Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A distributed stochastic optimization algorithm with gradient-tracking and distributed heavy-ball acceleration*

Bihao SUN[1], Jinhui HU[1], Dawen XIA[2], Huaqing LI†‡[1]

[1]*Chongqing Key Laboratory of Nonlinear Circuits and Intelligent Information Processing, College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China*

[2]*College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China*

†E-mail: huaqingli@swu.edu.cn

Received Nov. 8, 2020; Revision accepted Feb. 15, 2021; Crosschecked Apr. 1, 2021; Published online July 24, 2021

Abstract: Distributed optimization has been well developed in recent years due to its wide applications in machine learning and signal processing. In this paper, we focus on investigating distributed optimization to minimize a global objective. The objective is a sum of smooth and strongly convex local cost functions which are distributed over an undirected network of n nodes. In contrast to existing works, we apply a distributed heavy-ball term to improve the convergence performance of the proposed algorithm. To accelerate the convergence of existing distributed stochastic first-order gradient methods, a momentum term is combined with a gradient-tracking technique. It is shown that the proposed algorithm has better acceleration ability than GT-SAGA without increasing the complexity. Extensive experiments on real-world datasets verify the effectiveness and correctness of the proposed algorithm.

Key words: Distributed optimization; High-performance algorithm; Multi-agent system; Machine-learning problem; Stochastic gradient

https://doi.org/10.1631/FITEE.2000615 **CLC number:** TP14

1 Introduction

Distributed optimization has attracted much attention in many fields, such as wireless sensor networks (Yin et al., 2010; Cohen et al., 2017), machine learning (Xia and Wang, 2004; McMahan et al., 2017; Liu et al., 2019), and coordinated control (Han et al., 2015; Cheng B and Li, 2019). It is evident that practical problems based on distributed optimization can be modeled as the minimization of the global objective function, i.e., solving the problem over a connected network consisting of n agents cooperatively over a common variable $\boldsymbol{x}$:

$$\mathrm{P0}: \min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\boldsymbol{x}),$$

where f_i is only known by agent i to exchange information with its neighbor nodes over an undirected network.

1.1 Literature review

In the literature, distributed optimization has attracted widespread attention based mainly on the

‡ Corresponding author

* Project supported by the Open Research Fund Program of Data Recovery Key Laboratory of Sichuan Province, China (No. DRN2001), the National Natural Science Foundation of China (Nos. 61773321 and 61762020), the Science and Technology Top-Notch Talents Support Project of Colleges and Universities in Guizhou Province, China (No. QJHKY2016065), the Science and Technology Foundation of Guizhou Province, China (No. QKHJC20181083), and the Science and Technology Talents Fund for Excellent Young of Guizhou Province, China (No. QKHPTRC20195669)

ORCID: Bihao SUN, https://orcid.org/0000-0002-2790-2528; Huaqing LI, https://orcid.org/0000-0001-6310-8965

Lagrangian, (sub)gradient descent, and Newton's methods. As for Lagrangian methods, Boyd et al. (2011) proposed a classic distributed alternating direction method of multipliers (ADMM) on the basis of the dual-domain over augmented Lagrangian functions, and many other distributed algorithms (Ling and Tian, 2010; Erseghe et al., 2011; Cheng S et al., 2014; Ling and Ribeiro, 2014; Wang B et al., 2018; Zhang et al., 2019; Guan et al., 2020) have been proposed based on this method. Although distributed ADMM with a constant step-size has the advantage of a linear convergence rate for a strongly convex function, it requires a large amount of calculation because each node has to optimize a local problem at each iteration. Unlike the Lagrangian method, Newton's method (Bertsekas and Gafni, 1983; Wei et al., 2013) is a common method to solve unconstrained optimization problems due to the advantage of fast convergence. To simplify the calculation process of Newton's method, the quasi-Newton method (Eisen et al., 2017) approximates the inverse Hessian matrix or the Hessian matrix using a positive definite matrix. The (sub)gradient descent method includes dual average (Duchi et al., 2012; Yuan et al., 2013; Matthews et al., 2016), distributed gradient descent (DGD) (Mateos et al., 2010; Xu et al., 2015; Nedić et al., 2017a, 2017b; Xi and Khan, 2017; Xin et al., 2019c), and stochastic gradient descent (SGD) methods (Zinkevich et al., 2010). In the pioneering work of Tsitsiklis et al. (1986), a framework for analyzing distributed computing models was developed. Nedić and Ozdaglar (2009) applied this method to the distributed convex optimization problem in the network, and achieved convergence of the strongly convex non-smooth structure. The SGD algorithm uses only a set of data from the sample to perform gradient descent, which improves the training speed of samples and reduces the calculation cost. Stochastic average gradient (SAG) (Schmidt et al., 2017; Wang Z and Li, 2020), SAGA (Defazio et al., 2014), and SVRG (Johnson and Zhang, 2013; Tan et al., 2016) methods have been proposed because the SGD algorithm can reach only an optimal interval instead of an optimal value. At each iteration, only one randomly selected gradient of a subfunction is evaluated at a node, and a variance-reduced stochastic averaging gradient technique is applied to approximate the gradient of the local objective function. SAG calculates a random vector as the average value of the random gradient in the previous iterations, where in the k^{th} iteration, the algorithm stores derivatives to achieve exact convergence. In addition, based on SAG, the SAGA algorithm can directly support non-strongly convex problems, and is adaptive to any inherently strong convexity of the problem. Another algorithm, SVRG, which is performed in a loop, uses considerably amount of calculation to reduce the influence of noise, and is better than SGD at achieving more accurate convergence. GT-SAGA and GT-SVRG (Xin and Khan, 2020), which are based on SAGA and SVRG, respectively, achieve accelerated linear convergence by combining distributed stochastic gradient-tracking methods with variance-reduced techniques. In addition to the above methods (Johnson and Zhang, 2013; Defazio et al., 2014; Tan et al., 2016; Schmidt et al., 2017; Wang Z and Li, 2020), the $\mathcal{AB}$ method based on row and column randomization (Xin and Khan, 2018), its acceleration method, the $\mathcal{AB}m$ algorithm (Xin et al., 2020), and $\mathcal{S}$-$\mathcal{AB}$ algorithm (Xin et al., 2019b) have also made significant contributions to solving the directed network.

1.2 Motivations and contributions

We find that plenty of works are interested in solving large-scale optimization problems with numerous and complex local objective functions. However, in distributed settings, distributed algorithms with exact gradient need massive calculation. Therefore, plenty of methods such as SAG, SAGA, and SVRG have been proposed to reduce the cost of full evaluation and retain the advantage of fast convergence under strongly convex and smooth conditions. Based on these works, GT-SAGA (Xin and Khan, 2020) combined a gradient-tracking technique with a variance-reduced technique that accelerates linear convergence. By introducing the above method to our work, we propose a new algorithm with a faster linear convergence rate and provide the concise proof. We summarize the following contributions in this paper:

1. Aiming to accelerate the convergence of stochastic first-order gradient methods (Zinkevich et al., 2010; Johnson and Zhang, 2013; Defazio et al., 2014; Tan et al., 2016; Schmidt et al., 2017; Lan et al., 2018; Wang Z and Li, 2020), we incorporate the momentum term combining with a gradient-tracking technique to achieve an accelerated convergence rate over undirected networks.

2. Compared with the distributed gradient methods (Ling and Tian, 2010; Boyd et al., 2011; Erseghe et al., 2011; Ling and Ribeiro, 2014; Wang B et al., 2018; Xin and Khan, 2018; Zhang et al., 2019; Xin et al., 2020) with deterministic gradients, the proposed algorithm requires much less computation when facing large-scale dataset optimization problems, because SGD method can update the model parameters with just a single training dataset.

3. In this paper, we present rigorous theoretical analysis for the proposed algorithm in Section 3. In addition, extensive experiments are provided to verify the correctness of theoretical analysis, such as the distributed logistic regression experiment based on a real-world dataset, a signal processing oriented least-square experiment, and distributed quadratic programming.

Notations used in this paper are summarized as follows: lowercase bold letters denote vectors, while uppercase denotes a matrix. The Euclidean norm of a vector is denoted as $\|\cdot\|$, and $|||\cdot|||$ denotes the spectral norm of a matrix. $\mathbf{1}_n$ means the n-dimensional column vector with all ones. For arbitrary matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$, $\boldsymbol{A}\otimes\boldsymbol{B}$ denotes the Kronecker product. $\rho(\boldsymbol{A})$ is regarded as the spectral radius of $\boldsymbol{A}$.

2 Problem formulation and algorithm development

2.1 Distributed optimization problem

We consider n nodes communicating over an undirected network $\mathcal{G}$, so that it is capable of accessing a local cost function at each node i. The goal is to solve the following optimization problem with m nodes:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i\left(\boldsymbol{x}^i\right),\ f_i\left(\boldsymbol{x}^i\right) = \frac{1}{p_i}\sum_{h=1}^{p_i} f_i^h\left(\boldsymbol{x}^i\right),$$

where each local cost f_i is averaged by p_i constituent functions $\left\{f_i^h\right\}_{h=1}^{p_i}$. The number of local samples is retained by agent i, $i \in U$. The results of the proposed algorithm are on the basis of the following assumptions:

Assumption 1 The global objective function $f(\boldsymbol{x})$ is strongly convex with strong convexity parameter μ; i.e., $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, we have

$$\left(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\right)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}) \geq \mu\|\boldsymbol{x}-\boldsymbol{y}\|^2. \quad (1)$$

Assumption 2 Each local objective function f_i^h has Lipschitz continuous gradient with Lipschitz constant $L_f > 0$; i.e., $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, we have

$$\left\|\nabla f_i^h(\boldsymbol{x}) - \nabla f_i^h(\boldsymbol{y})\right\| \leq L_f\|\boldsymbol{x}-\boldsymbol{y}\|, \quad (2)$$

where $L_f > \mu > 0$.

Assumption 3 The weight matrix $\boldsymbol{W}$ is doubly stochastic and associated with the undirected network $\mathcal{G}$.

Let σ indicate the spectral norm of the matrix $\boldsymbol{W} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}$. According to Assumption 3, we have $\sigma < 1$, i.e., $\sigma = \left\|\boldsymbol{W} - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^{\mathrm{T}}\right\| < 1$. In addition, we denote $P := \max_i\{p_i\}$, $p := \min_i\{p_i\}$.

Remark 1 These three assumptions are standard in recent literature.

2.2 Algorithm development

Previous work on the DGD method updates two vectors $\boldsymbol{x}_k^i, \boldsymbol{y}_k^i \in \mathbb{R}^p$ at each node i, where $\boldsymbol{x}_k^i$ is the local estimate of the global minimizer and $\boldsymbol{y}_k^i$ is an auxiliary variable:

$$\begin{cases} \boldsymbol{x}_{k+1}^i = \sum_{j=1}^{m} w_{ij}\boldsymbol{x}_k^j - \alpha\boldsymbol{y}_k^i, \\ \boldsymbol{y}_{k+1}^i = \sum_{j=1}^{m} w_{ij}\boldsymbol{y}_k^j + \nabla f_i\left(\boldsymbol{x}_{k+1}^i\right) - \nabla f_i\left(\boldsymbol{x}_k^i\right). \end{cases} \quad (3)$$

Following Algorithm 1, we assume that the gradient of agent i at the k^{th} iteration is updated as shown in Fig. 1.

When $\boldsymbol{g}_k^i$ is updated, the $\nabla f_i^{t_k^i}\left(\boldsymbol{v}_{k+1}^{i,t_k^i}\right)$ entry in the gradient table is replaced with $\nabla f_i^{t_k^i}\left(\boldsymbol{x}_k^i\right)$, while others remain unchanged:

$$\nabla f_i^j\left(\boldsymbol{v}_{k+1}^{i,j}\right) \leftarrow \nabla f_i^j\left(\boldsymbol{v}_k^{i,j}\right),$$
$$j = \left\{j \mid j = 1, 2, \cdots, t_i,\ j \neq t_k^i\right\}.$$

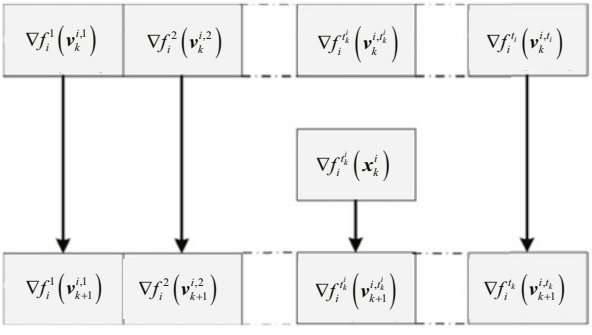

Fig. 1 Gradient updating of agent i in the k^{th} iteration

Therefore, one can obtain the updated formula as follows:

$$\sum_{j=1}^{p_i}\nabla f_i^j\left(\boldsymbol{v}_{k+1}^{i,j}\right)=\sum_{j=1}^{p_i}\nabla f_i^j\left(\boldsymbol{v}_k^{i,j}\right)+\nabla f_i^{t_{k+1}^i}\left(\boldsymbol{x}_k^i\right)-\nabla f_i^{t_{k+1}^i}\left(\boldsymbol{v}_k^{i,t_k^i}\right). \tag{4}$$

Algorithm 1 The proposed algorithm at each node i

Initialization: $\boldsymbol{x}_i^0$; $\boldsymbol{z}_{i,j}^1=\boldsymbol{x}_i^0$, $\forall j\in\{1,2,\cdots,p_i\}$; $\alpha>0$; $\{\tilde{w}_{ij}\}_{j=1}^n$; $\boldsymbol{y}_i^0=\boldsymbol{g}_i^0=\nabla f_i\left(\boldsymbol{x}_i^0\right)$

for $k=0,1,2,\cdots$ **do**

Choose t_{k+1}^i uniformly from local sample set $\{1,2,\cdots,p_i\}$ at random

Update the variable $\boldsymbol{g}_{k+1}^i$ as

$$\boldsymbol{g}_{k+1}^i=\nabla f_i^{t_{k+1}^i}\left(\boldsymbol{x}_{k+1}^i\right)-\nabla f_i^{t_{k+1}^i}\left(\boldsymbol{v}_{k+1}^{i,t_{k+1}^i}\right)+\frac{1}{p_i}\sum\nolimits_{j=1}^{p_i}\nabla f_i^j\left(\boldsymbol{v}_{k+1}^{i,j}\right)$$

Update the variable $\boldsymbol{y}_{k+1}^i$ as

$$\boldsymbol{y}_{k+1}^i=\sum\nolimits_{j=1}^{m}w_{ij}\boldsymbol{y}_k^j+\boldsymbol{g}_{k+1}^i-\boldsymbol{g}_k^i$$

if $j{=}t_{k+1}^i$ **then**

$\boldsymbol{v}_{k+1}^{i,j}=\boldsymbol{x}_k^{i,j}$

else

$\boldsymbol{v}_{k+1}^{i,j}=\boldsymbol{v}_k^{i,j}$

end if

end for

3 Convergence analysis of the proposed algorithm

We aim to introduce unified analysis for the proposed algorithm that relies on the following dynamical system, with $\boldsymbol{g}_0=\boldsymbol{y}_0$ and $\forall k\geq 0$:

$$\boldsymbol{x}_{k+1}=\boldsymbol{\mathcal{W}}\boldsymbol{x}_k-\alpha\boldsymbol{y}_k+\beta\left(\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right), \tag{5a}$$

$$\boldsymbol{y}_{k+1}=\boldsymbol{\mathcal{W}}\boldsymbol{y}_k+\boldsymbol{g}_{k+1}-\boldsymbol{g}_k, \tag{5b}$$

where $\boldsymbol{\mathcal{W}}=\boldsymbol{W}\otimes\boldsymbol{I}_n$ and

$$\boldsymbol{x}_k=\begin{bmatrix}\boldsymbol{x}_k^1\\ \boldsymbol{x}_k^2\\ \vdots\\ \boldsymbol{x}_k^m\end{bmatrix},\boldsymbol{y}_k=\begin{bmatrix}\boldsymbol{y}_k^1\\ \boldsymbol{y}_k^2\\ \vdots\\ \boldsymbol{y}_k^m\end{bmatrix},\boldsymbol{g}_k=\begin{bmatrix}\boldsymbol{g}_k^1\\ \boldsymbol{g}_k^2\\ \vdots\\ \boldsymbol{g}_k^m\end{bmatrix}.$$

3.1 Preliminaries

We denote $\boldsymbol{\mathcal{F}}_k$ as the history of the dynamical system generated by $\left\{t_s^i\right\}_{i=1,2,\cdots,m}^{s\leq k-1}$. To simplify the subsequent analysis, we use the following notations:

$$\bar{\boldsymbol{x}}_k=\frac{1}{m}\left(\boldsymbol{1}_m^{\mathrm{T}}\otimes\boldsymbol{I}_n\right)\boldsymbol{x}_k,$$

$$\bar{\boldsymbol{y}}_k=\frac{1}{m}\left(\boldsymbol{1}_m^{\mathrm{T}}\otimes\boldsymbol{I}_n\right)\boldsymbol{y}_k,$$

$$\bar{\boldsymbol{g}}_k=\frac{1}{m}\left(\boldsymbol{1}_m^{\mathrm{T}}\otimes\boldsymbol{I}_n\right)\boldsymbol{g}_k,$$

$$\nabla F\left(\boldsymbol{x}_k\right)=\left[\left(\nabla f_1\left(\boldsymbol{x}_k^1\right)\right)^{\mathrm{T}},\ldots,\left(\nabla f_m\left(\boldsymbol{x}_k^m\right)\right)^{\mathrm{T}}\right]^{\mathrm{T}},$$

$$\nabla\bar{F}\left(\boldsymbol{x}_k\right)=\frac{1}{m}\left(\boldsymbol{1}_m^{\mathrm{T}}\otimes\boldsymbol{I}_n\right)\nabla F\left(\boldsymbol{x}_k\right).$$

Lemma 1 $\forall k\geq 0$, it holds that

$$\mathbb{E}\left[\bar{\boldsymbol{y}}_k\left|\boldsymbol{\mathcal{F}}_k\right.\right]=\mathbb{E}\left[\bar{\boldsymbol{g}}_k\left|\boldsymbol{\mathcal{F}}_k\right.\right]=\frac{1}{m}\left(\boldsymbol{1}_m^{\mathrm{T}}\otimes\boldsymbol{I}_n\right)\nabla F\left(\boldsymbol{x}_k\right). \tag{6}$$

Proof Multiplying $\boldsymbol{1}_m^{\mathrm{T}}$ by both sides of dynamical iterating Eq. (5b) yields

$$\bar{\boldsymbol{y}}_{k+1}=\bar{\boldsymbol{y}}_k+\bar{\boldsymbol{g}}_{k+1}-\bar{\boldsymbol{g}}_k. \tag{7}$$

Then, recursively updating Eq. (7) reduces to

$$\bar{\boldsymbol{y}}_k=\bar{\boldsymbol{g}}_k. \tag{8}$$

Thus, we obtain

$$\mathbb{E}\left[\bar{\boldsymbol{y}}_k|\boldsymbol{\mathcal{F}}_k\right]=\mathbb{E}\left[\bar{\boldsymbol{g}}_k|\boldsymbol{\mathcal{F}}_k\right]. \tag{9}$$

According to Algorithm 1, there exists

$$\begin{aligned}&\mathbb{E}\left[\boldsymbol{g}_k^i|\boldsymbol{\mathcal{F}}_k\right]\\ =&\mathbb{E}\left[\nabla f_i^{t_{k+1}^i}\left(\boldsymbol{x}_k^i\right)-\nabla f_i^{t_{k+1}^i}\left(\boldsymbol{v}_k^i\right)+\nabla f_i\left(\boldsymbol{v}_k^i\right)\left|\boldsymbol{\mathcal{F}}_k\right.\right]\\ =&\frac{1}{p_i}\sum_{h=1}^{p_i}\nabla f_i^h\left(\boldsymbol{x}_k^i\right)-\frac{1}{p_i}\sum_{h=1}^{p_i}\nabla f_i^h\left(\boldsymbol{v}_k^i\right)+\nabla f_i\left(\boldsymbol{v}_k^i\right)\\ =&\frac{1}{p_i}\sum_{h=1}^{p_i}\nabla f_i^h\left(\boldsymbol{x}_k^i\right)-\frac{1}{p_i}\sum_{h=1}^{p_i}\nabla f_i^h\left(\boldsymbol{v}_k^i\right)+\frac{1}{p_i}\sum_{h=1}^{p_i}\nabla f_i^h\left(\boldsymbol{v}_k^i\right)\\ =&\frac{1}{p_i}\sum_{h=1}^{p_i}\nabla f_i^h\left(\boldsymbol{x}_k^i\right)\\ =&\nabla f_i\left(\boldsymbol{x}_k^i\right).\end{aligned}$$

Therefore, it is straightforward to obtain

$$\mathbb{E}\left[\bar{\boldsymbol{g}}_k|\boldsymbol{\mathcal{F}}_k\right]=\frac{1}{m}\left(\boldsymbol{1}_m\otimes\boldsymbol{I}_n\right)\nabla f\left(\boldsymbol{x}_k\right). \tag{10}$$

Combining Eqs. (9) and (10) completes the proof.

Lemma 2 Let Assumption 3 hold. $\forall\boldsymbol{x}\in\mathbb{R}^{mn}$, the inequality holds as follows:

$$\left\|\boldsymbol{\mathcal{W}}\boldsymbol{x}-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}\right\|\leq\sigma\left\|\boldsymbol{x}-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}\right\|, \tag{11}$$

where $0<\sigma<1$ is a constant.

Proof According to the definition of $\boldsymbol{\mathcal{W}}_\infty$, we know that $\boldsymbol{\mathcal{W}}_\infty = \boldsymbol{\mathcal{W}}\boldsymbol{\mathcal{W}}_\infty$ and $\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{W}}_\infty\| < 1$. Therefore,

$$\begin{aligned}&\|\boldsymbol{\mathcal{W}}\boldsymbol{x} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}\|\\ =&\|(\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{W}}_\infty)(\boldsymbol{x} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x})\|\\ \le&\|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{W}}_\infty\|\,\|\boldsymbol{x} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}\| = \sigma\,\|\boldsymbol{x} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}\|,\end{aligned}$$

where $\boldsymbol{\mathcal{W}}_\infty(\boldsymbol{x} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}) = \boldsymbol{0}$ is used in the inequality, and $\sigma = \|\boldsymbol{\mathcal{W}} - \boldsymbol{\mathcal{W}}_\infty\|$ is used to complete the proof.

Lemma 3 Let Assumption 2 hold. Considering dynamical system (5), $\forall k \ge 0$, it holds that

$$\left\|\nabla\bar{F}(\boldsymbol{x}_k) - \nabla f(\bar{\boldsymbol{x}}_k)\right\| \le \frac{L_f}{\sqrt{m}}\|\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|. \tag{12}$$

Proof Recalling the definition of $\nabla\bar{F}(\boldsymbol{x}_k)$, it holds that

$$\begin{aligned}&\left\|\nabla\bar{F}(\boldsymbol{x}_k) - \nabla f(\bar{\boldsymbol{x}}_k)\right\|\\ =&\left\|\frac{1}{m}\sum_{i=1}^m\nabla f_i(\boldsymbol{x}_k^i) - \frac{1}{m}\sum_{i=1}^m\nabla f_i(\bar{\boldsymbol{x}}_k)\right\|\\ =&\frac{1}{m}\left\|\sum_{i=1}^m\nabla f_i(\boldsymbol{x}_k^i) - \sum_{i=1}^m\nabla f_i(\bar{\boldsymbol{x}}_k)\right\|\\ =&\frac{1}{m}\sum_{i=1}^m\left\|\nabla f_i(\boldsymbol{x}_k^i) - \nabla f_i(\bar{\boldsymbol{x}}_k)\right\|\\ \le&\frac{L_f}{m}\sum_{i=1}^m\left\|\boldsymbol{x}_k^i - \bar{\boldsymbol{x}}_k\right\|\\ =&\frac{L_f}{\sqrt{m}}\|\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|,\end{aligned}$$

where the Lipschitz continuity is used in the inequality. The proof is completed.

Lemma 4 Suppose that Assumptions 1 and 2 hold. $\forall\boldsymbol{x} \in \mathbb{R}^{mn}$, if $0 < \alpha < 1/L_f$, it holds that

$$\|\boldsymbol{x} - \alpha\nabla f(\boldsymbol{x}) - \boldsymbol{x}^*\| \le (1 - \mu\alpha)\|\boldsymbol{x} - \boldsymbol{x}^*\|, \tag{13}$$

where $0 < \mu \le L_f$, $f(\boldsymbol{x})$ is the global objective function, and we define $\boldsymbol{x}^*$ as the global optimum, $\boldsymbol{x}^* = \boldsymbol{1}_m \otimes \tilde{\boldsymbol{x}}^* \in \mathbb{R}^{mn}$.

3.2 Auxiliary results

We analyze the convergence of the general dynamical system using the following four formulas:

(1) $\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|^2\right]$, the consensus error in the network;

(2) $\mathbb{E}\left[\|\bar{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}^*\|^2\right]$, the optimal gap;

(3) $\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2\right]$, the state difference;

(4) $\mathbb{E}\left[\|\boldsymbol{y}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\|^2\right]$, the gradient tracking error.

Lemma 5 Suppose that Assumption 3 holds. Considering the sequence $\{\boldsymbol{x}_k\}$ yielded by dynamical system (5), $\forall k \ge 0$, it holds that

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_{k+1}\|^2\right]\\ \le&\frac{1-\sigma^2}{2}\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|^2\right] + \frac{4\alpha^2}{1-\sigma^2}\mathbb{E}\left[\|\boldsymbol{y}_k - \boldsymbol{W}_\infty\boldsymbol{y}_k\|^2\right]\\ &+\frac{4\beta^2}{1-\sigma^2}\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2\right]\end{aligned} \tag{14}$$

and

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_{k+1}\|^2\right]\\ \le&2\sigma^2\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|^2\right] + 4\alpha^2\mathbb{E}\left[\|\boldsymbol{y}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\|^2\right]\\ &+4\beta^2\mathbb{E}\left[\|(\boldsymbol{x}_k - \boldsymbol{x}_{k-1})\|^2\right],\end{aligned} \tag{15}$$

where $0 < \mu \le L_f$, and $f(\boldsymbol{x})$ is the global objective function.

Proof Notice that $\boldsymbol{\mathcal{W}}_\infty\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{W}}\boldsymbol{\mathcal{W}}_\infty = \boldsymbol{\mathcal{W}}_\infty$ and $\|\boldsymbol{I}_{mn} - \boldsymbol{\mathcal{W}}_\infty\| = 1$. There exists

$$\begin{aligned}&\|\boldsymbol{x}_{k+1} - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_{k+1}\|^2\\ \le&(1+\eta)\|\boldsymbol{\mathcal{W}}\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|^2\\ &+(1+\eta^{-1})\left(2\alpha^2\|\boldsymbol{y}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\|^2 + 2\beta^2\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2\right),\end{aligned}$$

where Young's inequality and Lemma 2 are used. Setting η as $\frac{1-\sigma^2}{2\sigma^2}$ and 1 in the above inequality can lead to inequalities (14) and (15), respectively.

Lemma 6 Let Assumptions 1–3 hold. Considering the sequence $\{\boldsymbol{x}_k\}$ yielded by dynamical system (5), $\forall k \ge 0$, if $0 < \alpha \le 1/L_f$, we can obtain

$$\begin{aligned}&\mathbb{E}\left[m\|\bar{\boldsymbol{x}}_{k+1} - \tilde{\boldsymbol{x}}^*\|^2\right]\\ \le&4\alpha^2L_f^2\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|^2\right] + 4\beta^2\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2\right]\\ &+\frac{\alpha^2}{m}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{g}_k\|^2\right] + 2\mathbb{E}\left[m\|\bar{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}^*\|^2\right]\end{aligned} \tag{16}$$

and

$$\begin{aligned}&\mathbb{E}\left[m\|\bar{\boldsymbol{x}}_{k+1} - \tilde{\boldsymbol{x}}^*\|^2\right]\\ \le&(1-\mu\alpha)\,\mathbb{E}\left[m\|\bar{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}^*\|^2\right] + \frac{2\alpha L_f^2}{\mu}\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|^2\right]\\ &+\frac{2\beta^2}{\mu\alpha}\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2\right] + \frac{\alpha^2}{m^2}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{g}_k\|^2\right].\end{aligned} \tag{17}$$

Proof According to dynamical iterating Eq. (5a) and recalling the definition of $\bar{\boldsymbol{x}}_{k+1}$, we know that $\bar{\boldsymbol{x}}_{k+1}=\bar{\boldsymbol{x}}_k-\alpha\bar{\boldsymbol{y}}_k+\beta\left(\bar{\boldsymbol{x}}_k-\bar{x}_{k-1}\right)$:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_{k+1}-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\
=&\mathbb{E}\left[\left\|\bar{\boldsymbol{x}}_k-\alpha\bar{\boldsymbol{y}}_k+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\
=&\mathbb{E}\left[\left\|\alpha\left(\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\bar{\boldsymbol{g}}_k\right)+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)\right\|^2\right]\\
&+2\left\langle\bar{\boldsymbol{x}}_k-\alpha\nabla f(\bar{\boldsymbol{x}}_k)-\tilde{\boldsymbol{x}}^*,\ \alpha(\nabla f(\bar{\boldsymbol{x}}_k)-\bar{\boldsymbol{g}}_k)+\beta(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1})\right\rangle\\
&+\left\|\bar{\boldsymbol{x}}_k-\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\tilde{\boldsymbol{x}}^*\right\|^2,
\end{aligned}$$

where $\left[\bar{\boldsymbol{y}}_k\,|\boldsymbol{\mathcal{F}}_k\right]=\nabla\bar{F}\left(\boldsymbol{x}_k\right)$ is employed. Then we expand $\mathbb{E}\left[\left\|\alpha\left(\nabla f\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{g}}_k\right)\right)+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)\right\|\right]$ as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\alpha\left(\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\bar{\boldsymbol{g}}_k\right)+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)\right\|^2\right]\\
=&\left\|\alpha\left[\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\nabla\bar{F}\left(\boldsymbol{x}_k\right)\right]+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)\right\|^2\\
&+\alpha^2\mathbb{E}\left[\left\|\nabla\bar{F}\left(\boldsymbol{x}_k\right)-\bar{\boldsymbol{g}}_k\right\|^2\right]\\
\leq&\frac{2\alpha^2L_f^2}{m}\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2+\frac{2\beta^2}{m}\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\\
&+\frac{\alpha^2}{m^2}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|^2\right],
\end{aligned}$$

since

$$\begin{aligned}
&\left\langle\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\alpha\nabla\bar{F}\left(\boldsymbol{x}_k\right)+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right),\right.\\
&\left.\alpha\nabla\bar{F}\left(\boldsymbol{x}_k\right)-\alpha\bar{\boldsymbol{g}}_k\right\rangle=0.
\end{aligned}$$

Next, we simplify

$$\begin{aligned}
&2\left\langle\bar{\boldsymbol{x}}_k-\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\tilde{\boldsymbol{x}}^*,\right.\\
&\left.\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\alpha\bar{F}\left(\boldsymbol{x}_k\right)+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)\right\rangle\\
=&2\left\|\bar{\boldsymbol{x}}_k-\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\tilde{\boldsymbol{x}}^*\right\|\\
&\left\|\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\alpha\bar{F}\left(\boldsymbol{x}_k\right)+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)\right\|\\
\leq&\frac{1}{\lambda}\left\|\bar{\boldsymbol{x}}_k-\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\tilde{\boldsymbol{x}}^*\right\|^2\\
&+\lambda\left\|\alpha\nabla f\left(\bar{\boldsymbol{x}}_k\right)-\alpha\bar{F}\left(\boldsymbol{x}_k\right)+\beta\left(\bar{\boldsymbol{x}}_k-\bar{\boldsymbol{x}}_{k-1}\right)\right\|^2.
\end{aligned}$$

Setting λ as $(1-\mu\alpha)/(\mu\alpha)$ and 1 can lead to inequalities (16) and (17), respectively, which completes the proof.

Lemma 7 Considering the sequences $\{\boldsymbol{x}_k\}_{k\geq0}$ and $\{\boldsymbol{y}_k\}_{k\geq0}$ yielded by dynamical system (5), $\forall k\geq0$, if $0<\alpha\leq1/L_f$, the inequality holds as follows:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2\right]\\
\leq&\left(16+16\alpha^2L_f^2\right)\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\
&+16\alpha^2L_f^2\mathbb{E}\left[m\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\|^2\right]+16\alpha^2\mathbb{E}\left[\left\|\nabla F(\boldsymbol{x}_k)-\boldsymbol{g}_k\right\|^2\right]\\
&+16\alpha^2\mathbb{E}\left[\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|^2\right]+2\beta^2\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right].
\end{aligned}\tag{18}$$

Proof From dynamical iterating Eq. (5a), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2\right]\\
\leq&\mathbb{E}\left[\left\|(\boldsymbol{\mathcal{W}}-\boldsymbol{I}_{mn})(\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k)-\alpha\boldsymbol{y}_k+\beta(\boldsymbol{x}_k-\boldsymbol{x}_{k-1})\right\|^2\right]\\
\leq&2\mathbb{E}\left[\left\|\left(\boldsymbol{\mathcal{W}}-\boldsymbol{I}_{mn}\right)\left(\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right)-\alpha\boldsymbol{y}_k\right\|^2\right]\\
&+2\beta^2\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]\\
\leq&16\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]+4\alpha^2\mathbb{E}\left[\left\|\boldsymbol{y}_k\right\|^2\right]\\
&+2\beta^2\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right].
\end{aligned}\tag{19}$$

Let Assumption 2 hold. Then one can obtain

$$\begin{aligned}
&\left\|\boldsymbol{y}_k\right\|\\
\leq&\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|+\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|+L_f\left\|\boldsymbol{x}_k-\boldsymbol{x}^*\right\|\\
\leq&\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|+\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|\\
&+L_f\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|+\sqrt{m}L_f\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|,
\end{aligned}$$

where we used $\bar{\boldsymbol{y}}_k=\bar{\boldsymbol{g}}_k,\forall k\geq0$, for squaring the above inequality to obtain

$$\begin{aligned}
\left\|\boldsymbol{y}_k\right\|^2\leq&4\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|^2+4\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|^2\\
&+4L_f^2\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\|^2+4L_f^2m\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\|^2.
\end{aligned}\tag{20}$$

Therefore, inequality (18) can be proved by plugging inequality (20) into inequality (19).

Lemma 8 Let Assumptions 2 and 3 hold. Regarding the sequence $\{\boldsymbol{y}_k\}$ yielded by dynamical iterating Eq. (5b), $\forall k\geq0$, if $0<\alpha\leq1/\left(4\sqrt{2}L_f\right)$, it holds that

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{y}_{k+1}-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_{k+1}\right\|^2\right]\\
\leq&\left(\frac{1+\sigma^2}{2}+\frac{32\alpha^2L_f^2}{1-\sigma^2}\right)\mathbb{E}\left[\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|^2\right]\\
&+\frac{33L_f^2}{1-\sigma^2}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\
&+\frac{4L_f^2\beta^2}{1-\sigma^2}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]\\
&+\frac{4}{1-\sigma^2}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{x}_{k+1}\right)-\boldsymbol{g}_{k+1}\right\|^2\right]\\
&+\frac{L_f^2}{1-\sigma^2}\mathbb{E}\left[m\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\|^2\right]\\
&+\frac{4+32\alpha^2L_f^2}{1-\sigma^2}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|^2\right].
\end{aligned}\tag{21}$$

Proof From inequality (21), we can obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{y}_{k+1}-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{y}_{k+1}\right\|^2\right]\\=&\mathbb{E}\Big[\left\|\left(\boldsymbol{\mathcal{W}}-\boldsymbol{\mathcal{W}}_{\infty}\right)\left(\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{y}_k\right)\right.\\&\left.+\left(\boldsymbol{I}_{mn}-\boldsymbol{\mathcal{W}}_{\infty}\right)\left(\boldsymbol{g}_{k+1}+\boldsymbol{g}_k\right)\right\|^2\Big].\end{aligned}$$

Then we employ $\left\|\boldsymbol{I}_{mn}-\boldsymbol{\mathcal{W}}_{\infty}\right\|=1$ and Young's inequality to set η as $(1-\sigma^2)/(2\sigma^2)$ in the above equality, and obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{y}_{k+1}-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{y}_{k+1}\right\|^2\right]\\\leq&\frac{1+\sigma^2}{2}\mathbb{E}\left[\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{y}_k\right\|^2\right]\\&+\frac{2}{1-\sigma^2}\mathbb{E}\left[\left\|\boldsymbol{g}_{k+1}-\boldsymbol{g}_k\right\|^2\right].\end{aligned}\tag{22}$$

Next we expand $\mathbb{E}\left[\left\|\boldsymbol{g}_{k+1}-\boldsymbol{g}_k\right\|^2\right]$ as

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{g}_{k+1}-\boldsymbol{g}_k\right\|^2\right]\\\leq&2\mathbb{E}\Big[\left\|\boldsymbol{g}_{k+1}-\nabla F\left(\boldsymbol{x}_{k+1}\right)\right\|^2\Big]+2\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|^2\right]\\&+2\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{x}_{k+1}\right)-\nabla F\left(\boldsymbol{x}_k\right)\right\|^2\right]\\&+4\mathbb{E}\Big[\left\langle\boldsymbol{g}_{k+1}-\nabla F\left(\boldsymbol{x}_{k+1}\right),\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\rangle\Big]\\\leq&2L_f^2\mathbb{E}\left[\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2\right]+2\mathbb{E}\left[\left\|\boldsymbol{g}_{k+1}-\nabla F\left(\boldsymbol{x}_{k+1}\right)\right\|^2\right]\\&+2\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|^2\right],\end{aligned}\tag{23}$$

where in inequality (23), we use the fact that

$$\mathbb{E}\Big[\left\langle\boldsymbol{g}_{k+1}-\nabla F\left(\boldsymbol{x}_{k+1}\right),\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\rangle\Big]=0.$$

Using the bound of inequality (23) in inequality (22), we obtain inequality (21), which completes the proof.

Lemma 9 Considering $\{r_k\}$, the following holds:

$$\begin{aligned}\mathbb{E}\left[r_{k+1}\right]\leq&\left(1-\frac{1}{P}\right)\mathbb{E}\left[r_k\right]+\frac{2}{p}\mathbb{E}\left[m\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\|^2\right]\\&+\frac{2}{p}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{x}_k\right\|^2\right].\end{aligned}\tag{24}$$

Proof From Algorithm 1, if $j=t_{k+1}^i$, then $\boldsymbol{v}_{k+1}^{i,j}=\boldsymbol{x}_{k+1}^{i,j}$ and $\boldsymbol{v}_{k+1}^{i,j}=\boldsymbol{v}_k^{i,j}$. Then we can obtain

$$\begin{aligned}&\mathbb{E}\left[r_{k+1}\right]\\=&\frac{1}{p_i}\sum\nolimits_{j=1}^{p_i}\mathbb{E}\left[\left\|\boldsymbol{v}_{k+1}^{i,j}-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\=&\frac{1}{p_i}\sum\nolimits_{j=1}^{p_i}\mathbb{E}\left[\left(1-\frac{1}{p_i}\right)\left\|\boldsymbol{v}_{k+1}^{i,j}-\tilde{\boldsymbol{x}}^*\right\|^2+\frac{1}{p_i}\left\|\boldsymbol{x}_k^i-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\=&\left(1-\frac{1}{p_i}\right)r_k^i+\frac{1}{p_i}\left\|\boldsymbol{x}_k^i-\tilde{\boldsymbol{x}}^*\right\|^2\\\leq&\left(1-\frac{1}{P}\right)r_k^i+\frac{2}{p}\left\|\boldsymbol{x}_k^i-\bar{\boldsymbol{x}}_k\right\|^2+\frac{2}{p}\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\|^2.\end{aligned}$$

In the next lemma, we derive an upper bound on $\mathbb{E}\left[\left\|\boldsymbol{g}_k-\nabla F\left(\boldsymbol{x}_k\right)\right\|^2\right]$, and then define $r_k^i=\frac{1}{p_i}\sum_{j=1}^{p_i}\left\|\boldsymbol{v}_k^{i,j}-\tilde{\boldsymbol{x}}^*\right\|$ and $r_k=\sum_i^m r_k^i$, recalling that $P=\max_i\{p_i\}$ and $p=\min_i\{p_i\}$.

Lemma 10 Let Assumption 2 hold. Considering the sequence $\{\boldsymbol{g}_k\}$ yielded by dynamical system (5), $\forall k\geq 0$, the inequality is obtained as follows:

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{g}_k-\nabla F\left(\boldsymbol{x}_k\right)\right\|^2\right]\\\leq&4L_f^2\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{x}_k\right\|^2\right]\\&+4L_f^2\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\right]+2L_f^2\mathbb{E}\left[r_k\right].\end{aligned}\tag{25}$$

Proof Recalling Algorithm 1, we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{g}_k^i-\nabla f_i\left(\boldsymbol{x}_k^i\right)\right\|^2\right]\\=&\mathbb{E}\left[\left\|\nabla f_i^{q_k^i}\left(\boldsymbol{x}_k^i\right)-\nabla f_i^{q_k^i}\left(\boldsymbol{v}_k^i\right)-\nabla f_i\left(\boldsymbol{x}_k^i\right)\right.\right.\\&\left.\left.+\frac{1}{p_i}\sum_{j=1}^{p_i}\nabla f_i^j\left(\boldsymbol{v}_k^{i,j}\right)\right\|^2\right]\\\leq&\mathbb{E}\left[\left\|\nabla f_i^{t_k^i}\left(\boldsymbol{x}_k^i\right)-\nabla f_i^{t_k^i}\left(\boldsymbol{v}_k^i\right)\right\|^2\right]\\=&\frac{1}{p_i}\sum_{j=1}^{p_i}\left\|\nabla f_i^j\left(\boldsymbol{x}_k^i\right)-\nabla f_i^j(\tilde{\boldsymbol{x}}^*)+\nabla f_i^j(\tilde{\boldsymbol{x}}^*)-\nabla f_i^j\left(\boldsymbol{v}_k^i\right)\right\|^2\\\leq&2\left\|\nabla f_i\left(\boldsymbol{x}_k^i\right)-\nabla f_i\left(\tilde{\boldsymbol{x}}^*\right)\right\|^2+2\left\|\nabla f_i^j\left(\tilde{\boldsymbol{x}}^*\right)-\nabla f_i^j\left(\boldsymbol{v}_k^i\right)\right\|^2\\\leq&2L_f^2\left\|\boldsymbol{x}_k^i-\tilde{\boldsymbol{x}}^*\right\|^2+2L_f^2\left\|\boldsymbol{v}_k^i-\tilde{\boldsymbol{x}}^*\right\|^2\\\leq&4L_f^2\left\|\boldsymbol{x}_k^i-\bar{\boldsymbol{x}}_k\right\|^2+4L_f^2\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2+2L_f^2r_k^i.\end{aligned}$$

Corollary 1 Let Assumptions 2 and 3 hold. Considering the sequence $\{\boldsymbol{g}_k\}_{k\geq 0}$ yielded by dynamical system (5), $\forall k\geq 0$, we can obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{g}_{k+1}-\nabla F\left(\boldsymbol{x}_{k+1}\right)\right\|^2\right]\\\leq&\left(8L_f^2\sigma^2+\frac{4L_f^2}{p}+16L_f^2\alpha^2L_f^2\right)\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{x}_k\right\|^2\right]\\&+16L_f^2\alpha^2\mathbb{E}\Big[\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_{\infty}\boldsymbol{y}_k\right\|^2\Big]+32L_f^2\beta^2\mathbb{E}\Big[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\Big]\\&+\left(2+\frac{1}{p}\right)4L_f^2\mathbb{E}\Big[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\Big]+\left(1-\frac{1}{P}\right)2L_f^2\mathbb{E}[r_k]\\&+\frac{4L_f^2\alpha^2}{m}\mathbb{E}\left[\left\|\nabla F\left(\boldsymbol{x}_k\right)-\boldsymbol{g}_k\right\|^2\right].\end{aligned}\tag{26}$$

Proof From Lemma 10, it is straightforward to obtain

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{g}_{k+1}-\nabla F\left(\boldsymbol{x}_{k+1}\right)\right\|^2\right]\\ \leq&4L_f^2\mathbb{E}\left[\left\|\boldsymbol{x}_{k+1}-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_{k+1}\right\|^2\right]\\&+4L_f^2\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_{k+1}-\tilde{\boldsymbol{x}}^*\right\|^2\right]+2L_f^2\mathbb{E}\left[r_{k+1}\right].\end{aligned}$$

Then, we can expand the above inequality by plugging inequalities (15), (16), and (24), from Lemmas 5, 6, and 9, respectively, which completes the proof.

3.3 Main results

To summarize, the bound of the gradient variance is obtained to process the inequality, which is obtained from dynamical system (5), and then we start to derive the convergence rate of the proposed algorithm. First, the upper bounds on $\mathbb{E}\left[\left\|\boldsymbol{g}_k-\nabla F\left(\boldsymbol{x}_k\right)\right\|^2\right]$ and $\mathbb{E}\left[\left\|\boldsymbol{g}_{k+1}-\nabla F\left(\boldsymbol{x}_{k+1}\right)\right\|^2\right]$ are obtained from Lemma 10 and Corollary 1, respectively, and then $\forall k\geq 0$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{y}_{k+1}-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_{k+1}\right\|^2\right]\\ \leq&\left(32\sigma^2+\frac{16}{p}+55+\frac{2}{m}\right)\frac{L_f^2}{1-\sigma^2}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\&+\frac{132L_f^2\beta^2}{1-\sigma^2}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]\\&+\frac{L_f^2}{1-\sigma^2}\left(18-\frac{8}{P}+\frac{1}{m}\right)\mathbb{E}[r_k]\\&+\left(\frac{1+\sigma^2}{2}+\frac{96\alpha^2L_f^2}{1-\sigma^2}\right)\mathbb{E}\left[\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|^2\right]\\&+\frac{L_f^2}{1-\sigma^2}\left(53+\frac{16}{p}+\frac{2}{m}\right)\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\right].\end{aligned}\tag{27}$$

Second, we refine the following inequality by applying the upper bound on $\mathbb{E}\left[\left\|\boldsymbol{g}_k-\nabla F\left(\boldsymbol{x}_k\right)\right\|^2\right]$ from Lemma 10: $\forall k\geq 0$,

$$\begin{aligned}&\mathbb{E}\left[\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2\right]\\ \leq&\left(16+80\alpha^2L_f^2\right)\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\&+2\beta^2\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]+16\alpha^2\mathbb{E}\left[\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|^2\right]\\&+80\alpha^2L_f^2\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\right]+32\alpha^2L_f^2\mathbb{E}\left[r_k\right].\end{aligned}\tag{28}$$

Next, combining inequalities (17) and (25) yields

$$\begin{aligned}&\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_{k+1}-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\ \leq&\left(1-\mu\alpha\right)\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_{k+1}-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\&+\frac{2\alpha L_f^2}{\mu}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\&+\frac{2\beta^2}{\mu\alpha}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]+\frac{\alpha^2}{m}4L_f^2\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\&+4\frac{\alpha^2}{m}L_f^2\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\right]+2\frac{\alpha^2}{m}L_f^2\mathbb{E}\left[r_k\right]\\ \leq&\left(1-\mu\alpha+\frac{4L_f^2\alpha^2}{m}\right)\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\&+\alpha L_f^2\left(\frac{2}{\mu}+\frac{4\alpha}{m}\right)\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\&+\frac{2\beta^2}{\mu\alpha}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]+2\frac{\alpha^2}{m}L_f^2\mathbb{E}\left[r_k\right]\\ \leq&\left(1-\mu\alpha+\frac{4L_f^2\alpha^2}{m}\right)\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\&+\alpha L_f^2\left(\frac{2}{\mu}+\frac{4\alpha}{m}\right)\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\&+\frac{2\beta^2}{\mu\alpha}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]+\frac{2L_f^2\alpha^2}{m}\mathbb{E}\left[r_k\right].\end{aligned}$$

If $0<\alpha\leq 1/(2\mu)$, then $2/\mu+4\alpha/m\leq 4/\mu$; if $0<\alpha\leq \mu m/(8L^2)$, then $1-\mu\alpha+4L_f^2\alpha^2/m\leq 1-\mu\alpha/2$. Therefore, we can obtain the following inequality:

$$\begin{aligned}&\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_{k+1}-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\ \leq&\left(1-\frac{\mu\alpha}{2}\right)\mathbb{E}\left[m\left\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\right\|^2\right]\\&+\frac{4\alpha L_f^2}{\mu}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\&+\frac{2\beta^2}{\mu\alpha}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]+\frac{2L_f^2\alpha^2}{m}\mathbb{E}\left[r_k\right].\end{aligned}\tag{29}$$

Now we combine inequalities (14), (24), (27), and (28) with inequality (29) as a linear matrix to prepare for the subsequent derivation.

Proposition 1 Let Assumptions 1–3 hold. $\forall k\geq 1$, the following inequality holds entry-wise:

$$\boldsymbol{J}_{k+1}\leq\boldsymbol{G}_\alpha\boldsymbol{J}_k,\tag{30}$$

where $\boldsymbol{J}_k\in\mathbb{R}^5$ is given by

$$\boldsymbol{J}_k=\begin{bmatrix}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{x}_k\right\|^2\right]\\ \mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]\\ \mathbb{E}\left[\left\|\boldsymbol{y}_k-\boldsymbol{\mathcal{W}}_\infty\boldsymbol{y}_k\right\|^2\right]\\ \mathbb{E}\left[m\|\bar{\boldsymbol{x}}_k-\tilde{\boldsymbol{x}}^*\|^2\right]\\ \mathbb{E}\left[r_k\right]\end{bmatrix},$$

and $\boldsymbol{G}_\alpha \in \mathbb{R}^{5\times 5}$ is given at the bottom of this page, where $a_1 = 16/p + 55 + 2/m, a_2 = 16/p + 53 + 2/m, a_3 = 18 - 8/P + 1/m$. Obviously, it is feasible to obtain the range of α to satisfy $\rho(\boldsymbol{G}_\alpha) < 1$. To achieve this, we go to the next lemma.

Lemma 11 Let $\boldsymbol{H} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{x} \in \mathbb{R}^d$ be nonnegative matrix and positive vector, respectively. For $\gamma > 0$, if $\boldsymbol{H}\boldsymbol{x} \le \gamma\boldsymbol{x}$, then $\rho(\boldsymbol{H}) \le |||\boldsymbol{H}|||_\infty^{\boldsymbol{x}} \le \gamma$.

Theorem 1 Let Assumptions 1–3 hold. If stepsize α satisfies $0 < \alpha < \frac{p}{P}\frac{(1-\sigma^2)^2}{549QL_f}$, the proposed algorithm is linearly convergent.

Proof Let positive vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \delta_3, \delta_4, \delta_5]^{\mathrm{T}}$. The linear matrix inequality holds as follows:

$$\boldsymbol{G}_\alpha\boldsymbol{\delta} < \boldsymbol{\delta},$$

which can be equivalently written as the following inequalities:

$$0 < \frac{1-\sigma^2}{2} - \frac{4\beta^2}{1-\sigma^2}\frac{\delta_2}{\delta_1} - \frac{4\alpha^2}{1-\sigma^2}\frac{\delta_3}{\delta_1}, \tag{31}$$

$$\begin{aligned} 0 <& \left(1-2\beta^2\right) - \left(16+80\alpha^2L_f^2\right)\frac{\delta_1}{\delta_2} - 16\alpha^2\frac{\delta_3}{\delta_2} \\ & - 80\alpha^2L_f^2\frac{\delta_4}{\delta_2} - 32\alpha^2L_f^2\frac{\delta_5}{\delta_2}, \end{aligned} \tag{32}$$

$$\begin{aligned} 0 <& \left(\frac{1-\sigma^2}{2} - \frac{96\alpha^2L_f^2}{1-\sigma^2}\right)\delta_3 - \frac{L_f^2}{1-\sigma^2}\left(32\sigma^2+a_1\right)\delta_1 \\ & - \frac{L_f^2}{1-\sigma^2}a_2\delta_4 - \frac{L_f^2}{1-\sigma^2}a_3\delta_5, \end{aligned} \tag{33}$$

$$0 < \frac{\mu\alpha}{2} - \frac{4\alpha L_f^2}{\mu}\frac{\delta_1}{\delta_4} - \frac{2\beta^2}{\mu\alpha}\frac{\delta_2}{\delta_4} - \frac{2L_f^2\alpha^2}{m}\frac{\delta_5}{\delta_4}, \tag{34}$$

$$0 < \frac{1}{P} - \frac{2}{p}\frac{\delta_1}{\delta_5} - \frac{2}{p}\frac{\delta_4}{\delta_5}. \tag{35}$$

It is evident that for the right-hand sides (RHSs) of inequalities (31)–(35) to be positive, we can bound the range of α. First, we are supposed to fix vector $\boldsymbol{\delta}$, which is positive and independent of α and β. We can set $\delta_1 = 1$, $\delta_2 = 17$, and $Q = L_f/\mu$, and the following inequality is obtained from inequality (34):

$$\frac{2\beta^2}{\mu\alpha}\frac{\delta_2}{\delta_4} < \frac{\mu\alpha}{2} - \frac{4\alpha L_f^2}{\mu}\frac{\delta_1}{\delta_4} - \frac{2L_f^2\alpha^2}{m}\frac{\delta_5}{\delta_4}. \tag{36}$$

Because the RHS of inequality (36) is positive, it is straightforward that

$$\delta_4 > \frac{8L_f^2}{\mu^2}\delta_1 = 8Q^2. \tag{37}$$

Therefore, we set $\delta_4 = 80Q^2$, and obtain the following inequality because the RHS of inequality (35) is positive:

$$\delta_5 > \frac{2P}{p}\left(\delta_1+\delta_4\right) = \frac{2P}{p} + \frac{160P}{p}Q^2. \tag{38}$$

We denote $\delta_5 = 164PQ^2/p$. Then, from inequality (33), we obtain

$$\begin{aligned} & \frac{132L_f^2\beta^2}{1-\sigma^2}\delta_2 + \frac{96\alpha^2L_f^2}{1-\sigma^2}\delta_3 \\ <& \frac{1-\sigma^2}{2}\delta_3 - \left(\frac{105L_f^2}{1-\sigma^2}\delta_1 + \frac{71L_f^2}{1-\sigma^2}\delta_4 + \frac{19L_f^2}{1-\sigma^2}\delta_5\right). \end{aligned} \tag{39}$$

Because the RHS of inequality (39) is positive, the following inequality is obtained:

$$\delta_3 > \frac{2L_f^2}{\left(1-\sigma^2\right)^2}\left(105 + 5680Q^2 + \frac{3116PQ^2}{p}\right). \tag{40}$$

Because $105 + 5680Q^2 + 3116PQ^2/p < 8901PQ^2/p$, we can set $\delta_3 = \frac{18\,050L_f^2}{(1-\sigma^2)^2}\frac{PQ^2}{p}$. Until now, we have fixed values of δ_1–δ_5 as follows:

$$\begin{aligned} & \delta_1 = 1, \ \ \delta_2 = 17, \ \ \delta_3 = \frac{18\,050L_f^2}{\left(1-\sigma^2\right)^2}\frac{PQ^2}{p}, \\ & \delta_4 = 80Q^2, \ \ \delta_5 = \frac{164PQ^2}{p}. \end{aligned} \tag{41}$$

The range of α that satisfies forward inequalities can

$$\boldsymbol{G}_\alpha = \begin{bmatrix} \frac{1-\sigma^2}{2} & \frac{4\beta^2}{1-\sigma^2} & \frac{4\alpha^2}{1-\sigma^2} & 0 & 0 \\ 16+80\alpha^2L_f^2 & 2\beta^2 & 16\alpha^2 & 80\alpha^2L_f^2 & 32\alpha^2L_f^2 \\ \frac{L_f^2}{1-\sigma^2}\left(32\sigma^2+a_1\right) & \frac{132L_f^2\beta^2}{1-\sigma^2} & \frac{1+\sigma^2}{2} + \frac{96\alpha^2}{1-\sigma^2} & \frac{L_f^2}{1-\sigma^2}a_2 & \frac{L_f^2}{1-\sigma^2}a_3 \\ \frac{4\alpha L_f^2}{\mu} & \frac{2\beta^2}{\mu\alpha} & 0 & 1-\frac{\mu\alpha}{2} & \frac{2L_f^2\alpha^2}{m} \\ \frac{2}{p} & 0 & 0 & \frac{2}{p} & 1-\frac{1}{P} \end{bmatrix}.$$

thus be found. From inequality (31), we obtain

$$\begin{cases} \dfrac{4\beta^2}{1-\sigma^2}\delta_2 < \dfrac{1-\sigma^2}{2}\delta_1 - \dfrac{4\alpha^2}{1-\sigma^2}\delta_3, \\ \alpha^2 < \dfrac{\left(1-\sigma^2\right)^2}{8}\dfrac{\delta_1}{\delta_3}. \end{cases}$$

Then we have

$$\begin{aligned} \alpha &< \sqrt{\frac{\left(1-\sigma^2\right)^2}{8}\frac{\delta_1}{\delta_3}} = \sqrt{\frac{\left(1-\sigma^2\right)^2}{8}\frac{\left(1-\sigma^2\right)^2}{18\,050L_f^2}} \\ &= \frac{\left(1-\sigma^2\right)^2}{380}\frac{\sqrt{p}}{\sqrt{P}Q}. \end{aligned} \tag{42}$$

From inequality (32), we obtain

$$\begin{aligned} \alpha &< \sqrt{\frac{1}{80L_f^2\delta_1 + 16\delta_3 + 80L_f^2\delta_4 + 32L_f^2\delta_5}} \\ \Leftarrow \alpha &< \sqrt{\frac{1}{\left(80+16\frac{18\,050PQ^2}{\left(1-\sigma^2\right)^2p}+6400Q^2+\frac{5248PQ^2}{p}\right)L_f^2}} \\ \Leftarrow \alpha &< \sqrt{\frac{\left(1-\sigma^2\right)^2p}{300\,528PQ^2L_f^2}} \Leftrightarrow \alpha < \frac{1-\sigma^2}{12\sqrt{2087}}\frac{\sqrt{p}}{\sqrt{P}QL_f}. \end{aligned} \tag{43}$$

From inequality (33), we can obtain

$$\begin{aligned} \frac{96\alpha^2L_f^2}{1-\sigma^2}\delta_3 &< \frac{1-\sigma^2}{2}\delta_3 - \frac{L_f^2}{1-\sigma^2}\left(105\delta_1+71\delta_4+19\delta_5\right) \\ \Leftarrow \frac{96\alpha^2L_f^2}{1-\sigma^2}\delta_3 &< \frac{1-\sigma^2}{2}\delta_3 - \frac{L_f^2}{1-\sigma^2}\frac{8901PQ^2}{p} \\ \Leftarrow \frac{96\alpha^2L_f^2}{1-\sigma^2} &< \frac{1-\sigma^2}{2} - \frac{8901\left(1-\sigma^2\right)}{18\,050} = \frac{62\left(1-\sigma^2\right)}{9025} \\ \Leftrightarrow \alpha &< \frac{\sqrt{31}\left(1-\sigma^2\right)}{380\sqrt{3}L_f}. \end{aligned} \tag{44}$$

Finally, we can obtain the following inequality by applying inequality (34):

$$\frac{4L_f^2\alpha}{\mu m}\delta_5 < \delta_4 - \frac{8L_f^2}{\mu^2}\delta_1 - \frac{4\beta^2}{\mu^2\alpha^2}\delta_2.$$

Then we can obtain inequalities as follows:

$$\begin{cases} \alpha < \dfrac{\mu m}{4L_f^2}\dfrac{\delta_4}{\delta_5} - \dfrac{2m}{\mu}\dfrac{\delta_1}{\delta_5}, \\ \alpha < \dfrac{m}{4QL_f}\dfrac{80p}{164P} - \dfrac{2m}{\mu}\dfrac{p}{164PQ^2}, \\ \alpha < \dfrac{9}{82}\dfrac{p}{P}\dfrac{m}{QL_f}. \end{cases} \tag{45}$$

Therefore, α satisfies

$$\begin{aligned} \alpha < \bar{\alpha} \triangleq \min &\left\{ \frac{\left(1-\sigma^2\right)^2}{380}\frac{\sqrt{p}}{\sqrt{P}Q}, \frac{1-\sigma^2}{12\sqrt{2087}}\frac{\sqrt{p}}{\sqrt{P}QL_f}, \right. \\ &\left. \frac{\sqrt{31}\left(1-\sigma^2\right)}{380\sqrt{3}L_f}, \frac{m}{QL_f}\frac{9p}{82P} \right\} \\ \Leftarrow \alpha &< \frac{p}{P}\frac{\left(1-\sigma^2\right)^2}{549QL_f}. \end{aligned}$$

Next, we set $\alpha = \frac{p}{P}\frac{\left(1-\sigma^2\right)^2}{560QL_f}$ according to the previous range of α to obtain β. First, from inequality (31), we have

$$\begin{aligned} \frac{4\beta^2}{1-\sigma^2}\delta_2 &< \frac{1-\sigma^2}{2} - \frac{4}{1-\sigma^2}\frac{18\,050}{\left(1-\sigma^2\right)^2}\frac{p}{P}\frac{\left(1-\sigma^2\right)^4}{313\,600} \\ &= \frac{1-\sigma^2}{2} - \frac{p}{P}\frac{361\left(1-\sigma^2\right)}{1568} \\ \Leftarrow \frac{4\beta^2}{1-\sigma^2}\delta_2 &< \frac{9\times47\left(1-\sigma^2\right)}{1568}\frac{p}{P} \\ \Leftarrow \beta &< \frac{3\sqrt{47}\left(1-\sigma^2\right)}{56\sqrt{34}}\frac{\sqrt{p}}{\sqrt{P}}. \end{aligned} \tag{46}$$

Second, from inequality (32), we have

$$\begin{aligned} 2\beta^2\delta_2 &< \delta_2 - 16\delta_1 - 80L_f^2\delta_1\alpha^2 - 16\delta_3\alpha^2 \\ &\quad - 80L_f^2\delta_4\alpha^2 - 32L_f^2\delta_5\alpha^2 \\ \Leftarrow 2\beta^2\delta_2 &< 1 - \frac{p^2}{P^2}\frac{\left(1-\sigma^2\right)^4}{3920Q^2} - \frac{p}{P}\frac{361\left(1-\sigma^2\right)^2}{392} \\ &\quad - \frac{p^2}{P^2}\frac{47\left(1-\sigma^2\right)^4}{1568} - \frac{p}{P}\frac{82\left(1-\sigma^2\right)^4}{4900} \\ \Leftarrow \beta &< \frac{\sqrt{1259}\left(1-\sigma^2\right)}{280\sqrt{17}}\frac{\sqrt{p}}{\sqrt{P}}. \end{aligned} \tag{47}$$

Then from inequality (33), we have

$$\begin{aligned} \frac{132L_f^2\beta^2}{1-\sigma^2}\delta_2 &< \frac{1-\sigma^2}{2}\delta_3 - \frac{105L_f^2}{1-\sigma^2}\delta_1 - \frac{96\alpha^2L_f^2}{1-\sigma^2}\delta_3 \\ &\quad - \frac{71L_f^2}{1-\sigma^2}\delta_4 - \frac{19L_f^2}{1-\sigma^2}\delta_5 \\ \Leftarrow \frac{132L_f^2\beta^2}{1-\sigma^2}\delta_2 &< \frac{9025L_f^2}{1-\sigma^2}\frac{PQ^2}{p} - \frac{361}{196}\frac{PQ^2}{p}\frac{8904L_f^2}{1-\sigma^2} \\ \Leftrightarrow \beta &< \frac{\sqrt{59}}{\sqrt{1122}}\frac{\sqrt{P}Q}{\sqrt{p}}. \end{aligned} \tag{48}$$

Finally, from inequality (34), we have

$$\begin{cases} \dfrac{4\beta^2}{\mu^2\alpha^2}\delta_2 < \delta_4 - \dfrac{8L_f^2}{\mu^2}\delta_1 - \dfrac{4L_f^2\alpha}{\mu m}\delta_5, \\ \beta < \dfrac{\sqrt{69}\left(1-\sigma^2\right)^2}{560\sqrt{68}}\dfrac{p}{PQ}. \end{cases} \tag{49}$$

Therefore, we obtain the range of β from inequalities (46)–(49) as follows:

$$\beta<\bar{\beta}\triangleq\min\left\{\frac{3\sqrt{47p}\left(1-\sigma^2\right)}{56\sqrt{34P}},\frac{\sqrt{1259p}\left(1-\sigma^2\right)}{280\sqrt{17P}},\right.$$
$$\left.\frac{\sqrt{59}}{\sqrt{1122}}\frac{\sqrt{P}Q}{\sqrt{p}},\frac{\sqrt{69}\left(1-\sigma^2\right)^2}{560\sqrt{68}}\frac{p}{PQ}\right\},$$

which completes the proof.

Remark 2 It is worth emphasizing that this study does not establish explicit expression of the accelerated convergence rate from a theoretical point of view; the acceleration of the proposed algorithm can be demonstrated only from the practical point of view. The acceleration in this study can be observed through a series of experiments because this is still an open problem in recent literature (Xin et al., 2019a; Li HQ et al., 2020; Lü et al., 2020; Xin and Khan, 2020; Hu et al., 2021). Therefore, it would be worthwhile to analyze the acceleration from a theoretical point of view in future work.

4 Experimental results

In this section, we compare the proposed algorithm with other methods by solving some optimization problems, such as the distributed logistic regression problem, distributed least-squares regression problem, and distributed quadratic programming problem, to prove the proposed algorithm's practicability. The performances of all the tested algorithms are plotted by $\frac{1}{m}\sum_{i=1}^{m}\left\|\boldsymbol{x}_k^i-\tilde{\boldsymbol{x}}^*\right\|$.

4.1 Distributed logistic regression

This subsection illustrates the performance comparison between different algorithms. We set $N=\sum_{i=1}^{m}q_i$ as the number of samples and allocate q_i to m agents on average, i.e., $q_i=N/m$. The undirected network to distributed logistic regression problem is given as follows:

$$\tilde{\boldsymbol{x}}^*=\arg\min_{\tilde{\boldsymbol{x}}\in\mathbb{R}^n}\left(\frac{\gamma}{2}\|\tilde{\boldsymbol{x}}\|^2\right.$$
$$\left.+\sum_{i=1}^{m}\sum_{j=1}^{q_i}\log\left(1+\exp\left(-b_{ij}\boldsymbol{c}_{ij}^{\mathrm{T}}\tilde{\boldsymbol{x}}\right)\right)\right),\tag{50}$$

where $\boldsymbol{c}_{ij}\in\mathbb{R}^n$ is the j^{th} sample and $b_{ij}\in\{1,-1\}$ is the corresponding binary label. To avoid over-fitting, we add a regularization term. According to Eq. (50), the local objective function f_i corresponding to P0 is written as follows:

$$f_i\left(\tilde{\boldsymbol{x}}\right)=\frac{\gamma}{2}\|\tilde{\boldsymbol{x}}\|^2+\sum_{j=1}^{q_i}\log\left(1+\exp\left(-b_{ij}\boldsymbol{c}_{ij}^{\mathrm{T}}\tilde{\boldsymbol{x}}\right)\right).\tag{51}$$

We aim to solve the logistic regression problem by choosing the Wisconsin breast cancer (diagnostic) dataset in the UCI Machine Learning Repository (Dua and Graff, 2017). These samples are from a digitized image of a fine needle aspirate of a breast mass. By judging the average of distances from the center to points on the silhouette and the severity of concave portions of the contour, a patient's condition is predicted as malignant or benign. We randomly choose $N=500$ samples from the total of 683 dataset samples to train the discriminator $\tilde{\boldsymbol{x}}$ and suppose that samples are equally distributed to each agent; the rest of the samples are used for testing. Fig. 2 presents the performance comparison of the proposed algorithm, GT-SAGA, and EXTRA over an undirected network. All step-sizes are set as $\alpha=0.05$ and the heavy-ball momentum acceleration term $\beta=0.05$ in all the algorithms is the same. The accuracy, in the vertical axis of Fig. 3, is the ratio of the number of correct predictions to the total number of samples in the testing set.

4.2 Least-squares method of distributed signal processing

This subsection studies a least-squares problem provided in Li Z et al. (2019), which considers an undirected network of $m=10$ agents for an unknown signal $\tilde{\boldsymbol{x}}\in\mathbb{R}^n$ for the optimization problem:

$$\min_{\tilde{\boldsymbol{x}}\in\mathbb{R}^n}\tilde{f}\left(\tilde{\boldsymbol{x}}\right)\triangleq\sum_{i=1}^{m}\frac{1}{2}\|\boldsymbol{C}_i\tilde{\boldsymbol{x}}-\boldsymbol{d}_i\|^2,\tag{52}$$

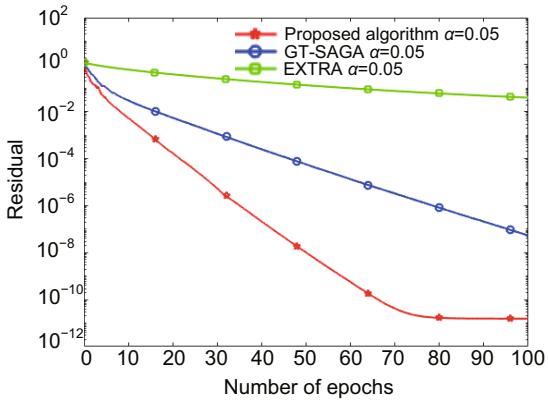

Fig. 2 Logistic regression over an undirected network

where the local objective function maintained by agent i $(i=1,2,\cdots,m)$ is given by

$$f_i(\tilde{\boldsymbol{x}})=\frac{1}{q_i}\sum\nolimits_{h=1}^{q_i}\left\|\boldsymbol{C}_i^h\tilde{\boldsymbol{x}}-\boldsymbol{d}_i^h\right\|^2. \tag{53}$$

From problem (52), we can see that $\boldsymbol{C}_i\in\mathbb{R}^{q_i\times n}$ is the sensing matrix and that $\boldsymbol{d}_i=\boldsymbol{C}_i\tilde{\boldsymbol{x}}+\boldsymbol{e}_i$ is the measurement, where $\boldsymbol{e}_i\in\mathbb{R}^{q_i}$ is the independent and identically distributed noise. When $q_i=600$ and $m=10$, Fig. 4 illustrates the proposed algorithm, GT-SAGA, and EXTRA algorithms over an undirected network to verify the acceleration of the proposed algorithm.

4.3 Distributed quadratic programming

In this subsection, the problem of quadratic programming can be resolved by an undirected network of $m=10$ agents as follows:

$$\min_{\tilde{\boldsymbol{x}}\in\mathbb{R}^n}\tilde{f}(\tilde{\boldsymbol{x}})=\sum_{i=1}^{m}\tilde{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{G}_i\tilde{\boldsymbol{x}}+\boldsymbol{c}_i^{\mathrm{T}}\tilde{\boldsymbol{x}}, \tag{54}$$

where matrix $\boldsymbol{G}_i\in\mathbb{R}^{n\times n}$ is diagonal and positive definite and $\boldsymbol{c}_i\in\mathbb{R}^n$ is a randomly generated vector.

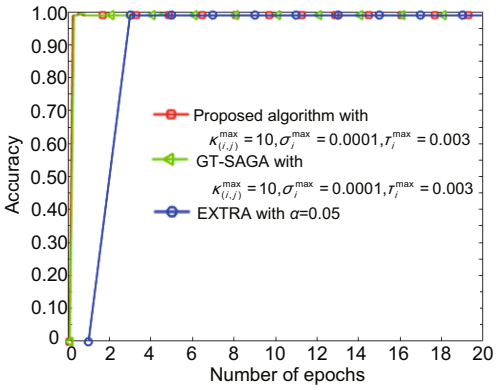

Fig. 3 Comparison of accuracy performance

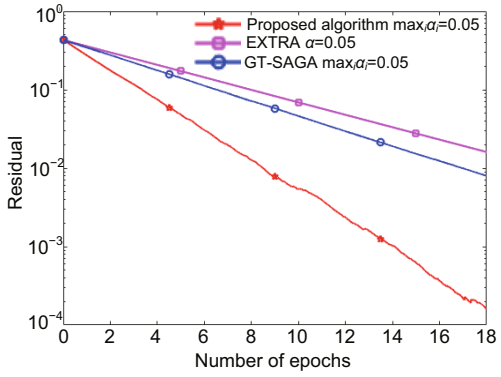

Fig. 4 Performance comparison over the least-squares method of distributed signal processing

We assume the dimension $n=10$. Then we set the step-size $\alpha=0.0008$ in all the algorithms and the momentum parameter $\beta=0.0008$. Fig. 5 shows the acceleration performance of the proposed algorithm compare to those of GT-SAGA and EXTRA over an undirected network.

Remark 3 In view of Tables 1–3, one can clearly see that in different experiments, there are fewer epochs of the proposed algorithm than those of the GT-SAGA and EXTRA algorithms under the same accuracy. This demonstrates that the proposed algorithm accelerates convergence. So, the acceleration of the proposed algorithm can be verified.

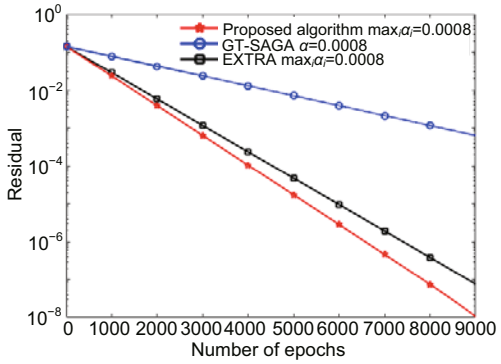

Fig. 5 Performance comparison over an undirected network when condition number $Q=500$

Table 1 Convergence performance comparison over logistic regression

Algorithm	Number of epochs		
	Accuracy=10^{-2}	10^{-4}	10^{-6}
Proposed algorithm	8	22	35
GT-SAGA	16	46	79
EXTRA	193	549	935

Table 2 Convergence performance comparison over least-squares

Algorithm	Number of epochs	
	Accuracy=10^{-1}	10^{-2}
Proposed algorithm	3	18
GT-SAGA	6	8
EXTRA	7	20

Table 3 Convergence performance comparison over the distributed quadratic programming

Algorithm	Number of epochs	
	Accuracy=10^{-1}	10^{-2}
Proposed algorithm	1472	2747
GT-SAGA	4412	8237
EXTRA	1657	3097

5 Conclusions

We presented a distributed stochastic algorithm which is capable of solving large-scale optimization problems over undirected networks. We showed that the proposed algorithm achieves accelerated linear convergence with a constant step-size α. Extensive experiments on real-world datasets illustrated that the performance of the proposed algorithm is superior to those of other comparable algorithms.

Contributors

Bihao SUN designed the research, processed the data, and drafted the manuscript. Jinhui HU, Dawen XIA, and Huaqing LI helped organize the manuscript and process the data. Bihao SUN and Huaqing LI revised and finalized the paper.

Compliance with ethics guidelines

Bihao SUN, Jinhui HU, Dawen XIA, and Huaqing LI declare that they have no conflict of interest.

References

Bertsekas D, Gafni E, 1983. Projected Newton methods and optimization of multicommodity flows. *IEEE Trans Autom Contr*, 28(12):1090-1096. https://doi.org/10.1109/TAC.1983.1103183

Boyd S, Parikh N, Chu E, et al., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends® Mach Learn*, 3(1):1-122. https://doi.org/10.1561/2200000016

Cheng B, Li ZK, 2019. Coordinated tracking control with asynchronous edge-based event-triggered communications. *IEEE Trans Autom Contr*, 64(10):4321-4328. https://doi.org/10.1109/TAC.2019.2895927

Cheng S, Chen MY, Wai RJ, et al., 2014. Optimal placement of distributed generation units in distribution systems via an enhanced multi-objective particle swarm optimization algorithm. *J Zhejiang Univ-Sci C (Comput & Electron)*, 15(4):300-311. https://doi.org/10.1631/jzus.C1300250

Cohen K, Nedić A, Srikant R, 2017. Distributed learning algorithms for spectrum sharing in spatial random access wireless networks. *IEEE Trans Autom Contr*, 62(6):2854-2869. https://doi.org/10.1109/TAC.2016.2626578

Defazio A, Bach F, Lacoste-Julien S, 2014. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. Proc 27th Int Conf on Neural Information Processing Systems, p.1646-1654.

Dua D, Graff C, 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Duchi JC, Agarwal A, Wainwright MJ, 2012. Dual averaging for distributed optimization: convergence analysis and network scaling. *IEEE Trans Autom Contr*, 57(3):592-606. https://doi.org/10.1109/TAC.2011.2161027

Eisen M, Mokhtari A, Ribeiro A, 2017. Decentralized quasi-Newton methods. *IEEE Trans Signal Process*, 65(10):2613-2628. https://doi.org/10.1109/TSP.2017.2666776

Erseghe T, Zennaro D, Dall'Anese E, et al., 2011. Fast consensus by the alternating direction multipliers method. *IEEE Trans Signal Process*, 59(11):5523-5537. https://doi.org/10.1109/TSP.2011.2162831

Guan L, Sun T, Qiao LB, et al., 2020. An efficient parallel and distributed solution to nonconvex penalized linear SVMs. *Front Inform Technol Electron Eng*, 21(4):587-603. https://doi.org/10.1631/FITEE.1800566

Han ZM, Lin ZY, Fu MY, et al., 2015. Distributed coordination in multi-agent systems: a graph Laplacian perspective. *Front Inform Technol Electron Eng*, 16(6):429-448. https://doi.org/10.1631/FITEE.1500118

Hu JH, Yan Y, Li HQ, et al., 2021. Convergence of an accelerated distributed optimisation algorithm over time-varying directed networks. *IET Contr Theory Appl*, 15(1):24-39. https://doi.org/10.1049/cth2.12022

Johnson R, Zhang T, 2013. Accelerating stochastic gradient descent using predictive variance reduction. Proc 26th Int Conf on Neural Information Processing Systems, p.315-323.

Lan Q, Qiao LB, Wang YJ, 2018. Stochastic extra-gradient based alternating direction methods for graph-guided regularized minimization. *Front Inform Technol Electron Eng*, 19(6):755-762. https://doi.org/10.1631/FITEE.1601771

Li HQ, Cheng HQ, Wang Z, et al., 2020. Distributed Nesterov gradient and heavy-ball double accelerated asynchronous optimization. *IEEE Trans Neur Netw Learn Syst*, in press. https://doi.org/10.1109/TNNLS.2020.3027381

Li Z, Shi W, Yan M, 2019. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Trans Signal Process*, 67(17):4494-4506. https://doi.org/10.1109/TSP.2019.2926022

Ling Q, Ribeiro A, 2014. Decentralized dynamic optimization through the alternating direction method of multipliers. *IEEE Trans Signal Process*, 62(5):1185-1197. https://doi.org/10.1109/TSP.2013.2295055

Ling Q, Tian Z, 2010. Decentralized sparse signal recovery for compressive sleeping wireless sensor networks. *IEEE Trans Signal Process*, 58(7):3816-3827. https://doi.org/10.1109/TSP.2010.2047721

Liu R, Sun WC, Hou T, et al., 2019. Block coordinate descentwith time perturbation for nonconvex nonsmooth problems in real-world studies. *Front Inform Technol Electron Eng*, 20(10):1390-1403. https://doi.org/10.1631/FITEE.1900341

Lü QG, Liao XF, Li HQ, et al., 2020. A Nesterov-like gradient tracking algorithm for distributed optimization over directed networks. *IEEE Trans Syst Man Cybern*, in press. https://doi.org/10.1109/TSMC.2019.2960770

Mateos G, Bazerque JA, Giannakis GB, 2010. Distributed sparse linear regression. *IEEE Trans Signal Process*, 58(10):5262-5276. https://doi.org/10.1109/TSP.2010.2055862

Matthews TP, Wang K, Li CP, et al., 2016. Nonlinear waveform inversion by use of the regularized dual averaging method for ultrasound computed tomography. Progress in Electromagnetic Research Symp, p.3948. https://doi.org/10.1109/PIERS.2016.7735487

McMahan B, Moore E, Ramage D, et al., 2017. Communication-efficient learning of deep networks from decentralized data. Proc 20th Int Conf on Artificial Intelligence and Statistics, p.1273-1282.

Nedić A, Ozdaglar A, 2009. Distributed subgradient methods for multi-agent optimization. *IEEE Trans Autom Contr*, 54(1):48-61. https://doi.org/10.1109/TAC.2008.2009515

Nedić A, Olshevsky A, Shi W, 2017a. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM J Optim*, 27(4):2597-2633. https://doi.org/10.1137/16M1084316

Nedić A, Olshevsky A, Shi W, et al., 2017b. Geometrically convergent distributed optimization with uncoordinated step-sizes. American Control Conf, p.3950-3955. https://doi.org/10.23919/ACC.2017.7963560

Schmidt M, Le Roux N, Bach F, 2017. Minimizing finite sums with the stochastic average gradient. *Math Program*, 162(1-2):83-112. https://doi.org/10.1007/s10107-016-1030-6

Tan CH, Ma SQ, Dai YH, et al., 2016. Barzilai-Borwein step size for stochastic gradient descent. Proc 30th Int Conf on Neural Information Processing Systems, p.685-693.

Tsitsiklis J, Bertsekas D, Athans M, 1986. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Trans Autom Contr*, 31(9):803-812. https://doi.org/10.1109/TAC.1986.1104412

Wang B, Jiang HY, Fang J, et al., 2018. A proximal ADMM for decentralized composite optimization. *IEEE Signal Process Lett*, 25(8):1121-1125. https://doi.org/10.1109/LSP.2018.2841648

Wang Z, Li HQ, 2020. Edge-based stochastic gradient algorithm for distributed optimization. *IEEE Trans Netw Sci Eng*, 7(3):1421-1430. https://doi.org/10.1109/TNSE.2019.2933177

Wei EM, Ozdaglar A, Jadbabaie A, 2013. A distributed Newton method for network utility maximization—I: algorithm. *IEEE Trans Autom Contr*, 58(9):2162-2175. https://doi.org/10.1109/TAC.2013.2253218

Xi CG, Khan UA, 2017. DEXTRA: a fast algorithm for optimization over directed graphs. *IEEE Trans Autom Contr*, 62(10):4980-4993. https://doi.org/10.1109/TAC.2017.2672698

Xia YS, Wang J, 2004. A one-layer recurrent neural network for support vector machine learning. *IEEE Trans Syst Man Cybern B*, 34(2):1261-1269. https://doi.org/10.1109/TSMCB.2003.822955

Xin R, Khan UA, 2018. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Contr Syst Lett*, 2(3):315-320. https://doi.org/10.1109/LCSYS.2018.2834316

Xin R, Khan UA, 2020. Distributed heavy-ball: a generalization and acceleration of first-order methods with gradient tracking. *IEEE Trans Autom Contr*, 65(6):2627-2633. https://doi.org/10.1109/TAC.2019.2942513

Xin R, Jakovetić D, Khan UA, 2019a. Distributed Nesterov gradient methods over arbitrary graphs. *IEEE Signal Process Lett*, 26(8):1247-1251. https://doi.org/10.1109/LSP.2019.2925537

Xin R, Sahu AK, Khan UA, et al., 2019b. Distributed stochastic optimization with gradient tracking over strongly-connected networks. Proc IEEE 58th Conf on Decision and Control, p.8353-8358. https://doi.org/10.1109/CDC40024.2019.9029217

Xin R, Xi CG, Khan UA, 2019c. FROST—fast row-stochastic optimization with uncoordinated step-sizes. *EURASIP J Adv Signal Process*, 2019(1):1. https://doi.org/10.1186/s13634-018-0596-y

Xin R, Khan UA, Kar S, 2020. Variance-reduced decentralized stochastic optimization with accelerated convergence. *IEEE Trans Signal Process*, 68:6255-6271. https://doi.org/10.1109/TSP.2020.3031071

Xu JM, Zhu SY, Soh YC, et al., 2015. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. Proc 54th IEEE Conf on Decision and Control, p.2055-2060. https://doi.org/10.1109/CDC.2015.7402509

Yin R, Zhang Y, Yu GD, et al., 2010. Centralized and distributed resource allocation in OFDM based multi-relay system. *J Zhejiang Univ-Sci C (Comput & Electron)*, 11(6):450-464. https://doi.org/10.1631/jzus.C0910405

Yuan DM, Ma Q, Wang Z, 2013. Distributed dual averaging method for solving saddle-point problems over multi-agent networks. Proc 32nd Chinese Control Conf, p.6868-6872.

Zhang CL, Ahmad M, Wang YQ, 2019. ADMM based privacy-preserving decentralized optimization. *IEEE Trans Inform Forens Secur*, 14(3):565-580. https://doi.org/10.1109/TIFS.2018.2855169

Zinkevich MA, Weimer M, Smola A, et al., 2010. Parallelized stochastic gradient descent. Proc 23rd Int Conf on Neural Information Processing Systems, p.2595-2603.

Huaqing LI received his BS degree in information and computing science in 2009 from Chongqing University of Posts and Telecommunications, Chongqing, China and his PhD degree in computer science and technology in 2013 from Chongqing University. From Sept. 2014 to Sept. 2015, he was a postdoctoral researcher at the School of Electrical and Information Engineering, The University of Sydney, Australia. From Nov. 2015 to Nov. 2016, he was a postdoctoral researcher at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently a professor at the College of Electronic and Information Engineering, Southwest University, Chongqing, China. His main research interests include nonlinear dynamics and control, multi-agent system, and distributed optimization. Prof. LI currently serves as a regional editor for *Neur Comput Appl*, an editorial board member for *IEEE Access*, and a corresponding expert for *Front Inform Technol Electron Eng*.