FULL LENGTH PAPER

Series A

Distributed stochastic gradient tracking methods

Shi Pu[1,2] · **Angelia Nedić**[3]

Abstract

In this paper, we study the problem of distributed multi-agent optimization over a network, where each agent possesses a local cost function that is smooth and strongly convex. The global objective is to find a common solution that minimizes the average of all cost functions. Assuming agents only have access to unbiased estimates of the gradients of their local cost functions, we consider a distributed stochastic gradient tracking method (DSGT) and a gossip-like stochastic gradient tracking method (GSGT). We show that, in expectation, the iterates generated by each agent are attracted to a neighborhood of the optimal solution, where they accumulate exponentially fast (under a constant stepsize choice). Under DSGT, the limiting (expected) error bounds on the distance of the iterates from the optimal solution decrease with the network size n, which is a comparable performance to a centralized stochastic gradient algorithm. Moreover, we show that when the network is well-connected, GSGT incurs lower communication cost than DSGT while maintaining a similar computational cost. Numerical example further demonstrates the effectiveness of the proposed methods.

Keywords Distributed optimization · Stochastic optimization · Convex programming · Communication networks

Mathematics Subject Classification 90C15 · 90C25 · 68Q25

Parts of the results appeared in the 57th IEEE Conference on Decision and Control [43].

✉ Shi Pu
  pushi@cuhk.edu.cn

Extended author information available on the last page of the article

1 Introduction

Consider a set of agents $\mathcal{N} = \{1, 2, \ldots, n\}$ connected over a network. Each agent has a local smooth and strongly convex cost function $f_i : \mathbb{R}^p \to \mathbb{R}$. The global objective is to locate $x \in \mathbb{R}^p$ that minimizes the average of all cost functions:

$$\min_{x \in \mathbb{R}^p} f(x) \left(= \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right). \tag{1}$$

Scenarios in which problem (1) is considered include distributed machine learning [12,15,35], multi-agent target seeking [8,42], and wireless networks [2,11,30], among many others.

To solve problem (1), we assume each agent i queries a stochastic oracle ($\mathcal{SO}$) to obtain noisy gradient samples of the form $g_i(x, \xi_i)$ that satisfies the following condition:

Assumption 1 For all $i \in \mathcal{N}$ and all $x \in \mathbb{R}^p$, each random vector $\xi_i \in \mathbb{R}^m$ is independent, and

$$\begin{aligned}
&\mathbb{E}_{\xi_i}[g_i(x, \xi_i) \mid x] = \nabla f_i(x), \\
&\mathbb{E}_{\xi_i}[\|g_i(x, \xi_i) - \nabla f_i(x)\|^2 \mid x] \leq \sigma^2 \quad \text{for some } \sigma > 0.
\end{aligned} \tag{2}$$

The above assumption of stochastic gradients holds true for many on-line distributed learning problems, where $f_i(x) = \mathbb{E}_{\xi_i}[F_i(x, \xi_i)]$ denotes the expected loss function agent i wishes to minimize, while independent samples ξ_i are gathered continuously over time. For another example, in simulation-based optimization, the gradient estimation often incurs noise that can be due to various sources, such as modeling and discretization errors, incomplete convergence, and finite sample size for Monte-Carlo methods [22].

Distributed algorithms dealing with problem (1) have been studied extensively in the literature [13,19,20,28,34,36,37,45,46,52,56]. Recently, there has been considerable interest in distributed implementation of stochastic gradient algorithms [3,5–7,9,10,14,18,24,26,32,40,41,48,51,54,55]. The literature has shown that distributed methods may compete with, or even outperform, their centralized counterparts under certain conditions [9,10,26,40,41]. For instance, in our recent work [40], we proposed a flocking-based approach for distributed stochastic optimization which beats a centralized gradient method in real-time assuming that all f_i are identical. However, to the best of our knowledge, there is no distributed stochastic gradient method addressing problem (1) that shows comparable performance with a centralized approach for optimizing the sum of smooth and strongly convex objective functions under Assumption 1 only. In particular, under constant stepsize policies none of the existing algorithms achieve an error bound that is decreasing in the network size n.

A distributed gradient tracking method was proposed in [13,34,46], where the agent-based auxiliary variables y_i were introduced to track the average gradients of f_i assuming accurate gradient information is available. It was shown that the method, with constant stepsize, generates iterates that converge linearly to the optimal solution.

Inspired by the approach, in this paper we consider a distributed stochastic gradient tracking method (DSGT). By comparison, in our proposed algorithm y_i are tracking the stochastic gradient averages of f_i. We are able to show that the iterates generated by each agent reach, in expectation, a neighborhood of the optimal point exponentially fast under a constant stepsize. Interestingly, with a sufficiently small stepsize, the limiting error bounds on the distance between the agent iterates and the optimal solution decrease in the network size, which is comparable to the performance of a centralized stochastic gradient algorithm.

Gossip-based communication protocols are popular choices for distributed computation due to their low communication costs [4,25,29,31]. In the second part of this paper, we consider a gossip-like stochastic gradient tracking algorithm (GSGT) where at each iteration, an agent wakes up uniformly randomly and communicates with one of her neighbors or updates by herself. Similar to DSGT, the method produces iterates that converge to a neighborhood of the optimal point exponentially fast under a sufficiently small constant stepsize. When the network of agents is well-connected (e.g., complete network, almost all regular graphs), GSGT is shown to employ a lower communication burden and similar computational cost when compared to DSGT.

1.1 Related work

We now briefly review the literature on (distributed) stochastic optimization. First of all, our work is related to the extensive literature in stochastic approximation (SA) methods dating back to the seminal works [21,49]. These works include the analysis of convergence (conditions for convergence, rates of convergence, suitable choice of stepsize) in the context of diverse noise models [23]. Assuming the objective function f is strongly convex with Lipschitz continuous gradients, the optimal rate of convergence for solving problem (1) has been shown to be $\mathcal{O}(1/k)$ under a diminishing SA stepsize where k denotes the iteration number [38]. With a constant stepsize $\alpha > 0$ that is sufficiently small, the iterates generated by a stochastic gradient method is attracted to an $\mathcal{O}(\alpha)$-neighborhood of the optimal solution exponentially fast (in expectation).

Distributed implementations of stochastic gradient methods have become increasingly popular in recent years. In [48], the authors considered minimizing a sum of (possibly nonsmooth) convex objective functions subject to a common convex constraint set. It was shown that when the means of the stochastic subgradient errors diminish, there is mean consensus among the agents and mean convergence to the optimum function value under SA stepsizes. The work [54] used two diminishing stepsizes to deal with communication noises and subgradient errors, respectively. Asymptotic convergence to the optimal set was established; for constant stepsizes asymptotic error bounds were derived. In [14], a distributed dual averaging method was proposed for minimizing (possibly nonsmooth) convex functions. Under a carefully chosen SA stepsize sequence, the method exhibits the convergence rate $\mathcal{O}(\frac{n \log(k)}{(1-\lambda_2(\mathbf{W}))\sqrt{k}})$, in which $\lambda_2(\mathbf{W})$ denotes the second largest singular value of the doubly stochastic mixing matrix $\mathbf{W}$. Paper [3] considered a projected stochastic gradient algorithm for solving non-convex optimization problems by combining a local stochastic gradient update and a gossip step. It was proved that consensus is asymptotically achieved in the

network and the solutions converge to the set of KKT points with SA stepsizes. A distributed online algorithm was devised and analyzed in [5] for solving dynamic optimization problems in noisy communication environments. Sufficient conditions were provided for almost sure convergence of the algorithm. In [55], the authors proposed an adaptive diffusion algorithm based on penalty methods. Under a constant stepsize α, it was shown that the expected distance between the optimal solution and that obtained at each node is bounded by $\mathcal{O}(\alpha)$. The work [9,10] further showed that under a sufficiently small stepsize and certain conditions on the stochastic gradients, distributed methods are able to achieve the same performance level as that of a centralized approach. Paper [7] considered the problem of distributed constrained convex optimization subject to multiple noise terms in both computation and communication stages. The authors utilized an augmented Lagrangian framework and established the almost sure convergence of the algorithm under a diminishing stepsize policy. In [32], a subgradient-push method was investigated for distributed optimization over time-varying directed graphs. When the objective function is strongly convex, the scheme exhibits the $\mathcal{O}(\frac{\ln k}{k})$ rate of convergence.

In a recent work [24], a class of decentralized first-order methods for nonsmooth and stochastic optimization was presented. The class was shown to exhibit the $\mathcal{O}(\frac{1}{k})$ (respectively, $\mathcal{O}(\frac{1}{\sqrt{k}})$) rate of convergence for minimizing the sum of strongly convex functions (respectively, general convex functions). The work in [1] considered a composite convex optimization problem with noisy gradient information and showed the $O(\frac{1}{\sqrt{k}})$ convergence rate using an ADMM-based approach. Paper [26] considered a decentralized stochastic gradient algorithm that achieves the $\mathcal{O}(\frac{1}{k} + \frac{1}{\sqrt{nk}})$ rate of convergence for minimizing the sum of non-convex functions. The rate is comparable to that of a centralized algorithm when k is large enough. At the same time, the communication cost for the decentralized approach is lower. Papers [40,41] also demonstrates the advantage of distributively implementing a stochastic gradient method assuming that all f_i are identical and sampling times are random and non-negligible. The work [51] utilized a time-dependent weighted mixing of stochastic subgradient updates to achieve the convergence rate of $\mathcal{O}(\frac{n\sqrt{n}}{(1-\lambda_2(\mathbf{W}))k})$ for minimizing the sum of (possibly nonsmooth) strongly convex functions. In [53], the authors considered a decentralized consensus-based algorithm with delayed gradient information. The method was shown to achieve the optimal $\mathcal{O}(\frac{1}{\sqrt{k}})$ rate of convergence for general convex functions. In [18], the $\mathcal{O}(\frac{1}{k})$ convergence rate was established for strongly convex costs and random networks.

1.2 Main contribution

Our main contribution is summarized as follows. Firstly, we propose a novel distributed stochastic gradient tracking method (DSGT) for optimizing the sum of smooth and strongly convex objective functions. We employ an auxiliary variable y_i for each agent that tracks the average stochastic gradients of the cost functions. We show that, under a constant stepsize choice, the algorithm is comparable to a centralized stochastic gradient scheme in terms of their convergence speeds and the ultimate

error bounds. In particular, the obtained error bound under DSGT decreases with the network size n, which has not been shown in the literature to the best of our knowledge. Moreover, assuming the gradient estimates are accurate, DSGT recovers the linear rate of convergence to the optimal solution [34,46], which is also a unique feature among other distributed stochastic gradient algorithms.

Secondly, with an SA stepsize $\alpha_k \to 0$, DSGT enjoys the optimal $\mathcal{O}(\frac{1}{k})$ rate of convergence to the optimal point. In addition, we characterize the dependency of the constant factors in the stepsize and the convergence rate on the properties of the mixing matrix as well as the characteristics of the objective functions, such as the strong convexity factor and the Lipschitz constant.

Thirdly, we introduce a gossip-like stochastic gradient tracking method that is efficient in communication. We show that, under a sufficiently small constant stepsize, GSGT also produces iterates that converge to a neighborhood of the optimal point exponentially fast. Again, when the gradient estimates are accurate, GSGT recovers the linear rate of convergence to the optimal solution. Compared to DSGT, we show that when the network is well-connected (e.g., complete network, almost all regular graphs), GSGT incurs lower communication cost than DSGT by a factor of $\mathcal{O}(\frac{|\mathcal{E}|}{n})$ ($|\mathcal{E}|$ denoting the number of edges in the network) while maintaining a similar computational cost.

Finally, we provide a numerical example that demonstrates the effectiveness of the proposed methods when contrasted with the centralized stochastic gradient algorithm and some existing variants of distributed stochastic gradient methods.

1.3 Notation and assumptions

Throughout the paper, vectors default to columns if not otherwise specified. Let each agent i hold a local copy $x_i \in \mathbb{R}^p$ of the decision variable and an auxiliary variable $y_i \in \mathbb{R}^p$. Their values at iteration/time k are denoted by $x_{i,k}$ and $y_{i,k}$, respectively. We let

$$\mathbf{x} := [x_1, x_2, \ldots, x_n]^\mathsf{T} \in \mathbb{R}^{n \times p}, \quad \mathbf{y} := [y_1, y_2, \ldots, y_n]^\mathsf{T} \in \mathbb{R}^{n \times p},$$

and

$$\overline{x} := \frac{1}{n}\mathbf{1}^\mathsf{T}\mathbf{x} \in \mathbb{R}^{1 \times p}, \quad \overline{y} := \frac{1}{n}\mathbf{1}^\mathsf{T}\mathbf{y} \in \mathbb{R}^{1 \times p},$$

where $\mathbf{1}$ denotes the vector with all entries equal to 1. We define an aggregate objective function of the local variables:

$$F(\mathbf{x}) := \sum_{i=1}^{n} f_i(x_i), \tag{3}$$

and let

$$\nabla F(\mathbf{x}) := [\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n)]^\mathsf{T} \in \mathbb{R}^{n \times p}.$$

In addition, denote

$$h(\mathbf{x}) := \frac{1}{n} \mathbf{1}^{\mathsf{T}} \nabla F(\mathbf{x}) \in \mathbb{R}^{1 \times p},$$

$$\boldsymbol{\xi} := [\xi_1, \xi_2, \ldots, \xi_n]^{\mathsf{T}} \in \mathbb{R}^{n \times m},$$

$$G(\mathbf{x}, \boldsymbol{\xi}) := [g_1(x_1, \xi_1), g_2(x_2, \xi_2), \ldots, g_n(x_n, \xi_n)]^{\mathsf{T}} \in \mathbb{R}^{n \times p}.$$

The inner product of two vectors a, b of the same dimension is denoted by $\langle a, b \rangle$. For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times p}$, we let $\langle \mathbf{A}, \mathbf{B} \rangle$ be the Frobenius inner product. We use $\| \cdot \|$ to denote the 2-norm of vectors; for matrices, $\| \cdot \|$ represents the Frobenius norm. The spectral radius of a square matrix $\mathbf{M}$ is denoted by $\rho(\mathbf{M})$.

We make the following standing assumption on the individual objective functions f_i.

Assumption 2 Each $f_i : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex with L-Lipschitz continuous gradients, i.e., for any $x, x' \in \mathbb{R}^p$,

$$\langle \nabla f_i(x) - \nabla f_i(x'), x - x' \rangle \geq \mu \|x - x'\|^2,$$

$$\|\nabla f_i(x) - \nabla f_i(x')\| \leq L \|x - x'\|.$$

We note that, under Assumption 2, problem (1) has a unique solution denoted by $x^* \in \mathbb{R}^{1 \times p}$.

A graph is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2, \ldots, n\}$ is the set of vertices (nodes) and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the set of edges connecting vertices. We assume agents communicate in an undirected graph, i.e., $(i, j) \in \mathcal{E}$ iff (if and only if) $(j, i) \in \mathcal{E}$. For each agent i, let $\mathcal{N}_i = \{j \mid j \neq i, (i, j) \in \mathcal{E}\}$ be its set of neighbors. The cardinality of $\mathcal{N}_i$, denoted by $\deg(i)$, is referred to as agent i's degree. We consider the following condition regarding the interaction graph of agents.

Assumption 3 The graph $\mathcal{G}$ corresponding to the network of agents is undirected and connected, i.e., there exists a path between any two agents.

1.4 Organization of the paper

The paper is organized as follows. In Sect. 2, we introduce the distributed stochastic gradient tracking method and present its main convergence results. We perform analysis in Sect. 3. In Sect. 4 we propose the gossip-like stochastic gradient tracking method. A numerical example is provided in Sect. 5 to illustrate our theoretical findings. Section 6 concludes the paper.

2 A distributed stochastic gradient tracking method (DSGT)

We consider the following distributed stochastic gradient tracking method: At each step $k \in \mathbb{N}$, every agent i independently implements the following two steps:

$$x_{i,k+1} = \sum_{j=1}^{n} w_{ij}(x_{j,k} - \alpha y_{j,k}),$$

$$y_{i,k+1} = \sum_{j=1}^{n} w_{ij} y_{j,k} + g_i(x_{i,k+1}, \xi_{i,k+1}) - g_i(x_{i,k}, \xi_{i,k}), \qquad (4)$$

where w_{ij} are nonnegative weights and $\alpha > 0$ is a constant stepsize. Agent i and j are connected iff $w_{ij}, w_{ji} > 0$. The iterates are initiated with an arbitrary $x_{i,0}$ and $y_{i,0} = g_i(x_{i,0}, \xi_{i,0})$ for all $i \in \mathcal{N}$. We can also write (4) in the following compact form:

$$\mathbf{x}_{k+1} = \mathbf{W}(\mathbf{x}_k - \alpha \mathbf{y}_k),$$
$$\mathbf{y}_{k+1} = \mathbf{W}\mathbf{y}_k + G(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - G(\mathbf{x}_k, \boldsymbol{\xi}_k), \qquad (5)$$

where $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n \times n}$ denotes the coupling matrix of agents. We assume that $\mathbf{W}$ satisfies the following condition.

Assumption 4 Nonnegative coupling matrix $\mathbf{W}$ is doubly stochastic, i.e., $\mathbf{W}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\mathsf{T}\mathbf{W} = \mathbf{1}^\mathsf{T}$. In addition, $w_{ii} > 0$ for some $i \in \mathcal{N}$.

In the subsequent analysis, we will frequently use the following result, which is a direct implication of Assumptions 3 and 4 (see [46] Section II-B).

Lemma 1 *Let Assumptions 3 and 4 hold, and let ρ_w denote the spectral norm of the matrix $\mathbf{W} - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$. Then, $\rho_w < 1$ and*

$$\|\mathbf{W}\omega - \mathbf{1}\overline{\omega}\| \le \rho_w \|\omega - \mathbf{1}\overline{\omega}\|$$

for all $\omega \in \mathbb{R}^{n \times p}$, where $\overline{\omega} = \frac{1}{n}\mathbf{1}^\mathsf{T}\omega$.

Algorithm (4) is closely related to the schemes considered in [13,34,46], in which auxiliary variables $y_{i,k}$ were introduced to track the average $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_{i,k})$. This design ensures that the algorithms achieve linear convergence under a constant stepsize choice. Correspondingly, under our approach $y_{i,k}$ are (approximately) tracking $\frac{1}{n}\sum_{i=1}^{n} g_i(x_{i,k}, \xi_{i,k})$. To see why this is the case, note that $\overline{y}_k = \frac{1}{n}\mathbf{1}^\mathsf{T}\mathbf{y}_k$. Since $y_{i,0} = g(x_{i,0}, \xi_{i,0}), \forall i$, by induction we have

$$\overline{y}_k = \frac{1}{n}\mathbf{1}^\mathsf{T} G(\mathbf{x}_k, \boldsymbol{\xi}_k) = \frac{1}{n}\sum_{i=1}^{n} g_i(x_{i,k}, \xi_{i,k}), \forall k. \qquad (6)$$

It will be shown that $\mathbf{y}_k$ is close to $\mathbf{1}\overline{y}_k$ in expectation when k is sufficiently large. As a result, $y_{i,k}$ are (approximately) tracking $\frac{1}{n}\sum_{i=1}^{n} g_i(x_{i,k}, \xi_{i,k})$.

It is worth noting that compared to the standard distributed subgradient methods [36], DSGT incurs two times the communication and storage costs per iteration.

2.1 Main results

Main convergence properties of DSGT are covered in the following theorem.

Theorem 1 *Let $\Gamma > 1$ be arbitrarily chosen. Suppose Assumptions 1 and 2 hold and the stepsize α satisfies*

$$\alpha \leq \min\left\{ \frac{(1-\rho_w^2)}{12\rho_w L}, \frac{(1-\rho_w^2)^2}{2\sqrt{\Gamma}L \max\{6\rho_w \|\mathbf{W}-\mathbf{I}\|, 1-\rho_w^2\}}, \frac{(1-\rho_w^2)}{3\rho_w^{2/3}L}\left[\frac{\mu^2}{L^2}\frac{(\Gamma-1)}{\Gamma(\Gamma+1)}\right]^{1/3} \right\}. \tag{7}$$

Then $\sup_{l\geq k} \mathbb{E}[\|\bar{x}_l - x^*\|^2]$ and $\sup_{l\geq k} \mathbb{E}[\|\mathbf{x}_l - \mathbf{1}\bar{x}_l\|^2]$, respectively, converge to $\limsup_{k\to\infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2]$ and $\limsup_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]$ at the linear rate $\mathcal{O}(\rho(\mathbf{A})^k)$, where $\rho(\mathbf{A}) < 1$ is the spectral radius of the matrix $\mathbf{A}$ given by

$$\mathbf{A} = \begin{bmatrix} 1-\alpha\mu & \frac{\alpha L^2}{\mu n}(1+\alpha\mu) & 0 \\ 0 & \frac{1}{2}(1+\rho_w^2) & \alpha^2 \frac{(1+\rho_w^2)\rho_w^2}{(1-\rho_w^2)} \\ 2\alpha n L^3 & \left(\frac{1}{\beta}+2\right)\|\mathbf{W}-\mathbf{I}\|^2 L^2 + 3\alpha L^3 & \frac{1}{2}(1+\rho_w^2) \end{bmatrix},$$

where $\beta = \frac{1-\rho_w^2}{2\rho_w^2} - 4\alpha L - 2\alpha^2 L^2$. Furthermore,

$$\limsup_{k\to\infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2] \leq \frac{(\Gamma+1)}{\Gamma}\frac{\alpha\sigma^2}{\mu n} + \left(\frac{\Gamma+1}{\Gamma-1}\right)\frac{4\alpha^2 L^2(1+\alpha\mu)(1+\rho_w^2)\rho_w^2}{\mu^2 n(1-\rho_w^2)^3}M_\sigma, \tag{8}$$

and

$$\limsup_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] \leq \left(\frac{\Gamma+1}{\Gamma-1}\right)\frac{4\alpha^2(1+\rho_w^2)\rho_w^2(2\alpha^2 L^3\sigma^2 + \mu M_\sigma)}{\mu(1-\rho_w^2)^3}, \tag{9}$$

where

$$M_\sigma := \left[3\alpha^2 L^2 + 2(\alpha L + 1)(n+1)\right]\sigma^2. \tag{10}$$

Remark 1 The first term on the right-hand side of (8) can be interpreted as the error caused by stochastic gradients only, since it does not depend on the network topology. The second term on the right hand side of (8) and the bound in (9) are network dependent, and they increase with ρ_w (larger ρ_w indicates worse network connectivity).

In light of (8) and (9), we have

$$\limsup_{k\to\infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2] = \alpha\mathcal{O}\left(\frac{\sigma^2}{\mu n}\right) + \frac{\alpha^2}{(1-\rho_w)^3}\mathcal{O}\left(\frac{L^2\sigma^2}{\mu^2}\right),$$

and

$$\limsup_{k\to\infty} \frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] = \frac{\alpha^2}{(1-\rho_w)^3}\mathcal{O}\left(\sigma^2\right) + \frac{\alpha^4}{(1-\rho_w)^3}\mathcal{O}\left(\frac{L^3\sigma^2}{\mu n}\right).$$

Let $\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2]$ measure the average quality of solutions obtained by all the agents. We have

$$\limsup_{k\to\infty} \frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2] = \limsup_{k\to\infty}\mathbb{E}[\|\overline{x}_k - x^*\|^2] + \limsup_{k\to\infty}\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2]$$

$$= \alpha\mathcal{O}\left(\frac{\sigma^2}{\mu n}\right) + \frac{\alpha^2}{(1-\rho_w)^3}\mathcal{O}\left(\frac{L^2\sigma^2}{\mu^2}\right), \tag{11}$$

which is decreasing in the network size n when α is sufficiently small, i.e., when

$$\alpha\frac{\sigma^2}{\mu n} \sim \frac{\alpha^2}{(1-\rho_w)^3}\frac{L^2\sigma^2}{\mu^2}, \quad \text{or} \quad \alpha \sim \frac{\mu}{L^2}\frac{(1-\rho_w)^3}{n}.$$

The spectral gap $1 - \rho_w$ depends on the graph topology as well as the choices of weights $[w_{ij}]$. For example, suppose the Lazy Metropolis rule is adopted (see [33]). In this case, if the communication graph is a 1) path or ring graph, then $1-\rho_w = \mathcal{O}(1/n^2)$; 2) lattice, then $1 - \rho_w = \mathcal{O}(1/n)$; 3) geometric random graph, then $1 - \rho_w = \mathcal{O}(1/n\log n)$; 4) Erdős-Rényi random graph, then $1 - \rho_w = \mathcal{O}(1)$; 5) complete graph, then $1 - \rho_w = 1$; 6) any connected undirected graph, then $1 - \rho_w = \mathcal{O}(1/n^2)$.

From the above argument, the condition $\alpha \sim \frac{\mu}{L^2}\frac{(1-\rho_w)^3}{n}$ is in general more strict than (7) in Theorem 1, which requires $\alpha \sim \frac{(1-\rho_w)^2}{L}$ (when $1 - \rho_w \leq (\frac{\mu}{L})^{2/3}$). Such a difference suggests that when implementing DSGT in practice, it can be advantageous to use a larger stepsize in the early stage to achieve a faster convergence speed (see Corollary 1 below) and then switch to smaller stepsizes for more accuracy on the final solutions. □

Remark 2 Under a centralized stochastic gradient (CSG) algorithm in the form of

$$x_{k+1} = x_k - \alpha\frac{1}{n}\sum_{i=1}^{n} g_i(x_k, \xi_{i,k}), \quad k \in \mathbb{N}, \tag{12}$$

we would obtain

$$\limsup_{k\to\infty}\mathbb{E}[\|x_k - x^*\|^2] = \alpha\mathcal{O}\left(\frac{\sigma^2}{\mu n}\right).$$

It can be observed that DSGT is comparable with CSG in their ultimate error bounds (up to constant factors) with sufficiently small stepsizes. □

As shown in Theorem 1, the convergence rate of DSGT is determined by the spectral radius $\rho(\mathbf{A}) < 1$. In the corollary below we provide an upper bound of $\rho(\mathbf{A})$.

Corollary 1 *Under the conditions in Theorem* 1, *assuming in addition that the stepsize* α *also satisfies*[1]

$$\alpha \leq \frac{(\Gamma + 1)}{\Gamma} \frac{(1 - \rho_w^2)}{8\mu}, \tag{13}$$

we have

$$\rho(\mathbf{A}) \leq 1 - \left(\frac{\Gamma - 1}{\Gamma + 1}\right) \alpha\mu.$$

Corollary 1 implies that, for sufficiently small stepsizes, the distributed gradient tracking method has a comparable convergence speed to that of a centralized scheme (in which case the linear rate is $\mathcal{O}((1 - 2\alpha\mu)^k)$).

In the next theorem, we show DSGT achieves the $\mathcal{O}(\frac{1}{k})$ rate of convergence under a diminishing stepsize policy.

Theorem 2 *Let Assumptions* 1 *and* 2 *hold. Consider the method in* (4) *where* α *is replaced with the time-varying stepsize* α_k *given by* $\alpha_k := \theta/(m + k)$, *where* $\theta > 1/\mu$ *and* m *satisfies*

$$\begin{cases} m > \max\left\{\frac{\theta}{2}(\mu + L), \frac{4\theta L \rho_w^2 + 2\theta L \rho_w \sqrt{1 + 3\rho_w^2}}{1 - \rho_w^2}\right\}, \\ \frac{(1 - \rho_w^2)^2}{\theta^2(1 + \rho_w^2)\rho_w^2}\left[\frac{(1 - \rho_w^2)}{2} - \frac{2m+1}{(m+1)^2}\right] > \frac{1}{(\theta\mu - 1)}\left(\frac{1}{\mu} + \frac{\theta}{m}\right)\frac{4\theta^2 L^5}{m^3} + \frac{2C}{m^2}, \end{cases} \tag{14}$$

with $C = \left[\left(\frac{1 - \rho_w^2}{2\rho_w^2} - \frac{4\theta L}{m} - \frac{2\theta^2 L^2}{m^2}\right)^{-1} + 2\right] \|\mathbf{W} - \mathbf{I}\|^2 L^2 + \frac{3\theta L^3}{m}$. *Then for all* $k \geq 0$, *we have*

$$\mathbb{E}[\|\bar{x}_k - x^*\|^2] \leq \frac{2\theta^2\sigma^2}{n(\theta\mu - 1)(m + k)} + \frac{\mathcal{O}_k(1)}{(m + k)^{\theta\mu}} + \frac{\mathcal{O}_k(1)}{(m + k)^2}, \tag{15a}$$

$$\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] \leq \frac{\mathcal{O}_k(1)}{(m + k)^2}, \tag{15b}$$

where $\mathcal{O}_k(1)$ *denotes some constant that does not depend on* k.

Remark 3 From (15), noting that $\theta\mu > 1$, we have

$$\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2] = \mathbb{E}[\|\bar{x}_k - x^*\|^2] + \frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] = \mathcal{O}_k\left(\frac{1}{k}\right).$$

In particular, $\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2]$ asymptotically converges to 0 at the rate $\frac{2\theta^2\sigma^2}{n(\theta\mu - 1)k}$, which does not depend on the spectral norm ρ_w and matches the convergence rate (up to constant factors) for centralized stochastic gradient methods (see [38,47]).[2] □

---

[1] This condition is weaker than (7) when $\rho_w^2 \geq \frac{\Gamma}{\Gamma + 1}\frac{2\mu}{3L}$.

[2] It is worth noting that the transient time for $\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2]$ to approach the asymptotic convergence rate depends on the network topology, and its dependence on n can be significant.

Remark 4 Under a well connected network ($\rho_w \simeq 0$), $m > \frac{\theta}{2}(\mu + L)$ suffices. When $\rho_w \simeq 1$, the lower bound of m is in the order of $\mathcal{O}((1 - \rho_w)^{-2})$. □

Theorem 2 implies the following result if we choose $\alpha_k = \theta/(k + 1)$.

Corollary 2 *Under Assumptions 1 and 2 and stepsize policy* $\alpha_k := \theta/(k + 1)$ *for some* $\theta > 1/\mu$, *we have for all* $k \geq m$ *where* m *satisfies condition* (14),

$$\mathbb{E}[\|\overline{x}_k - x^*\|^2] \leq \frac{\tilde{U}}{k}, \quad \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] \leq \frac{\tilde{X}}{k^2}, \quad \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] \leq \tilde{Y},$$

where $\tilde{U}$, $\tilde{X}$, *and* $\tilde{Y}$ *are some positive constants.*

3 Analysis

In this section, we prove Theorem 1 by studying the evolution of $\mathbb{E}[\|\overline{x}_k - x^*\|^2]$, $\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2]$ and $\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2]$. Our strategy is to bound the three expressions in terms of linear combinations of their past values, in which way we establish a linear system of inequalities. This approach is different from those employed in [34,46], where the analyses pertain to the examination of $\|\overline{x}_k - x^*\|$, $\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|$ and $\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|$. Such distinction is due to the stochastic gradients $g_i(x_{i,k}, \xi_{i,k})$ whose variances play a crucial role in deriving the main inequalities.

We first introduce some lemmas that will be used later in the analysis. Denote by $\mathcal{F}_k$ the σ-algebra generated by $\{\boldsymbol{\xi}_0, \ldots, \boldsymbol{\xi}_{k-1}\}$, and define $\mathbb{E}[\cdot \mid \mathcal{F}_k]$ as the conditional expectation given $\mathcal{F}_k$.

Lemma 2 *Under Assumption 1, recalling that* $h(\mathbf{x}) = \frac{1}{n}\mathbf{1}^{\mathsf{T}}\nabla F(\mathbf{x})$, *we have for all* $k \geq 0$,

$$\mathbb{E}\left[\|\overline{y}_k - h(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k\right] \leq \frac{\sigma^2}{n}. \tag{16}$$

□

Proof By the definitions of $\overline{y}_k$ and $h(\mathbf{x}_k)$,

$$\mathbb{E}\left[\|\overline{y}_k - h(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\|g_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k})\|^2 \mid \mathcal{F}_k\right] \leq \frac{\sigma^2}{n}.$$

□

Lemma 3 *Under Assumption 2, we have for all* $k \geq 0$,

$$\|\nabla f(\overline{x}_k) - h(\mathbf{x}_k)\| \leq \frac{L}{\sqrt{n}}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|. \tag{17}$$

If in addition $\alpha < 2/(\mu + L)$, then

$$\|x - \alpha \nabla f(x) - x^*\| \leq (1 - \alpha\mu)\|x - x^*\|, \ \forall x \in \mathbb{R}^p.$$

Proof See [46] Lemma 10 for reference. □

In the following lemma, we establish bounds on $\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2$ and on the conditional expectations of $\|\overline{x}_{k+1} - x^*\|^2$ and $\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2$, respectively.

Lemma 4 *Suppose Assumptions* 1 *and* 2 *hold and* $\alpha < 2/(\mu + L)$. *We have the following inequalities:*

$$\mathbb{E}[\|\overline{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \leq (1 - \alpha\mu)\|\overline{x}_k - x^*\|^2$$
$$+ \frac{\alpha L^2}{\mu n}(1 + \alpha\mu)\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{\alpha^2\sigma^2}{n}, \tag{18}$$

$$\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2 \leq \frac{(1 + \rho_w^2)}{2}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2$$
$$+ \alpha^2 \frac{(1 + \rho_w^2)\rho_w^2}{(1 - \rho_w^2)}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2, \tag{19}$$

and for any $\beta > 0$,

$$\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2 \mid \mathcal{F}_k] \leq \left(1 + 4\alpha L + 2\alpha^2 L^2 + \beta\right)\rho_w^2 \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k]$$
$$+ \left(\frac{1}{\beta}\|\mathbf{W}-\mathbf{I}\|^2 L^2 + 2\|\mathbf{W}-\mathbf{I}\|^2 L^2 + 3\alpha L^3\right)\|\mathbf{x}_k-\mathbf{1}\overline{x}_k\|^2 + 2\alpha n L^3\|\overline{x}_k - x^*\|^2 + M_\sigma. \tag{20}$$

Proof See Appendix 7.1. □

3.1 Proof of Theorem 1

Taking full expectation on both sides of (18)–(20), we obtain the following linear system of inequalities

$$\begin{bmatrix} \mathbb{E}[\|\overline{x}_{k+1} - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2] \\ \mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2] \end{bmatrix} \leq \mathbf{A} \begin{bmatrix} \mathbb{E}[\|\overline{x}_k - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] \\ \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] \end{bmatrix} + \begin{bmatrix} \frac{\alpha^2\sigma^2}{n} \\ 0 \\ M_\sigma \end{bmatrix}, \tag{21}$$

where the inequality is to be taken component-wise, and the entries of the matrix $\mathbf{A} = [a_{ij}]$ are given by

$$\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = \begin{bmatrix} 1 - \alpha\mu \\ 0 \\ 2\alpha n L^3 \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} = \begin{bmatrix} \frac{\alpha L^2}{\mu n}(1 + \alpha\mu) \\ \frac{1}{2}(1 + \rho_w^2) \\ \left(\frac{1}{\beta} + 2\right)\|\mathbf{W} - \mathbf{I}\|^2 L^2 + 3\alpha L^3 \end{bmatrix},$$

$$\begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix} = \begin{bmatrix} 0 \\ \alpha^2 \frac{(1+\rho_w^2)\rho_w^2}{(1-\rho_w^2)} \\ \left(1 + 4\alpha L + 2\alpha^2 L^2 + \beta\right)\rho_w^2 \end{bmatrix},$$

and M_σ is given in (10). Therefore, by induction we have

$$\begin{bmatrix} \mathbb{E}[\|\overline{x}_k - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] \\ \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] \end{bmatrix} \leq \mathbf{A}^k \begin{bmatrix} \mathbb{E}[\|\overline{x}_0 - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_0 - \mathbf{1}\overline{x}_0\|^2] \\ \mathbb{E}[\|\mathbf{y}_0 - \mathbf{1}\overline{y}_0\|^2] \end{bmatrix} + \sum_{l=0}^{k-1} \mathbf{A}^l \begin{bmatrix} \frac{\alpha^2\sigma^2}{n} \\ 0 \\ M_\sigma \end{bmatrix}. \qquad (22)$$

If the spectral radius of $\mathbf{A}$ satisfies $\rho(\mathbf{A}) < 1$, then $\mathbf{A}^k$ converges to $\mathbf{0}$ at the linear rate $\mathcal{O}(\rho(\mathbf{A})^k)$ (see [17]), in which case $\sup_{l \geq k} \mathbb{E}[\|\overline{x}_l - x^*\|^2]$, $\sup_{l \geq k} \mathbb{E}[\|\mathbf{x}_l - \mathbf{1}\overline{x}_l\|^2]$ and $\sup_{l \geq k} \mathbb{E}[\|\mathbf{y}_l - \mathbf{1}\overline{y}_l\|^2]$ all converge to a neighborhood of 0 at the linear rate $\mathcal{O}(\rho(\mathbf{A})^k)$. The next lemma provides conditions that ensure $\rho(\mathbf{A}) < 1$.

Lemma 5 *Let* $\mathbf{S} = [s_{ij}] \in \mathbb{R}^{3 \times 3}$ *be a nonnegative, irreducible matrix with* $s_{ii} < \lambda^*$ *for some* $\lambda^* > 0$ *for all* $i = 1, 2, 3$. *Then* $\rho(\mathbf{S}) < \lambda^*$ *iff* $\det(\lambda^* \mathbf{I} - \mathbf{S}) > 0$.

Proof See Appendix 7.2. □

We now derive the conditions such that $\rho(\mathbf{A}) < 1$. Suppose α and β meet the following relations:[3]

$$a_{33} = \left(1 + 4\alpha L + 2\alpha^2 L^2 + \beta\right)\rho_w^2 = \frac{1 + \rho_w^2}{2} < 1, \qquad (23)$$

$$a_{23}a_{32} = \alpha^2 \frac{(1+\rho_w^2)\rho_w^2}{(1-\rho_w^2)}\left[\left(\frac{1}{\beta} + 2\right)\|\mathbf{W} - \mathbf{I}\|^2 L^2 + 3\alpha L^3\right] \leq \frac{1}{\Gamma}(1 - a_{22})(1 - a_{33}), \qquad (24)$$

and

$$a_{12}a_{23}a_{31} = \frac{2\alpha^4 L^5(1 + \alpha\mu)}{\mu}\frac{(1 + \rho_w^2)}{(1 - \rho_w^2)}\rho_w^2 \leq \frac{1}{\Gamma + 1}(1 - a_{11})[(1 - a_{22})(1 - a_{33}) - a_{23}a_{32}]. \qquad (25)$$

Then,

---

[3] Matrix $\mathbf{A}$ in Theorem 1 corresponds to such a choice of α and β.

$$\det(\mathbf{I} - \mathbf{A}) = (1 - a_{11})(1 - a_{22})(1 - a_{33}) - (1 - a_{11})a_{23}a_{32} - a_{12}a_{23}a_{31}$$
$$\geq \frac{\Gamma}{(\Gamma + 1)}(1 - a_{11})[(1 - a_{22})(1 - a_{33}) - a_{23}a_{32}]$$
$$\geq \left(\frac{\Gamma - 1}{\Gamma + 1}\right)(1 - a_{11})(1 - a_{22})(1 - a_{33}) > 0.$$

Given that $a_{11}, a_{22}, a_{33} < 1$, in light of Lemma 5, we have $\rho(\mathbf{A}) < 1$. In addition, by denoting $B := [\frac{\alpha^2\sigma^2}{n}, 0, M_\sigma]^\mathsf{T}$ and $[(\mathbf{I} - \mathbf{A})^{-1}B]_j$ the j-th element of the vector $[(\mathbf{I} - \mathbf{A})^{-1}B]$, we obtain from (22) that

$$\limsup_{k\to\infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2] \leq [(\mathbf{I} - \mathbf{A})^{-1}B]_1$$
$$= \frac{1}{\det(\mathbf{I} - \mathbf{A})}\left\{[(1 - a_{22})(1 - a_{33}) - a_{23}a_{32}]\frac{\alpha^2\sigma^2}{n} + a_{12}a_{23}M_\sigma\right\}$$
$$\leq \frac{(\Gamma + 1)}{\Gamma}\frac{\alpha\sigma^2}{\mu n} + \left(\frac{\Gamma + 1}{\Gamma - 1}\right)\frac{a_{12}a_{23}M_\sigma}{(1 - a_{11})(1 - a_{22})(1 - a_{33})}$$
$$= \frac{(\Gamma + 1)}{\Gamma}\frac{\alpha\sigma^2}{\mu n} + \left(\frac{\Gamma + 1}{\Gamma - 1}\right)\frac{\alpha^3 L^2(1 + \alpha\mu)(1 + \rho_w^2)\rho_w^2 M_\sigma}{\mu n(1 - \rho_w^2)(1 - a_{11})(1 - a_{22})(1 - a_{33})}$$
$$= \frac{(\Gamma + 1)}{\Gamma}\frac{\alpha\sigma^2}{\mu n} + \left(\frac{\Gamma + 1}{\Gamma - 1}\right)\frac{4\alpha^2 L^2(1 + \alpha\mu)(1 + \rho_w^2)\rho_w^2}{\mu^2 n(1 - \rho_w^2)^3}M_\sigma, \tag{26}$$

and

$$\limsup_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] \leq [(\mathbf{I} - \mathbf{A})^{-1}B]_2 = \frac{1}{\det(\mathbf{I} - \mathbf{A})}\left[a_{23}a_{31}\frac{\alpha^2\sigma^2}{n} + a_{23}(1 - a_{11})M_\sigma\right]$$
$$\leq \left(\frac{\Gamma + 1}{\Gamma - 1}\right)\frac{a_{23}}{(1 - a_{11})(1 - a_{22})(1 - a_{33})}\left(2\alpha n L^3\frac{\alpha^2\sigma^2}{n} + \alpha\mu M_\sigma\right)$$
$$= \frac{4(\Gamma + 1)\alpha^2(1 + \rho_w^2)\rho_w^2(2\alpha^2 L^3\sigma^2 + \mu M_\sigma)}{(\Gamma - 1)\mu(1 - \rho_w^2)^3}.$$

It remains to show that (23)–(25) are satisfied under condition (7). By (23), we need

$$\beta = \frac{1 - \rho_w^2}{2\rho_w^2} - 4\alpha L - 2\alpha^2 L^2.$$

Since $\alpha \leq \frac{1 - \rho_w^2}{12\rho_w L}$ by (7), we know that

$$\beta \geq \frac{1 - \rho_w^2}{2\rho_w^2} - \frac{1 - \rho_w^2}{3\rho_w} - \frac{(1 - \rho_w^2)^2}{72\rho_w^2} \geq \frac{11(1 - \rho_w^2)}{72\rho_w^2} \geq \frac{1 - \rho_w^2}{8\rho_w^2} > 0. \tag{27}$$

Condition (24) leads to the inequality below:

$$\alpha^2 \frac{(1 + \rho_w^2)\rho_w^2}{(1 - \rho_w^2)} \left[\left(\frac{1}{\beta} + 2 \right) \|\mathbf{W} - \mathbf{I}\|^2 L^2 + 3\alpha L^3 \right] \leq \frac{(1 - \rho_w^2)^2}{4\Gamma}.$$

By (27), we only need

$$\alpha^2 \left[\frac{(2 + 6\rho_w^2)}{(1 - \rho_w^2)} \|\mathbf{W} - \mathbf{I}\|^2 L^2 + \frac{(1 - \rho_w^2)}{4\rho_w^2} L^2 \right] \leq \frac{(1 - \rho_w^2)^3}{4\Gamma(1 + \rho_w^2)\rho_w^2}.$$

The preceding inequality is equivalent to

$$\alpha \leq \frac{(1 - \rho_w^2)^2}{L\sqrt{\Gamma(1 + \rho_w^2)}\sqrt{4\rho_w^2(2 + 6\rho_w^2)\|\mathbf{W} - \mathbf{I}\|^2 + (1 - \rho_w^2)^2}},$$

implying that it is sufficient to have

$$\alpha \leq \frac{(1 - \rho_w^2)^2}{2\sqrt{\Gamma}L \max(6\rho_w\|\mathbf{W} - \mathbf{I}\|, 1 - \rho_w^2)}.$$

To see that relation (25) holds, consider a stronger condition

$$\frac{2\alpha^4 L^5(1 + \alpha\mu)}{\mu} \frac{(1 + \rho_w^2)}{(1 - \rho_w^2)}\rho_w^2 \leq \frac{(\Gamma - 1)}{\Gamma(\Gamma + 1)}(1 - a_{11})(1 - a_{22})(1 - a_{33}),$$

or equivalently,

$$\frac{2\alpha^3 L^5(1 + \alpha\mu)}{\mu^2} \frac{(1 + \rho_w^2)}{(1 - \rho_w^2)}\rho_w^2 \leq \frac{(\Gamma - 1)}{4\Gamma(\Gamma + 1)}(1 - \rho_w^2)^2.$$

It suffices that

$$\alpha \leq \frac{(1 - \rho_w^2)}{3\rho_w^{2/3}L} \left[\frac{\mu^2}{L^2} \frac{(\Gamma - 1)}{\Gamma(\Gamma + 1)} \right]^{1/3}. \tag{28}$$

3.2 Proof of Corollary 1

We derive an upper bound of $\rho(\mathbf{A})$ under conditions (7) and (13). Note that the characteristic function of $\mathbf{A}$ is given by

$$\det(\lambda\mathbf{I} - \mathbf{A}) = (\lambda - a_{11})(\lambda - a_{22})(\lambda - a_{33}) - (\lambda - a_{11})a_{23}a_{32} - a_{12}a_{23}a_{31}.$$

Since $\det(\mathbf{I} - \mathbf{A}) > 0$ and $\det(\max\{a_{11}, a_{22}, a_{33}\}\mathbf{I} - \mathbf{A}) = \det(a_{11}\mathbf{I} - \mathbf{A}) < 0$, we have $\rho(\mathbf{A}) \in (a_{11}, 1)$. By (24) and (25),

$$\det(\lambda\mathbf{I} - \mathbf{A}) \geq (\lambda - a_{11})(\lambda - a_{22})(\lambda - a_{33}) - (\lambda - a_{11})a_{23}a_{32}$$

$$-\frac{1}{\Gamma+1}(1-a_{11})[(1-a_{22})(1-a_{33})-a_{23}a_{32}]$$

$$\geq (\lambda-a_{11})(\lambda-a_{22})(\lambda-a_{33}) - \frac{1}{\Gamma}(\lambda-a_{11})(1-a_{22})(1-a_{33})$$

$$-\frac{(\Gamma-1)}{\Gamma(\Gamma+1)}(1-a_{11})(1-a_{22})(1-a_{33}).$$

Suppose $\lambda = 1 - \epsilon$ for some $\epsilon \in (0, \alpha\mu)$, satisfying

$$\det(\lambda\mathbf{I}-\mathbf{A}) \geq \frac{1}{4}(\alpha\mu-\epsilon)\left(1-\rho_w^2-2\epsilon\right)^2$$

$$-\frac{1}{4\Gamma}(\alpha\mu-\epsilon)(1-\rho_w^2)^2 - \frac{(\Gamma-1)\alpha\mu}{4\Gamma(\Gamma+1)}(1-\rho_w^2)^2 \geq 0,$$

or equivalently,

$$\frac{(\alpha\mu-\epsilon)}{\alpha\mu}\left[\frac{\left(1-\rho_w^2-2\epsilon\right)^2}{(1-\rho_w^2)^2} - \frac{1}{\Gamma}\right] \geq \frac{(\Gamma-1)}{\Gamma(\Gamma+1)}.$$

It suffices that

$$\epsilon \leq \left(\frac{\Gamma-1}{\Gamma+1}\right)\alpha\mu.$$

To see why this is the case, note that $\frac{(\alpha\mu-\epsilon)}{\alpha\mu} \geq \frac{2}{\Gamma+1}$, and under (13),

$$\epsilon \leq \left(\frac{\Gamma-1}{\Gamma+1}\right)\frac{(\Gamma+1)}{\Gamma}\frac{(1-\rho_w^2)}{8\mu}\mu = \frac{(\Gamma-1)(1-\rho_w^2)}{8\Gamma}.$$

As a result,

$$1-\rho_w^2-2\epsilon \geq 1-\rho_w^2 - \frac{(\Gamma-1)(1-\rho_w^2)}{4\Gamma} = \frac{(3\Gamma+1)(1-\rho_w^2)}{4\Gamma}.$$

We then have

$$\frac{(\alpha\mu-\epsilon)}{\alpha\mu}\left[\frac{\left(1-\rho_w^2-2\epsilon\right)^2}{(1-\rho_w^2)^2} - \frac{1}{\Gamma}\right] \geq \frac{2}{(\Gamma+1)}\left[\frac{(3\Gamma+1)^2}{16\Gamma^2} - \frac{1}{\Gamma}\right] = \frac{1}{\Gamma(\Gamma+1)}\left[\frac{(3\Gamma+1)^2}{8\Gamma} - 2\right]$$

$$= \frac{1}{\Gamma(\Gamma+1)}\left(\Gamma-1+\frac{\Gamma}{8}+\frac{1}{8\Gamma}-\frac{1}{4}\right) \geq \frac{(\Gamma-1)}{\Gamma(\Gamma+1)}.$$

Denote

$$\tilde{\lambda} = 1 - \left(\frac{\Gamma-1}{\Gamma+1}\right)\alpha\mu.$$

Then $\det(\tilde{\lambda}\mathbf{I} - \mathbf{A}) \geq 0$ so that $\rho(\mathbf{A}) \leq \tilde{\lambda}$.

3.3 Proof of Theorem 2

Similar to (21), under stepsize policy $\alpha_k = \theta/(m+k)$ where $m > \frac{\theta}{2}(\mu + L)$, we have

$$
\begin{bmatrix}
\mathbb{E}[\|\overline{x}_{k+1} - x^*\|^2] \\
\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2] \\
\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2]
\end{bmatrix}
\leq \mathbf{A}_k
\begin{bmatrix}
\mathbb{E}[\|\overline{x}_k - x^*\|^2] \\
\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] \\
\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2]
\end{bmatrix}
+
\begin{bmatrix}
\frac{\alpha_k^2 \sigma^2}{n} \\
0 \\
M_k
\end{bmatrix},
\tag{29}
$$

where

$$
\mathbf{A}_k =
\begin{bmatrix}
1 - \alpha_k \mu & \frac{\alpha_k L^2}{\mu n}(1 + \alpha_k \mu) & 0 \\
0 & \frac{1}{2}(1 + \rho_w^2) & \alpha_k^2 \frac{(1+\rho_w^2)\rho_w^2}{(1-\rho_w^2)} \\
2\alpha_k n L^3 & \left(\frac{1}{\beta_k} + 2\right)\|\mathbf{W} - \mathbf{I}\|^2 L^2 + 3\alpha_k L^3 & \frac{1}{2}(1 + \rho_w^2)
\end{bmatrix},
$$

$\beta_k = \frac{1-\rho_w^2}{2\rho_w^2} - 4\alpha_k L - 2\alpha_k^2 L^2 > 0$, and

$$
M_k := \left[3\alpha_k^2 L^2 + 2(\alpha_k L + 1)(n + 1)\right]\sigma^2.
\tag{30}
$$

The condition $\beta_k > 0$ is satisfied when

$$
\beta_0 = \frac{1 - \rho_w^2}{2\rho_w^2} - \frac{4\theta L}{m} - \frac{2\theta^2 L^2}{m^2} > 0,
$$

or equivalently,

$$
m > \frac{4\theta L \rho_w^2 + 2\theta L \rho_w \sqrt{1 + 3\rho_w^2}}{1 - \rho_w^2}.
\tag{31}
$$

We first show that $\mathbb{E}[\|\overline{x}_k - x^*\|^2] \leq \hat{U}/(m+k)$ and $\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] \leq \hat{X}/(m+k)^2$ for some $\hat{U}, \hat{X} > 0$ by induction. Denote $U_k := \mathbb{E}[\|\overline{x}_k - x^*\|^2]$, $X_k := \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2]$, and $Y_k := \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2]$. Suppose for some $k \geq 0$,

$$
U_k \leq \hat{U}/(m+k), \quad X_k \leq \hat{X}/(m+k)^2, \quad Y_k \leq \hat{Y}.
\tag{32}
$$

We want to prove that

$$
U_{k+1} \leq \frac{(1 - \alpha_k \mu)\hat{U}}{(m+k)} + \frac{\alpha_k L^2}{\mu n}\frac{(1 + \alpha_k \mu)\hat{X}}{(m+k)^2} + \frac{\alpha_k^2 \sigma^2}{n} \leq \frac{\hat{U}}{(m+k+1)},
\tag{33a}
$$

$$
X_{k+1} \leq \frac{(1 + \rho_w^2)\hat{X}}{2(m+k)^2} + \alpha_k^2 \frac{(1 + \rho_w^2)\rho_w^2 \hat{Y}}{(1 - \rho_w^2)} \leq \frac{\hat{X}}{(m+k+1)^2},
\tag{33b}
$$

$$Y_{k+1} \leq \frac{2\alpha_k n L^3 \hat{U}}{(m+k)} + \frac{C\hat{X}}{(m+k)^2} + \frac{(1+\rho_w^2)\hat{Y}}{2} + M_0 \leq \hat{Y}, \tag{33c}$$

where $C = (\frac{1}{\beta_0} + 2)\|\mathbf{W} - \mathbf{I}\|^2 L^2 + 3\alpha_0 L^3$. Plugging in $\alpha_k = \theta/(m+k)$, it suffices to show

$$\hat{U} \geq \frac{1}{n(\theta\mu - 1)}\left[\left(\frac{\theta L^2}{\mu m} + \frac{\theta^2 L^2}{m^2}\right)\hat{X} + \theta^2 \sigma^2\right], \tag{34a}$$

$$\hat{Y} \leq \frac{(1-\rho_w^2)}{\theta^2(1+\rho_w^2)\rho_w^2}\left[\frac{(1-\rho_w^2)}{2} - \frac{2m+1}{(m+1)^2}\right]\hat{X}, \tag{34b}$$

$$\hat{Y} \geq \frac{2}{(1-\rho_w^2)}\left(\frac{2\theta n L^3 \hat{U}}{m^2} + \frac{C\hat{X}}{m^2} + \left[\frac{3\theta^2 L^2}{m^2} + 2\left(\frac{\theta L}{m} + 1\right)(n+1)\right]\sigma^2\right). \tag{34c}$$

Let

$$\hat{U} := \max\left\{\frac{1}{n(\theta\mu - 1)}\left[\left(\frac{\theta L^2}{\mu m} + \frac{\theta^2 L^2}{m^2}\right)\hat{X} + \theta^2 \sigma^2\right], m\|\bar{x}_0 - x^*\|^2\right\},$$

condition (34) admits a solution iff

$$\frac{(1-\rho_w^2)^2}{\theta^2(1+\rho_w^2)\rho_w^2}\left[\frac{(1-\rho_w^2)}{2} - \frac{2m+1}{(m+1)^2}\right] > \frac{1}{(\theta\mu - 1)}\left(\frac{1}{\mu} + \frac{\theta}{m}\right)\frac{4\theta^2 L^5}{m^3} + \frac{2C}{m^2}, \tag{35}$$

in which case $\hat{X}$ and $\hat{Y}$ are lower bounded under constraints $\{(34b), (34c), \hat{X} \geq m^2 X_0, \hat{Y} \geq Y_0\}$. Specifically, $\hat{X}$ can be chosen as follows:

$$\hat{X} := \max\left\{\frac{C_3}{C_1 - C_2}, \frac{C_5}{C_1 - C_4}, m^2 X_0, \frac{Y_0}{C_1}\right\},$$

where

$$C_1 = \frac{(1-\rho_w^2)}{\theta^2(1+\rho_w^2)\rho_w^2}\left[\frac{(1-\rho_w^2)}{2} - \frac{2m+1}{(m+1)^2}\right], \tag{36a}$$

$$C_2 = \frac{2C}{(1-\rho_w^2)m^2}, \tag{36b}$$

$$C_3 = \frac{2}{(1-\rho_w^2)}\left\{\frac{2\theta n L^3\|\bar{x}_0 - x^*\|^2}{m} + \left[\frac{3\theta^2 L^2}{m^2} + 2\left(\frac{\theta L}{m} + 1\right)(n+1)\right]\sigma^2\right\}, \tag{36c}$$

$$C_4 = \frac{2}{(1-\rho_w^2)}\left[\frac{2\theta^2 L^5}{m^3(\theta\mu - 1)}\left(\frac{1}{\mu} + \frac{\theta}{m}\right) + \frac{C}{m^2}\right], \tag{36d}$$

$$C_5 = \frac{2}{(1-\rho_w^2)}\left[\frac{2\theta^3 L^3}{m^2(\theta\mu - 1)} + \frac{3\theta^2 L^2}{m^2} + 2\left(\frac{\theta L}{m} + 1\right)(n+1)\right]\sigma^2. \tag{36e}$$

Noticing that relation (32) holds trivially when $k = 0$, the induction is complete.

We further improve the bound on U_k (inspired by [39,44]). From (29),

$$U(k+1) \leq (1 - \alpha_k \mu) U_k + \frac{2\alpha_k L^2}{\mu n} X_k + \frac{\alpha_k^2 \sigma^2}{n}.$$

Since $X_k \leq \hat{X}/(m+k)^2$,

$$U(k+1) \leq \left(1 - \frac{\theta\mu}{m+k}\right) U_k + \frac{2\theta L^2 \hat{X}}{\mu n (m+k)^3} + \frac{\theta^2 \sigma^2}{n(m+k)^2}.$$

Hence

$$U_k \leq \prod_{t=0}^{k-1} \left(1 - \frac{\theta\mu}{m+t}\right) U_0 + \sum_{t=0}^{k-1} \left(\prod_{j=t+1}^{k-1} \left(1 - \frac{\theta\mu}{m+j}\right)\right) \left(\frac{2\theta L^2 \hat{X}}{\mu n (m+t)^3} + \frac{\theta^2 \sigma^2}{n(m+t)^2}\right).$$

In light of Lemma 4.1 in [44],

$$\prod_{t=0}^{k-1} \left(1 - \frac{\theta\mu}{m+t}\right) \leq \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}}, \quad \prod_{j=t+1}^{k-1} \left(1 - \frac{\theta\mu}{m+j}\right) \leq \frac{(m+t+1)^{\theta\mu}}{(m+k)^{\theta\mu}}.$$

Then,

$$
\begin{aligned}
U_k &\leq \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}} U_0 + \sum_{t=0}^{k-1} \frac{(m+t+1)^{\theta\mu}}{(m+k)^{\theta\mu}} \left(\frac{2\theta L^2 \hat{X}}{\mu n (m+t)^3} + \frac{\theta^2 \sigma^2}{n(m+t)^2}\right) \\
&= \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}} U_0 + \frac{1}{(m+k)^{\theta\mu}} \left(\frac{2\theta L^2 \hat{X}}{\mu n} \sum_{t=0}^{k-1} \frac{(m+t+1)^{\theta\mu}}{(m+t)^3} + \frac{\theta^2 \sigma^2}{n} \sum_{t=0}^{k-1} \frac{(m+t+1)^{\theta\mu}}{(m+t)^2}\right) \\
&\leq \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}} U_0 + \frac{1}{(m+k)^{\theta\mu}} \left(\frac{4\theta L^2 \hat{X}}{\mu n} \sum_{t=0}^{k-1} (m+t)^{\theta\mu-3} + \frac{2\theta^2 \sigma^2}{n} \sum_{t=0}^{k-1} (m+t)^{\theta\mu-2}\right).
\end{aligned}
$$

Note that

$$\sum_{t=0}^{k-1} (m+t)^{\theta\mu-3} \leq \int_{t=-1}^{k} (m+t)^{\theta\mu-3} dt \leq \max\left\{\frac{1}{\theta\mu-2}(m+k)^{\theta\mu-2}, \frac{1}{2-\theta\mu}(m-1)^{\theta\mu-2}\right\},$$

$$\sum_{t=0}^{k-1} (m+t)^{\theta\mu-2} \leq \int_{t=-1}^{k} (m+t)^{\theta\mu-2} dt \leq \frac{1}{\theta\mu-1}(m+k)^{\theta\mu-1}.$$

We conclude that

$$U_k \leq \frac{2\theta^2 \sigma^2}{n(\theta\mu-1)(m+k)} + \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}} U_0$$

$$+ \frac{4\theta L^2 \hat{X}}{\mu n} \max\left\{ \frac{1}{(\theta\mu - 2)(m + k)^2}, \frac{(m-1)^{\theta\mu - 2}}{(2 - \theta\mu)(m + k)^{\theta\mu}} \right\}$$

$$\leq \frac{2\theta^2\sigma^2}{n(\theta\mu - 1)(m + k)} + \frac{\mathcal{O}_k(1)}{(m + k)^{\theta\mu}} + \frac{\mathcal{O}_k(1)}{(m + k)^2}.$$

4 A gossip-like stochastic gradient tracking method (GSGT)

In this section, we consider a gossip-like stochastic gradient tracking method (GSGT): Initialize with an arbitrary $x_{i,0}$ and $y_{i,0} = g_i(x_{i,0}, \xi_{i,0})$ for all $i \in \mathcal{N}$. At each round $k \in \mathbb{N}$, agent $i_k \in \mathcal{N}$ wakes up with probability $1/n$. Then i_k either communicates with one of its neighbors j_k (with probability $\pi_{i_k j_k}$) or not (with probability $\pi_{i_k i_k} = 1 - \sum_{j \in \mathcal{N}_{i_k}} \pi_{i_k j}$). In the former situation, the update rule for $i \in \{i_k, j_k\}$ is as follows,

$$x_{i,k+1} = \frac{1}{2}(x_{i_k,k} + x_{j_k,k}) - \alpha y_{i,k}, \tag{37a}$$

$$y_{i,k+1} = \frac{1}{2}(y_{i_k,k} + y_{j_k,k}) + g_i(x_{i,k+1}, \xi_{i,k+1}) - g_i(x_{i,k}, \xi_{i,k}), \tag{37b}$$

and for $i \notin \{i_k, j_k\}$, $x_{i,k+1} = x_{i,k}$, $y_{i,k+1} = y_{i,k}$, and $\xi_{i,k+1} = \xi_{i,k}$.[4] In the latter situation, agent i_k performs update based on its own information:

$$x_{i_k,k+1} = x_{i_k,k} - 2\alpha y_{i_k,k}, \tag{38a}$$

$$y_{i_k,k+1} = y_{i_k,k} + g_{i_k}(x_{i_k,k+1}, \xi_{i_k,k+1}) - g_{i_k}(x_{i_k,k}, \xi_{i_k,k}), \tag{38b}$$

while no action is taken by agent $i \neq i_k$. For ease of analysis, we denote $j_k = i_k$ in this case, and let $\mathbb{1}_k$ be the indicator function for the event $\{j_k \neq i_k\}$, i.e., $\mathbb{1}_k = 1$ iff $j_k \neq i_k$.

The use of stepsize 2α instead of α in (38a) can be understood as follows. At each iteration, GSGT performs two gradient updates within the network. This can be achieved either by two different agents respectively updating their solutions, or by one agent using a doubled stepsize. The method is different from a standard gossip algorithm where exactly two agents update at each round. This difference allows us to design the probabilities π_{ij} with more flexibility. In particular, it is possible to construct a doubly stochastic probability matrix $\mathbf{\Pi} = [\pi_{ij}]$ for any graph $\mathcal{G}$ under Assumption 3.

We can present GSGT in the following compact matrix form, in which we adopt the notation previously used.

$$\mathbf{x}_{k+1} = \mathbf{W}_k \mathbf{x}_k - \alpha \mathbf{D}_k \mathbf{y}_k, \tag{39a}$$

$$\mathbf{y}_{k+1} = \mathbf{W}_k \mathbf{y}_k + \tilde{\mathbf{D}}_k(G(\mathbf{x}_{k+1}, \xi_{k+1}) - G(\mathbf{x}_k, \xi_k)), \tag{39b}$$

---

[4] In practice, this means agent i holds vectors x_i, y_i and $g_i(x_i, \xi_i)$ if it does not wake up.

where the random coupling matrix $\mathbf{W}_k$ is defined as

$$\mathbf{W}_k := \mathbf{I} - \frac{(e_{i_k} - e_{j_k})(e_{i_k} - e_{j_k})^\mathsf{T}}{2},$$

in which $e_i = [0 \cdots 0\, 1\, 0 \cdots]^\mathsf{T} \in \mathbb{R}^{n \times 1}$ is a unit vector with the ith component equal to 1. By definition, each $\mathbf{W}_k$ is symmetric and doubly stochastic. The matrices $\mathbf{D}_k$ and $\tilde{\mathbf{D}}_k$ are diagonal with their i_kth and j_kth diagonal entries equal to 1 and all other entries equal to 0 if $j_k \neq i_k$, otherwise the i_kth entry of $\mathbf{D}_k$ (respectively, $\tilde{\mathbf{D}}_k$) equals 2 (respectively, 1) while all other entries equal 0.

We assume the following condition on the probability matrix $\mathbf{\Pi}$:

Assumption 5 Nonnegative matrix $\mathbf{\Pi}$ is doubly stochastic.

Let

$$\bar{\mathbf{W}} := \mathbb{E}[\mathbf{W}_k^\mathsf{T} \mathbf{W}_k]. \tag{40}$$

It can be shown that (see [4])

$$\bar{\mathbf{W}} = \left(1 - \frac{1}{n}\right)\mathbf{I} + \frac{\mathbf{\Pi} + \mathbf{\Pi}^\mathsf{T}}{2n}, \tag{41}$$

which is doubly stochastic.

Lemma 6 *Let Assumptions 3 and 5 hold, and let $\rho_{\bar{w}}$ denote the spectral norm of the matrix $\bar{\mathbf{W}} - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$. Then, $\rho_{\bar{w}} \in [1 - 2/n, 1)$.*

Proof Since $\bar{W}$ is doubly stochastic, $\rho_{\bar{w}} < 1$ follows from Lemma 1. To see $\rho_{\bar{w}} \geq 1 - 2/n$, note that $\rho(\frac{\mathbf{\Pi}+\mathbf{\Pi}^\mathsf{T}}{2n}) = \frac{1}{n}$, and 1 is an algebraically simple eigenvalue of $\frac{\mathbf{\Pi}+\mathbf{\Pi}^\mathsf{T}}{2}$ and $\frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$. We have

$$\rho_{\bar{w}} = 1 - \frac{1}{n} + \frac{1}{n}\lambda_2\left(\frac{\mathbf{\Pi} + \mathbf{\Pi}^\mathsf{T}}{2}\right), \tag{42}$$

where $\lambda_2(\frac{\mathbf{\Pi}+\mathbf{\Pi}^\mathsf{T}}{2})$ is the second largest eigenvalue of $\frac{\mathbf{\Pi}+\mathbf{\Pi}^\mathsf{T}}{2}$.

Since $\lambda_2(\frac{\mathbf{\Pi}+\mathbf{\Pi}^\mathsf{T}}{2}) \in [-1, 1]$, we conclude that $\rho_{\bar{w}} \geq 1 - 2/n$. □

Before proceeding, it is worth noting that for GSGT, we still have the following relation:

$$\bar{y}_k = \frac{1}{n}\mathbf{1}^\mathsf{T} G(\mathbf{x}_k, \boldsymbol{\xi}_k), \forall k. \tag{43}$$

4.1 Main results

We present the main convergence results of GSGT in the following theorem.

Theorem 3 *Let $\Gamma > 1$ be arbitrarily chosen. Suppose Assumptions 1–5 hold, and assume that the stepsize α satisfies*

$$\alpha \le \frac{2n(1-\rho_{\bar{w}})}{\sqrt{\Gamma}L} \{[27(2\eta+3)Qn + 16(8\eta+9)] Q(1-\rho_{\bar{w}}) + 48(6\eta+1)(8\eta+3) + 96Q(1-\rho_{\bar{w}})\}^{-1/2}, \tag{44}$$

where $\eta = \frac{1}{n(1-\rho_{\bar{w}})}$ and $Q = L/\mu$. Then $\sup_{l \ge k} \mathbb{E}[\|\bar{x}_l - x^*\|^2]$ and $\sup_{l \ge k} \mathbb{E}[\|\mathbf{x}_l - \mathbf{1}\bar{x}_l\|^2]$, respectively, converge to $\lim\sup_{k\to\infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2]$ and $\lim\sup_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]$ at the linear rate $\mathcal{O}(\rho(\mathbf{A}_g)^k)$, where $\rho(\mathbf{A}_g) < 1$ is the spectral radius of the matrix $\mathbf{A}_g$ given by

$$\mathbf{A}_g = \begin{bmatrix} 1 - \frac{2\alpha\mu}{n} & \frac{2\alpha L^2}{\mu n^2}\left(1 + \frac{2\alpha\mu}{n}\right) & \frac{4\alpha^2}{n^3} \\ 8\alpha^2 L^2 & \frac{1}{2}(1+\rho_{\bar{w}}) & \frac{2\alpha}{n}\left(\frac{1}{\beta_1} + \alpha\right) \\ 8\alpha^2 L^4 + 4\alpha L^3 & \frac{L^2}{n}\left(4 + \frac{2}{\beta_2} + 8\alpha^2 L^2 + 4\alpha L\right) & \frac{1}{2}(1+\rho_{\bar{w}}) \end{bmatrix},$$

in which $\beta_1 = \frac{n(1-\rho_{\bar{w}})}{4\alpha} - 4L^2$ and $\beta_2 = \frac{n(1-\rho_{\bar{w}})}{4} - 2\alpha L - 2\alpha^2 L^2$. In addition,

$$\lim\sup_{k\to\infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2] \le \frac{\Gamma}{(\Gamma-1)} \frac{\sigma^2}{n^2} \left[\frac{20\alpha}{\mu(1-\rho_{\bar{w}})} + \frac{42(6\eta+1)\alpha^2 L^2}{\mu^2(1-\rho_{\bar{w}})^2} \right], \tag{45}$$

and

$$\lim\sup_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] \le \frac{4\Gamma\sigma^2}{(\Gamma-1)(1-\rho_{\bar{w}})^2} \left[\frac{9(6\eta+1)\alpha^2}{n} + \frac{72\alpha^3 L^2}{\mu n^2} \right]. \tag{46}$$

Remark 5 Notice that $\eta = \frac{1}{n(1-\rho_{\bar{w}})}$ and $1 - \rho_{\bar{w}} \le \frac{2}{n}$ from Lemma 6. We can see from (45) and (46) that

$$\lim\sup_{k\to\infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2] = \frac{\alpha}{(1-\rho_{\bar{w}})} \mathcal{O}\left(\frac{\sigma^2}{\mu n^2}\right) + \frac{\alpha^2}{(1-\rho_{\bar{w}})^3} \mathcal{O}\left(\frac{L^2\sigma^2}{\mu^2 n^3}\right), \tag{47}$$

and

$$\lim\sup_{k\to\infty} \frac{1}{n} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] = \frac{\alpha^2}{(1-\rho_{\bar{w}})^3} \mathcal{O}\left(\frac{\sigma^2}{n^3}\right) + \frac{\alpha^3}{(1-\rho_{\bar{w}})^2} \mathcal{O}\left(\frac{L^2\sigma^2}{\mu n^3}\right). \tag{48}$$

Since

$$\frac{L^2\sigma^2}{\mu^2 n^3} \ge \frac{\sigma^2}{n^3},$$

and by (44),

$$\frac{L^2\sigma^2}{\mu^2 n^3} \geq \alpha(1-\rho_{\bar{w}})\frac{L^2\sigma^2}{\mu n^3},$$

the second term on the right-hand side of (47) dominates the two terms on the right-hand side of (48). Thus, we have

$$\limsup_{k\to\infty}\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2] = \limsup_{k\to\infty}\mathbb{E}[\|\bar{x}_k - x^*\|^2] + \limsup_{k\to\infty}\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]$$

$$= \frac{\alpha}{(1-\rho_{\bar{w}})}\mathcal{O}\left(\frac{\sigma^2}{\mu n^2}\right) + \frac{\alpha^2}{(1-\rho_{\bar{w}})^3}\mathcal{O}\left(\frac{L^2\sigma^2}{\mu^2 n^3}\right). \quad (49)$$

$\square$

The corollary below provides an upper bound for $\rho(\mathbf{A}_g)$.

Corollary 3 *Under the conditions in Theorem 2 where $\Gamma > 3/2$, we have*

$$\rho(\mathbf{A}_g) \leq 1 - \frac{(2\Gamma - 3)}{\Gamma}\frac{\alpha\mu}{n}.$$

Remark 6 Compared to DSGT, the convergence speed of GSGT is slower than DSGT under the same stepsize α (see Corollary 1). This is due to the fact that in GSGT, only two agents update their iterates at each iteration. $\square$

4.2 Performance comparison between DSGT and GSGT

In this section, we compare the performances of the two proposed algorithms in terms of their required computation and communication efforts for achieving an ϵ-solution (with constant stepsizes), that is, we compute the number of stochastic gradient computations and communications needed to obtain $\frac{1}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2] \leq \epsilon$. Without loss of generality, for each method we first choose stepsize α such that $\frac{1}{n}\limsup_{k\to\infty}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}x^*\|^2] \leq \epsilon/2$ and then compute the number of iterations K such that $\frac{1}{n}\mathbb{E}[\|\mathbf{x}_K - \mathbf{1}x^*\|^2] \leq \epsilon$.

For DSGT, when ϵ is small enough, we have $\alpha = \mathcal{O}(\frac{n\mu\epsilon}{\sigma^2})$ from (11). Then, noting that $\sup_{l\geq k}\mathbb{E}[\|\mathbf{x}_l - \mathbf{1}x^*\|^2]$ converges linearly at the rate $\mathcal{O}(\rho(\mathbf{A})^k)$ where $\rho(\mathbf{A}) = 1 - \mathcal{O}(\alpha\mu)$, we obtain the number of required iterations:

$$K_d = \mathcal{O}\left(\frac{\ln(\frac{1}{\epsilon})}{\alpha\mu}\right) = \mathcal{O}\left(\frac{\sigma^2}{n\mu^2}\frac{\ln(\frac{1}{\epsilon})}{\epsilon}\right).$$

In K_d iterations, the number of stochastic gradient computations is $N_d = nK_d = \mathcal{O}\left(\frac{\sigma^2}{\mu^2}\frac{\ln(\frac{1}{\epsilon})}{\epsilon}\right)$ and the number of communications is $N_d^c = 2|\mathcal{E}|K_d = \mathcal{O}\left(\frac{|\mathcal{E}|}{n}\frac{\sigma^2}{\mu^2}\frac{\ln(\frac{1}{\epsilon})}{\epsilon}\right)$ where $|\mathcal{E}|$ stands for the number of edges in the graph.

For GSGT we need $\alpha = \mathcal{O}(\frac{n^2 \mu \epsilon (1-\rho_{\bar{w}})}{\sigma^2})$ from (49). Given that $\rho(\mathbf{A}_g) = 1 - \mathcal{O}(\frac{\alpha \mu}{n})$, the number of required iterations K_g can be calculated as follows:

$$K_g = \mathcal{O}\left(\frac{n \ln(\frac{1}{\epsilon})}{\alpha \mu}\right) = \mathcal{O}\left(\frac{\sigma^2}{n(1-\rho_{\bar{w}})\mu^2} \frac{\ln(\frac{1}{\epsilon})}{\epsilon}\right).$$

In K_g iterations, the number of gradient computations and communications are both bounded by $N_g = N_g^c = 2K_g = \mathcal{O}(\frac{1}{n(1-\rho_{\bar{w}})} \frac{\sigma^2}{\mu^2} \frac{\ln(\frac{1}{\epsilon})}{\epsilon})$.

Suppose the Metropolis rule is applied to define the weights π_{ij} [50]. We first compare the number of stochastic gradient computations for DSGT and GSGT, respectively. Noticing that $1 - \rho_{\bar{w}} \leq \frac{2}{n}$ by Lemma 6, N_g is at most in the same order of $N_d = \mathcal{O}(\frac{\sigma^2}{\mu^2} \frac{\ln(\frac{1}{\epsilon})}{\epsilon})$, which happens when $1 - \rho_{\bar{w}} = \mathcal{O}(\frac{1}{n})$. Given that $1 - \rho_{\bar{w}} = (1 - \lambda_2(\mathbf{\Pi}))/n$ from (42), we have $1 - \rho_{\bar{w}} = \mathcal{O}(\frac{1}{n})$ for complete networks, almost all regular graphs [16], among others.

We then compare the number of required communications $N_d^c = \mathcal{O}(\frac{|\mathcal{E}|}{n} \frac{\sigma^2}{\mu^2} \frac{\ln(\frac{1}{\epsilon})}{\epsilon})$ and $N_g^c = \mathcal{O}(\frac{1}{n(1-\rho_{\bar{w}})} \frac{\sigma^2}{\mu^2} \frac{\ln(\frac{1}{\epsilon})}{\epsilon})$. When $1 - \rho_{\bar{w}} = \mathcal{O}(\frac{1}{n})$, we have $N_g^c = \mathcal{O}(\frac{\sigma^2}{\mu^2} \frac{\ln(\frac{1}{\epsilon})}{\epsilon})$. By contrast, N_d^c is $\mathcal{O}(\frac{|\mathcal{E}|}{n})$ times larger than N_g^c. In particular when $|\mathcal{E}| = \mathcal{O}(n^2)$ (e.g., complete network), the number of communications for GSGT is $\mathcal{O}(n)$ times smaller than that of DSGT.

4.3 Proof of Theorem 3

We first derive a linear system of inequalities regarding $\mathbb{E}[\|\bar{x}_k - x^*\|^2]$, $\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]$, $\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2]$ and their values in the last iteration.

Lemma 7 *Suppose Assumptions 1–5 hold and the stepsize satisfies $\alpha < n/(\mu + L)$. Then, we have the following inequalities:*

$$\mathbb{E}[\|\bar{x}_{k+1} - x^*\|^2] \leq \left(1 - \frac{2\alpha\mu}{n}\right) \mathbb{E}[\|\bar{x}_k - x^*\|^2] + \frac{2\alpha L^2}{\mu n^2}\left(1 + \frac{2\alpha\mu}{n}\right) \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]$$
$$+ \frac{4\alpha^2}{n^3} \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \frac{4\alpha^2 \sigma^2}{n^3}. \tag{50}$$

For any $\beta_1, \beta_2 > 0$,

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2] \leq \left(\rho_{\bar{w}} + \frac{2\alpha}{n}\beta_1 + \frac{8\alpha^2 L^2}{n}\right) \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]$$
$$+ \frac{2\alpha}{n}\left(\frac{1}{\beta_1} + \alpha\right) \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2]$$
$$+ 8\alpha^2 L^2 \mathbb{E}[\|\bar{x}_k - x^*\|^2] + \frac{4\alpha^2 \sigma^2}{n}, \tag{51}$$

$$\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2] \leq \left(\rho_{\bar{w}} + \frac{2}{n}\beta_2 + \frac{4\alpha L}{n} + \frac{4\alpha^2 L^2}{n}\right)\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2]$$

$$+ \frac{L^2}{n}\left(4 + \frac{2}{\beta_2} + 8\alpha^2 L^2 + 4\alpha L\right)\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2]$$

$$+ (8\alpha^2 L^4 + 4\alpha L^3)\mathbb{E}[\|\overline{x}_k - x^*\|^2] + \frac{(4\alpha^2 L^2 + 2\alpha L)\sigma^2}{n}$$

$$+ 4(\alpha L + 1)\sigma^2. \tag{52}$$

Proof See Appendix 7.3. □

In light of Lemma 7, we have the following linear system of inequalities:

$$\begin{bmatrix} \mathbb{E}[\|\overline{x}_{k+1} - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2] \\ \mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2] \end{bmatrix} \leq \mathbf{A}_g \begin{bmatrix} \mathbb{E}[\|\overline{x}_k - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] \\ \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] \end{bmatrix} + \begin{bmatrix} \frac{4\alpha^2\sigma^2}{n^3} \\ \frac{4\alpha^2\sigma^2}{n} \\ M_g \end{bmatrix},$$

where

$$\mathbf{A}_g = [b_{ij}] = \begin{bmatrix} 1 - \frac{2\alpha\mu}{n} & \frac{2\alpha L^2}{\mu n^2}\left(1 + \frac{2\alpha\mu}{n}\right) & \frac{4\alpha^2}{n^3} \\ 8\alpha^2 L^2 & \rho_{\bar{w}} + \frac{2\alpha}{n}\beta_1 + \frac{8\alpha^2 L^2}{n} & \frac{2\alpha}{n}\left(\frac{1}{\beta_1} + \alpha\right) \\ 8\alpha^2 L^4 + 4\alpha L^3 & \frac{L^2}{n}\left(4 + \frac{2}{\beta_2} + 8\alpha^2 L^2 + 4\alpha L\right) & \rho_{\bar{w}} + \frac{2}{n}\beta_2 + \frac{4\alpha L}{n} + \frac{4\alpha^2 L^2}{n} \end{bmatrix},$$

and $M_g = \frac{(4\alpha^2 L^2 + 2\alpha L)\sigma^2}{n} + 4(\alpha L + 1)\sigma^2$. Suppose $\alpha, \beta_1, \beta_2 > 0$ satisfy

$$b_{22} = \rho_{\bar{w}} + \frac{2\alpha}{n}\beta_1 + \frac{8\alpha^2 L^2}{n} = \frac{1 + \rho_{\bar{w}}}{2}, \tag{53}$$

$$b_{33} = \rho_{\bar{w}} + \frac{2}{n}\beta_2 + \frac{4\alpha L}{n} + \frac{4\alpha^2 L^2}{n} = \frac{1 + \rho_{\bar{w}}}{2}, \tag{54}$$

and

$$\det(\mathbf{I} - \mathbf{A}_g) = (1 - b_{11})(1 - b_{22})(1 - b_{33}) - b_{12}b_{23}b_{31} - b_{13}b_{21}b_{32}$$
$$- (1 - b_{11})b_{23}b_{32} - (1 - b_{22})b_{13}b_{31}$$
$$- (1 - b_{33})b_{12}b_{21} \geq (1 - 1/\Gamma)(1 - b_{11})(1 - b_{22})(1 - b_{33}) > 0. \tag{55}$$

Then, by Lemma 5, the spectral radius of $\mathbf{A}_g$ is smaller than 1, and we have

$$\begin{bmatrix} \mathbb{E}[\|\overline{x}_k - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] \\ \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] \end{bmatrix} \leq \mathbf{A}_g^k \begin{bmatrix} \mathbb{E}[\|\overline{x}_0 - x^*\|^2] \\ \mathbb{E}[\|\mathbf{x}_0 - \mathbf{1}\overline{x}_0\|^2] \\ \mathbb{E}[\|\mathbf{y}_0 - \mathbf{1}\overline{y}_0\|^2] \end{bmatrix} + \sum_{l=0}^{k-1} \mathbf{A}_g^l \begin{bmatrix} \frac{4\alpha^2\sigma^2}{n^3} \\ \frac{4\alpha^2\sigma^2}{n} \\ M_g \end{bmatrix}. \tag{56}$$

Hence, $\sup_{l \geq k} \mathbb{E}[\|\bar{x}_l - x^*\|^2]$, $\sup_{l \geq k} \mathbb{E}[\|\mathbf{x}_l - \mathbf{1}\bar{x}_l\|^2]$ and $\sup_{l \geq k} \mathbb{E}[\|\mathbf{y}_l - \mathbf{1}\bar{y}_l\|^2]$ all converge to a neighborhood of 0 at the linear rate $\mathcal{O}(\rho(\mathbf{A}_g)^k)$. Moreover,

$$
\begin{bmatrix} \limsup_{k \to \infty} \mathbb{E}[\|\bar{x}_k - x^*\|^2] \\ \limsup_{k \to \infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] \\ \limsup_{k \to \infty} \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] \end{bmatrix} \leq (\mathbf{I} - \mathbf{A}_g)^{-1} \begin{bmatrix} \frac{4\alpha^2\sigma^2}{n^3} \\ \frac{2\alpha^2\sigma^2}{n} \\ M_g \end{bmatrix} = \frac{1}{\det(\mathbf{I} - \mathbf{A})}
$$

$$
\cdot \begin{bmatrix} (1-b_{22})(1-b_{33}) - b_{23}b_{32} & b_{13}b_{32} + b_{12}(1-b_{33}) & b_{12}b_{23} + b_{13}(1-b_{22}) \\ b_{23}b_{31} + b_{21}(1-b_{33}) & (1-b_{11})(1-b_{33}) - b_{13}b_{31} & b_{13}b_{21} + b_{23}(1-b_{11}) \\ b_{21}b_{32} + b_{31}(1-b_{22}) & b_{12}b_{31} + b_{32}(1-b_{11}) & (1-b_{11})(1-b_{22}) - b_{12}b_{21} \end{bmatrix}
$$

$$
\begin{bmatrix} \frac{4\alpha^2\sigma^2}{n^3} \\ \frac{2\alpha^2\sigma^2}{n} \\ M_g \end{bmatrix}. \tag{57}
$$

We now show (53), (54), and (57) are satisfied under condition (44). First, relation (44) implies that

$$
4\alpha^2 L^2 \leq \frac{n(1 - \rho_{\bar{w}})}{12}, \tag{58}
$$

$$
2\alpha L + 2\alpha^2 L^2 \leq \frac{n(1 - \rho_{\bar{w}})}{12}. \tag{59}
$$

Therefore, from (53) and (54) we have

$$
\beta_1 = \frac{n(1 - \rho_{\bar{w}})}{4\alpha} - 4\alpha L^2 \geq \frac{n(1 - \rho_{\bar{w}})}{6\alpha} > 0, \tag{60}
$$

$$
\beta_2 = \frac{n(1 - \rho_{\bar{w}})}{4} - 2\alpha L - 2\alpha^2 L^2 \geq \frac{n(1 - \rho_{\bar{w}})}{6} > 0. \tag{61}
$$

By (58)–(61) and the fact that $\rho_{\bar{w}} \geq 1 - 2/n$ obtained from (41), we have

$$
b_{12} \leq \frac{2\alpha L^2}{\mu n^2}\left(1 + \frac{1}{8}\right) \leq \frac{9\alpha L^2}{4\mu n^2}, \tag{62a}
$$

$$
b_{23} \leq \frac{2\alpha^2}{n}(6\eta + 1), \tag{62b}
$$

$$
b_{31} = \alpha L^3(8\alpha L + 4) \leq 6\alpha L^3, \tag{62c}
$$

$$
b_{32} \leq \frac{L^2}{n}\left(12\eta + \frac{9}{2}\right). \tag{62d}
$$

Then, for relation (55) to hold, it is sufficient that

$$
\frac{1}{\Gamma}\frac{\alpha\mu}{2n}(1 - \rho_{\bar{w}})^2 \geq \frac{27(6\eta+1)\alpha^4 L^5}{\mu n^3} + \frac{48(8\eta + 3)\alpha^4 L^4}{n^4} + \frac{6(6\eta + 1)(8\eta + 3)\alpha^3\mu L^2}{n^3}
$$

$$
+ \frac{12\alpha^3 L^3}{n^3}(1 - \rho_{\bar{w}}) + \frac{9\alpha^3 L^4}{\mu n^2}(1 - \rho_{\bar{w}}).
$$

In light of (58), $\alpha L \leq n(1 - \rho_{\bar{w}})/24$. We only need

$$\frac{1}{\Gamma}\frac{\mu n^2}{2}(1 - \rho_{\bar{w}})^2 \geq \frac{9(6\eta + 1)\alpha^2 L^4 n}{8\mu}(1 - \rho_{\bar{w}}) + 2(8\eta + 3)\alpha^2 L^3(1 - \rho_{\bar{w}})$$
$$+ 6(6\eta + 1)(8\eta + 3)\alpha^2\mu L^2 + 12\alpha^2 L^3(1 - \rho_{\bar{w}})$$
$$+ \frac{9\alpha^2 L^4 n}{\mu}(1 - \rho_{\bar{w}}),$$

which gives

$$\alpha \leq \frac{2n(1 - \rho_{\bar{w}})}{\sqrt{\Gamma}L}\{[27(2\eta + 3)Qn + 16(8\eta + 9)]Q(1 - \rho_{\bar{w}})$$
$$+ 48(6\eta + 1)(8\eta + 3) + 96Q(1 - \rho_{\bar{w}})\}^{-1/2}.$$

We now derive the bounds for $\limsup_{k\to\infty} \mathbb{E}[\|\bar{x}_{k+1} - x^*\|^2]$ and $\limsup_{k\to\infty} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2]$. By (57) and (62),

$$\limsup_{k\to\infty} \mathbb{E}[\|\bar{x}_{k+1} - x^*\|^2] \leq \frac{\Gamma}{(\Gamma - 1)(1 - b_{11})(1 - b_{22})(1 - b_{33})}$$
$$\cdot \left\{[(1 - b_{22})(1 - b_{33}) - b_{23}b_{32}]\frac{4\alpha^2\sigma^2}{n^3} + [b_{13}b_{32} + b_{12}(1 - b_{33})]\frac{2\alpha^2\sigma^2}{n}\right.$$
$$\left.+ [b_{12}b_{23} + b_{13}(1 - b_{22})]M_g\right\}$$
$$\leq \frac{2\Gamma n}{(\Gamma - 1)\alpha\mu(1 - \rho_{\bar{w}})^2}\left\{\frac{\alpha^2\sigma^2(1 - \rho_{\bar{w}})^2}{n^3} + \left[\frac{6\alpha^2 L^2}{n^4}(8\eta + 3) + \frac{9\alpha L^2}{8\mu n^2}(1 - \rho_{\bar{w}})\right]\frac{2\alpha^2\sigma^2}{n}\right.$$
$$\left.+ \left[\frac{9(6\eta + 1)\alpha^3 L^2}{2\mu n^3} + \frac{2\alpha^2}{n^3}(1 - \rho_{\bar{w}})\right]\frac{9}{2}\sigma^2\right\}$$
$$\leq \frac{2\Gamma n\sigma^2}{(\Gamma - 1)\alpha\mu(1 - \rho_{\bar{w}})^2}\left[\frac{10\alpha^2(1 - \rho_{\bar{w}})}{n^3} + \frac{21(6\eta + 1)\alpha^3 L^2}{\mu n^3}\right]$$
$$= \frac{\Gamma}{(\Gamma - 1)}\frac{\sigma^2}{n^2}\left[\frac{20\alpha}{\mu(1 - \rho_{\bar{w}})} + \frac{42(6\eta + 1)\alpha^2 L^2}{\mu^2(1 - \rho_{\bar{w}})^2}\right].$$

$$\limsup_{k\to\infty} \mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] \leq \frac{\Gamma}{(\Gamma - 1)(1 - b_{11})(1 - b_{22})(1 - b_{33})}$$
$$\cdot \left\{[b_{23}b_{31} + b_{21}(1 - b_{33})]\frac{4\alpha^2\sigma^2}{n^3} + [(1 - b_{11})(1 - b_{33}) - b_{13}b_{31}]\frac{2\alpha^2\sigma^2}{n}\right.$$
$$\left.+ [b_{13}b_{21} + b_{23}(1 - b_{11})]M_g\right\}$$
$$\leq \frac{2\Gamma n}{(\Gamma - 1)\alpha\mu(1 - \rho_{\bar{w}})^2}\left\{\left[\frac{12(6\eta + 1)\alpha^3 L^3}{n} + 4\alpha^2 L^2(1 - \rho_{\bar{w}})\right]\frac{4\alpha^2\sigma^2}{n^3}\right.$$
$$\left.+ \left[\frac{\alpha\mu(1 - \rho_{\bar{w}})}{n} - \frac{24\alpha^3 L^3}{n^3}\right]\frac{2\alpha^2\sigma^2}{n} + \left[\frac{32\alpha^4 L^2}{n^3} + \frac{4(6\eta + 1)\alpha^3\mu}{n^2}\right]\frac{17}{4}\sigma^2\right\}$$

$$\leq \frac{2\Gamma n\sigma^2}{(\Gamma - 1)\alpha\mu(1 - \rho_{\bar{w}})^2} \left[\frac{18(6\eta + 1)\alpha^3\mu}{n^2} + \frac{136\alpha^4 L^2}{n^3} \right]$$

$$= \frac{4\Gamma\sigma^2}{(\Gamma - 1)(1 - \rho_{\bar{w}})^2} \left[\frac{9(6\eta + 1)\alpha^2}{n} + \frac{72\alpha^3 L^2}{\mu n^2} \right].$$

4.4 Proof of Corollary 3

The characteristic function of $\mathbf{A}_g$ is

$$\det(\lambda\mathbf{I} - \mathbf{A}_g) = (\lambda - b_{11})(\lambda - b_{22})(\lambda - b_{33}) - b_{12}b_{23}b_{31} - b_{13}b_{21}b_{32} - (\lambda - b_{11})b_{23}b_{32}$$
$$- (\lambda - b_{22})b_{13}b_{31} - (\lambda - b_{33})b_{12}b_{21}. \tag{63}$$

By (55),

$$\det(\lambda\mathbf{I} - \mathbf{A}_g) \geq (\lambda - b_{11})(\lambda - b_{22})(\lambda - b_{33}) + (1 - \lambda)b_{23}b_{32}$$
$$+ (1 - \lambda)b_{13}b_{31} + (1 - \lambda)b_{12}b_{21}$$
$$- \frac{1}{\Gamma}(1 - b_{11})(1 - b_{22})(1 - b_{33})$$
$$\geq (\lambda - b_{11})(\lambda - b_{22})(\lambda - b_{33})$$
$$- \frac{1}{\Gamma}(1 - b_{11})(1 - b_{22})(1 - b_{33}). \tag{64}$$

Let $\lambda = 1 - \epsilon$ for some $\epsilon \in (0, 2\alpha\mu/n)$ that satisfies

$$\det(\lambda\mathbf{I} - \mathbf{A}_g) \geq \left(\frac{2\alpha\mu}{n} - \epsilon \right) \left[\frac{1 - \rho_{\bar{w}}}{2} - \epsilon \right]^2 - \frac{1}{\Gamma} \frac{2\alpha\mu}{n} \frac{(1 - \rho_{\bar{w}})^2}{4} \geq 0.$$

Under condition (44), it suffices that

$$\epsilon \leq \frac{(2\Gamma - 3)}{\Gamma} \frac{\alpha\mu}{n}.$$

Denote $\tilde{\lambda} = 1 - \frac{(2\Gamma - 3)}{\Gamma}\frac{\alpha\mu}{n}$. We have $\det(\tilde{\lambda}\mathbf{I} - \mathbf{A}_g) \geq 0$, and therefore $\rho(\mathbf{A}_g) \leq \tilde{\lambda}$.

5 Numerical example

In this section, we provide a numerical example to illustrate our theoretic findings. Consider the *on-line* Ridge regression problem, i.e.,

$$\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \left(= \mathbb{E}_{u_i, v_i} \left[\left(u_i^\mathsf{T} x - v_i \right)^2 + \rho\|x\|^2 \right] \right), \tag{65}$$

where $\rho > 0$ is a penalty parameter. For each agent i, samples in the form of (u_i, v_i) are gathered continuously with $u_i \in \mathbb{R}^p$ representing the features and $v_i \in \mathbb{R}$ being

the observed outputs. We assume that each $u_i \in [0.3, 0.4]^p$ is uniformly distributed, and v_i is drawn according to $v_i = u_i^\mathsf{T} \tilde{x}_i + \varepsilon_i$, where $\tilde{x}_i$ are predefined parameters evenly located in $[0, 10]^p$, and ε_i are independent Gaussian noises with mean 0 and variance 1. Given a pair (u_i, v_i), agent i can compute an estimated gradient of $f_i(x)$: $g_i(x, u_i, v_i) = 2(u_i^\mathsf{T} x - v_i)u_i + 2\rho x$, which is unbiased. Problem (65) has a unique solution x^* given by $x^* = (\sum_{i=1}^n \mathbb{E}_{u_i}[u_i u_i^\mathsf{T}] + n\rho\mathbf{I})^{-1} \sum_{i=1}^n \mathbb{E}_{u_i}[u_i u_i^\mathsf{T}]\tilde{x}_i$.

In addition to DSGT, GSGT and CSG, we consider the following distributed stochastic gradient (DSG) algorithm, which is similar to the ones studied in [18,26]:

$$\mathbf{x}_{k+1} = \mathbf{W}\mathbf{x}_k - \alpha G(\mathbf{x}_k, \boldsymbol{\xi}_k). \tag{66}$$

Noticing that some existing algorithms for deterministic distributed optimization can also be adapted to the stochastic gradient setting, e.g., EXTRA [52] and decentralized ADMM (DLM) [27], we also include them in our experiments for comparison.

In the experiments, we consider 3 instances with $p = 20$ and $n \in \{10, 25, 100\}$, respectively. Under each instance, we let $\mathbf{x}_0 = \mathbf{0}$ and the penalty parameter $\rho = 0.1$. For the distributed methods, we assume that n agents constitute a random network, in which each two agents are linked with probability 0.4. The Metropolis rule is applied to define the weights w_{ij} (and π_{ij}) where applicable [50]:
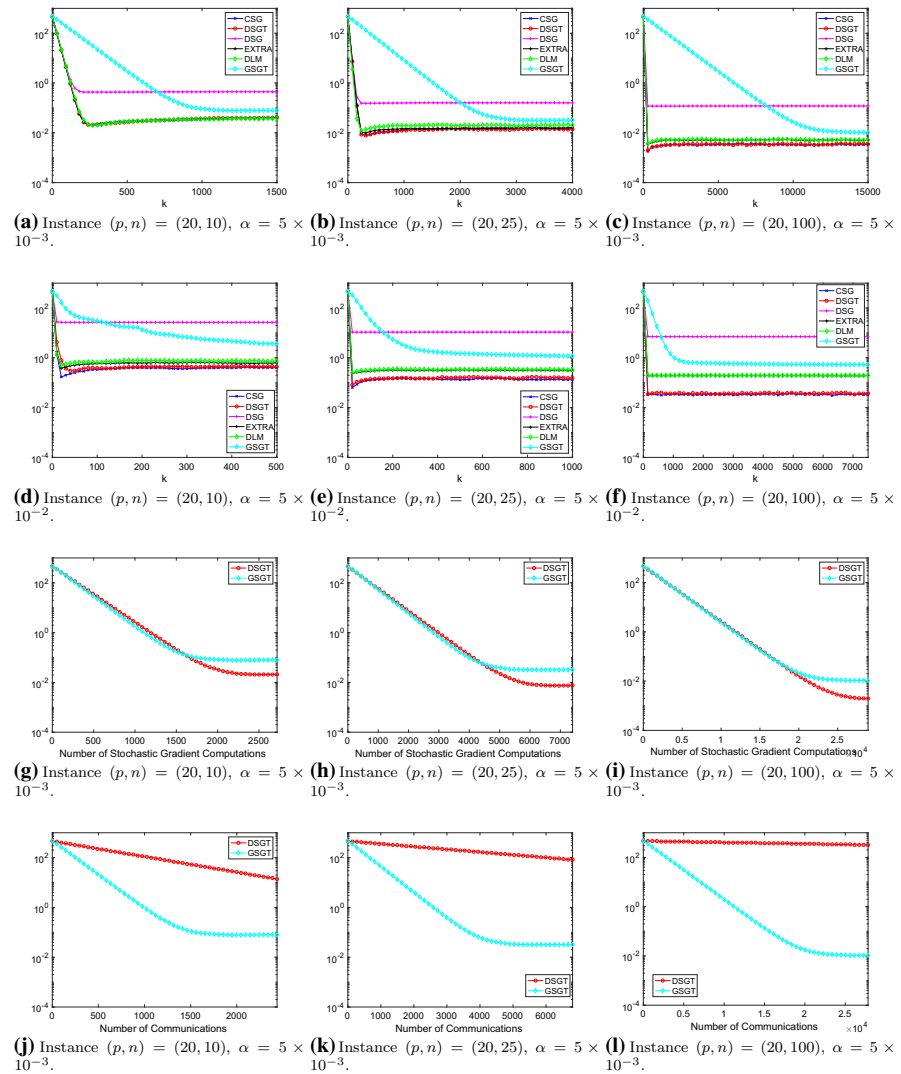
$$w_{ij} = \begin{cases} 1/\max\{\deg(i), \deg(j)\} & \text{if } i \in \mathcal{N}_i, \\ 1 - \sum_{j \in \mathcal{N}_i} w_{ij} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

For EXTRA, we choose $\tilde{\mathbf{W}} = \frac{\mathbf{I}+\mathbf{W}}{2}$ as recommended by [52]. For DLM, we tune the free parameters to make its convergence speed comparable to the other algorithms. In each instance, we use two different stepsizes $\alpha = 5 \times 10^{-3}$ and $\alpha = 5 \times 10^{-2}$, respectively. We run the simulations 50 times for DSGT, CSG, DSG, EXTRA and DLM and 100 times for GSGT and average the results to approximate the expected errors.

In Fig. 1a–f, we compare the average performances of DSGT, GSGT, CSG, DSG, EXTRA and DLM with the same parameters. It can be seen that DSGT and CSG are comparable in their convergence speeds as well as the ultimate error bounds (almost indistinguishable). EXTRA and DLM are worse than DSGT and CSG in their final error bounds. The performance gap increases with the network size and the stepsize. GSGT is slower as expected but still reaches a comparable error level under small stepsize $\alpha = 5 \times 10^{-3}$. In addition, the error bounds for DSGT, GSGT and CSG decrease in n as expected from our theoretical analysis. The performance of DSG is not favorable given its largest final errors.[5]

In Fig. 1g–i (respectively, j–l), we further compare the solutions obtained under DSGT and GSGT with the same number of stochastic gradient evaluations (respectively, inter-node communications) under small stepsize $\alpha = 5 \times 10^{-3}$. We see the two

---

[5] DSG still holds the advantage over DSGT in the early stage for achieving a similar convergence speed with lower communication and storage costs.

(a) Instance $(p, n) = (20, 10)$, $\alpha = 5 \times 10^{-3}$.

(b) Instance $(p, n) = (20, 25)$, $\alpha = 5 \times 10^{-3}$.

(c) Instance $(p, n) = (20, 100)$, $\alpha = 5 \times 10^{-3}$.

(d) Instance $(p, n) = (20, 10)$, $\alpha = 5 \times 10^{-2}$.

(e) Instance $(p, n) = (20, 25)$, $\alpha = 5 \times 10^{-2}$.

(f) Instance $(p, n) = (20, 100)$, $\alpha = 5 \times 10^{-2}$.

(g) Instance $(p, n) = (20, 10)$, $\alpha = 5 \times 10^{-3}$.

(h) Instance $(p, n) = (20, 25)$, $\alpha = 5 \times 10^{-3}$.

(i) Instance $(p, n) = (20, 100)$, $\alpha = 5 \times 10^{-3}$.

(j) Instance $(p, n) = (20, 10)$, $\alpha = 5 \times 10^{-3}$.

(k) Instance $(p, n) = (20, 25)$, $\alpha = 5 \times 10^{-3}$.

(l) Instance $(p, n) = (20, 100)$, $\alpha = 5 \times 10^{-3}$.

Fig. 1 Performance comparison between DSGT, GSGT, CSG, DSG, EXTRA [52] and DLM [27] for on-line Ridge regression. For CSG, the plots show $\|x_k - x^*\|^2$. For the other methods, the plots show $\frac{1}{n} \sum_{i=1}^{n} \|x_{i,k} - x^*\|^2$.

methods are comparable in their speeds of convergence w.r.t. the number of gradient evaluations. However, GSGT is much faster than DSGT assuming the same number of communications. These numerical results verified our arguments in Sect. 4.2.

6 Conclusions and future work

This paper considers distributed multi-agent optimization over a network, where each agent only has access to inexact gradients of its local cost function. We propose a distributed stochastic gradient tracking method (DSGT) and show that the iterates obtained by each agent, using a constant stepsize value, reach a neighborhood of the optimum (in expectation) exponentially fast. More importantly, in a limit, the error bounds for the distances between the iterates and the optimal solution decrease in the network size, which is comparable with the performance of a centralized stochastic gradient algorithm. With a diminishing stepsize, the method exhibits the optimal $\mathcal{O}(1/k)$ rate of convergence. In the second part of this paper, we discuss a gossip-like stochastic gradient tracking method (GSGT) that is communication-efficient. Under a well-connected interaction graph, we show GSGT requires fewer communications than DSGT to reach an ϵ error level. Finally, we provide a numerical example that demonstrates the effectiveness of both algorithms. In our future work, we will deal with directed and/or time-varying interaction graphs among agents. We also plan to explore other more flexible randomized algorithms such as broadcast-based protocols with possible transmission failures.

7 Appendix

7.1 Proof of Lemma 4

By (4),

$$\overline{x}_{k+1} = \overline{x}_k - \alpha \overline{y}_k. \tag{67}$$

It follows that

$$\|\overline{x}_{k+1} - x^*\|^2 = \|\overline{x}_k - \alpha \overline{y}_k - x^*\|^2 = \|\overline{x}_k - x^*\|^2 - 2\alpha \langle \overline{x}_k - x^*, \overline{y}_k \rangle + \alpha^2 \|\overline{y}_k\|^2. \tag{68}$$

Notice that $\mathbb{E}[\overline{y}_k \mid \mathcal{F}_k] = h(\mathbf{x}_k)$, and

$$\mathbb{E}[\|\overline{y}_k\|^2 \mid \mathcal{F}_k] = \mathbb{E}[\|\overline{y}_k - h(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] + \|h(\mathbf{x}_k)\|^2.$$

We have

$$\begin{aligned}
\mathbb{E}[\|\overline{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] &= \|\overline{x}_k - x^*\|^2 - 2\alpha \langle \overline{x}_k - x^*, h(\mathbf{x}_k) \rangle \\
&\quad + \alpha^2 \mathbb{E}[\|\overline{y}_k - h(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] + \alpha^2 \|h(\mathbf{x}_k)\|^2 \\
&\leq \|\overline{x}_k - x^*\|^2 - 2\alpha \langle \overline{x}_k - x^*, h(\mathbf{x}_k) \rangle + \alpha^2 \|h(\mathbf{x}_k)\|^2 + \frac{\alpha^2 \sigma^2}{n},
\end{aligned} \tag{69}$$

where the inequality follows from Lemma 2. Denote $\lambda = 1 - \alpha\mu$. In light of Lemma 3,

$$
\begin{aligned}
\mathbb{E}&[\|\overline{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_k] \\
&\leq \|\overline{x}_k - x^*\|^2 - 2\alpha\langle \overline{x}_k - x^*, \nabla f(\overline{x}_k)\rangle + 2\alpha\langle \overline{x}_k - x^*, \nabla f(\overline{x}_k) - h(\mathbf{x}_k)\rangle \\
&\quad + \alpha^2\|\nabla f(\overline{x}_k) - h(\mathbf{x}_k)\|^2 + \alpha^2\|\nabla f(\overline{x}_k)\|^2 - 2\alpha^2\langle\nabla f(\overline{x}_k), \nabla f(\overline{x}_k) - h(\mathbf{x}_k)\rangle + \frac{\alpha^2\sigma^2}{n} \\
&= \|\overline{x}_k - \alpha\nabla f(\overline{x}_k) - x^*\|^2 + \alpha^2\|\nabla f(\overline{x}_k) - h(\mathbf{x}_k)\|^2 + \frac{\alpha^2\sigma^2}{n} \\
&\quad + 2\alpha\langle\overline{x}_k - \alpha\nabla f(\overline{x}_k) - x^*, \nabla f(\overline{x}_k) - h(\mathbf{x}_k)\rangle \\
&\leq \lambda^2\|\overline{x}_k - x^*\|^2 + 2\alpha\lambda\|\overline{x}_k - x^*\|\|\nabla f(\overline{x}_k) - h(\mathbf{x}_k)\| + \alpha^2\|\nabla f(\overline{x}_k) - h(\mathbf{x}_k)\|^2 + \frac{\alpha^2\sigma^2}{n} \\
&\leq \lambda^2\|\overline{x}_k - x^*\|^2 + \frac{2\alpha\lambda L}{\sqrt{n}}\|\overline{x}_k - x^*\|\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \frac{\alpha^2 L^2}{n}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{\alpha^2\sigma^2}{n} \\
&\leq \lambda^2\|\overline{x}_k - x^*\|^2 + \alpha\left(\lambda^2\mu\|\overline{x}_k - x^*\|^2 + \frac{L^2}{\mu n}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2\right) + \frac{\alpha^2 L^2}{n}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{\alpha^2\sigma^2}{n} \\
&= \lambda^2\left(1 + \alpha\mu\right)\|\overline{x}_k - x^*\|^2 + \frac{\alpha L^2}{\mu n}\left(1 + \alpha\mu\right)\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{\alpha^2\sigma^2}{n} \\
&\leq (1 - \alpha\mu)\|\overline{x}_k - x^*\|^2 + \frac{\alpha L^2}{\mu n}\left(1 + \alpha\mu\right)\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{\alpha^2\sigma^2}{n}.
\end{aligned}
$$

Relation (19) follows from the following argument:

$$
\begin{aligned}
\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2 &= \|W\mathbf{x}_k - \alpha W\mathbf{y}_k - \mathbf{1}\overline{x}_k + \alpha\mathbf{1}\overline{y}_k\|^2 \\
&\leq \|W\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 - 2\alpha\langle W\mathbf{x}_k - \mathbf{1}\overline{x}_k, W\mathbf{y}_k - \mathbf{1}\overline{y}_k\rangle + \alpha^2\|W\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \\
&\leq \rho_w^2\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \alpha\rho_w^2\left[\frac{(1 - \rho_w^2)}{2\alpha\rho_w^2}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{2\alpha\rho_w^2}{(1 - \rho_w^2)}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2\right] \\
&\quad + \alpha^2\rho_w^2\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \leq \frac{(1 + \rho_w^2)}{2}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \alpha^2\frac{(1 + \rho_w^2)\rho_w^2}{(1 - \rho_w^2)}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2,
\end{aligned}
$$

where we used Lemma 1.

To prove (20), we need some preparations first. For ease of exposition we will write $G_k := G(\mathbf{x}_k, \boldsymbol{\xi}_k)$ and $\nabla_k := \nabla F(\mathbf{x}_k)$ for short. From (5) and Lemma 1,

$$
\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2 &= \|W\mathbf{y}_k + G_{k+1} - G_k - \mathbf{1}\overline{y}_k + \mathbf{1}(\overline{y}_k - \overline{y}_{k+1})\|^2 \\
&= |W\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \|G_{k+1} - G_k\|^2 + n\|\overline{y}_k - \overline{y}_{k+1}\|^2 + 2\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, G_{k+1} - G_k\rangle \\
&\quad + 2\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, \mathbf{1}(\overline{y}_k - \overline{y}_{k+1})\rangle + 2\langle G_{k+1} - G_k, \mathbf{1}(\overline{y}_k - \overline{y}_{k+1})\rangle \\
&= \|W\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \|G_{k+1} - G_k\|^2 - n\|\overline{y}_k - \overline{y}_{k+1}\|^2 + 2\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, G_{k+1} - G_k\rangle \\
&\leq \rho_w^2\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \|G_{k+1} - G_k\|^2 + 2\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, G_{k+1} - G_k\rangle.
\end{aligned}
$$

Notice that

$$
\begin{aligned}
\mathbb{E}[\|G_{k+1} - G_k\|^2 \mid \mathcal{F}_k] &= \mathbb{E}[\|\nabla_{k+1} - \nabla_k\|^2 \mid \mathcal{F}_k] \\
&\quad + 2\mathbb{E}[\langle \nabla_{k+1} - \nabla_k, G_{k+1} - \nabla_{k+1} - G_k + \nabla_k \rangle \mid \mathcal{F}_k] \\
&\quad + \mathbb{E}[\|G_{k+1} - \nabla_{k+1} - G_k + \nabla_k\|^2 \mid \mathcal{F}_k] \\
&\le \mathbb{E}[\|\nabla_{k+1} - \nabla_k\|^2 \mid \mathcal{F}_k] + 2\mathbb{E}[\langle \nabla_{k+1}, -G_k + \nabla_k \rangle \mid \mathcal{F}_k] + 2n\sigma^2
\end{aligned}
$$

by Assumption 1, and

$$
\begin{aligned}
\mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, G_{k+1} - G_k \rangle \mid \mathcal{F}_k] &= \mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, \nabla_{k+1} - G_k \rangle \mid \mathcal{F}_k] \\
&= \mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, \nabla_{k+1} - \nabla_k \rangle \mid \mathcal{F}_k] + \mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, -G_k + \nabla_k \rangle \mid \mathcal{F}_k].
\end{aligned}
$$

We have

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2 \mid \mathcal{F}_k] &\le \rho_w^2 \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k] + \mathbb{E}[\|\nabla_{k+1} - \nabla_k\|^2 \mid \mathcal{F}_k] \\
&\quad + 2\mathbb{E}[\langle \nabla_{k+1}, -G_k + \nabla_k \rangle \mid \mathcal{F}_k] + 2\mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, \nabla_{k+1} - \nabla_k \rangle \mid \mathcal{F}_k] \\
&\quad + 2\mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, -G_k + \nabla_k \rangle \mid \mathcal{F}_k] + 2n\sigma^2.
\end{aligned} \tag{70}
$$

Two additional lemmas are in hand.

Lemma 8

$$
\mathbb{E}[\langle \nabla_{k+1}, -G_k + \nabla_k \rangle \mid \mathcal{F}_k] \le \alpha L n \sigma^2.
$$

Proof From (4),

$$
\begin{aligned}
\nabla f_i(x_{i,k+1}) &= \nabla f_i \left(\sum_{j=1}^{n} w_{ij}(x_{j,k} - \alpha y_{j,k}) \right) \\
&= \nabla f_i \left(\sum_{j=1}^{n} w_{ij} x_{j,k} - \alpha \sum_{j=1}^{n} w_{ij} \left[\sum_{l=1}^{n} w_{jl} y_{l,k-1} + g_j(x_{j,k}, \xi_{j,k}) - g_j(x_{j,k-1}, \xi_{j,k-1}) \right] \right) \\
&= \nabla f_i \left(\sum_{j=1}^{n} w_{ij} x_{j,k} - \alpha \sum_{j=1}^{n} w_{ij} \sum_{l=1}^{n} w_{jl} y_{l,k-1} - \alpha \sum_{j=1}^{n} w_{ij} g_j(x_{j,k}, \xi_{j,k}) \right. \\
&\quad \left. + \alpha \sum_{j=1}^{n} w_{ij} g_j(x_{j,k-1}, \xi_{j,k-1}) \right).
\end{aligned}
$$

In light of Assumption 2,

$$
\left\| \nabla f_i(x_{i,k+1}) - \nabla f_i \left(\sum_{j=1}^n w_{ij} x_{j,k} - \alpha \sum_{j=1}^n w_{ij} \sum_{l=1}^n w_{jl} y_{l,k-1} - \alpha \sum_{j \neq i}^n w_{ij} g_j(x_{j,k}, \xi_{j,k}) \right. \right.
$$
$$
\left. \left. - \alpha w_{ii} \nabla f_i(x_{i,k}) + \alpha \sum_{j=1}^n w_{ij} g_j(x_{j,k-1}, \xi_{j,k-1}) \right) \right\| \leq \alpha L \| g_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k}) \|.
$$
(71)

Then,

$$
\mathbb{E}[\langle \nabla f_i(x_{i,k+1}), -g_i(x_{i,k}, \xi_{i,k}) + \nabla f_i(x_{i,k}) \rangle \mid \mathcal{F}_k]
$$
$$
\leq \alpha L \mathbb{E}[\| g_i(x_{i,k}, \xi_{i,k}) - \nabla f_i(x_{i,k}) \|^2 \mid \mathcal{F}_k] \leq \alpha L \sigma^2.
$$
(72)

The desired result then follows. □

Lemma 9

$$
\mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, -G_k + \nabla_k \rangle \mid \mathcal{F}_k] \leq \sigma^2.
$$
(73)

Proof By (4), we have

$$
\mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, -G_k + \nabla_k \rangle \mid \mathcal{F}_k]
$$
$$
= \sum_{i=1}^n \mathbb{E}\left[\left\langle \sum_{j=1}^n w_{ij} y_{j,k} - \overline{y}_k, \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \right\rangle \middle| \mathcal{F}_k \right].
$$

On one hand,

$$
\mathbb{E}[\langle y_{j,k}, \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \rangle \mid \mathcal{F}_k]
$$
$$
= \mathbb{E}\left[\left\langle \sum_{n=1}^n w_{jn} y_{n,k-1} + g_j(x_{j,k}, \xi_{j,k}) - g_j(x_{j,k-1}, \xi_{j,k-1}), \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \right\rangle \middle| \mathcal{F}_k \right]
$$
$$
= \mathbb{E}\left[\langle g_j(x_{j,k}, \xi_{j,k}), \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \rangle \middle| \mathcal{F}_k \right],
$$

which gives

$$
\mathbb{E}\left[\left\langle \sum_{j=1}^n w_{ij} y_{j,k}, \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \right\rangle \middle| \mathcal{F}_k \right]
$$
$$
= \mathbb{E}[\langle w_{ii} g_i(x_{i,k}, \xi_{i,k}), \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \rangle \mid \mathcal{F}_k] \leq 0.
$$

On the other hand,

$$\mathbb{E}[\langle \overline{y}_k, \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \rangle \mid \mathcal{F}_k]$$

$$= \mathbb{E}\left[\left\langle \frac{1}{n}\sum_{j=1}^{n} g_j(x_{j,k}, \xi_{j,k}), \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \right\rangle \middle| \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[\left\langle \frac{1}{n} g_i(x_{i,k}, \xi_{i,k}), \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \right\rangle \middle| \mathcal{F}_k\right].$$

We have

$$\mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, -G_k + \nabla_k \rangle \mid \mathcal{F}_k]$$

$$\leq -\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\langle g_i(x_{i,k}, \xi_{i,k}), \nabla f_i(x_{i,k}) - g_i(x_{i,k}, \xi_{i,k}) \rangle \mid \mathcal{F}_k] \leq \sigma^2. \quad (74)$$

$\square$

By (70), Lemma 8 and Lemma 9, we obtain

$$\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2 \mid \mathcal{F}_k] \leq \rho_w^2 \mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k] + \mathbb{E}[\|\nabla_{k+1} - \nabla_k\|^2 \mid \mathcal{F}_k]$$

$$+ 2\mathbb{E}[\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, \nabla_{k+1} - \nabla_k \rangle \mid \mathcal{F}_k] + 2(n + \alpha L n + 1)\sigma^2. \quad (75)$$

Now we bound $\|\nabla_{k+1} - \nabla_k\|^2$ and $\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, \nabla_{k+1} - \nabla_k \rangle$. First, by Assumption 2 and Lemma 1,

$$\|\nabla_{k+1} - \nabla_k\|^2 \leq L^2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 = L^2 \|W\mathbf{x}_k - \mathbf{x}_k - \alpha W\mathbf{y}_k\|^2$$

$$= L^2 \|(\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \mathbf{1}\overline{x}_k) - \alpha W\mathbf{y}_k\|^2$$

$$= \|\mathbf{W} - \mathbf{I}\|^2 L^2 \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 - 2\alpha L^2 \langle (\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \mathbf{1}\overline{x}_k), W\mathbf{y}_k \rangle + \alpha^2 L^2 \|W\mathbf{y}_k\|^2$$

$$= \|\mathbf{W} - \mathbf{I}\|^2 L^2 \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 - 2\alpha L^2 \langle (\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \mathbf{1}\overline{x}_k), W\mathbf{y}_k - \mathbf{1}\overline{y}_k \rangle$$

$$+ \alpha^2 L^2 \|W\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \alpha^2 n L^2 \|\overline{y}_k\|^2$$

$$\leq \|\mathbf{W} - \mathbf{I}\|^2 L^2 \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + 2\alpha \|\mathbf{W} - \mathbf{I}\|L^2 \rho_w \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|$$

$$+ \alpha^2 L^2 \rho_w^2 \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \alpha^2 n L^2 \|\overline{y}_k\|^2$$

$$\leq 2\|\mathbf{W} - \mathbf{I}\|^2 L^2 \|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + 2\alpha^2 L^2 \rho_w^2 \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \alpha^2 n L^2 \|\overline{y}_k\|^2.$$

Second,

$$\langle W\mathbf{y}_k - \mathbf{1}\overline{y}_k, \nabla_{k+1} - \nabla_k \rangle \leq L\rho_w \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|\|(\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \mathbf{1}\overline{x}_k) - \alpha W\mathbf{y}_k\|$$

$$\leq \|\mathbf{W} - \mathbf{I}\|L\rho_w \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \alpha L\rho_w \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|\|W\mathbf{y}_k - \mathbf{1}\overline{y}_k + \mathbf{1}\overline{y}_k\|$$

$$\leq \|\mathbf{W} - \mathbf{I}\|L\rho_w \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \alpha L\rho_w^2 \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \alpha\sqrt{n}L\rho_w \|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|\|\overline{y}_k\|.$$

Notice that

$$\|\overline{y}_k\| \leq \|\overline{y}_k - h(\mathbf{x}_k)\| + \|h(\mathbf{x}_k) - \nabla f(\overline{x}_k)\| + \|\nabla f(\overline{x}_k)\|$$
$$\leq \|\overline{y}_k - h(\mathbf{x}_k)\| + \frac{L}{\sqrt{n}}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + L\|\overline{x}_k - x^*\|.$$

We have

$$\sqrt{n}L\rho_w\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|\|\overline{y}_k\| \leq \sqrt{n}L\rho_w\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|$$
$$\left(\|\overline{y}_k - h(\mathbf{x}_k)\| + \frac{L}{\sqrt{n}}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + L\|\overline{x}_k - x^*\|\right)$$
$$\leq L\rho_w^2\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + nL\|\overline{y}_k - h(\mathbf{x}_k)\|^2 + L^3\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{1}{2}nL^3\|\overline{x}_k - x^*\|^2,$$

and

$$\|\overline{y}_k\|^2 \leq 3\|\overline{y}_k - h(\mathbf{x}_k)\|^2 + \frac{3L^2}{n}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + 3L^2\|\overline{x}_k - x^*\|^2.$$

By (75) and the above relations,

$$\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2 \mid \mathcal{F}_k] \leq \rho_w^2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k]$$
$$+ 2\|\mathbf{W} - \mathbf{I}\|^2 L^2\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + 2\alpha^2 L^2\rho_w^2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k]$$
$$+ \alpha^2 n L^2\left(3\mathbb{E}[\|\overline{y}_k - h(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] + \frac{3L^2}{n}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + 3L^2\|\overline{x}_k - x^*\|^2\right)$$
$$+ 2\left(\|\mathbf{W} - \mathbf{I}\|L\rho_w\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + \alpha L\rho_w^2\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2\right)$$
$$+ 2\left(\alpha L\rho_w^2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k] + \alpha n L\mathbb{E}[\|\overline{y}_k - h(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k]\right.$$
$$+ \alpha L^3\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{1}{2}\alpha n L^3\|\overline{x}_k - x^*\|^2\right)$$
$$+ 2(n + \alpha L n + 1)\sigma^2 n$$
$$\leq \left(\rho_w^2 + 4\alpha L\rho_w^2 + 2\alpha^2 L^2\rho_w^2\right)\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k]$$
$$+ \left(\beta\rho_w^2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k] + \frac{1}{\beta}\|\mathbf{W} - \mathbf{I}\|^2 L^2\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2\right)$$
$$+ \left(2\|\mathbf{W}-\mathbf{I}\|^2 L^2 + 2\alpha L^3 + 3\alpha^2 L^4\right)\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \left(3\alpha^2 n L^4 + \alpha n L^3\right)\|\overline{x}_k - x^*\|^2$$
$$+ \left[3\alpha^2 L^2 + 2\alpha L + 2(n + \alpha L n + 1)\right]\sigma^2$$
$$= \left(1 + 4\alpha L + 2\alpha^2 L^2 + \beta\right)\rho_w^2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k]$$

$$+ \left(\frac{1}{\beta} \|\mathbf{W} - \mathbf{I}\|^2 L^2 + 2\|\mathbf{W} - \mathbf{I}\|^2 L^2 + 3\alpha L^3 \right) \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2$$
$$+ 2\alpha n L^3 \|\bar{x}_k - x^*\|^2 + M_\sigma \tag{76}$$

for any $\beta > 0$.

7.2 Proof of Lemma 5

The characteristic function of $\mathbf{S}$ is given by

$$g(\lambda) := \det(\lambda \mathbf{I} - \mathbf{S}) = (\lambda - s_{11})(\lambda - s_{22})(\lambda - s_{33}) - a_{23}a_{32}(\lambda - s_{11}) - a_{13}a_{31}(\lambda - s_{22})$$
$$- a_{12}a_{21}(\lambda - s_{33}) - a_{12}a_{23}a_{31} - a_{13}a_{32}a_{21}. \tag{77}$$

Necessity is trivial since $\det(\lambda^* \mathbf{I} - \mathbf{S}) \leq 0$ implies $g(\lambda) = 0$ for some $\lambda \geq \lambda^*$. We now show $\det(\lambda^* \mathbf{I} - \mathbf{S}) > 0$ is also a sufficient condition. Given that $g(\lambda^*) = \det(\lambda^* \mathbf{I} - \mathbf{S}) > 0$,

$$(\lambda^* - s_{11})(\lambda^* - s_{22})(\lambda^* - s_{33}) > a_{23}a_{32}(\lambda^* - s_{11}) + a_{13}a_{31}(\lambda^* - s_{22}) + a_{12}a_{21}(\lambda^* - s_{33}).$$

It follows that

$$\begin{aligned} \gamma_1(\lambda^* - s_{22})(\lambda^* - s_{33}) &> a_{23}a_{32} \\ \gamma_2(\lambda^* - s_{11})(\lambda^* - s_{33}) &> a_{13}a_{31} \\ \gamma_3(\lambda^* - s_{11})(\lambda^* - s_{22}) &> a_{12}a_{21} \end{aligned} \tag{78}$$

for some $\gamma_1, \gamma_2, \gamma_3 > 0$ with $\gamma_1 + \gamma_2 + \gamma_3 \leq 1$. Consider

$$g'(\lambda) = (\lambda - s_{22})(\lambda - s_{33}) + (\lambda - s_{11})(\lambda - s_{33}) + (\lambda - s_{11})(\lambda - s_{22}) - a_{23}a_{32} - a_{13}a_{31} - a_{12}a_{21}.$$

We have $g'(\lambda) > 0$ for $\lambda \in (-\infty, -\lambda^*] \cup [\lambda^*, +\infty)$. Notice that

$$g(-\lambda^*) \leq -(\lambda^* + s_{11})(\lambda^* + s_{22})(\lambda^* + s_{33}) + a_{23}a_{32}(1 + s_{11})$$
$$+ a_{13}a_{31}(\lambda^* + s_{22}) + a_{12}a_{21}(\lambda^* + s_{33}) < 0.$$

All the real roots of $g(\lambda) = 0$ lie in the interval $(-\lambda^*, \lambda^*)$. By the Perron-Frobenius theorem, $\rho(\mathbf{S}) \in \mathbb{R}$ is an eigenvalue of $\mathbf{S}$. We conclude that $\rho(\mathbf{S}) < \lambda^*$.

7.3 Proof of Lemma 7

First we prove relation (50). In light of (39a), we have

$$\bar{x}_{k+1} = \frac{1}{n} \left[(x_{i_k,k} + x_{j_k,k}) - \alpha(y_{i_k,k} + y_{j_k,k}) \right] + \frac{1}{n} \sum_{i \neq \{i_k, j_k\}} x_{i,k} = \bar{x}_k - \frac{\alpha}{n}(y_{i_k,k} + y_{j_k,k}). \tag{79}$$

Then,

$$\|\overline{x}_{k+1} - x^*\|^2 = \left\|\overline{x}_k - \frac{\alpha}{n}(y_{i_k,k} + y_{j_k,k}) - x^*\right\|^2$$

$$= \|\overline{x}_k - x^*\|^2 - \frac{2\alpha}{n}\langle \overline{x}_k - x^*, y_{i_k,k} + y_{j_k,k}\rangle + \frac{\alpha^2}{n^2}\|y_{i_k,k} + y_{j_k,k}\|^2.$$

Taking conditional expectations of $(y_{i_k,k} + y_{j_k,k})$ and $\|y_{i_k,k} + y_{j_k,k}\|^2$ w.r.t. the random selections of i_k and j_k, we get

$$\mathbb{E}[y_{i_k,k} + y_{j_k,k} \mid \mathcal{F}_{k+1}] = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,k} + \sum_{j=1}^{n}\pi_{ij}y_{j,k}\right) = 2\overline{y}_k, \qquad (80)$$

and

$$\mathbb{E}[\|y_{i_k,k} + y_{j_k,k}\|^2 \mid \mathcal{F}_{k+1}] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{ij}\|y_{i,k} + y_{j,k}\|^2$$

$$= \frac{2}{n}\sum_{i=1}^{n}\|y_{i,k}\|^2 + \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{ij}\langle y_{i,k}, y_{j,k}\rangle$$

$$= \frac{2}{n}\|\mathbf{y}_k\|^2 + \frac{2}{n}\mathbf{y}_k^\mathsf{T}\mathbf{\Pi}\mathbf{y}_k = \frac{2}{n}\|\mathbf{y}_k\|^2 + \frac{2}{n}(\mathbf{y}_k - \mathbf{1}\overline{y}_k)^\mathsf{T}\mathbf{\Pi}(\mathbf{y}_k - \mathbf{1}\overline{y}_k) + 2\|\overline{y}_k\|^2$$

$$\leq \frac{2}{n}(1 + \rho_\pi)\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + 4\|\overline{y}_k\|^2$$

$$\leq \frac{4}{n}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + 4\|\overline{y}_k\|^2, \qquad (81)$$

where $\rho_\pi < 1$ denotes the spectral norm of $\mathbf{\Pi} - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T}$. It follows that

$$\mathbb{E}[\|\overline{x}_{k+1} - x^*\|^2 \mid \mathcal{F}_{k+1}] \leq \|\overline{x}_k - x^*\|^2 - \frac{4\alpha}{n}\langle \overline{x}_k - x^*, \overline{y}_k\rangle + \frac{4\alpha^2}{n^2}\|\overline{y}_k\|^2 + \frac{4\alpha^2}{n^3}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2.$$
$$(82)$$

Noticing that $\mathbb{E}[\overline{y}_k] = \mathbb{E}[h(\mathbf{x}_k)]$, from Lemma 2 we have

$$\mathbb{E}[\|\overline{y}_k\|^2] = \mathbb{E}[\|\overline{y}_k - h(\mathbf{x}_k)\|^2 + 2\langle \overline{y}_k - h(\mathbf{x}_k), h(\mathbf{x}_k)\rangle + \|h(\mathbf{x}_k)\|^2 \mid \mathcal{F}_k] \leq \frac{\sigma^2}{n} + \mathbb{E}[\|h(\mathbf{x}_k)\|^2].$$
$$(83)$$

Then from (82) we obtain

$$\mathbb{E}[\|\overline{x}_{k+1} - x^*\|^2]$$

$$\leq \mathbb{E}[\|\overline{x}_k - x^*\|^2] - \frac{4\alpha}{n}\mathbb{E}[\langle \overline{x}_k - x^*, h(\mathbf{x}_k)\rangle] + \frac{4\alpha^2}{n^2}\mathbb{E}[\|h(\mathbf{x}_k)\|^2] + \frac{4\alpha^2\sigma^2}{n^3} + \frac{4\alpha^2}{n^3}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2]$$

$$= \mathbb{E}[\|\bar{x}_k - x^*\|^2] - \frac{4\alpha}{n}\mathbb{E}[\langle \bar{x}_k - x^*, \nabla f(\bar{x}_k)\rangle] - \frac{4\alpha}{n}\mathbb{E}[\langle \bar{x}_k - x^*, h(\mathbf{x}_k) - \nabla f(\bar{x}_k)\rangle]$$

$$+ \frac{4\alpha^2}{n^2}\mathbb{E}[\|h(\mathbf{x}_k) - \nabla f(\bar{x}_k)\|^2] + \frac{4\alpha^2}{n^2}\mathbb{E}[\|\nabla f(\bar{x}_k)\|^2] + \frac{8\alpha^2}{n^2}\mathbb{E}[\langle h(\mathbf{x}_k) - \nabla f(\bar{x}_k), \nabla f(\bar{x}_k)\rangle]$$

$$+ \frac{4\alpha^2}{n^3}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \frac{4\alpha^2\sigma^2}{n^3}$$

$$= \mathbb{E}\left[\left\|\bar{x}_k - \frac{2\alpha}{n}\nabla f(\bar{x}_k) - x^*\right\|^2\right] - \frac{4\alpha}{n}\mathbb{E}[\langle \bar{x}_k - \frac{2\alpha}{n}\nabla f(\bar{x}_k) - x^*, h(\mathbf{x}_k) - \nabla f(\bar{x}_k)\rangle]$$

$$+ \frac{4\alpha^2}{n^2}\mathbb{E}[\|h(\mathbf{x}_k) - \nabla f(\bar{x}_k)\|^2] + \frac{4\alpha^2}{n^3}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2 \mid \mathcal{F}_k^-] + \frac{4\alpha^2\sigma^2}{n^3}.$$

Since $\alpha < n/(\mu + L)$, we know from Lemma 3 that

$$\left\|\bar{x}_k - \frac{2\alpha}{n}\nabla f(\bar{x}_k) - x^*\right\|^2 \leq \left(1 - \frac{2\alpha\mu}{n}\right)^2 \|\bar{x}_k - x^*\|^2,$$

and

$$\|h(\mathbf{x}_k) - \nabla f(\bar{x}_k)\| \leq \frac{L}{\sqrt{n}}\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|.$$

It follows that

$$\mathbb{E}[\|\bar{x}_{k+1} - x^*\|^2]$$

$$\leq \left(1 - \frac{2\alpha\mu}{n}\right)^2 \mathbb{E}[\|\bar{x}_k - x^*\|^2] + \frac{4\alpha}{n}\left(1 - \frac{2\alpha\mu}{n}\right)\frac{L}{\sqrt{n}}\mathbb{E}[\|\bar{x}_k - x^*\| \|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|]$$

$$+ \frac{4\alpha^2 L^2}{n^3}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] + \frac{4\alpha^2}{n^3}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \frac{4\alpha^2\sigma^2}{n^3}$$

$$\leq \left(1 - \frac{2\alpha\mu}{n}\right)^2 \mathbb{E}[\|\bar{x}_k - x^*\|^2] + \frac{2\alpha}{n}\left(\left(1 - \frac{2\alpha\mu}{n}\right)^2 \mu\mathbb{E}[\|\bar{x}_k - x^*\|^2] + \frac{L^2}{\mu n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]\right)$$

$$+ \frac{4\alpha^2 L^2}{n^3}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2] + \frac{4\alpha^2}{n^3}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \frac{4\alpha^2\sigma^2}{n^3}$$

$$\leq \left(1 - \frac{2\alpha\mu}{n}\right)\mathbb{E}[\|\bar{x}_k - x^*\|^2] + \frac{2\alpha L^2}{\mu n^2}\left(1 + \frac{2\alpha\mu}{n}\right)\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2]$$

$$+ \frac{4\alpha^2}{n^3}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\bar{y}_k\|^2] + \frac{4\alpha^2\sigma^2}{n^3}.$$

To bound the consensus error $\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2]$, note that from (39a) and (79) we have

$$\|\mathbf{x}_{k+1} - \mathbf{1}\bar{x}_{k+1}\|^2 = \left\|\mathbf{W}_k\mathbf{x}_k - \alpha\mathbf{D}_k\mathbf{y}_k - \mathbf{1}\bar{x}_k + \frac{\alpha}{n}\mathbf{1}(y_{i_k,k} + y_{j_k,k})\right\|^2$$

$$= \|\mathbf{W}_k\mathbf{x}_k - \mathbf{1}\bar{x}_k\|^2 - 2\alpha\langle \mathbf{W}_k\mathbf{x}_k - \mathbf{1}\bar{x}_k, \mathbf{D}_k\mathbf{y}_k - \frac{1}{n}\mathbf{1}(y_{i_k,k} + y_{j_k,k})\rangle$$

$$+ \alpha^2\left\|\mathbf{D}_k\mathbf{y}_k - \frac{1}{n}\mathbf{1}(y_{i_k,k} + y_{j_k,k})\right\|^2$$

$$= \mathbf{tr}\left[(\mathbf{x}_k - \mathbf{1}\overline{x}_k)^\mathsf{T}\mathbf{W}_k^\mathsf{T}\mathbf{W}_k(\mathbf{x}_k - \mathbf{1}\overline{x}_k)\right] - 2\alpha\langle\mathbf{W}_k\mathbf{x}_k - \mathbf{1}\overline{x}_k, \mathbf{D}_k\mathbf{y}_k\rangle$$

$$+ \alpha^2\left\|\mathbf{D}_k\mathbf{y}_k - \frac{1}{n}\mathbf{1}(y_{i_k,k} + y_{j_k,k})\right\|^2. \tag{84}$$

By Lemma 6, the conditional expectation of $\mathbf{tr}\left[(\mathbf{x}_k - \mathbf{1}\overline{x}_k)^\mathsf{T}\mathbf{W}_k^\mathsf{T}\mathbf{W}_k(\mathbf{x}_k - \mathbf{1}\overline{x}_k)\right]$ can be bounded below,

$$\mathbb{E}[\mathbf{tr}[(\mathbf{x}_k - \mathbf{1}\overline{x}_k)^\mathsf{T}\mathbf{W}_k^\mathsf{T}\mathbf{W}_k(\mathbf{x}_k - \mathbf{1}\overline{x}_k)] \mid \mathcal{F}_{k+1}] = \mathbf{tr}\left[(\mathbf{x}_k - \mathbf{1}\overline{x}_k)^\mathsf{T}\tilde{\mathbf{W}}(\mathbf{x}_k - \mathbf{1}\overline{x}_k)\right]$$

$$\leq \rho_{\overline{w}}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2. \tag{85}$$

In view of the structure of $\mathbf{D}_k$, we rewrite $2\langle\mathbf{W}_k\mathbf{x}_k - \mathbf{1}\overline{x}_k, \mathbf{D}_k\mathbf{y}_k\rangle$ as follows:

$$2\langle\mathbf{W}_k\mathbf{x}_k - \mathbf{1}\overline{x}_k, \mathbf{D}_k\mathbf{y}_k\rangle$$
$$= 2\left\langle\frac{1}{2}(x_{i_k,k} + x_{j_k,k}) - \overline{x}_k, y_{i_k,k} + y_{j_k,k}\right\rangle = \langle x_{i_k,k} + x_{j_k,k} - 2\overline{x}_k, y_{i_k,k} + y_{j_k,k}\rangle.$$

Note that

$$\mathbb{E}[\langle x_{i_k,k} - \overline{x}_k, y_{i_k,k}\rangle \mid \mathcal{F}_{k+1}] = \frac{1}{n}\sum_{i=1}^{n}\langle x_{i,k} - \overline{x}_k, y_{i,k}\rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n}\langle x_{i,k} - \overline{x}_k, y_{i,k} - \overline{y}_k\rangle = \frac{1}{n}\langle\mathbf{x}_k - \mathbf{1}\overline{x}_k, \mathbf{y}_k - \mathbf{1}\overline{y}_k\rangle,$$

$$\mathbb{E}[\langle x_{j_k,k} - \overline{x}_k, y_{j_k,k}\rangle \mid \mathcal{F}_{k+1}] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{ij}\langle x_{j,k} - \overline{x}_k, y_{j,k}\rangle$$

$$= \frac{1}{n}\sum_{j=1}^{n}\langle x_{j,k} - \overline{x}_k, y_{j,k}\rangle = \frac{1}{n}\langle\mathbf{x}_k - \mathbf{1}\overline{x}_k, \mathbf{y}_k - \mathbf{1}\overline{y}_k\rangle,$$

$$\mathbb{E}[\langle x_{i_k,k} - \overline{x}_k, y_{j_k,k}\rangle \mid \mathcal{F}_{k+1}] = \frac{1}{n}\sum_{i=1}^{n}\langle x_{i,k} - \overline{x}_k, \sum_{j=1}^{n}\pi_{ij}y_{j,k}\rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle x_{i,k} - \overline{x}_k, \pi_{ij}(y_{j,k} - \overline{y}_k)\rangle$$

$$= \frac{1}{n}\mathbf{tr}\left[(\mathbf{x}_k - \mathbf{1}\overline{x}_k)^\mathsf{T}\mathbf{\Pi}(\mathbf{y}_k - \mathbf{1}\overline{y}_k)\right],$$

and similarly,

$$\mathbb{E}[\langle x_{j_k,k} - \overline{x}_k, y_{i_k,k}\rangle \mid \mathcal{F}_{k+1}] = \frac{1}{n}\mathbf{tr}\left[(\mathbf{x}_k - \mathbf{1}\overline{x}_k)^\mathsf{T}\mathbf{\Pi}(\mathbf{y}_k - \mathbf{1}\overline{y}_k)\right].$$

The following inequality holds:

$$2\mathbb{E}[\langle \mathbf{W}_k\mathbf{x}_k - \mathbf{1}\overline{x}_k, \mathbf{D}_k\mathbf{y}_k\rangle \mid \mathcal{F}_{k+1}] = \frac{2}{n}\langle \mathbf{x}_k - \mathbf{1}\overline{x}_k, \mathbf{y}_k - \mathbf{1}\overline{y}_k\rangle + \frac{2}{n}\mathbf{tr}\left[(\mathbf{x}_k - \mathbf{1}\overline{x}_k)^{\mathsf{T}}\mathbf{\Pi}(\mathbf{y}_k - \mathbf{1}\overline{y}_k)\right]$$

$$\leq \frac{4}{n}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|. \tag{86}$$

The last term in Eq. (84) can be bounded in the following way:

$$\left\|\mathbf{D}_k\mathbf{y}_k - \frac{1}{n}\mathbf{1}(y_{i_k,k} + y_{j_k,k})\right\|^2 = \|\mathbf{D}_k\mathbf{y}_k\|^2 - \frac{2}{n}\langle \mathbf{D}_k\mathbf{y}_k, \mathbf{1}(y_{i_k,k} + y_{j_k,k})\rangle + \frac{1}{n}\|y_{i_k,k} + y_{j_k,k}\|^2$$

$$\leq 2(\|y_{i_k,k}\|^2 + \|y_{j_k,k}\|^2) - \frac{2}{n}\|y_{i_k,k} + y_{j_k,k}\|^2 + \frac{1}{n}\|y_{i_k,k} + y_{j_k,k}\|^2$$

$$\leq 2(\|y_{i_k,k}\|^2 + \|y_{j_k,k}\|^2). \tag{87}$$

In view of inequalities (85)–(87), from (84) we obtain

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2 \mid \mathcal{F}_{k+1}] \leq \rho_{\overline{w}}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{4\alpha}{n}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\| + \frac{4\alpha^2}{n}\|\mathbf{y}_k\|^2$$

$$\leq \rho_{\overline{w}}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{2\alpha}{n}\left(\beta_1\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{1}{\beta_1}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2\right) + \frac{4\alpha^2}{n}\|\mathbf{y}_k\|^2$$

$$= \left(\rho_{\overline{w}} + \frac{2\alpha}{n}\beta_1\right)\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{2\alpha}{n}\frac{1}{\beta_1}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \frac{4\alpha^2}{n}\left(\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + n\|\overline{y}_k\|^2\right)$$

$$= \left(\rho_{\overline{w}} + \frac{2\alpha}{n}\beta_1\right)\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \frac{2\alpha}{n}\left(\frac{1}{\beta_1} + 2\alpha\right)\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + 4\alpha^2\|\overline{y}_k\|^2,$$

where $\beta_1 > 0$ is arbitrary. Notice that by relation (83) and Lemma 3,

$$\mathbb{E}[\|\overline{y}_k\|^2] \leq \frac{\sigma^2}{n} + \mathbb{E}[\|h(\mathbf{x}_k) - \nabla f(\overline{x}_k) + \nabla f(\overline{x}_k)\|^2]$$

$$\leq \frac{\sigma^2}{n} + \mathbb{E}\left[\left(\frac{L}{\sqrt{n}}\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\| + L\|\overline{x}_k - x^*\|\right)^2\right]$$

$$\leq \frac{\sigma^2}{n} + \frac{2L^2}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] + 2L^2\mathbb{E}[\|\overline{x}_k - x^*\|^2]. \tag{88}$$

We obtain

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{1}\overline{x}_{k+1}\|^2] \leq \left(\rho_{\overline{w}} + \frac{2\alpha}{n}\beta_1 + \frac{8\alpha^2 L^2}{n}\right)\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2]$$

$$+ \frac{2\alpha}{n}\left(\frac{1}{\beta_1} + \alpha\right)\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2]$$

$$+ 8\alpha^2 L^2\mathbb{E}[\|\overline{x}_k - x^*\|^2] + \frac{4\alpha^2\sigma^2}{n},$$

which is exactly inequality (51).

Finally we prove inequality (52). From the update rule (39b),

$$
\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2 &= \|\mathbf{W}_k\mathbf{y}_k + \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k - \mathbf{1}\overline{y}_k + \mathbf{1}\overline{y}_k - \mathbf{1}\overline{y}_{k+1}\|^2 \\
&= \|\mathbf{W}_k\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \|\tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\|^2 + n\|\overline{y}_k - \overline{y}_{k+1}\|^2 \\
&\quad + 2\langle \mathbf{W}_k\mathbf{y}_k - \mathbf{1}\overline{y}_k, \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\rangle \\
&\quad + 2\langle \mathbf{W}_k\mathbf{y}_k - \mathbf{1}\overline{y}_k, \mathbf{1}\overline{y}_k - \mathbf{1}\overline{y}_{k+1}\rangle + 2\langle \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k, \mathbf{1}\overline{y}_k - \mathbf{1}\overline{y}_{k+1}\rangle \\
&= \mathbf{tr}\left[(\mathbf{y}_k - \mathbf{1}\overline{y}_k)^{\mathsf{T}}\mathbf{W}_k^{\mathsf{T}}\mathbf{W}_k(\mathbf{y}_k - \mathbf{1}\overline{y}_k)\right] + \|\tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\|^2 + n\|\overline{y}_k - \overline{y}_{k+1}\|^2 \\
&\quad + 2\langle \mathbf{W}_k\mathbf{y}_k - \mathbf{1}\overline{y}_k, \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\rangle + 2\langle \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k, \mathbf{1}\overline{y}_k - \mathbf{1}\overline{y}_{k+1}\rangle.
\end{aligned}
\tag{89}
$$

Denote $g_{i,k} := g_i(x_{i,k}, \xi_{i,k})$ for short. Notice that

$$
2\langle \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k, \mathbf{1}\overline{y}_k - \mathbf{1}\overline{y}_{k+1}\rangle = 2\langle g_{i_k,k+1} - g_{i_k,k} + \mathbb{1}_k(g_{j_k,k+1} - g_{j_k,k}), \overline{y}_k - \overline{y}_{k+1}\rangle,
$$

and

$$
\begin{aligned}
\overline{y}_k - \overline{y}_{k+1} &= \frac{1}{n}\left(y_{i_k,k} - y_{i_k,k+1} + \mathbb{1}_k(y_{j_k,k} - y_{j_k,k+1})\right) \\
&= -\frac{1}{n}\left(g_{i_k,k+1} - g_{i_k,k} + \mathbb{1}_k(g_{j_k,k+1} - g_{j_k,k})\right).
\end{aligned}
$$

We have

$$
2\langle \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k, \mathbf{1}\overline{y}_k - \mathbf{1}\overline{y}_{k+1}\rangle = -2n\|\overline{y}_k - \overline{y}_{k+1}\|^2.
$$

Hence relation (89) leads to

$$
\begin{aligned}
\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2 &\le \mathbf{tr}\left[(\mathbf{y}_k - \mathbf{1}\overline{y}_k)^{\mathsf{T}}\mathbf{W}_k^{\mathsf{T}}\mathbf{W}_k(\mathbf{y}_k - \mathbf{1}\overline{y}_k)\right] + \|\tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\|^2 \\
&\quad + 2\langle \mathbf{W}_k\mathbf{y}_k - \mathbf{1}\overline{y}_k, \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\rangle.
\end{aligned}
\tag{90}
$$

Now we analyze the three terms on the right-hand side of (90), respectively. First,

$$
\mathbb{E}[\mathbf{tr}[(\mathbf{y}_k - \mathbf{1}\overline{y}_k)^{\mathsf{T}}\mathbf{W}_k^{\mathsf{T}}\mathbf{W}_k(\mathbf{y}_k - \mathbf{1}\overline{y}_k)] \mid \mathcal{F}_{k+1}] \le \rho_{\tilde{w}}\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2.
\tag{91}
$$

Second,

$$
\|\tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\|^2 = \|g_{i_k,k+1} - g_{i_k,k}\|^2 + \mathbb{1}_k\|g_{j_k,k+1} - g_{j_k,k}\|^2.
\tag{92}
$$

In light of Assumptions 1 and 2, we can bound $\mathbb{E}[\|g_{i_k,k+1} - g_{i_k,k}\|^2]$ and $\mathbb{E}[\mathbb{1}_k\|g_{j_k,k+1} - g_{j_k,k}\|^2]$:

$$
\mathbb{E}[\|g_{i_k,k+1} - g_{i_k,k}\|^2] = \mathbb{E}[\|\nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k})\|^2]
$$

$$+ 2\mathbb{E}[\langle \nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k}), g_{i_k,k+1} - \nabla f_{i_k}(x_{i_k,k+1}) - g_{i_k,k} + \nabla f_{i_k}(x_{i_k,k})\rangle]$$
$$+ \mathbb{E}[\|g_{i_k,k+1} - \nabla f_{i_k}(x_{i_k,k+1}) - g_{i_k,k} + \nabla f_{i_k}(x_{i_k,k})\|^2]$$
$$= \mathbb{E}[\|\nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k})\|^2] + 2\mathbb{E}[\langle \nabla f_{i_k}(x_{i_k,k+1}), -g_{i_k,k} + \nabla f_{i_k}(x_{i_k,k})\rangle]$$
$$+ \mathbb{E}[\|g_{i_k,k+1} - \nabla f_{i_k}(x_{i_k,k+1})\|^2] + \mathbb{E}[\|g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k})\|^2]$$
$$\leq L^2\mathbb{E}[\|x_{i_k,k+1} - x_{i_k,k}\|^2] + 2\mathbb{E}[\langle \nabla f_{i_k}(x_{i_k,k+1}), -g_{i_k,k} + \nabla f_{i_k}(x_{i_k,k})\rangle] + 2\sigma^2. \tag{93}$$

Similarly,

$$\mathbb{E}[\mathbb{1}_k\|g_{j_k,k+1} - g_{j_k,k}\|^2] \leq L^2\mathbb{E}[\mathbb{1}_k\|x_{j_k,k+1} - x_{j_k,k}\|^2]$$
$$+ 2\mathbb{E}[\mathbb{1}_k\langle \nabla f_{j_k}(x_{j_k,k+1}), -g_{j_k,k} + \nabla f_{j_k}(x_{j_k,k})\rangle] + 2\sigma^2. \tag{94}$$

To further bound $\mathbb{E}[\|g_{i_k,k+1} - g_{i_k,k}\|^2 + \mathbb{1}_k\|g_{j_k,k+1} - g_{j_k,k}\|^2]$, we introduce the following lemma.

Lemma 10

$$\mathbb{E}[\langle \nabla f_{i_k}(x_{i_k,k+1}), -g_{i_k,k} + \nabla f_{i_k}(x_{i_k,k})\rangle] + \mathbb{E}[\mathbb{1}_k\langle \nabla f_{j_k}(x_{j_k,k+1}), -g_{j_k,k}$$
$$+ \nabla f_{j_k}(x_{j_k,k})\rangle] \leq 2\alpha L\sigma^2.$$

The proof is similar to that of Lemma 8 and is omitted here. Equation (92) together with (93), (94) and Lemma 10 leads to the following inequality:

$$\mathbb{E}[\|\tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\|^2]$$
$$\leq L^2\mathbb{E}[\|x_{i_k,k+1} - x_{i_k,k}\|^2] + L^2\mathbb{E}[\mathbb{1}_k\|x_{j_k,k+1} - x_{j_k,k}\|^2] + 4(\alpha L + 1)\sigma^2$$
$$= L^2\mathbb{E}[\mathbb{1}_k(\|x_{i_k,k+1} - x_{i_k,k}\|^2 + \|x_{j_k,k+1} - x_{j_k,k}\|^2)]$$
$$+ L^2\mathbb{E}[(1 - \mathbb{1}_k)\|x_{i_k,k+1} - x_{i_k,k}\|^2] + 4(\alpha L + 1)\sigma^2. \tag{95}$$

Note that from (37a) and (38a), we have

$$\mathbb{1}_k(\|x_{i_k,k+1} - x_{i_k,k}\|^2 + \|x_{j_k,k+1} - x_{j_k,k}\|^2)$$
$$= \mathbb{1}_k\left[\left\|\frac{1}{2}(x_{i_k,k} + x_{j_k,k}) - \alpha y_{i_k,k} - x_{i_k,k}\right\|^2 + \left\|\frac{1}{2}(x_{i_k,k} + x_{j_k,k}) - \alpha y_{j_k,k} - x_{j_k,k}\right\|^2\right]$$
$$= \frac{1}{4}\mathbb{1}_k\left(\|x_{j_k,k} - x_{i_k,k} - 2\alpha y_{i_k,k}\|^2 + \|x_{i_k,k} - x_{j_k,k} - 2\alpha y_{j_k,k}\|^2\right)$$
$$= \frac{1}{2}\mathbb{1}_k\left[\|x_{i_k,k} - x_{j_k,k}\|^2 + 2\alpha^2\left(\|y_{i_k,k}\|^2 + \|y_{j_k,k}\|^2\right) + 2\alpha\langle x_{i_k,k} - x_{j_k,k}, y_{i_k,k} - y_{j_k,k}\rangle\right],$$

and

$$(1 - \mathbb{1}_k)\|x_{i_k,k+1} - x_{i_k,k}\|^2 = 4\alpha^2(1 - \mathbb{1}_k)\|y_{i_k,k}\|^2.$$

Then by (95),

$$
\begin{aligned}
\mathbb{E}[\|\tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\|^2] \\
\leq L^2 \mathbb{E}\left[\mathbb{1}_k \left(\frac{1}{2}\|x_{i_k,k} - x_{j_k,k}\|^2 + \alpha^2 \left(\|y_{i_k,k}\|^2 + \|y_{j_k,k}\|^2\right)\right.\right. \\
\left.+ \alpha\langle x_{i_k,k} - x_{j_k,k}, y_{i_k,k} - y_{j_k,k}\rangle\right)\right] \\
+ 4\alpha^2 L^2 \mathbb{E}[(1 - \mathbb{1}_k)\|y_{i_k,k}\|^2] + 4(\alpha L + 1)\sigma^2 \\
\leq L^2 \mathbb{E}\left[\mathbb{1}_k \left(\|x_{i_k,k} - x_{j_k,k}\|^2 + 2\alpha^2 \left(\|y_{i_k,k}\|^2 + \|y_{j_k,k}\|^2\right)\right)\right] \\
+ 4\alpha^2 L^2 \mathbb{E}[(1 - \mathbb{1}_k)\|y_{i_k,k}\|^2] + 4(\alpha L + 1)\sigma^2 \\
\leq \frac{L^2}{n} \mathbb{E}\left[2\|\mathbf{I} - \mathbf{\Pi}\|\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + 4\alpha^2\|\mathbf{y}_k\|^2\right] + 4(\alpha L + 1)\sigma^2 \\
\leq \frac{4L^2}{n} \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2 + \alpha^2\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 + \alpha^2 n\|\overline{y}_k\|^2\right] + 4(\alpha L + 1)\sigma^2. \quad (96)
\end{aligned}
$$

For the last term on the right-hand side of (90), note that

$$
2\langle \mathbf{W}_k \mathbf{y}_k - \mathbf{1}\overline{y}_k, \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\rangle = \langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, g_{i_k,k+1} - g_{i_k,k} + \mathbb{1}_k(g_{j_k,k+1} - g_{j_k,k})\rangle.
$$

By Assumption 1, we have

$$
\begin{aligned}
2\mathbb{E}[\langle \mathbf{W}_k \mathbf{y}_k - \mathbf{1}\overline{y}_k, \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\rangle] \\
= \mathbb{E}[\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, \nabla f_{i_k}(x_{i_k,k+1}) - g_{i_k,k} + \mathbb{1}_k(\nabla f_{j_k}(x_{j_k,k+1}) - g_{j_k,k})\rangle]. \quad (97)
\end{aligned}
$$

The following lemma is useful.

Lemma 11

$$
\mathbb{E}[\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k}) + \mathbb{1}_k(g_{j_k,k} - \nabla f_{j_k}(x_{j_k,k}))\rangle] \geq 0.
$$

Proof See Appendix 7.4. □

In light of (97) and Lemma 11,

$$
\begin{aligned}
2\mathbb{E}[\langle \mathbf{W}_k \mathbf{y}_k - \mathbf{1}\overline{y}_k, \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\rangle] \\
\leq \mathbb{E}[\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, \nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k}) + \mathbb{1}_k[\nabla f_{j_k}(x_{j_k,k+1}) - \nabla f_{j_k}(x_{j_k,k})]\rangle] \\
= \mathbb{E}[\mathbb{1}_k\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, \nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k}) + \nabla f_{j_k}(x_{j_k,k+1}) - \nabla f_{j_k}(x_{j_k,k})\rangle] \\
+ 2\mathbb{E}[(1 - \mathbb{1}_k)\langle y_{i_k,k} - \overline{y}_k, \nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k})\rangle]. \quad (98)
\end{aligned}
$$

Notice that from (37a), (38a) and Assumption 2,

$$
\mathbb{1}_k\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, \nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k}) + \nabla f_{j_k}(x_{j_k,k+1}) - \nabla f_{j_k}(x_{j_k,k})\rangle
$$

$$\leq L\mathbb{1}_k(\|y_{i_k,k} - \overline{y}_k\| + \|y_{j_k,k} - \overline{y}_k\|)(\|x_{i_k,k+1} - x_{i_k,k}\| + \|x_{j_k,k+1} - x_{j_k,k}\|)$$

$$= \frac{L}{2}\mathbb{1}_k(\|y_{i_k,k} - \overline{y}_k\| + \|y_{j_k,k} - \overline{y}_k\|)(\|x_{j_k,k} - x_{i_k,k} - 2\alpha y_{i_k,k}\| + \|x_{i_k,k} - x_{j_k,k} - 2\alpha y_{j_k,k}\|)$$

$$= L\mathbb{1}_k(\|y_{i_k,k} - \overline{y}_k\| + \|y_{j_k,k} - \overline{y}_k\|)\|x_{i_k,k} - x_{j_k,k}\|$$

$$\quad + \alpha L\mathbb{1}_k(\|y_{i_k,k} - \overline{y}_k\| + \|y_{j_k,k} - \overline{y}_k\|)(\|y_{i_k,k}\| + \|y_{j_k,k}\|),$$

and

$$(1 - \mathbb{1}_k)\langle y_{i_k,k} - \overline{y}_k, \nabla f_{i_k}(x_{i_k,k+1}) - \nabla f_{i_k}(x_{i_k,k})\rangle \leq L(1 - \mathbb{1}_k)\|y_{i_k,k} - \overline{y}_k\|\|x_{i_k,k+1} - x_{i_k,k}\|$$

$$= 2\alpha L(1 - \mathbb{1}_k)\|y_{i_k,k} - \overline{y}_k\|\|y_{i_k,k}\|.$$

We have from (98) that

$$2\mathbb{E}[\langle \mathbf{W}_k\mathbf{y}_k - \mathbf{1}\overline{y}_k, \tilde{\mathbf{D}}_k G_{k+1} - \tilde{\mathbf{D}}_k G_k\rangle]$$

$$\leq L\mathbb{E}[\mathbb{1}_k(\|y_{i_k,k} - \overline{y}_k\| + \|y_{j_k,k} - \overline{y}_k\|)\|x_{i_k,k} - x_{j_k,k}\|]$$

$$\quad + \alpha L\mathbb{E}[\mathbb{1}_k(\|y_{i_k,k} - \overline{y}_k\| + \|y_{j_k,k} - \overline{y}_k\|)(\|y_{i_k,k}\| + \|y_{j_k,k}\|)]$$

$$\quad + 4\alpha L\mathbb{E}[(1 - \mathbb{1}_k)\|y_{i_k,k} - \overline{y}_k\|\|y_{i_k,k}\|)]$$

$$\leq \mathbb{E}\left[\mathbb{1}_k\beta_2(\|y_{i_k,k} - \overline{y}_k\|^2 + \|y_{j_k,k} - \overline{y}_k\|^2) + \frac{L^2\mathbb{1}_k}{2\beta_2}\|x_{i_k,k} - x_{j_k,k}\|^2\right]$$

$$\quad + \alpha L\mathbb{E}\left[\mathbb{1}_k(\|y_{i_k,k} - \overline{y}_k\|^2 + \|y_{j_k,k} - \overline{y}_k\|^2 + \|y_{i_k,k}\|^2 + \|y_{j_k,k}\|^2)\right]$$

$$\quad + 2\alpha L\mathbb{E}[(1 - \mathbb{1}_k)(\|y_{i_k,k} - \overline{y}_k\|^2 + \|y_{i_k,k}\|^2)]$$

$$\leq \frac{2}{n}\beta_2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] + \frac{L^2\|\mathbf{I} - \mathbf{\Pi}\|}{n}\frac{1}{\beta_2}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2]$$

$$\quad + \frac{2\alpha L}{n}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] + \frac{2\alpha L}{n}\mathbb{E}[\|\mathbf{y}_k\|^2]$$

$$= \frac{2}{n}\beta_2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] + \frac{2L^2}{n}\frac{1}{\beta_2}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] + \frac{4\alpha L}{n}\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] + 2\alpha L\mathbb{E}[\|\overline{y}_k\|^2],$$
(99)

for any $\beta_2 > 0$. In light of (91), (96) and (99), we obtain by (90) that

$$\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2]$$

$$\leq \left(\rho_{\overline{w}} + \frac{2}{n}\beta_2 + \frac{4\alpha L}{n}\right)\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] + \frac{2L^2}{n}\frac{1}{\beta_2}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] + 2\alpha L\mathbb{E}[\|\overline{y}_k\|^2]$$

$$\quad + \frac{4L^2}{n}\left[\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] + \alpha^2\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2] + \alpha^2 n\mathbb{E}[\|\overline{y}_k\|^2]\right] + 4(\alpha L + 1)\sigma^2$$

$$= \left(\rho_{\overline{w}} + \frac{2}{n}\beta_2 + \frac{4\alpha L}{n} + \frac{4\alpha^2 L^2}{n}\right)\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2 \mid \mathcal{F}_k] + \frac{L^2}{n}\left(4 + \frac{2}{\beta_2}\right)\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2]$$

$$\quad + (4\alpha^2 L^2 + 2\alpha L)\mathbb{E}[\|\overline{y}_k\|^2] + 4(\alpha L + 1)\sigma^2.$$

Since by (88),

$$\mathbb{E}[\|\overline{y}_k\|^2] \le \frac{\sigma^2}{n} + \frac{2L^2}{n}\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] + 2L^2\mathbb{E}[\|\overline{x}_k - x^*\|^2].$$

We conclude that

$$\mathbb{E}[\|\mathbf{y}_{k+1} - \mathbf{1}\overline{y}_{k+1}\|^2] \le \left(\rho_{\overline{w}} + \frac{2}{n}\beta_2 + \frac{4\alpha L}{n} + \frac{4\alpha^2 L^2}{n}\right)\mathbb{E}[\|\mathbf{y}_k - \mathbf{1}\overline{y}_k\|^2]$$

$$+ \frac{L^2}{n}\left(4 + \frac{2}{\beta_2} + 8\alpha^2 L^2 + 4\alpha L\right)\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}\overline{x}_k\|^2] + (8\alpha^2 L^4 + 4\alpha L^3)\mathbb{E}[\|\overline{x}_k - x^*\|^2]$$

$$+ \frac{(4\alpha^2 L^2 + 2\alpha L)\sigma^2}{n} + 4(\alpha L + 1)\sigma^2.$$

7.4 Proof of Lemma 11

The following relation holds:

$$\mathbb{E}[\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k}) + \mathbb{1}_k(g_{j_k,k} - \nabla f_{j_k}(x_{j_k,k}))\rangle]$$
$$= \mathbb{E}[\mathbb{1}_k\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k}) + g_{j_k,k} - \nabla f_{j_k}(x_{j_k,k})\rangle]$$
$$+ 2\mathbb{E}[(1 - \mathbb{1}_k)\langle y_{i_k,k} - \overline{y}_k, g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k})\rangle].$$

From the updating rules (37b) and (38b), we have

$$\mathbb{E}[\mathbb{1}_k\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k}) + g_{j_k,k} - \nabla f_{j_k}(x_{j_k,k})\rangle]$$
$$= \left(1 - \frac{2}{n}\right)\mathbb{E}[\mathbb{1}_k\langle(g_{i_k,k} + g_{j_k,k}), g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k}) + g_{j_k,k} - \nabla f_{j_k}(x_{j_k,k})\rangle]$$
$$= \left(1 - \frac{2}{n}\right)\mathbb{E}[\mathbb{1}_k\|g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k})\|^2 + \mathbb{1}_k\|g_{j_k,k} - \nabla f_{j_k}(x_{j_k,k})\|^2] \ge 0,$$

and

$$\mathbb{E}[(1 - \mathbb{1}_k)\langle y_{i_k,k} - \overline{y}_k, g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k})\rangle] = \left(1 - \frac{1}{n}\right)\mathbb{E}[(1 - \mathbb{1}_k)\|g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k})\|^2] \ge 0.$$

Hence

$$\mathbb{E}[\langle y_{i_k,k} + y_{j_k,k} - 2\overline{y}_k, g_{i_k,k} - \nabla f_{i_k}(x_{i_k,k}) + \mathbb{1}_k(g_{j_k,k} - \nabla f_{j_k}(x_{j_k,k}))\rangle] \ge 0.$$

References

1. Aybat, N.S., Wang, Z., Lin, T., Ma, S.: Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. IEEE Trans. Autom. Control **63**(1), 5–20 (2017)

2. Baingana, B., Mateos, G., Giannakis, G.B.: Proximal-gradient algorithms for tracking cascades over social networks. IEEE J. Sel. Top. Signal Process. **8**(4), 563–575 (2014)
3. Bianchi, P., Jakubowicz, J.: Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. IEEE Trans. Autom. Control **58**(2), 391–405 (2013)
4. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Randomized gossip algorithms. IEEE Trans. Inf. Theory **52**(6), 2508–2530 (2006)
5. Cavalcante, R.L., Stanczak, S.: A distributed subgradient method for dynamic convex optimization problems under noisy information exchange. IEEE J. Sel. Top. Signal Process. **7**(2), 243–256 (2013)
6. Chatzipanagiotis, N., Dentcheva, D., Zavlanos, M.M.: An augmented Lagrangian method for distributed optimization. Math. Program. **152**(1–2), 405–434 (2015)
7. Chatzipanagiotis, N., Zavlanos, M.M.: A distributed algorithm for convex constrained optimization under noise. IEEE Trans. Autom. Control **61**(9), 2496–2511 (2016)
8. Chen, J., Sayed, A.H.: Diffusion adaptation strategies for distributed optimization and learning over networks. IEEE Trans. Signal Process. **60**(8), 4289–4305 (2012)
9. Chen, J., Sayed, A.H.: On the learning behavior of adaptive networks—-part I: transient analysis. IEEE Trans. Inf. Theory **61**(6), 3487–3517 (2015)
10. Chen, J., Sayed, A.H.: On the learning behavior of adaptive networks—part II: performance analysis. IEEE Trans. Inf. Theory **61**(6), 3518–3548 (2015)
11. Cohen, K., Nedić, A., Srikant, R.: Distributed learning algorithms for spectrum sharing in spatial random access wireless networks. IEEE Trans. Autom. Control **62**(6), 2854–2869 (2017)
12. Cohen, K., Nedić, A., Srikant, R.: On projected stochastic gradient descent algorithm with weighted averaging for least squares regression. IEEE Trans. Autom. Control **62**(11), 5974–5981 (2017)
13. Di Lorenzo, P., Scutari, G.: Next: in-network nonconvex optimization. IEEE Trans. Signal Inf. Process. Over Netw. **2**(2), 120–136 (2016)
14. Duchi, J.C., Agarwal, A., Wainwright, M.J.: Dual averaging for distributed optimization: convergence analysis and network scaling. IEEE Trans. Autom. Control **57**(3), 592–606 (2012)
15. Forrester, A.I., Sóbester, A., Keane, A.J.: Multi-fidelity optimization via surrogate modelling. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. The Royal Society, vol. 463, pp. 3251–3269 (2007)
16. Friedman, J., Kahn, J., Szemeredi, E.: On the second eigenvalue of random regular graphs. In: Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing. ACM, pp. 587–598 (1989)
17. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1990)
18. Jakovetic, D., Bajovic, D., Sahu, A.K., Kar, S.: Convergence rates for distributed stochastic optimization over random networks. In: 2018 IEEE Conference on Decision and Control (CDC), pp. 4238–4245. IEEE (2018)
19. Jakovetić, D., Xavier, J., Moura, J.M.: Fast distributed gradient methods. IEEE Trans. Autom. Control **59**(5), 1131–1146 (2014)
20. Kia, S.S., Cortés, J., Martínez, S.: Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. Automatica **55**, 254–264 (2015)
21. Kiefer, J., Wolfowitz, J., et al.: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. **23**(3), 462–466 (1952)
22. Kleijnen, J.P.: Design and Analysis of Simulation Experiments, vol. 20. Springer, Berlin (2008)
23. Kushner, H., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications, vol. 35. Springer, Berlin (2003)
24. Lan, G., Lee, S., Zhou, Y.: Communication-efficient algorithms for decentralized and stochastic optimization. Math. Program. **180**, 237–284 (2020). https://doi.org/10.1007/s10107-018-1355-4
25. Lee, S., Nedić, A.: Asynchronous gossip-based random projection algorithms over networks. IEEE Trans. Autom. Control **61**(4), 953–968 (2016)
26. Lian, X., Zhang, C., Zhang, H., Hsieh, C.J., Zhang, W., Liu, J.: Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp. 5336–5346 (2017)
27. Ling, Q., Shi, W., Wu, G., Ribeiro, A.: Dlm: decentralized linearized alternating direction method of multipliers. IEEE Trans. Signal Process. **63**(15), 4051–4064 (2015)
28. Lobel, I., Ozdaglar, A., Feijer, D.: Distributed multi-agent optimization with state-dependent communication. Math. Program. **129**(2), 255–284 (2011)

29. Lu, J., Tang, C.Y., Regier, P.R., Bow, T.D.: Gossip algorithms for convex consensus optimization over networks. IEEE Trans. Autom. Control **56**(12), 2917–2923 (2011)
30. Mateos, G., Giannakis, G.B.: Distributed recursive least-squares: stability and performance analysis. IEEE Trans. Signal Process. **60**(7), 3740–3754 (2012)
31. Mathkar, A.S., Borkar, V.S.: Nonlinear gossip. SIAM J. Control Optim. **54**(3), 1535–1557 (2016)
32. Nedić, A., Olshevsky, A.: Stochastic gradient-push for strongly convex functions on time-varying directed graphs. IEEE Trans. Autom. Control **61**(12), 3936–3947 (2016)
33. Nedić, A., Olshevsky, A., Rabbat, M.G.: Network topology and communication–computation tradeoffs in decentralized optimization. Proc. IEEE **106**(5), 953–976 (2018)
34. Nedic, A., Olshevsky, A., Shi, W.: Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM J. Optim. **27**(4), 2597–2633 (2017)
35. Nedić, A., Olshevsky, A., Uribe, C.A.: Fast convergence rates for distributed non-bayesian learning. IEEE Trans. Autom. Control **62**(11), 5538–5553 (2017)
36. Nedić, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control **54**(1), 48–61 (2009)
37. Nedić, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control **55**(4), 922–938 (2010)
38. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2009)
39. Olshevsky, A., Paschalidis, I.C., Spiridonoff, A.: Robust asynchronous stochastic gradient-push: asymptotically optimal and network-independent performance for strongly convex functions. arXiv preprint arXiv:1811.03982 (2018)
40. Pu, S., Garcia, A.: A flocking-based approach for distributed stochastic optimization. Oper. Res. **1**, 267–281 (2018)
41. Pu, S., Garcia, A.: Swarming for faster convergence in stochastic optimization. SIAM J. Control Optim. **56**(4), 2997–3020 (2018)
42. Pu, S., Garcia, A., Lin, Z.: Noise reduction by swarming in social foraging. IEEE Trans. Autom. Control **61**(12), 4007–4013 (2016)
43. Pu, S., Nedić, A.: A distributed stochastic gradient tracking method. In: 2018 IEEE Conference on Decision and Control (CDC). IEEE, pp. 963–968 (2018)
44. Pu, S., Olshevsky, A., Paschalidis, I.C.: A sharp estimate on the transient time of distributed stochastic gradient descent. arXiv preprint arXiv:1906.02702 (2019)
45. Pu, S., Shi, W., Xu, J., Nedic, A.: Push–pull gradient methods for distributed optimization in networks. IEEE Trans. Autom. Control (2020). https://doi.org/10.1109/TAC.2020.2972824
46. Qu, G., Li, N.: Harnessing smoothness to accelerate distributed optimization. IEEE Trans. Control Netw. Syst. **5**(3), 1245–1260 (2017)
47. Rakhlin, A., Shamir, O., Sridharan, K.: Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Conference on International Conference on Machine Learning. Omnipress, pp. 1571–1578 (2012)
48. Ram, S.S., Nedić, A., Veeravalli, V.V.: Distributed stochastic subgradient projection algorithms for convex optimization. J. Optim. Theory Appl. **147**(3), 516–545 (2010)
49. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)
50. Sayed, A.H.: Adaptive networks. Proc. IEEE **102**(4), 460–497 (2014)
51. Sayin, M.O., Vanli, N.D., Kozat, S.S., Başar, T.: Stochastic subgradient algorithms for strongly convex optimization over distributed networks. IEEE Trans. Netw. Sci. Eng. **4**(4), 248–260 (2017)
52. Shi, W., Ling, Q., Wu, G., Yin, W.: Extra: an exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. **25**(2), 944–966 (2015)
53. Sirb, B., Ye, X.: Decentralized consensus algorithm with delayed and stochastic gradients. SIAM J. Optim. **28**(2), 1232–1254 (2018)
54. Srivastava, K., Nedic, A.: Distributed asynchronous constrained stochastic optimization. IEEE J. Sel. Top. Signal Process. **5**(4), 772–790 (2011)
55. Towfic, Z.J., Sayed, A.H.: Adaptive penalty-based distributed stochastic convex optimization. IEEE Trans. Signal Process. **62**(15), 3924–3938 (2014)
56. Tsitsiklis, J., Bertsekas, D., Athans, M.: Distributed asynchronous deterministic and stochastic gradient optimization algorithms. IEEE Trans. Autom. Control **31**(9), 803–812 (1986)

Affiliations

Shi Pu[1,2] [iD] **· Angelia Nedić[3]**

Angelia Nedić
Angelia.Nedich@asu.edu

1 Institute for Data and Decision Analytics, The Chinese University of Hong Kong, Shenzhen, China

2 Shenzhen Research Institute of Big Data, Shenzhen, China

3 School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287, USA