$

For $s\in(0,1)$ and $s>1$, there holds $1/s^\lambda\le1/s^{\lambda_M}$ and $1/s^\lambda\le1$, respectively. Then it follows from (7) that

$$C\le\int_{t_0}^{1}\frac{e^{\mu^*s}}{s^{\lambda_M}}\mathrm{d}s+\int_{1}^{\frac{\lambda_M}{\mu^*-\phi}}e^{\mu^*s}\mathrm{d}s=:C_M.\tag{8}$$

As seen from (8), C_M depends only on μ^*, λ_m and λ_M. By (6) and (8), we have

$$\int_{t_0}^{t}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s\le\frac{e^{\mu t}}{\phi t^\lambda}+C_M.\tag{9}$$

Evidently, $e^{\mu t}/t^\lambda$ monotonically goes to infinity over $t\in(\frac{\lambda_M}{\mu^*},\infty)$ as t goes to infinity. Thus, we can construct $T\ge\max(\frac{\lambda_M}{\mu^*},1)$ such that

$$\frac{e^{\mu T}}{T^\lambda}\ge\frac{e^{\mu^*T}}{T^{\lambda_M}}\ge C_M.\tag{10}$$

It is worth noting from (10) that the construction of such T is only dependent of μ^*, λ_M, C_M, and thereby μ^*, λ_m, λ_M in total. Based on (9) and (10), it can be concluded

$$\int_{t_0}^{t}\frac{e^{\mu s}}{s^\lambda}\mathrm{d}s\le\left(1+\frac{1}{\phi}\right)\frac{e^{\mu t}}{t^\lambda}$$

for all $t\ge T$. We let $q=1+1/\phi$ that only depends on ϕ, and thereby μ^*, λ_m. The same conclusion can be shown to hold for $t_0\ge1$. This completes the proof. □

Lemma 8. *Consider a continuously differentiable function $g:\mathbb{R}^{\ge t_0}\to\mathbb{R}^{\ge0}$ for some $t_0>0$. If there exist continuous functions $\gamma:\mathbb{R}^{\ge t_0}\to\mathbb{R}^+$ and $\beta:\mathbb{R}^{\ge t_0}\to\mathbb{R}^+$ satisfying $\dot g(t)\le-\gamma(t)g(t)+\beta(t)$, then $g(t)\le e^{-\int_{t_0}^t\gamma(s)\mathrm{d}s}g(t_0)+\int_{t_0}^te^{-\int_s^t\gamma(r)\mathrm{d}r}\beta(s)\mathrm{d}s$. Furthermore, if $\int_{t_0}^\infty\gamma(t)\mathrm{d}t=\infty$ holds, the following statements hold:*

(i) $\lim_{t\to\infty}\frac{\beta(t)}{\gamma(t)}=0$ implies $\lim_{t\to\infty}g(t)=0$.
(ii) $\limsup_{t\to\infty}\frac{\beta(t)}{\gamma(t)}<\infty$ implies that $\{g(t)\}_{t\ge t_0}$ is bounded.

Proof. The proof of the inequality of $g(t)$ follows from Grönwall's Inequality (Grönwall, 1919). Now we prove the two statements in the following:
(i) Suppose the conditions $\int_{t_0}^\infty\gamma(t)\mathrm{d}t=\infty$ and $\lim_{t\to\infty}\frac{\beta(t)}{\gamma(t)}=0$ hold. Evidently, the term $u(t):=\exp(-\int_{t_0}^t\gamma(s)\mathrm{d}s)g(0)$ goes to zero as t goes to infinity. Then we see $k(t)=\int_{t_0}^t\exp(-\int_s^t\gamma(r)\mathrm{d}r)\beta(s)\mathrm{d}s$. Since for a sufficiently small $\epsilon>0$, there exists $t^*>t_0$ such that $\frac{\beta(t)}{\gamma(t)}<\epsilon$ for all $t>t^*$. Define $\xi=\max_{t_0\le t\le t^*}\frac{\beta(t)}{\gamma(t)}$. Then for all $t>t^*$, there holds $k(t)<\xi\int_{t_0}^{t^*}\mathrm{d}(\exp(-\int_s^t\gamma(r)\mathrm{d}r))$

$+\epsilon\int_{t^*}^{t}\mathrm{d}(\exp(-\int_s^t\gamma(r)\mathrm{d}r)) = \xi\exp(-\int_{t^*}^{t}\gamma(r)\mathrm{d}r)(1-\exp(-\int_{t_0}^{t^*}\gamma(r)\mathrm{d}r)) + \epsilon(1-\exp(-\int_{t_0}^{t}\gamma(r)\mathrm{d}r)) < \xi\exp(-\int_{t^*}^{t}\gamma(r)\mathrm{d}r) + \epsilon$. Since $\exp(-\int_{t^*}^{t}\gamma(r)\mathrm{d}r)$ goes to zero as t goes to infinity, one has $\lim_{t\to\infty}k(t) = 0$. Then we have $\lim_{t\to\infty}g(t) = 0$.
(ii) Suppose the conditions $\int_{t_0}^{\infty}\gamma(t)\mathrm{d}t = \infty$ and $\limsup_{t\to\infty}\frac{\beta(t)}{\gamma(t)} < \infty$ hold. Then there exist $B > 0$ and $\hat{t} > t_0$ such that $\frac{\beta(t)}{\gamma(t)} < B$ for all $t > \hat{t}$. Similarly, the limit of the term $u(t) = \exp(-\int_{t_0}^{t}\gamma(s)\mathrm{d}s)$ $g(0)$ is zero as t goes to infinity, i.e., given $B > 0$, there exists $t_u > t_0$ such that $u(t) < B$ for all $t > t_u$. Also we have $k(t) < B\int_{t_0}^{t}\exp(-\int_s^t\gamma(r)\mathrm{d}r)\gamma(s)\mathrm{d}s < B$ for $t > \hat{t}$. Let $t_M := \max\{\hat{t}, t_u\}$. Hence, $g(t) < 2B$ for $t > t_M$. Since $g(t)$ is continuous, we have $g(t) < \max\{B_1, 2B\}$ for all $t \geq t_0$, where $B_1 = \max_{t_0\leq t\leq t_M} g(t)$, i.e., $\{g(t)\}_{t\geq t_0}$ is bounded. □

Lemma 9. *Consider the flow (2) and the underlying communication graph $\mathcal{G}_{\sigma(t)}$. Suppose there exists $M > 0$ such that $\|\mathbf{x}(t)\| \leq M$ for all $t \geq t_0$. Suppose $\mathcal{G}_\sigma(t)$ is uniformly jointly connected. Let $\mathbf{x}_i(t)$ for all i denote the state held by node i of $\mathcal{G}_{\sigma(t)}$. Define $\Phi(t) = \max_{1\leq i,j\leq N}\|\mathbf{x}_i(t)-\mathbf{x}_j(t)\|$ and a continuous function $\alpha : \mathbb{R}^{\geq t_0} \to \mathbb{R}^+$. If $\int_{t_0}^{\infty}\alpha^2(t)\mathrm{d}t < \infty$, then $\int_{t_0}^{\infty}\alpha(t)\Phi(t)\mathrm{d}t < \infty$.*

Proof. By Shi and Johansson (2013), we know that there exist $C_1 > 0, C_2 > 0$ such that for all $k \geq 0$ and $kC_1 \leq t-t_0 \leq (k+1)C_1$, the following inequalities hold

$$\Phi(t) \leq \Phi(kC_1) + C_2\int_{kC_1+t_0}^{(k+1)C_1+t_0}\alpha(t)\mathrm{d}t \tag{11}$$

$$\Phi((k+1)C_1+t_0) \leq \beta\Phi(kC_1+t_0) + C_2\int_{kC_1+t_0}^{(k+1)C_1+t_0}\alpha(t)\mathrm{d}t \tag{12}$$

with $\beta \in (0, 1)$. Define $\omega_k := \int_{kC_1+t_0}^{(k+1)C_1+t_0}\alpha(t)\mathrm{d}t$ and $\alpha^* := \sup_{t\geq t_0}\alpha(t)$. Then the proof is completed by the following inequalities

$$\begin{aligned}
&\int_{t_0}^{\infty}\alpha(t)\Phi(t)\mathrm{d}t = \sum_{k=0}^{\infty}\int_{kC_1+t_0}^{(k+1)C_1+t_0}\alpha(t)\Phi(t)\mathrm{d}t \\
&\overset{(a)}{\leq}\sum_{k=0}^{\infty}\int_{kC_1+t_0}^{(k+1)C_1+t_0}\alpha(t)\Big(\Phi(kC_1+t_0) + C_2\int_{kC_1+t_0}^{(k+1)C_1+t_0}\alpha(s)\mathrm{d}s\Big)\mathrm{d}t \\
&= \sum_{k=0}^{\infty}\omega_k\Phi(kC_1+t_0) + C_2\sum_{k=0}^{\infty}\Big(\int_{kC_1+t_0}^{(k+1)C_1+t_0}\alpha(t)\mathrm{d}t\Big)^2 \\
&\overset{(b)}{\leq}\sum_{k=0}^{\infty}\omega_k\Phi(kC_1+t_0) + C_1C_2\int_{t_0}^{\infty}\alpha^2(t)\mathrm{d}t \\
&\overset{(c)}{\leq}\sum_{k=1}^{\infty}\omega_k\Big(\beta^k\Phi(t_0) + C_2\sum_{r=1}^{k}\beta^{k-r}\omega_{r-1}\Big) + \omega_0\Phi(t_0) \\
&\quad + C_1C_2\int_{t_0}^{\infty}\alpha^2(t)\mathrm{d}t,
\end{aligned}$$

where (a) is from (11), (b) is due to Cauchy–Schwarz inequality, and (c) is from (12). This allows us to further conclude

$$\begin{aligned}
&\int_{t_0}^{\infty}\alpha(t)\Phi(t)\mathrm{d}t \leq \alpha^*C_1\Phi(t_0)\sum_{k=1}^{\infty}\beta^k + \frac{C_2}{2}\cdot \\
&\sum_{k=1}^{\infty}\sum_{r=1}^{k}\beta^{k-r}(\omega_k^2+\omega_{r-1}^2) + \omega_0\Phi(t_0) + C_1C_2\int_{t_0}^{\infty}\alpha^2(t)\mathrm{d}t \\
&\leq \frac{\alpha^*\beta C_1\Phi(t_0)}{1-\beta} + \frac{C_2}{1-\beta}\sum_{k=1}^{\infty}\omega_k^2 + \omega_0\Phi(t_0) \\
&\quad + C_1C_2\int_{t_0}^{\infty}\alpha^2(t)\mathrm{d}t \\
&= \Big(\frac{C_2}{1-\beta} + C_1C_2\Big)\int_{t_0}^{\infty}\alpha^2(t)\mathrm{d}t + \Big(\frac{\alpha^*\beta C_1}{1-\beta} + \omega_0\Big)\Phi(t_0),
\end{aligned}$$

which completes the proof of the lemma. □

Lemma 10. *Consider a nonnegative sequence $\{u_k\}_{k=0}^{\infty}$ satisfying $u_{k+1} \leq qu_k + d_k$, where $0 \leq q < 1$ and $d_k = \mathcal{O}(\frac{1}{k^\lambda})$ with $0 < \lambda < 1$. Then $u_k = \mathcal{O}(\frac{1}{k^\lambda})$.*

Proof. Since $d_k = \mathcal{O}(\frac{1}{k^\lambda})$, there exist $c, k_0 > 0$ such that $d_k \leq \frac{c}{k^\lambda}$ for all $k \geq k_0$. Introduce $v_k = k^\lambda u_k$. Based on the definition of u_k, we have for $k > k_0$, $v_k \leq q^{k-k_0}\big(\frac{k}{k_0}\big)^\lambda v_{k_0} + c\sum_{\kappa=k_0}^{k-1}q^{k-\kappa-1}\big(\frac{k}{\kappa}\big)^\lambda$. By performing ratio test on the series at right-hand-side above, we find that it converges as k goes to infinity. Hence, $v_k = \mathcal{O}(1)$. This in turn gives $u_k = k^{-\lambda}v_k = \mathcal{O}(k^{-\lambda})$. □

Appendix B. Proof of Theorem 2

The proof starts by showing that $\mathbf{x}(t)$ is bounded. Consider $Q_K(\mathbf{x}, t) = \mathbf{x}^\top\mathbf{M}(t)\mathbf{x} = K\sum_{\{i,j\}\in\mathcal{E}}[\mathbf{A}]_{ij}\|\mathbf{x}_j-\mathbf{x}_i\|^2 + \alpha(t)\sum_{i=1}^{N}|\mathbf{h}_i^\top\mathbf{x}_i|^2$ with $\mathbf{x} \neq 0$. Then clearly $Q_K(\mathbf{x}, t) \geq 0$ and the equality holds only if $\mathbf{x}_i = \mathbf{x}_j$ for any i, j and $\mathbf{h}_i^\top\mathbf{x}_i = 0$ for all i. Because $\mathrm{rank}(\mathbf{H}) = m$ by hypothesis, there does not exist $\mathbf{x} \neq 0$ such that $Q_K(\mathbf{x}, t) = 0$, i.e., $Q_K(\mathbf{x}, t) > 0$ for $\mathbf{x} \neq 0$. Therefore, $\mathbf{M}(t)$ is positive-definite for all t. Similarly, $\mathbf{P} := \mathbf{L}\otimes\mathbf{I}_m + \tilde{\mathbf{H}}$ is also positive-definite. Under Assumption 1(ii), we know that there exists sufficiently large t^* such that $\alpha(t) < K$ for all $t \geq t^*$. By Theorem 4.2.2 in Horn and Johnson (2012), we know $Q_K(\mathbf{x}, t) \geq \alpha(t)\mathbf{x}^\top\mathbf{P}\mathbf{x} \geq \alpha(t)\sigma_{\mathrm{m}}(\mathbf{P})\|\mathbf{x}\|^2$ for any $\mathbf{x}$ and all $t \geq t^*$. Let $h(t) = \|\mathbf{x}(t)\|^2$. Then $\frac{\mathrm{d}}{\mathrm{d}t}h(t) \leq -2\alpha(t)\sigma_{\mathrm{m}}(\mathbf{P})\|\mathbf{x}(t)\|^2 + 2\alpha(t)\|\mathbf{x}(t)\|\|\mathbf{z}_H\|$ for $t \geq t^*$. Consider

$$\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{h(t)} \leq -\alpha(t)\sigma_{\mathrm{m}}(\mathbf{P})\sqrt{h(t)} + \alpha(t)\|\mathbf{z}_H\|,\ \ t \geq t^*. \tag{13}$$

By Lemma 8(ii), identifying $g(t)$ with $\sqrt{h(t)}$, we have that $\sqrt{h(t)} = \|\mathbf{x}(t)\|$ is bounded for $t \geq t^*$. Due to the continuity of $\mathbf{x}(t)$, $\|\mathbf{x}(t)\|$ is bounded for all $t \geq t_0$.

For the second step of the proof, we first denote $\bar{\mathbf{x}}(t) := \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i(t)$ and $\bar{\mathbf{x}}^\diamond(t) := \mathbf{1}_N\otimes\bar{\mathbf{x}}(t)$. By simple calculation, it can be shown that $\dot{\bar{\mathbf{x}}}^\diamond(t) = \mathbf{1}_N\otimes(\frac{1}{N}\sum_{i=1}^{N}\dot{\mathbf{x}}_i(t)) = -\mathbf{1}_N\otimes(\frac{\alpha(t)}{2N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i))$. Then by Horn and Johnson (2012)

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 = 2\langle\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t),\ \dot{\mathbf{x}}(t)-\dot{\bar{\mathbf{x}}}^\diamond(t)\rangle \\
&= 2\langle\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t), -K(\mathbf{L}\otimes\mathbf{I}_m)(\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t))\rangle + \beta(t) \\
&\leq -2\sigma_2(\mathbf{L})K\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 + \beta(t),
\end{aligned} \tag{14}$$

where $\beta(t) = 2\alpha(t)\langle\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t), \mathbf{z}_H - \tilde{\mathbf{H}}\mathbf{x}(t) + \mathbf{1}_N\otimes(\frac{1}{2N}\sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i(t)))\rangle$. Under Assumption 1(ii) and by the claim that $\|\mathbf{x}(t)\|$ is bounded, we know that $\lim_{t\to\infty}\beta(t) = 0$. By Lemma 8 (i), $\lim_{t\to\infty}\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 = 0$, i.e., the dynamical system (3) achieves a consensus.

Now we turn to the last step of the proof and analyze the relationship between $\bar{\mathbf{x}}(t)$ and the optimal point $\mathbf{y}^*$. Let

$$\omega(t) = \frac{\alpha(t)}{N}\Big\langle\bar{\mathbf{x}}(t)-\mathbf{y}^*, \nabla f(\bar{\mathbf{x}}(t)) - \sum_{i=1}^{N}\nabla f_i(\mathbf{x}_i(t))\Big\rangle.$$

By Lemma 6, $f(\mathbf{y})$ is $2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})$-strongly convex, and there holds

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{dt}}\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2 = 2\langle\bar{\mathbf{x}}(t)-\mathbf{y}^*, \dot{\bar{\mathbf{x}}}(t)\rangle\\ &= -\frac{\alpha(t)}{N}\langle\bar{\mathbf{x}}(t)-\mathbf{y}^*, \nabla f(\bar{\mathbf{x}}(t))\rangle + \omega(t)\\ &\leq -\frac{\alpha(t)}{N}(f(\bar{\mathbf{x}}(t))-f(\mathbf{y}^*)+\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2)+\omega(t) \quad (15)\\ &\leq -\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})\alpha(t)}{N}\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2+\omega(t). \quad (16)\end{aligned}$$

Since $\lim_{t\to\infty}(\bar{\mathbf{x}}(t)-\mathbf{x}_i(t)) = 0$, namely $\lim_{t\to\infty}(\nabla f(\bar{\mathbf{x}}(t)) - \sum_{i=1}^N \nabla f_i(\mathbf{x}_i(t))) = 0$, we have $\lim_{t\to\infty}\frac{\omega(t)}{\alpha(t)} = \frac{1}{N}\lim_{t\to\infty}\langle\bar{\mathbf{x}}(t)-\mathbf{y}^*, \nabla f(\bar{\mathbf{x}}(t))-\sum_{i=1}^N\nabla f_i(\mathbf{x}_i(t))\rangle = 0$ by the boundedness of $\|\mathbf{x}(t)\|$. Therefore, $\lim_{t\to\infty}\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2 = 0$ by Assumption 1 (i) and Lemma 8 (i), i.e., (3) reaches a consensus and finally all nodes hold the value of the least-squares solution to (1), which completes the proof. □

Appendix C. Proof of Theorem 3

We continue to use the definitions of $\beta(t), \bar{\mathbf{x}}(t), \bar{\mathbf{x}}^\diamond(t), \omega(t)$ in the proof of Theorem 2.

(i) Let $\alpha(t) = \mathcal{O}(\frac{1}{t})$. Due to the boundedness of $\|\mathbf{x}(t)\|$ proved by (13), there exist $c_\beta, M_0 > 0$ such that

$$\begin{aligned}\beta(t) &\leq c_\beta t^{-1}\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\| \quad (17)\\ &\leq c_\beta M_0 t^{-1} \quad (18)\end{aligned}$$

for all $t > t_0$. By applying Lemma 8 to (14) and based on (18), one has for all $t > t_0$

$$\begin{aligned}&\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2\\ &\leq c_0 e^{-2\sigma_2(\mathbf{L})Kt} + c_\beta M_0\int_{t_0}^t \frac{e^{2\sigma_2(\mathbf{L})K(s-t)}}{s}\mathrm{d}s \quad (19)\end{aligned}$$

with $c_0 = \|\mathbf{x}(t_0)-\bar{\mathbf{x}}^\diamond(t_0)\|^2$. Clearly, (19) with Lemma 7 yields that there exist $q > 0, T_1 > t_0$ satisfying

$$\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 \leq c_0 e^{-2\sigma_2(\mathbf{L})Kt} + M_0 c_\beta q t^{-1} \quad (20)$$

for all $t \geq T_1$. Let $\bar{M} = \max(M_0, 1)$. It is evident that $\bar{M} \geq \bar{M}^e$ for all $0 < e \leq 1$. Introduce $T_2 \geq 2c_\beta q$ satisfying $c_0 e^{-2\sigma_2(\mathbf{L})Kt} \leq (2c_\beta q t^{-1})^2\bar{M}/2$ for all $t \geq T_2$. Then it follows from (20) that

$$\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 \leq 2c_\beta q t^{-1}\bar{M} \quad (21)$$

for all $t \geq T := \max(T_1, T_2)$. It can be noticed that (17) shows $\beta(t)$ is bounded by a function of $\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|$. Hence (21) leads to a tighter bound of $\beta(t)$ than (18): $\beta(t) = \mathcal{O}(t^{-\frac{3}{2}})$. In detail, one has from (17) and (21)

$$\beta(t) \leq c_\beta(2c_\beta q)^{\frac{1}{2}} t^{-\frac{3}{2}}\bar{M}^{\frac{1}{2}} \leq c_\beta(2c_\beta q)^{\frac{1}{2}} t^{-\frac{3}{2}}\bar{M} \quad (22)$$

for all $t \geq T$. Again we apply Lemmas 7 and 8 to (14) and based on (22), we have

$$\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 \leq c_0 e^{-2\sigma_2(\mathbf{L})Kt} + \frac{\bar{M}}{2}(2c_\beta q t^{-1})^{\frac{3}{2}} \quad (23)$$

for all $t \geq T$. Thus by (23) and the definition of T, the following inequality also holds for all $t \geq T$

$$\begin{aligned}\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 &\leq \frac{\bar{M}}{2}(2c_\beta q t^{-1})^2 + \frac{\bar{M}}{2}(2c_\beta q t^{-1})^{\frac{3}{2}}\\ &\leq (2c_\beta q t^{-1})^{\frac{3}{2}}\bar{M}\end{aligned}$$

Based on (17), by recursively applying Lemmas 7 and 8 on (14) with constantly updated upper bounds of $\beta(t)$ initialized by (18), we can obtain a sequence of bounds on $\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2$ as following:

$$\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 \leq (2c_\beta q t^{-1})^{a_r}\bar{M},\ r = 1, 2, \ldots \quad (24)$$

for all $t \geq T$, where $a_{r+1} = \frac{1}{2}a_r + 1,\ a_1 = 1$. Such q, T are well defined by letting $\mu^* = 2\sigma_2(\mathbf{L})K, \lambda_m = 1, \lambda_M = 2$ in Lemma 7. Clearly, a_r in (24) monotonically goes to 2 as r go to infinity and $1 \leq a_r < 2$ for all r. Then there holds

$$\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|^2 = \mathcal{O}(t^{-2}). \quad (25)$$

From the Cauchy–Schwarz inequality and (25)

$$\begin{aligned}\omega(t) &= \frac{2\alpha(t)}{N}(\bar{\mathbf{x}}(t)-\mathbf{y}^*)^\top\sum_{i=1}^N \mathbf{h}_i\mathbf{h}_i^\top(\bar{\mathbf{x}}(t)-\mathbf{x}_i(t))\\ &\leq \rho\alpha(t)\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|\|\mathbf{x}(t)-\bar{\mathbf{x}}^\diamond(t)\|\\ &= \mathcal{O}\big(t^{-2}\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|\big), \quad (26)\end{aligned}$$

where $\rho := \max\{2N^{-\frac{1}{2}}\|\mathbf{h}_i\|^2 : i \in \mathcal{V}\}$. We apply Lemma 8 to (16) using the bound in (26) and obtain

$$\begin{aligned}\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2 &= \mathcal{O}\Big(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}\Big) + \mathcal{O}\Big(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}\Big)\\ &\cdot\int_{t_0}^t \mathcal{O}\Big(s^{\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}-2}\cdot\|\bar{\mathbf{x}}(s)-\mathbf{y}^*\|\Big)\mathrm{d}s. \quad (27)\end{aligned}$$

Depending on whether $s^{\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}-2}\cdot\|\bar{\mathbf{x}}(s)-\mathbf{y}^*\| = \mathcal{O}(s^{-1})$, the integral part in (27) falls into two different function classes. Therefore, we will discuss the bound of $\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2$ in two cases (a) and (b).

(a) We assume $\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H}) \neq N$. Define a set $\mathscr{U} \subset [1, 2)$ with $\mathscr{U} := \big\{\sum_{i=1}^r\big(\frac{1}{2}\big)^{i-1} : r = 1, 2, \ldots\big\}\bigcup\{2\}$. We will see the proof of (a) can be achieved under two complementary scenarios.

[Scenario 1] Suppose $\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N} \in \mathbb{R}^+\setminus\mathscr{U}$. From (27) with the fact $\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\| = \mathcal{O}(1)$

$$\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2 = \mathcal{O}\big(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}} + t^{-1}\big). \quad (28)$$

Define two sequences $\{b_r\}_{r=1,2,\ldots}$ and $\{\hat{b}_r\}_{r=1,2,\ldots}$ with $b_{r+1} = \frac{1}{2}b_r - 1,\ b_1 = -\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}$ and $\hat{b}_{r+1} = \frac{1}{2}\hat{b}_r - 1,\ \hat{b}_1 = -1$. Direct verification shows $b_r \neq -\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}$, $\forall r \geq 2$ and $\hat{b}_r \neq -\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}$, $\forall r \geq 1$. It is evident that they guarantee that no integral of $\mathcal{O}(s^{-1})$ arises from the following iteration process. Clearly

$$\begin{aligned}\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2 &\overset{(a)}{=} \mathcal{O}\Big(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}\Big) + \mathcal{O}\Big(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}\Big)\\ &\cdot\int_{t_0}^t\mathcal{O}\Big(s^{\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}-2}\cdot\big(s^{-\frac{\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}+s^{-\frac{1}{2}}\big)\Big)\mathrm{d}s\\ &\overset{(b)}{=}\mathcal{O}\left(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}} + \left|\frac{\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}-1\right|^{-1}\cdot\right.\\ &\left. t^{-\frac{\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}-1} + \left|\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}-\frac{3}{2}\right|^{-1}t^{-\frac{3}{2}}\right), \quad (29)\end{aligned}$$

where (a) comes from (27) and (28), and (b) is obtained by direct calculation. We apply a sufficiently large positive integer ζ of the recursions as from (28) to (29) and obtain the following bound:

$$\begin{aligned}&\|\bar{\mathbf{x}}(t)-\mathbf{y}^*\|^2\\ &= \mathcal{O}\left(t^{b_1} + \sum_{r=2}^{\zeta}\prod_{l=2}^{r}\left|b_l+\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}\right|^{-1}t^{b_r}\right.\\ &\left.+\prod_{l=2}^{\zeta}\left|\hat{b}_l+\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}\right|^{-1}t^{\hat{b}_\zeta}\right). \quad (30)\end{aligned}$$

Next we show that $b_1 > b_2 > \cdots$ if $b_1 > -2$, and $b_1 < b_2 < \cdots < -2$ otherwise. We first suppose $b_r > -2$. Then it is clear that both $b_{r+1} - b_r < 0$ and $b_{r+1} > -2$ hold. Hence, if $b_1 > -2$, b_r is a strictly decreasing sequence and asymptotically goes to -2. Similarly, it can be shown that b_r is a strictly increasing sequence and goes to -2 if $b_1 < -2$. Therefore, if $b_1 > -2$, then with (30) we have $\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}(t^{b_1})$. Otherwise, $\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}(t^{\hat{b}_\zeta})$. Due to the arbitrariness of ζ, $\epsilon = \hat{b}_\zeta + 2$ can be sufficiently close to zero. In conclusion, one has for any $0 < \epsilon < 2$, there holds

$$\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}\big(t^{\max(-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}, \epsilon-2)}\big). \tag{31}$$

[Scenario 2] Suppose $\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N} \in \mathscr{U}$. For ease of presentation, we define $\hat{b}_0 = 0$. Then there exists $r^* \in \{1, 2, \ldots\}$ such that $\hat{b}_{r^*-1} = 2 - \frac{4\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}$ and $\hat{b}_{r^*} = -\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}$. Similar to the process of obtaining (31), we apply r^* rounds of iterations based on (27), and arrive at

$$\begin{aligned}\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 &= \mathcal{O}\left(\sum_{r=1}^{r^*} t^{b_r}\right) + \mathcal{O}\left(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}\right) \\ &\quad \cdot \int_{t_0}^{t} \mathcal{O}\left(s^{\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}-2} \cdot s^{\frac{1}{2}\hat{b}_{r^*-1}}\right)\mathrm{d}s \\ &= \mathcal{O}\left(\sum_{r=1}^{r^*} t^{b_r} + t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}} \log t\right).\end{aligned} \tag{32}$$

Noticing the fact that the scenario hypothesis $\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N} \in [1, 2)$, we claim that there exists $\delta \in (0, 2 - 2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})/N)$ such that

$$\log t = \mathcal{O}(t^{\delta}). \tag{33}$$

Then it follows from (32) and (33)

$$\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}\left(\sum_{r=1}^{r^*} t^{b_r} + t^{\delta - \frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}\right). \tag{34}$$

Define a sequence $\{d_r\}_{r=1,2,\ldots}$ with $d_{r+1} = \frac{1}{2}d_r - 1,\ d_0 = \delta - \frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}$. Then it can be easily verified that

$$d_1 < -2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})/N < d_0, \tag{35}$$

which implies that there is no element in $\{d_r\}_{r=1,2,\ldots}$ equal to $-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}$. Now we continue the iteration from (34), during which process (35) guarantees no integral of $\mathcal{O}(s^{-1})$ arises. With any sufficiently large ζ, ζ iterations indicate that the following bound holds

$$\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}\left(\sum_{r=1}^{\zeta+r^*} t^{b_r} + t^{d_\zeta}\right) = \mathcal{O}\left(t^{-\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N}}\right). \tag{36}$$

(b) We assume $\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H}) = N$. Similarly, (27) gives

$$\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}(t^{-2} + t^{-1}). \tag{37}$$

Starting from (37) and based on (27), we obtain

$$\begin{aligned}\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 &= \mathcal{O}\big(t^{-2}\big) + \mathcal{O}\big(t^{-2}\big)\int_{t_0}^{t}\mathcal{O}\big(s^{-1} + s^{-\frac{1}{2}}\big)\mathrm{d}s \\ &= \mathcal{O}\big(t^{-2}\big) + \mathcal{O}\big(t^{-2}\log t\big) + \mathcal{O}\big(t^{-\frac{3}{2}}\big).\end{aligned} \tag{38}$$

Again, we repeat the process from (37) to (38) recursively for $\zeta > 0$ times and obtain

$$\begin{aligned}&\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 \\ &= \mathcal{O}\left(t^{-2} + t^{-2}\sum_{r=1}^{\zeta}(\log t)^{-\hat{b}_r} + t^{\hat{b}_\zeta}\right) \\ &= \mathcal{O}\left(t^{-2}\left(\sum_{r=1}^{\zeta}(\log t)^{-\hat{b}_r} + t^{\hat{b}_\zeta+2}\right)\right).\end{aligned} \tag{39}$$

Since $\sum_{r=1}^{\zeta}(\log t)^{-\hat{b}_r} = \mathcal{O}(t^{\hat{b}_\zeta+2})$ for any $\zeta > 0$, we have from (39)

$$\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}(t^{\hat{b}_\zeta}) = \mathcal{O}(t^{\epsilon-2}), \tag{40}$$

for any sufficiently small $\epsilon > 0$. In conclusion, the proof of (i) can be achieved by (31), (36) and (40).

(ii) Let $\alpha(t) = \mathcal{O}(\frac{1}{t^\lambda})$. Next we will apply the similar method in (17)–(25) to the analysis of this case, while omitting the coefficients for simplicity. Immediately there holds

$$\beta(t) = \mathcal{O}(t^{-\lambda}). \tag{41}$$

Starting from (41), similar recursive applications of Lemmas 7 and 8 to (14) result in

$$\begin{aligned}\|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 &= \int_{t_0}^{t}\mathcal{O}\left(\frac{e^{2\sigma_2(\mathbf{L})K(s-t)}}{s^\lambda}\right)\mathrm{d}s = \mathcal{O}\left(\frac{1}{t^\lambda}\right) \\ \|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 &= \int_{t_0}^{t}\mathcal{O}\left(\frac{e^{2\sigma_2(\mathbf{L})K(s-t)}}{s^{\frac{3}{2}\lambda}}\right)\mathrm{d}s = \mathcal{O}\left(\frac{1}{t^{\frac{3}{2}\lambda}}\right) \\ &\cdots \\ \|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 &= \mathcal{O}\left(\frac{1}{t^{2\lambda}}\right).\end{aligned} \tag{42}$$

It follows from (42) and the fact $\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\| = \mathcal{O}(1)$ that

$$\omega(t) = \mathcal{O}\big(\alpha(t)\|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|\big) = \mathcal{O}(t^{-2\lambda}). \tag{43}$$

With (43) inserted in (16), Lemma 8 and simple change of variables yield

$$\begin{aligned}\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 &= \int_{t_0}^{t}\mathcal{O}\left(\frac{e^{\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N(1-\lambda)}(s^{1-\lambda}-t^{1-\lambda})}}{s^{2\lambda}}\right)\mathrm{d}s \\ &= \int_{t_0^{1-\lambda}}^{t^{1-\lambda}}\mathcal{O}\left(\frac{e^{\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N(1-\lambda)}(s-t^{1-\lambda})}}{s^{\frac{\lambda}{1-\lambda}}}\right)\mathrm{d}s.\end{aligned} \tag{44}$$

Clearly, one obtains by applying Lemma 7 to (44)

$$\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 = \mathcal{O}(t^{-\lambda}). \tag{45}$$

Again starting from (45), finite recursive applications of Lemmas 7 and 8 on (16) give

$$\begin{aligned}\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 &= \mathcal{O}\left(\frac{e^{\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})}{N(1-\lambda)}(s-t^{1-\lambda})}}{s^{\frac{\frac{3}{2}\lambda}{1-\lambda}}}\right)\mathrm{d}s = \mathcal{O}\big(t^{-\frac{3}{2}\lambda}\big) \\ &\cdots \\ \|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 &= \mathcal{O}\big(t^{\epsilon-2\lambda}\big)\end{aligned}$$

for any $0 < \epsilon < 2\lambda$, which completes the proof of (b). □

Appendix D. Proof of Theorem 4

Denote the averaged state at time t by $\bar{\mathbf{x}}(t) = \sum_{i=1}^{N}\mathbf{x}_i(t)/N$ and $\bar{\mathbf{x}}^\diamond(t) = \mathbf{1}_N \otimes \bar{\mathbf{x}}(t)$. Denote $h(t) = \|\mathbf{x}(t)\|^2$. Let $\mathbf{L}_{\sigma(t)}$ be the Laplacian of the graph $\mathcal{G}_{\sigma(t)} \in \mathcal{Q}^*$. Let $\mathbf{P}_{\sigma(t)} = \mathbf{L}_{\sigma(t)} \otimes \mathbf{I}_m + \tilde{\mathbf{H}}$. By a minor variant of a step in the proof of Theorem 2, one has $\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{h(t)} \le -\alpha(t)\sigma_{\mathrm{m}}(\mathbf{P}_{\sigma(t)})\sqrt{h(t)} + \alpha(t)\|\mathbf{z}_H\|,\ t \ge t^*$. Since $|\mathcal{Q}^*| < \infty$, the quantity $\min_{t \ge t_0}\sigma_{\mathrm{m}}(\mathbf{P}_{\sigma(t)}) = \sigma_{\mathrm{m}}^*$ is well-defined and positive. Then it follows $\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{h(t)} \le -\alpha(t)\sigma_{\mathrm{m}}^*\sqrt{h(t)} + \alpha(t)\|\mathbf{z}_H\|,\ t \ge t^*$. Thus a conclusion can be drawn that $\|\mathbf{x}(t)\|$ is bounded. Similarly, $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 \le -2\sigma_2(\mathbf{L}_{\sigma(t)})K\|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 + \beta(t)$, where

$\beta(t) = 2\alpha(t)\langle \mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t), \mathbf{z}_H - \tilde{\mathbf{H}}\mathbf{x}(t) + \mathbf{1}_N \otimes (\frac{1}{2N}\sum_{i=1}^N \nabla f_i(\mathbf{x}_i))\rangle$. Then we select $\sigma_2^* = \min_{t\ge t_0} \sigma_2(\mathbf{L}_{\sigma(t)})$ so that $\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 \le -2\sigma_2^* K\|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 + \beta(t)$. Similarly, by Lemma 8 and the fact that $\lim_{t\to\infty}\beta(t) = 0$, we can conclude $\lim_{t\to\infty}\|\mathbf{x}(t) - \bar{\mathbf{x}}^\diamond(t)\|^2 = 0$, i.e., the system (3) achieves a consensus over switching networks. Next we prove that the consensus value is exactly the least-squares solution of (1). Let $\mathbf{y}^* \in \mathcal{Y}_{\mathrm{LS}}$. Recall in (15) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 \le -\frac{\alpha(t)}{N}(f(\bar{\mathbf{x}}(t)) - f(\mathbf{y}^*)) + \omega(t), \tag{46}$$

where $\omega(t) = \frac{\alpha(t)}{N}\langle \bar{\mathbf{x}}(t) - \mathbf{y}^*, \sum_{i=1}^N \mathbf{h}_i\mathbf{h}_i^\top(\mathbf{x}_i(t) - \bar{\mathbf{x}}(t))\rangle$. By simple calculation and the fact that $\|\mathbf{x}(t)\|$ is bounded, it can be obtained that $|\omega(t)| \le \frac{\alpha(t)\Phi(t)}{N}\|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|\sum_{i=1}^N \|\mathbf{h}_i\mathbf{h}_i^\top\| = \mathcal{O}(\alpha(t)\Phi(t))$, where $\Phi(t) = \max_{1\le i,j\le N}\|\mathbf{x}_i(t) - \mathbf{x}_j(t)\|$. By Lemma 9, $\int_{t_0}^\infty |\omega(t)|\,\mathrm{d}t < \infty$, which implies $\int_{t_0}^\infty \omega(t)\mathrm{d}t < \infty$. Note that the constantly connected graph considered in this theorem is clearly uniformly jointly connected. Based on (46), we have

$$\begin{aligned}&\frac{1}{N}\int_{t_0}^t \alpha(s)(f(\bar{\mathbf{x}}(s)) - f(\mathbf{y}^*))\mathrm{d}s\\ &\le \|\bar{\mathbf{x}}(t_0) - \mathbf{y}^*\|^2 - \|\bar{\mathbf{x}}(t) - \mathbf{y}^*\|^2 + \int_{t_0}^t \omega(s)\mathrm{d}s.\end{aligned} \tag{47}$$

Since $\mathbf{x}(t)$ is bounded and $\int_{t_0}^\infty \omega(t)\mathrm{d}t < \infty$, the right-hand side of (47) is less than infinity, which implies $\int_{t_0}^\infty \alpha(s)(f(\bar{\mathbf{x}}(s)) - f(\mathbf{y}^*))\mathrm{d}s < \infty$. Since $\int_{t_0}^\infty \alpha(s)\mathrm{d}s = \infty$, $\liminf_{s\to\infty}(f(\bar{\mathbf{x}}(s)) - f(\mathbf{y}^*)) = 0$. Since the states $\mathbf{x}_i(t)$ for all i are bounded, we can find a sequence $\{s_k\}_{k\ge 0}$ such that $\lim_{k\to\infty} f(\bar{\mathbf{x}}(s_k)) = f(\mathbf{y}^*)$. By Bolzano–Weierstrass theorem, we select $\{s_{k_r}\}_{r\ge 0}$ as a subsequence of $\{s_k\}_{k\ge 0}$ such that $\lim_{r\to\infty}\bar{\mathbf{x}}(s_{k_r}) = \hat{\mathbf{y}}$ for some $\hat{\mathbf{y}}$. It is obvious that $f(\hat{\mathbf{y}}) = f(\mathbf{y}^*)$, i.e. $\hat{\mathbf{y}} \in \mathcal{Y}$ is also an optimal solution. Moreover, by replacing $\mathbf{y}^*$ with $\hat{\mathbf{y}}$ in (46), we have by the convexity of the function f

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\bar{\mathbf{x}}(t) - \hat{\mathbf{y}}\|^2 \le \omega(t) \le |\omega(t)|\,. \tag{48}$$

In order to prove by contradiction that $\|\bar{\mathbf{x}}(t) - \hat{\mathbf{y}}\|^2$ is convergent, we suppose, by the boundedness of $\bar{\mathbf{x}}(t)$, that there exist sequences $\{t_{s_k}\}$, $\{t_{r_k}\}$ satisfying that $l_1 = \lim_{k\to\infty}\|\bar{\mathbf{x}}(t_{s_k}) - \hat{\mathbf{y}}\|^2$, $l_2 = \lim_{k\to\infty}\|\bar{\mathbf{x}}(t_{r_k}) - \hat{\mathbf{y}}\|^2$, respectively and $l_1 \ne l_2$. We also assume, without loss of generality, $l_1 - l_2 = \epsilon_0 > 0$. Then by (48) we have $l_1 - l_2 = \lim_{k\to\infty}\int_{t_{r_k}}^{t_{s_k}} \frac{\mathrm{d}}{\mathrm{d}t}\|\bar{\mathbf{x}}(t) - \hat{\mathbf{y}}\|^2\mathrm{d}t \le \lim_{k\to\infty}\int_{t_{r_k}}^{t_{s_k}}|\omega(t)|\,\mathrm{d}t$. Since $\int_{t_0}^\infty |\omega(t)|\,\mathrm{d}t < \infty$ as proved above, it can be concluded that $\lim_{k\to\infty}\int_{t_{r_k}}^{t_{s_k}}|\omega(t)|\,\mathrm{d}t = 0$, i.e., there exists $k_0 > 0$ such that $\int_{t_{r_k}}^{t_{s_k}}|\omega(t)|\,\mathrm{d}t < \epsilon_0$ for all $k > k_0$. This implies $l_1 - l_2 < \epsilon_0$, which is contradictory to the assumption that $l_1 - l_2 = \epsilon_0$. Hence $\|\bar{\mathbf{x}}(t) - \hat{\mathbf{y}}\|^2$ is convergent. Since it has been shown that there exists a sequence $\{s_r\}_{r\ge 0}$ such that $\lim_{r\to\infty}\bar{\mathbf{x}}(s_r) = \hat{\mathbf{y}}$, we have $\lim_{t\to\infty}\|\bar{\mathbf{x}}(t) - \hat{\mathbf{y}}\|^2 = 0$. Due to the fact that the network achieves a consensus, there holds $\lim_{t\to\infty}\mathbf{x}_i(t) = \hat{\mathbf{y}}$ for all $i \in \mathcal{V}$.

Appendix E. Proof of Theorem 5

We continue to use the notations $\mathbf{x}(k)$, $\bar{\mathbf{x}}(k)$ in the continuous analysis. We first show the stability of (4). Introduce $\mathbf{W} \in \mathbb{R}^{N\times N}$ with $[\mathbf{W}]_{ij} = w_{ij}$. Then (4) can be written compactly as

$$\mathbf{x}(k+1) = (\mathbf{W}\otimes\mathbf{I} - \eta(k)\tilde{\mathbf{H}})\mathbf{x}(k) + \eta(k)\mathbf{z}_H. \tag{49}$$

Since $\mathbf{W}$ is stochastic, there holds $\mathbf{v}^\top(\mathbf{W}\otimes\mathbf{I})\mathbf{v} \le \|\mathbf{v}\|^2$ for any $\mathbf{v} \ne 0 \in \mathbb{R}^{Nm}$ with the form $\mathbf{v}^\top = [\mathbf{v}_1^\top \ \ldots \ \mathbf{v}_N^\top]$, $\mathbf{v}_i \in \mathbb{R}^m$, and the equality holds if and only if $\mathbf{v}_1 = \cdots = \mathbf{v}_N$. Because $\mathbf{H}$ has full column rank, by the continuity of eigenvalues of one matrix with respect to its components, we can see that for all sufficiently large k, the largest eigenvalue of the matrix $\mathbf{W}\otimes\mathbf{I} - \eta(k)\tilde{\mathbf{H}}$ is not greater than $1 - \eta(k)\mathbf{v}^\top\tilde{\mathbf{H}}\mathbf{v}/2$, where $\mathbf{v}$ is the unit eigenvector of the matrix $\mathbf{W}\otimes\mathbf{I}$ with the eigenvalue one. Consequently, it follows from (49) that for all sufficiently large k, $\|\mathbf{x}(k+1)\|^2 \le (1 - \eta(k)\mathbf{v}^\top\tilde{\mathbf{H}}\mathbf{v}/2)^2\|\mathbf{x}(k)\|^2 + \eta^2(k)\mathbf{z}_H^\top\mathbf{z}_H + 2\eta(k)\|\mathbf{z}_H\|\|\mathbf{x}(k)\|$. By the previous inequality, we can show by contradiction that $\{\mathbf{x}(k)\}$ is bounded. Now we assume $B = \max_{k\ge 0, i\in\mathcal{V}}\|\nabla f_i(\mathbf{x}_i(k))\|$. Define $\psi(k) = \max_{i,j\in\mathcal{V}}\|\mathbf{x}_i(k) - \mathbf{x}_j(k)\|$. Then with Lemma 3 in Hajnal and Bartlett (1958), there exists $0 < \delta < 1$ such that for all l

$$\psi((l+1)n) \le \delta\psi(ln) + \eta(ln)nB/2. \tag{50}$$

Then we can apply Lemma 10 to (50) to get $\psi(ln) = \mathcal{O}(\eta(ln))$. Furthermore, because $\psi(k+1) \le \psi(k) + \eta(k)B/2$ for all k, we have that $\psi(k) = \mathcal{O}(\eta(k))$. Let $\bar{\mathbf{x}}(k) = \sum_{i=1}^N \mathbf{x}_i(k)/N$. Further, one has $\sum_{i=1}^N\|\mathbf{x}_i(k) - \bar{\mathbf{x}}(k)\| \le N\psi(k) = \mathcal{O}(\eta(k))$. Define $u(k) = \|\bar{\mathbf{x}}(k) - \mathbf{y}^*\|$. Then along (4) there holds

$$\begin{aligned}u^2(k+1) = u^2(k) &+ \frac{\eta^2(k)}{4N^2}\Big\|\sum_{i=1}^N \nabla f_i(\mathbf{x}_i(k))\Big\|^2\\ &- \frac{\eta(k)}{N}\langle\bar{\mathbf{x}}(k) - \mathbf{y}^*, \sum_{i=1}^N \nabla f_i(\mathbf{x}_i(k))\rangle.\end{aligned} \tag{51}$$

Let $h^* = \max_i\|\mathbf{h}_i\|^2$. We continue to study the last two terms in (51) as follows.

$$\begin{aligned}&\frac{\eta^2(k)}{4N^2}\Big\|\sum_{i=1}^N \nabla f_i(\mathbf{x}_i(k))\Big\|^2 = \frac{\eta^2(k)}{4N^2}\Big\|\sum_{i=1}^N\big(\nabla f_i(\mathbf{x}_i(k))\\ &- \nabla f_i(\mathbf{y}^*)\big)\Big\|^2 \overset{(a)}{\le} h^*\eta^2(k)\sum_{i=1}^N\big\|\mathbf{x}_i(k) - \mathbf{y}^*\big\|^2\\ &\overset{(b)}{\le} h^*\eta^2(k)\sum_{i=1}^N\big\|\mathbf{x}_i(k) - \bar{\mathbf{x}}(k)\big\|^2 + Nh^*\eta^2(k)u^2(k)\\ &\overset{(c)}{=} Nh^*\eta^2(k)u^2(k) + \mathcal{O}(\eta^4(k)),\end{aligned} \tag{52}$$

where (a) and (b) are derived from norm inequalities, and (c) is obtained from the derived bound. Using the same trick in (16), one can obtain

$$\begin{aligned}&-\frac{\eta(k)}{N}\langle\bar{\mathbf{x}}(k) - \mathbf{y}^*, \sum_{i=1}^N \nabla f_i(\mathbf{x}_i(k))\rangle\\ &\le -\frac{2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})\eta(k)}{N}u^2(k) + \mathcal{O}(\eta^2(k)u(k)).\end{aligned} \tag{53}$$

With (51)–(53), one can find that for all sufficiently large k

$$\begin{aligned}&u^2(k+1) \le \big(1 - 2\sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})\eta(k)/N + Nh^*\eta^2(k)\big)u^2(k)\\ &+ \mathcal{O}\big(\eta^2(k)(u(k) + \eta^2(k))\big) \le \big(1 - \sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})\eta(k)/N\big)\\ &\cdot u^2(k) + \mathcal{O}\big(\eta^2(k)u(k) + \eta^4(k)\big).\end{aligned} \tag{54}$$

It follows from the boundedness of $\{\mathbf{x}(k)\}$ that $u(k) = \mathcal{O}(1)$. Hence, (54) becomes

$$u^2(k+1) \le \big(1 - \sigma_{\mathrm{m}}(\mathbf{H}^\top\mathbf{H})\eta(k)/N\big)u^2(k) + \mathcal{O}(\eta^2(k)). \tag{55}$$

The application of Lemma 5, Chapter 2 in Polyak (2010) to (55) shows $u(k) = \mathcal{O}(\eta^{\frac{1}{2}}(k))$. Inserting $u(k) = \mathcal{O}(\eta^{\frac{1}{2}}(k))$ into (54), we see that (55) also holds by replacing $\mathcal{O}(\eta^2(k))$ with $\mathcal{O}(\eta^{\frac{5}{2}}(k))$. Applying Lemma 5, Chapter 2 in Polyak (2010) again, we can get $u(k) = \mathcal{O}(\eta^{\frac{3}{4}}(k))$. The conclusion follows by repeating the previous procedure infinitely many times. This completes the proof.

References

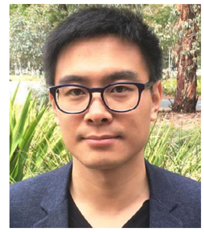
Anderson, B., Mou, S., Morse, A. S., & Helmke, U. (2015). Decentralized gradient algorithm for solution of a linear equation. *Numerical Algebra, Control and Optimization*, *6*(3), 319–328.

Cattivelli, F. S., Lopes, C. G., & Sayed, A. H. (2008). Diffusion recursive least-squares for distributed estimation over adaptive networks. *IEEE Transactions on Signal Processing*, *56*(5), 1865–1877.
Chen, S., Billings, S. A., & Luo, W. (1989). Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, *50*(5), 1873–1896.
Childs, A. M. (2009). Universal computation by quantum walk. *Physical Review Letters*, *102*(18), Article 180501.
Gharesifard, B., & Cortés, J. (2014). Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, *59*(3), 781–786.
Golub, G. H., & Van Loan, C. F. (1980). An analysis of the total least squares problem. *SIAM Journal on Numerical Analysis*, *17*(6), 883–893.
Grönwall, T. H. (1919). Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, *20*(4), 292–296.
Hajnal, J., & Bartlett, M. S. (1958). Weak ergodicity in non-homogeneous Markov chains. *Mathematical Proceedings of the Cambridge Philosophical Society*, *54*(2), 233–246.
Helmke, U., & Moore, J. B. (2012). *Optimization and dynamical systems*. Springer S. & B. Media.
Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.
Jadbabaie, A., Lin, J., & Morse, A. S. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, *48*(6), 988–1001.
Kar, S., & Moura, J. M. (2011). Gossip and distributed Kalman filtering: Weak consensus under weak detectability. *IEEE Transactions on Signal Processing*, *59*(4), 1766–1784.
Lines, L. R., & Treitel, S. (1984). Tutorial: A review of least-squares inversion and its application to geophysical problems. *Geophysical Prospecting*, *32*(2), 159–196.
Liu, Y., Lageman, C., Anderson, B. D. O., & Shi, G. (2019). An Arrow-Hurwicz-Uzawa type flow as least squares solver for network linear equations. *Automatica*, *100*, 187–193.
Liu, Y., Lou, Y., Anderson, B. D. O., & Shi, G. (2017). Network flows as least squares solvers for linear equations. In *56th IEEE Conference on Decision and Control* (pp. 1046–1051).
Liu, J., Morse, A. S., Nedic, A., & Basar, T. (2014). Stability of a distributed algorithm for solving linear algebraic equations. In *53rd IEEE Conference on Decision and Control* (pp. 3707–3712).
Liu, J., Mou, S., & Morse, A. S. (2013). An asynchronous distributed algorithm for solving a linear algebraic equation. In *52nd IEEE Conference on Decision and Control* (pp. 5409–5414).
Ljung, L. (2017). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, *22*(4), 551–575.
Lu, J., & Tang, C. Y. (2009). Distributed asynchronous algorithms for solving positive definite linear equations over networks-Part i: Agent networks. *IFAC Proceedings Volumes*, *42*(20), 252–257.
Lynch, N. A. (1996). *Distributed algorithms*. Elsevier.
Mesbahi, M., & Egerstedt, M. (2010). *Graph theoretic methods in multiagent networks*. Princeton Uni. Press.
Mou, S., Liu, J., & Morse, A. S. (2015). A distributed algorithm for solving a linear algebraic equation. *IEEE Transactions on Automatic Control*, *60*(11), 2863–2878.
Mou, S., Morse, A., & S (2013). A fixed-neighbor, distributed algorithm for solving a linear algebraic equation. In *European Control Conference* (pp. 2269–2273).
Nedić, A., & Olshevsky, A. (2015). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, *60*(3), 601–615.
Nedić, A., Olshevsky, A., & Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, *27*(4), 2597–2633.
Nedić, A., & Ozdaglar, A. (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, *54*(1), 48–61.
Nedić, A., Ozdaglar, A., & Parrilo, P. A. (2010). Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, *55*(4), 922–938.
Olfati-Saber, R., & Murray, R. M. (2004). Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, *49*(9), 1520–1533.
Orvieto, A., & Lucchi, A. (2019). Shadowing properties of optimization algorithms. *Advances in Neural Information Processing Systems*, 12671–12682.
Polyak, B. T. (2010). *Introduction to optimization*. New York: Inc. Publications Division.
Qi, B., Hou, Z., Li, L., Dong, D., Xiang, G., & Guo, G. (2013). Quantum state tomography via linear regression estimation. *Scientific Reports*, *3*, 3496.
Rabbat, M., Nowak, R., & Bucklew, J. (2005). Robust decentralized source localization via averaging. In *Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing*, (pp. v–1057).
Shi, G., Anderson, B. D. O., & Helmke, U. (2017). Network flows that solve linear equations. *IEEE Transactions on Automatic Control*, *62*(6), 2659–2674.
Shi, G., & Johansson, K. H. (2013). Robust consensus for continuous-time multiagent dynamics. *SIAM Journal on Control and Optimization*.
Su, W., Boyd, S., & Candes, E. (2016). A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research (JMLR)*, *17*(153), 1–43.
Sun, Y., Scutari, G., & Palomar, D. (2016). Distributed nonconvex multiagent optimization over time-varying networks. In *IEEE Asilomar Conference on Signals, Systems and Computers* (pp. 788–794).
Touri, B., & Gharesifard, B. (2015). Continuous-time distributed convex optimization on time-varying directed networks. In *54th IEEE Conference on Decision and Control* (pp. 724–729).
Tsitsiklis, J. N. (1984). *Problems in decentralized decision making and computation*. Massachusetts Inst of Tech Cambridge Lab for Information & Decision Systems.
Tsitsiklis, J. N., & Bertsekas, D. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, *31*(9), 803–812.
Wang, J., & Elia, N. (2009). Distributed solution of linear equations over unreliable networks. In *American Control Conference* (pp. 6471–6476).
Wang, J., & Elia, N. (2010). Control approach to distributed optimization. In *48th Annual Allerton Conference on Communication, Control, and Computing* (pp. 557–561).
Wang, J., & Elia, N. (2012). Distributed least square with intermittent communications. In *American Control Conference* (pp. 6479–6484).
Wang, J., & Elia, N. (2014). Solving systems of linear equations by distributed convex optimization in the presence of stochastic uncertainty. *IFAC Proceedings Volumes*, *47*(3), 1210–1215.

Yang Liu received the B.Eng. degree from the Department of Microelectronics, Tsinghua University, Beijing, China, in 2015, and the Ph.D. degree in electrical engineering from the Research School of Engineering, College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia, in 2019. He is currently a Senior Researcher with the Tencent Cloud Product Department, Tencent, Shenzhen, China. His research interests include distributed computation and optimization and privacy-aware machine learning.

Youcheng Lou received the B.Sc. degree in mathematics and applied mathematics from Zhengzhou University, Zhengzhou, China, in 2008 and the Ph.D. degree in complex systems and control from the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS), Beijing, China, in 2013. From 2013 to 2018, he was a Postdoctoral Researcher with the National Center for Mathematics and Interdisciplinary Sciences, AMSS, CAS, where from 2016 to 2018, he was also a Postdoctoral Fellow with the Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, funded by Hong Kong Scholars Program. From 2018 to 2020, he was an assistant professor with the AMSS, CAS. He is currently an associate professor with the AMSS, CAS. His current research areas include microeconomics, behavioral finance and distributed optimization.

Brian D.O. Anderson was born in Sydney, Australia, and educated at Sydney University in mathematics and electrical engineering, with Ph.D. in electrical engineering from Stanford University in 1966. Following graduation, he joined the faculty at Stanford University and worked in Vidar Corporation of Mountain View, California, as a staff consultant. He then returned to Australia to become a department chair in electrical engineering at the University of Newcastle. From there, he moved to the Australian National University in 1982, as the first engineering professor at that university. He is now Emeritus Professor at the Australian National University, Distinguished Professor in Hangzhou Dianzi University and Distinguished Researcher in Data61-CSIRO, formerly NICTA. During his period in academia, he spent significant time working for the Australian Government, with this service including membership of the Prime Minister's Science Council under the chairmanship of three prime ministers. He also served on advisory boards or boards of various companies, including the board of the world's major supplier of cochlear implants, Cochlear Corporation, where he was a director for ten years. His awards include the IFAC Quazza Medal in 1999, IEEE Control Systems Award of 1997, the 2001 IEEE James H. Mulligan, Jr. Education Medal, and the Bode Prize of the IEEE Control System Society in 1992, as well as IEEE and other best paper prizes, including one from Automatica. He is a Fellow of the Australian Academy of Science, the Australian Academy of Technological Sciences and Engineering, the Royal Society (London), and a foreign member of the US National Academy of Engineering. He holds honorary doctorates from a number of universities, including Université Catholique de Louvain, Belgium, and ETH, Zürich. He served as IFAC President from 1990 to 1993, having had earlier periods in various IFAC roles, including editor of Automatica. He was also President of the Australian Academy of Science from 1998 to 2002. His current research interests are in distributed control, social networks and econometric modeling.



Guodong Shi received the B.Sc. degree in mathematics and applied mathematics from the School of Mathematics, Shandong University, Jinan, China in 2005, and the Ph.D. degree in systems theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China in 2010. From 2010 to 2014, he was a Postdoctoral Researcher at the ACCESS Linnaeus Centre, KTH Royal Institute of Technology, Stockholm, Sweden. From 2014 to 2018, he was with the Research School of Engineering, The Australian National University, Canberra, Australia as a Lecturer and then Senior Lecturer, and a Future Engineering Research Leadership Fellow. Since 2019 he has been with the Australian Center for Field Robotics, The University of Sydney, NSW, Australia as a Senior Lecturer. His research interests include distributed control systems, quantum networking and decisions, and social opinion dynamics.