Distributed Stochastic Gradient Tracking Algorithm With Variance Reduction for Non-Convex Optimization

Xia Jiang, Xianlin Zeng, Member, IEEE, Jian Sun, Senior Member, IEEE, and Jie Chen, Fellow, IEEE

Abstract—This article proposes a distributed stochastic algorithm with variance reduction for general smooth non-convex finite-sum optimization, which has wide applications in signal processing and machine learning communities. In distributed setting, a large number of agents are connected by an underlying network. Each agent computes local stochastic gradient and communicates with its neighbors to seek for the global optimum. In this article, we develop a modified variance reduction technique to deal with the variance introduced by stochastic gradients. Combining gradient tracking and variance reduction techniques, this article proposes a distributed stochastic algorithm, gradient tracking algorithm with variance reduction (GT-VR), to solve large-scale non-convex finite-sum optimization over multiagent networks. A complete and rigorous proof shows that the GT-VR algorithm converges to the first-order stationary points with $O(1/k)$ convergence rate. In addition, we provide the complexity analysis of the proposed algorithm. Compared with some existing first-order methods, the proposed algorithm has a lower $O(1/M + 1/t)$ gradient complexity under some mild condition. By comparing state-of-the-art algorithms and GT-VR in numerical simulations, we verify the efficiency of the proposed algorithm.

Index Terms—Complexity analysis, distributed algorithm, non-convex finite-sum optimization, stochastic gradient, variance reduction.

I. INTRODUCTION

As big data are becoming increasingly important, distributed finite-sum optimization has received extensive attention from researchers in signal processing, control, and machine learning communities [1]–[6]. In distributed finite-sum optimization, large-scale signal information or training samples are allocated to different nodes, and each node updates variable by local data and obtains global optimal estimate through communication with neighbors. When the dataset is located in a decentralized setting or contains private information, it is infeasible to transmit the dataset over networks and handle it at a centralized node [4], [7]–[11]. In addition, for functions with large-size local data, computing the local gradient using the whole local dataset becomes practically difficult. Due to the abovementioned reasons, distributed stochastic first-order algorithms are preferable as they own a low computation complexity without the calculation of Hessian matrices and are easy to analyze.

Distributed stochastic gradient algorithms are often a combination of average consensus steps between neighbors and local stochastic gradients, which have been popular in machine learning tasks [12]–[14]. With the similar design idea, considerable works have been studied for more complex optimization problems and various multiagent networks in recent years, e.g., distributed stochastic gradient projection algorithms [15], distributed stochastic mirror descent [16], distributed stochastic primal-dual algorithm over random networks with imperfect communications [17], and stochastic gradient-push over time-varying directed graphs [18]. However, the performance of distributed stochastic gradient algorithm is generally limited by two components. One is the local variance introduced by stochastic gradient at each agent, and the other is the heterogeneous datasets between different agents. To handle the local variance, many variance reduction techniques have been proposed to reduce storage space and computation complexity, such as stochastic average gradient algorithm (SAGA) [19], stochastic variance reduced gradient (SVRG) [20], Stochastic Recursive grAdient algoritHm (SARAH) [21], and asynchronous variance reduced algorithm (Asyn-VR) [22]. Various variance-reduction techniques have been applied to the design of decentralized algorithms for convex (strongly convex) problems recently [23]–[25]. In particular, the integration of gradient tracking and SVRG was proposed in [25] to obtain a linear convergence rate for smooth and strongly convex optimization. To achieve robustness to heterogeneous environments, some works develop distributed bias-correction techniques, such as gradient tracking [26], [27], exact first-order algorithm (EXTRA) [28], and primal-dual principles [29], [30]. Integrating variance reduction and bias-correction techniques, efficient distributed algorithms with linear convergence rate arise for strongly convex finite-sum optimization [23]. However, the
applicability of these distributed methods for non-convex optimization remains unclear.

A large-scale non-convex optimization has wide applications including logistic regression with non-convex regularization and neural networks training. When the cost functions are non-convex, the design and theoretical analysis of efficient algorithms become difficult due to the lack of good properties of convexity. Very recent works have proposed distributed variance-reduced methods for non-convex finite-sum problems. Sun et al. [31] proposed decentralized gradient estimation and tracking (D-GET) for decentralized non-convex finite-sum minimization, which considers a local SARAH-type variance-reduced technique and gradient tracking. However, as is pointed in [32], the D-GET does not have a network-independent gradient complexity. Gradient tracking with SARAH (GT-SARAH) proposed in [32] has achieved a near-optimal total gradient computation complexity at the cost of twice communication rounds of the D-GET algorithm. Xin et al. [33] proposed a gradient tracking with SAGA (GT-SAGA) algorithm by combining the SAGA and gradient tracking techniques. However, SAGA needs additional storage space compared with SVRG. Inspired by the SVRG technique, we propose a distributed stochastic first-order algorithm with low network-independent gradient complexity for large-scale finite-sum non-convex optimization in this article.

The contributions of this article are summarized as follows.

1) For general smooth non-convex optimization, we propose a novel distributed stochastic iterative algorithm, gradient tracking algorithm with variance reduction (GT-VR), by combining gradient tracking and variance reduction techniques. This is the first work to combine the SVRG technique [20] with a Bernoulli distribution to solve the non-convex optimization, while the original SVRG technique in [20] only proves the convergence for convex optimization.

2) By linear matrix inequality techniques, we prove that the proposed GT-VR algorithm converges to the first-order stationary points with\( O(1/k) \) convergence rate.

3) The proposed algorithm has a lower gradient complexity compared with some state-of-the-art algorithms. To be specific, compared with distributed algorithms decentralized stochastic gradient tracking (DSGT) [35], D2 [37], and decentralized stochastic gradient descent (DSGD) [34], whose gradient complexity is \( O(v^2 \epsilon^{-2}) \), the proposed algorithm has a lower network-independent gradient complexity \( O(P M \epsilon^{-1}) \) under some mild condition. Comparative experimental results of these algorithms and GT-VR also verify the efficiency of the proposed algorithm.

The remainder of this article is organized as follows. Mathematical notations and some stochastic properties are given in Section II. The problem description and distributed stochastic algorithm are provided in Section III. The convergence properties of proposed methods are provided in Section IV and are analyzed theoretically in Section V. The efficiency of the distributed algorithms is verified by simulations in Section VI, and the conclusion is made in Section VII.

II. NOTATIONS AND PRELIMINARIES

A. Mathematical Notations

We denote \( \mathbb{R} \) as the set of real numbers, \( \mathbb{R}^+ \) as the set of positive real numbers, \( \mathbb{Z}^+ \) as the set of positive integers, and \( \mathbb{R}^n \) as the set of \( n \)-dimensional real column vectors, respectively. All vectors in this article are column vectors, unless otherwise noted. \( I_n \) denotes an \( n \times 1 \) vector with all elements of 1, \( 0_n \) denotes a \( d \times 1 \) vector with all elements of 0, and \( I_d \) denotes a \( d \times d \) identity matrix. The notation \( \otimes \) denotes the Kronecker product, and \( \max\{ \cdots \} \) denotes the maximum element in the set \{ \cdots \}. For a real vector \( o \), \( \| o \| \) is the Euclidean norm. For a differentiable function \( f(x) \), its gradient is represented by \( \nabla f(x) \). In the following, subscript \( i \) refers to the \( i \)th agent, e.g., \( x_i \) means a local variable of agent \( i \).

We fix a rich enough probability space \( (\Omega, \mathcal{F}, \mathbb{P}) \), where all random variables in discussion are properly defined, and \( \mathbb{E}[\cdot] \) denotes the expectation operator with respect to the probability measure \( \mathbb{P} \). Let \( A \) be an event in \( \mathcal{F} \) with nonzero probability, and \( X \) be a random variable. The conditional expectation of \( X \) given \( A \) is denoted by \( \mathbb{E}[X|A] \). For an event \( A \in \mathcal{F} \), its indicator function is denoted as \( I_A \). We use \( \sigma(\cdot) \) to denote the \( \sigma \) – algebra generated by the random variables and sets in its argument. For a matrix \( A \), \( d(A) \) denotes its spectral radius.

B. Stochastic Theory

For conditional expectation, there is one basic property, which is useful in the subsequent analysis and is stated as the following.

**Proposition 1** [38, Sec. 4.1]: Let \( X \), \( Y \), and \( X_i (1 \leq i \leq n) \) be random variables, and \( \mathbb{E}[X|Y] < \infty \), \( \mathbb{E}[X_i|Y] < \infty \) (1 \( \leq i \leq n \)).

1) \( \mathbb{E}[\sum_{i=1}^{n} a_i X_i|Y] = \sum_{i=1}^{n} a_i \mathbb{E}[X_i|Y] \) (a.s.), where \( a_i \) is a constant.

2) If \( X \) and \( Y \) are mutually independent, then \( \mathbb{E}[X|Y] = \mathbb{E}[X] \).

3) \( \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \).

**Bernoulli Distribution**: The Bernoulli \((P)\) distribution is the discrete probability distribution of a random variable, which takes the value 1 with probability \( P \) and the value 0 with probability \( 1 - P \). If \( X \) is a random variable with the Bernoulli \((P)\) distribution, then

\[
\mathbb{P}(X = 1) = P, \quad \mathbb{P}(X = 0) = 1 - P. \tag{1}
\]

III. PROBLEM DESCRIPTION AND DISTRIBUTED SOLVER DESIGN

In this article, we aim to solve the following distributed finite-sum optimization problem over a connected multiagent network:

\[
\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x), \quad f_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} f_{i,j}(x) \tag{2}
\]

where \( f_i : \mathbb{R}^d \to \mathbb{R} \) is the local differentiable objective function of agent \( i \), composed as the average of \( m_i \) component.
costs \(\{f_i, v_i\}_{i=1}^m\), \(n\) is the number of agents, \(m_i\) is the number of local samples, and \(\sum_{i=1}^n m_i = M\) is the total number of samples in the network. The connected multiagent network containing \(n\) agents is denoted by \(G(V, E, W)\), where \(V = \{1, \ldots, n\}\), \(E = V \times V\), and \(W\) is the adjacent matrix associated with \(G\). In distributed setting, each agent handles local information and communicates with its neighbors over the network \(G\) to solve (2) cooperatively.

Remark 2: This formulation of optimization problem (2) is widely adopted in empirical risk minimization [39], [40], where each local cost function \(f_i\) is considered as an empirical risk computed over \(m_i\) local data samples. Compared with the parameter-server-type machine learning system with a fusion center [41], [42], distributed algorithms for (2) can preserve data privacy, improve the computation efficiency, and enhance network robustness. Furthermore, in many emerging applications, such as collaborative filtering, federated learning, distributed beamforming, and dictionary learning, the data are naturally collected in a distributed setting, and it is impossible to transfer the distributed data to a central location [31]. Therefore, decentralized computation has sparked considerable interest in both academia and industry.

Next, we design a distributed stochastic algorithm for the general smooth non-convex optimization (2). There are two challenges in the distributed stochastic algorithm design. One challenge is the slow convergence due to the variance of stochastic gradients by asymptotically estimating the local full gradient \(\nabla f_i\) based on randomly selected samples from the local dataset of agent \(i\). The other is the difference between local and global objective functions, i.e., \(\nabla f_i(x^*), \forall i \in \{1, \ldots, n\}\), for the global optimum \(x^*\). This issue may be handled by the popular gradient tracking technique that introduces a local gradient estimator to track the global gradient.

By combining the distributed gradient tracking [43] with a variance reduction technique, we propose the first-order GT-VR algorithm. The complete implementation of GT-VR is summarized in Algorithm 1. The local gradient estimator \(v_i^{k+1}\) is updated by

\[ v_i^{k+1} = \nabla f_i(x_i^{k+1} - \nabla f_i(x_i^{k+1}) + \nabla f_i(x_i^{k+1}). \]  

In addition, in GT-VR, we introduce the gradient tracking technique to achieve the global gradient tracking in distributed optimization.

Remark 3: The variance reduction technique taken in GT-VR is a modification of the well-known SVRG technique. In both techniques, the entire local full gradient \(\nabla f_i(x_i^k)\) needs to be computed with a certain probability, and the local gradient estimator \(v_i^k\) is updated by (3). The only main difference is that in GT-VR, \(t_i^k\) is updated following the Bernoulli distribution, while in SVRG, it is updated periodically. Denote \(\mathcal{F}_k\) as the history of the dynamical system defined by \(\sigma((x_i^t, v_i^{t+1})_{t=1}^{k-1})\). Note that in SVRG, each local gradient estimator \(v_i^k\) is an unbiased estimator of the local gradient \(\nabla f_i(x_i^k)\) given \(\mathcal{F}_k\) [20], whereas it does not hold in our proposed algorithm GT-VR. This modification is vital for the convergence of proposed algorithm for the smooth non-convex optimization.

Algorithm 1 GT-VR Updating at Each Agent \(i\)

1: Initialize: \(x_i^0, \tau_i^0 = x_i^0; \eta; \{w_{ir}\}_{r=1}^M; y_i^0 = \nabla f_i(x_i^0)\).
2: for \(k = 1, 2, \ldots\) do
3: Update the local estimate of the solution:
   \[ x_i^{k+1} = \sum_{r=1}^n w_{ir}(x_i^k - \eta y_i^k); \]
4: Select \(t_i^{k+1}\) at random from the Bernoulli(\(P\)) distribution.
   If \(t_i^{k+1} = 1\), \(t_i^{k+1} = x_i^{k+1}\), and otherwise, \(t_i^{k+1} = t_i^k\).
5: Select \(s_i^{k+1}\) uniformly at random from \(\{1, \ldots, m_i\}\).
6: Update the local stochastic gradient estimator by (3);
7: Update the local gradient tracker:
   \[ y_i^{k+1} = \sum_{r=1}^n w_{ir}(y_i^k + v_i^{k+1} - v_i^k); \]
8: end for

Remark 4: Compared with distributed deterministic optimization, which needs to compute the entire local gradient \(\nabla f_i\) at each iteration, the proposed distributed stochastic first-order algorithm using sampled batch data to compute stochastic gradient is more suitable for the training and processing of large-scale data.

Compared with GT-SAGA [33], the proposed algorithm does not need to store the value of gradient and saves more storage space. Compared with the two-timescale hybrid algorithm GT-SARAH [32], the proposed algorithm is one single-timescale randomized gradient algorithm. In addition, at each iteration, there are only two communication rounds with neighbors at each agent. However, GT-SARAH has a near-optimal gradient computational complexity, which is better than the proposed algorithm.

IV. CONVERGENCE RESULT

In this section, we provide the convergence analysis of the proposed algorithm GT-VR with some mild assumptions.

Assumption 5:

1) The gradient of cost function \(f_{i,j}\) is \(L\)-Lipschitz continuous, i.e., for some \(L > 0\)
   \[ \|\nabla f_{i,j}(x) - \nabla f_{i,j}(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d. \]
2) The underlying multiagent network \(G\) is connected, and the adjacent matrix \(W\) is a doubly stochastic matrix, where \(w_{ii} > 0\) for all \(i \in V\), and \(w_{ij} > 0\) if \((i, j) \in E\) for \(i, j \in V\).
3) The family \(\{l_i^k, s_i^k : i \in V, k \geq 1\}\) of random variables in the proposed algorithm is independent.

In Assumption 5, 1) and 2) are common assumptions in distributed optimization, and 1) guarantees that local batch objective functions \(\{f_{i,j}\}_{i=1}^m\) and the global objective function \(f\) are \(L\) smooth. In machine learning, the cost function of logistic regression is often a smooth non-convex sigmoid function, and the activation functions between layers of deep neural networks are usually nonlinear continuously differentiable. Hence, non-convex optimization with continuously

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
differentiable functions is common and important in machine learning. The adjacent matrix satisfying (2) holds for the family of undirected graphs and weight-balanced directed graphs. In addition, in Assumption 5, 2) guarantees that the radius of the network $\rho$ satisfies

$$\rho \triangleq \sup_{x \in \mathcal{X}} \frac{\|W(x - n^{-1}1_n1_n^T x)\|}{\|x - n^{-1}1_n1_n^T x\|} < 1. \quad (4)$$

In Assumption 5, 3) is standard in the design of stochastic algorithm and is practical in application.

For the convenience of analysis, we define several auxiliary quantities as follows:

$$\begin{align*}
x^k &= \begin{bmatrix} x^k_1 \\ \vdots \\ x^k_n \end{bmatrix}, \\
y^k &= \begin{bmatrix} y^k_1 \\ \vdots \\ y^k_n \end{bmatrix}, \\
v^k &= \begin{bmatrix} v^k_1 \\ \vdots \\ v^k_n \end{bmatrix}, \\
r^k &= \begin{bmatrix} r^k_1 \\ \vdots \\ r^k_n \end{bmatrix}, \\
\Delta f(x^k) &= \begin{bmatrix} \nabla f_1(x^k_1) \\ \vdots \\ \nabla f_n(x^k_n) \end{bmatrix}
\end{align*}$$

where $x^k = (1/n)\sum_{i=1}^n x^k_i$, $y^k = (1/n)\sum_{i=1}^n y^k_i$, $\bar{v}^k = (1/n)\sum_{i=1}^n v^k_i$, $\bar{x}^k = 1_n \otimes \bar{x}^k \in \mathbb{R}^n$, and $\bar{v}^k = 1_n \otimes \bar{v}^k \in \mathbb{R}^n$.

Then, the proposed algorithm GT-VR satisfies

$$\begin{align*}
x^{k+1} &= (W \otimes I_n)(x^k - \bar{v}^k) \\
y^{k+1} &= (W \otimes I_n)(y^k + v^{k+1} - \bar{v}^k) \\
\end{align*}$$

and

$$\tilde{x}^k = \tilde{v}^k, \quad \tilde{x}^{k+1} = \tilde{x}^k - \eta \tilde{v}^k \quad (7)$$

where the doubly stochastic property of $W$ is used to derive (7).

Define $\tilde{\eta} = \min\{(1 - 3\rho^2)/((16\rho^2L^2 + 32\rho^2L^2 + 2(1-P)((1 + \rho)/\rho)5L), (1/6L), ((1 - (4/3) + (8/9)P)\rho^2(2T)))\}$, where $T \triangleq 16L^2 + (8/3) + (16/3)(1 + \rho)(1 - P)\rho^2L^2 + 32\rho(16/9 + 16(1 - P)(1 + P + (2(\rho + 1))/\rho) )\rho^2 \in \mathbb{R}$, and $\epsilon_3$ is a positive number.

Then, the convergence result of the proposed algorithm GT-VR is covered in Theorem 6.

**Theorem 6:** Suppose Assumption 5 holds. Let $\rho^2 < (1/3)$, $1 - (3\rho^2)/((1 + (1/\rho))(2/9) + 1 + \rho)) < P < 1$, $0 < \eta < \tilde{\eta}$, and $f^* \triangleq \inf_{x \in \mathbb{R}} f(x) > -\infty$. Then, $f(\tilde{x}^k)$ converges, $(1/k)\sum_{t=1}^k \mathbb{E}||\nabla f(\tilde{x}^t)||^2 = O(1/k)$, $(1/k)\sum_{t=1}^k \mathbb{E}||x^t - \tilde{x}^t||^2 = O(1/k)$, and $(1/k)\sum_{t=1}^k \mathbb{E}||y^t - \nabla f(\tilde{x}^t)||^2 = O(1/k)$.

**Remark 7:** For the convergence analysis of the proposed distributed algorithm in Theorem 6, the consensus performance of local variables in different agents is measured by the convergence of $(1/k)\sum_{t=1}^k \mathbb{E}||x^t - \tilde{x}^t||^2$. The convergence rate of $(1/k)\sum_{t=1}^k \mathbb{E}||\nabla f(\tilde{x}^t)||^2$ is used to measure the convergence of the first-order gradients in the non-convex optimization.

**Remark 8:** Theorem 6 implies that the proposed GT-VR algorithm converges to the first-order stationary points with $O(1/k)$ convergence rate. Compared with GT-HSOGD [36], which converges sublinearly at a rate of $O(1/k)$ up to a steady-state error, the proposed GT-VR has no steady-state error.

**Remark 9:** In the algorithm design, the lower bound of $P$ is negatively correlated with $\rho$. Therefore, when the upper bound of $\rho$ is known (i.e., $\rho^2 < 1/3$), we can get the range of $P$. In addition, the radius of network $\rho$ can be estimated by distributed information according to the study [44], which is not the focus of our research.

Then, we present the complexity of GT-VR in the following sense.

**Definition 10** [34]: The algorithm GT-VR is said to achieve an $\epsilon$-accurate stationary point of $f$ in $k$ iterations if

$$\frac{1}{k} \sum_{i=1}^k \mathbb{E}||\nabla f(\tilde{x}^i)||^2 \leq \epsilon. \quad (8)$$

**Iteration complexity of the proposed GT-VR** is the smallest number of iterations $k$ required to find an $\epsilon$-accurate stationary point defined in (8). The gradient complexity is the total number of component gradient computations $\nabla f_i(\cdot)$ across all nodes to find an $\epsilon$-accurate stationary point defined in (8). In one round of communication, each node $i \in [n]$ communicates one $d$-dimensional vector with its neighbors. Then, the communication complexity is defined as the total rounds of communications over the multiagent network required to achieve an $\epsilon$-accurate stationary solution defined in (8).

Based on the results in Theorem 6, the iteration complexity, gradient computation complexity, and communication complexity of GT-VR are established in the following corollary.

**Corollary 11:** Suppose Assumption 5 holds. Let $\rho^2 < (1/3)$, $1 - (3\rho^2)/((1 + (1/\rho))(2/9) + 1 + \rho)) < P < 1$, and

$$0 < \eta < \tilde{\eta} \quad (9)$$

where $\tilde{\eta} = \min\{(1 - 3\rho^2)/(3\rho^2L))\}$. Then, the following hold.

1. GT-VR achieves an $\epsilon$-accurate stationary point of $f$ with $O(\eta^{-2}\epsilon^{-1})$ iterations.
2. GT-VR achieves an $\epsilon$-accurate stationary point of $f$ in $O(PM\eta^{-2}\epsilon^{-1})$ gradient computations across all agents.
3. GT-VR achieves an $\epsilon$-accurate stationary point of $f$ with $O(\sum_{i=1}^n N_i\eta^{-2}\epsilon^{-1})$ communication rounds across all agents, where $N_i$ denotes the number of neighbors of agent $i$.

**Remark 12:** In large-scale sample case, the gradient complexity is $O(PM\eta^{-2}\epsilon^{-1})$. If $\eta = (1/6L)$, the gradient complexity is $O(PM\epsilon^{-1})$, which is network-independent. The gradient complexities of some existing algorithms are summarized in Table I. If the accuracy $\epsilon$ is chosen small enough, the gradient complexity of the proposed algorithm GT-VR is lower than the gradient complexity of the existing algorithms [34], [35], [37], [45]. In addition, the proposed algorithm has a constant step size, alleviating the slow convergence caused by diminishing step size.
V. Theoretical Analysis

In this section, we present the proofs for Theorem 6 and Corollary 11. The analysis framework is general and may be applied to other distributed algorithms based on variance reduction and gradient tracking. Recall that $\mathcal{F}^k$ is the history of the dynamical system, defined by

$$\mathcal{F}^k \triangleq \sigma (\{x^i, v^i\}_{i=1}^{k-1})$$

The convergence analysis is roughly divided into three steps. First, the first step is to prove the boundness of $\mathbb{E}[\|x^k - \tilde{x}^k\|^2]$, $\mathbb{E}[\|\bar{x}^k - x^k\|^2]$, and $\mathbb{E}[\|\bar{v}^k - v^k\|^2]$ by constructing a linear matrix inequality. Second, the second step is to prove the boundness of $\sum_{i=1}^n \mathbb{E}[\|x^i - \tilde{x}^i\|^2]$, $\sum_{i=1}^n \mathbb{E}[\|r^i - \tilde{r}^i\|^2]$, and $\sum_{i=1}^n \mathbb{E}[\|y^i - \bar{y}^i\|^2]$ by recursion. Third, we prove the boundness of $\sum_{i=1}^n \mathbb{E}[\|\bar{v}^i - v^i\|^2]$ and the convergence of $f(\bar{x}^k)$.

At first, we provide a standard result for the adjacent matrix $W$ satisfying Assumption 5, which will be frequently used in the subsequent analysis.

**Lemma 13 (43):** Suppose Assumption 5 holds. For any $x_1, \ldots, x_n \in \mathbb{R}^d$, we have

$$\|W(x - \tilde{x})\| \leq \rho \|x - \tilde{x}\|$$

where $\rho$ is the radius of the network.

In the following proposition, we present some useful properties of local stochastic gradient estimator $h_i^k$, local gradient $\nabla f_i(x_i^k)$, and $\nabla f_i(\tilde{x}^k)$, which will be used to bound $\mathbb{E}[\|r^k - \tilde{r}^k\|^2]$ and to prove Theorem 6.

**Proposition 14:** Suppose Assumption 5 holds. Then

$$\begin{align*}
\mathbb{E}[\|v^k - \nabla f(x^k)\|^2] & \leq 2L^2\mathbb{E}[\|x^k - \tilde{x}^k\|^2] + 2L^2\mathbb{E}[\|r^k - \tilde{r}^k\|^2] \\
& \leq \frac{1}{n} \sum_{i=1}^n \left( (b_i - \nabla f_i(x_i^k)) \right) ^2
\end{align*}
$$

and

$$\begin{align*}
\frac{1}{4} \mathbb{E}[\|\tilde{v}^k\|^2] & \leq \mathbb{E}[\|\nabla f(\bar{x}^k)\|^2] + \frac{2L^2}{n} \\
& \times \sum_{i=1}^n \left( 3\mathbb{E}[\|x_i^k - \tilde{x}_i^k\|^2] + 2\mathbb{E}[\|r_i^k - \tilde{r}_i^k\|^2] \right).
\end{align*}
$$

**Proof:** See Appendix A.

With Proposition 14, in the following lemma, we establish bounds on $\mathbb{E}[\|x^k - \tilde{x}^k\|^2]$, $\mathbb{E}[\|r^k - \tilde{r}^k\|^2]$, and $\mathbb{E}[\|y^k - \bar{y}^k\|^2]$, respectively.

**Proposition 15:** Suppose Assumption 5 holds. We have the following inequalities:

1) $\mathbb{E}[\|x^k - \tilde{x}^k\|^2] \leq 2p^2\mathbb{E}[\|x^k - \tilde{x}^k\|^2] + 2p^2\eta^2\mathbb{E}[\|y^k - \bar{y}^k\|^2]$

2) $\mathbb{E}[\|r^k - \tilde{r}^k\|^2] \leq \left( 2p^2P + (1 - P) \left( \eta^2 + \frac{\eta}{\beta} \right) 12L^2 \right) \mathbb{E}[\|\bar{x}^k - x^k\|^2] + 2p^2\eta^2P\mathbb{E}[\|y^k - \bar{y}^k\|^2]$

where $P \triangleq \frac{\mathbb{E}[\|r^k - \tilde{r}^k\|^2]}{\mathbb{E}[\|x^k - \tilde{x}^k\|^2]}$.

Proof: Let $\beta = (\rho/\eta)$ and $\eta L \leq (1/6)$. With Lemma 15, we have

$$u_{k+1} \leq C_1 2p\eta^2 + C_2 \frac{\eta^2}{\beta} 2p^2 0 \left[ \begin{array}{c} u_k \\ C_{2u} \end{array} \right]$$

where $u_k \triangleq \frac{\mathbb{E}[\|r^k - \tilde{r}^k\|^2]}{\mathbb{E}[\|x^k - \tilde{x}^k\|^2]}$, $C_1 = 2p^2 + ((12 + 8\beta)/9)\rho^4$, $C_2 = 16p^2L^2 + ((8\beta + 16(\rho + 1)(1 - P))/3\beta)\rho^2L^2 + (32 + 32\beta)\rho^2L^2$, $C_3 = (16/9)(1 + ((9\beta^2 + 11\rho + 2)/\rho)(1 - P))\rho^2L^2$, $C_{2u} = 2p^2P + (1/3)(1 - P)(1 + (1/\rho))$.
Define a positive vector \( \epsilon \) converges to zero at the linear rate \( JIANG \).

From (20) over iterations and (21), we obtain

\[
\begin{align*}
\text{Proposition 18:} & \quad \text{Suppose Assumption 5 holds. If the step size satisfies } \eta \leq (1/6L), \text{ then} \\
& \quad \mathbb{E} f(\bar{x}^{k+1}) \leq \mathbb{E} f(\bar{x}^k) - \frac{\eta}{3} \mathbb{E} \| \nabla f(\bar{x}^k) \|^2 \\
& \quad + \frac{2L}{3n} \mathbb{E} \| x^k - \bar{x}^k \|^2 + \frac{4L}{9n} \mathbb{E} \| r^k - \bar{x}^k \|^2. 
\end{align*}
\]

Proof: By \( \bar{x}^{k+1} = \bar{x}^k - \eta \tilde{v}^k \) and the L smoothness of the function f, we have

\[
f(\bar{x}^{k+1}) \leq f(\bar{x}^k) - \eta \| \nabla f(\bar{x}^k) \|^2 + \frac{\eta L}{2} \| \tilde{v}^k \|^2 \\
- \eta \nabla f(\bar{x}^k) - \frac{\eta L}{2} \left( \frac{1}{n} \sum_{i=1}^{n} (v_i^k - \nabla f_i(\bar{x}^k)) \right) \\
\leq f(\bar{x}^k) - \frac{\eta}{2} \| \nabla f(\bar{x}^k) \|^2 + \frac{\eta^2 L}{2} \| \tilde{v}^k \|^2 \\
+ \frac{\eta}{2} \left( \frac{1}{n} \sum_{i=1}^{n} (v_i^k - \nabla f_i(\bar{x}^k)) \right)^2.
\]

Taking the conditional expectation with respect to \( \mathcal{F}^{k+1} \)

\[
\mathbb{E}[f(\bar{x}^{k+1})|\mathcal{F}^{k+1}] \leq f(\bar{x}^k) - \frac{\eta}{2} \| \nabla f(\bar{x}^k) \|^2 + \frac{\eta^2 L}{2} \| \tilde{v}^k \|^2 \\
+ \frac{\eta^2 L}{2} \left( \frac{1}{n} \sum_{i=1}^{n} (v_i^k - \nabla f_i(\bar{x}^k)) \right)^2.
\]

Then, by (11) and (12), we have

\[
\mathbb{E}[f(\bar{x}^{k+1})|\mathcal{F}^{k+1}] \\
\leq f(\bar{x}^k) - \frac{\eta}{2} \| \nabla f(\bar{x}^k) \|^2 \\
+ \frac{\eta^2 L}{2} \left( \| \nabla f(\bar{x}^k) \|^2 + \frac{2L^2}{n} (3\| x^k - \bar{x}^k \|^2 + 2\| r^k - \bar{x}^k \|^2) \right) \\
+ \frac{\eta}{2n} (6L^2 \| x^k - \bar{x}^k \|^2 + 4L^2 \| r^k - \bar{x}^k \|^2) \\
= f(\bar{x}^k) - \left( \frac{\eta}{2} - \frac{\eta^2 L}{2} \right) \| \nabla f(\bar{x}^k) \|^2 + \frac{3\eta L^2 + 6\eta^2 L^3}{n} \\
\times \| x^k - \bar{x}^k \|^2 + \frac{4\eta^2 L^2 + 2\eta L^2}{n} \| r^k - \bar{x}^k \|^2.
\]

Substituting \( \eta L \leq (1/6) \) to the abovementioned inequality and taking the total expectation on both sides of (24), we obtain (23).

The following lemma will be used to establish the convergence of GT-VR.

**Lemma 19 [47]:** Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space, and \( (\mathcal{F}_t)_{t \in \mathbb{Z}^+} \) be a filtration. Let \( U^t, \xi^t, \) and \( \zeta^t \) be nonnegative \( \mathcal{F}_t \) measurable random variables for \( t \in \mathbb{Z}^+ \), such that

\[
\mathbb{E}[U^{t+1} | \mathcal{F}_t] \leq U^t + \zeta^t - \xi^t \quad \forall t = 1, 2, \ldots
\]

Then, \( U^t \) converges to a random variable and \( \sum_{t=1}^{\infty} \zeta^t < +\infty \) almost surely on the event \( \{ \sum_{t=1}^{\infty} \zeta^t < +\infty \} \). Now, we are ready to provide the Proof of Theorem 6.
Proof of Theorem 6:
Proof: By Proposition 18 and (22), we have

\[ 0 \leq f(\bar{x}^1) - f^* - \eta \sum_{i=1}^{k} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] \]
\[ + \left( \frac{3L}{3n} + 4L \right) \left( \frac{1}{5} - \frac{1}{3} \rho^2 \right) + \frac{C_4 + C_4' \sum_{i=1}^{k} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2]}{1 - 3\rho^2} \]
\[ = f(\bar{x}^1) - f^* - \frac{\eta}{3} \sum_{i=1}^{k} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] \]
\[ + \frac{10L}{27} \frac{\rho^2 R_0}{1 - 3\rho^2} \]
\[ = f(\bar{x}^1) - f^* - C \sum_{i=1}^{k} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] + \frac{10L}{27} \frac{\rho^2 R_0}{1 - 3\rho^2} \tag{26} \]

Since

\[ 0 < \eta < \frac{(1 - 3\rho^2)}{(16\rho^2 L^2 + 32\rho^2 L^2 + 2(1 - P)\frac{\rho^2}{\rho})} \]
\[ C = \eta \left( \frac{10L}{27} \frac{1}{1 - 3\rho^2} \left( \frac{16\rho^2 \eta L^2 + 32(1 - P)\rho^2 \eta L^2}{\rho^2} \right) \right) \]
\[ = \eta \left( \frac{10L}{27} \frac{1}{1 - 3\rho^2} \right) \]
\[ > 0. \tag{27} \]

It follows from (26) that:

\[ \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] \leq \frac{1}{kC} \left( f(\bar{x}^1) - f^* + \frac{10L}{27} \frac{\rho^2 R_0}{1 - 3\rho^2} \right) \]
\[ \leq \frac{1}{kC} \left( f(\bar{x}^1) - f^* + \frac{10L}{27} \frac{\rho^2 R_0}{1 - 3\rho^2} \right) \leq \epsilon \tag{28} \]

which implies that \((1/k) \sum_{i=1}^{k} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] = O(1/k)\) holds in Theorem 6.

Now, by (22) and (28), we see that

\[ \max \left\{ \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[\|x^i - \bar{x}^i\|^2], \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[\|y^i - \bar{x}^i\|^2] \right\} \]
\[ \leq \frac{1}{k} \left( \frac{R_0}{1 - 3\rho^2} + \frac{C_4 + C_4'}{1 - 3\rho^2} \left( f(\bar{x}^1) - f^* + \frac{10L}{27} \frac{\rho^2 R_0}{1 - 3\rho^2} \right) \right) \]
\[ = \frac{1}{k} \left( \frac{C_4 + C_4'}{1 - 3\rho^2} \left( f(\bar{x}^1) - f^* \right) \right) \]
\[ + \left( \frac{10L}{27} \frac{\rho^2 R_0}{1 - 3\rho^2} \right) \]
\[ \leq \frac{2}{k} \sum_{i=1}^{k} \mathbb{E}[\|y^i - \bar{y}^i\|^2 + \mathbb{E}[\|\bar{y}^i - \nabla f(\bar{x}^i)\|^2] \leq \frac{2}{k} \sum_{i=1}^{k} \mathbb{E}[\|y^i - \bar{y}^i\|^2 + \mathbb{E}[\|\bar{y}^i - \nabla f(\bar{x}^i)\|^2] \]

In addition, by (11), we have

\[ \frac{1}{k} \sum_{i=1}^{k} \mathbb{E}[\|y^i - \nabla f(\bar{x}^i)\|^2] \leq \frac{2}{k} \sum_{i=1}^{k} \mathbb{E}[\|y^i - \bar{y}^i\|^2 + \mathbb{E}[\|\bar{y}^i - \nabla f(\bar{x}^i)\|^2] \]

which implies that \((1/k) \sum_{i=1}^{k} \mathbb{E}[\|y^i - \nabla f(\bar{x}^i)\|^2] = O(1/k)\) in Theorem 6 holds. Finally, by Proposition 18 and Lemma 19, we see that \(f(\bar{x}^i)\) converges and \(\sum_{i=1}^{k} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] < +\infty\). In addition, because \(\|\nabla f(\bar{x}^i)\|^2\) is non-negative, we have \(\lim_{i \to \infty} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] = 0\). Therefore, combining the fact that \(f(\bar{x}^i)\) converges and \(\lim_{i \to \infty} \mathbb{E}[\|\nabla f(\bar{x}^i)\|^2] = 0\), we obtain that the algorithm converges to critical points of the non-convex optimization problem.

With the convergence result in Theorem 6 and the definition of \(\epsilon\)-accurate stationary point in Definition 10, we present the analysis of Corollary 11 in the following.

Proof of Corollary 11:
Proof: By Definition 10 and (28), it is clear that to find an \(\epsilon\)-accurate stationary point, it is sufficient to find the iterations \(k\), such that

\[ \frac{1}{kC} \left( f(\bar{x}^1) - f^* + \frac{10L}{27} \frac{\rho^2 R_0}{1 - 3\rho^2} \right) \leq \epsilon \tag{29} \]

where \(C = \epsilon\). If \(\eta\) satisfies (9), \(C \leq \epsilon\). If \(\eta\) satisfies (9), \(\eta\) satisfies (9). The inequality (29) holds. Therefore, the iteration complexity of GT-VR is \(O((1/\eta)(f(\bar{x}^1) - f^* + R_0/\eta))\).

For the gradient computations, since there are \(Pm_i + 2\) gradient computations at each iteration of agent \(i\), the number of gradient computations across all agents is the iteration complexity multiplied by \(\sum_{i=1}^{n} (Pm_i + 2)\). Since \(\sum_{i=1}^{n} m_i = M\), the computational complexity is \(O((1/\eta)(Pm_i + 2))\).

At each iteration, each agent \(i\) communicates twice with its neighbors. Let \(N_i\) denote the number of neighbors of agent \(i\). Then, the communication complexity across all agents is \(O((\sum_{i=1}^{n} (N_i/\eta))(f(\bar{x}^1) - f^* + R_0/\eta))\).

Thus, when the iteration \(k\) satisfies

\[ \frac{9}{\eta} \left( f(\bar{x}^1) - f^* + \frac{10R_0}{9n\eta} \right) \leq \frac{k}{C} \tag{30} \]

we have the inequality (29) holds. Therefore, the iteration complexity of GT-VR is \(O((1/\eta)(f(\bar{x}^1) - f^* + R_0/\eta))\).

VI. SIMULATION
To verify the efficacy of the proposed algorithm, we consider the classical binary classification problem, which is to find one optimal predictor \(x \in \mathbb{R}^d\) on a popular logistic regression learning model. We compare the proposed

| Datasets | #samples | #classes |
|----------|----------|----------|
| agricola | 32561    | 123      |
| w8a      | 64700    | 300      |
| covtype   | 581012   | 2        |
algorithm with recently proposed algorithms GT-SAGA [33], GT-SARAH [32], and D-GET [31]. The learning model is to optimize the following problem:

\[
\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(x) \\
\quad f_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} \left( \frac{1}{1 + \exp(l_{ij}a_{ij}^T x)} + \lambda_1 \|x\|^2 \right) \quad (31)
\]

where \( a_{ij} \in \mathbb{R}^d, l_{ij} \in \{-1, 1\}, \) and \( (a_{ij}, l_{ij})_{j=1}^{m_i} \) denote the set of training samples of agent \( i \).

In this experiment, we use the publicly available real datasets,\(^1\) which are summarized in Table II. All algorithms are applied over a ten-agent undirected connected network with a doubly stochastic adjacent matrix \( W \) to solve (31). Meanwhile, \( \lambda_1 \triangleq 5 \times 10^{-4} \). For the proposed GT-VR algorithm, the probability \( P \) is taken as 0.3, and the step size is taken as 0.1. Note that the ranges of step size and possibility provided in Theorem 6 are rigorous theoretical results. In practice, we can adjust them according to the convergence performance. For comparison, the local variables in all algorithms are initialized as the same value, and all the algorithms take the same step sizes. The simulation codes are provided at https://github.com/managerjiang/GT_VR_Simulation.

Define \( D(\bar{x}) = \sum_{i=1}^{10} \lambda_i \sum_{j=1}^{10} w_{ij} (x_i - \bar{x}_j) \). The trajectory of \( D(\bar{x}) \) converging to zero implies that the variable estimates of different agents achieve consensus. For different datasets, the trajectories of cost function \( f \) and \( D(\bar{x}) \) are shown in Fig. 1. We observe that for all datasets, the algorithms all have good consensus performance. For the trajectories of cost function, the algorithm GT-VR decays faster than the state-of-the-art algorithms D-GET, GT-SAGA, and GT-SARAH, especially for the w8a dataset, demonstrating the excellent iteration complexity of GT-VR.

VII. CONCLUSION

Focusing on distributed non-convex stochastic optimization, this article has developed a novel variance-reduced distributed stochastic first-order algorithm over undirected and weight-balanced directed graphs by combining gradient tracking and variance reduction. The variance reduction technique makes use of Bernoulli distribution to handle the variance by stochastic gradients. The proposed algorithm converges with \( O(1/k) \) rate and has lower iteration complexity compared with some existing excellent algorithms. In comparative simulations, the proposed algorithm converges faster than GT-SAGA, GT-SARAH, and D-GET algorithms. To tackle dynamic and complex environment, one future research direction is to extend the distributed stochastic algorithms to solve online optimization problems over time-varying multiagent systems.

APPENDIX

A. Proof of Proposition 14

Proof:

1) For convenience, we define \( A^k \triangleq \sigma (\cup_{i=1}^{10} \sigma (l_i^k), F^k) \) and clearly \( F^k \subseteq A^k \). By the tower property of the conditional expectation, we have

\[
\mathbb{E}[\|v_i^k - \nabla f_i(x_i^k)\|^2|F^k] = \mathbb{E}[\mathbb{E}[\|v_i^k - \nabla f_i(x_i^k)\|^2|A^k]|F^k]. \quad (32)
\]

For \( \mathbb{E}[\|v_i^k - \nabla f_i(x_i^k)\|^2|A^k] \), we have

\[
\mathbb{E}[\|v_i^k - \nabla f_i(x_i^k)\|^2|A^k] \\
= \mathbb{E}[\|\nabla f_i(x_i^k) - \nabla f_i(x_i^k)(\tau_i^k)\|^2|A^k] \\
- \mathbb{E}[\|\nabla f_i(x_i^k) - \nabla f_i(x_i^k)\|^2|A^k] \\
\leq \frac{1}{m_i} \sum_{j=1}^{m_i} \|\nabla f_i(x_i^k) - \nabla f_i(x_i^k)\|^2 |A^k| \\
= 2L^2 \|x_i^k - \bar{x}^k\|^2 + 2L^2 \|\tau_i^k - \bar{x}^k\|^2 \quad (33)
\]

---

\(^1\)a9a, w8a, and covtype.binary are from the website www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/
where the first inequality is from the standard conditional variance decomposition
\[
\mathbb{E}\left[\left|a_k^i - \mathbb{E}[a_k^i | A^i]\right|^2 | A^k\right] = \mathbb{E}\left[\left|a_k^i\right|^2 | A^i\right] - \mathbb{E}\left[\mathbb{E}[a_k^i | A^i]\right]^2 \leq \mathbb{E}\left[\left|a_k^i\right|^2 | A^i\right],
\]
with \(a_k^i = \nabla f_i(x_i^k) - \nabla f_i(x_i^k)\). Substituting the abovementioned inequality to (32), we have
\[
\mathbb{E}\left[\left|a_k^i - \nabla f_i(x_i^k)\right|^2 | \mathcal{F}^k\right] \leq 2L^2\mathbb{E}\left[\left|x_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right] + 2L^2\mathbb{E}\left[\left|\bar{x}_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right].
\]
(34)
The proof follows by summing (34) over \(i\) and taking the total expectation on both sides.
2) With the result in 1), we have
\[
\mathbb{E}\left[\left\{\frac{1}{n} \sum_{i=1}^{n} \left( a_k^i - \nabla f_i(\bar{x}_i^k) \right)^2 \right\} | \mathcal{F}^k\right] \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left|a_k^i - \nabla f_i(\bar{x}_i^k)\right|^2 | \mathcal{F}^k\right] + \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left|\nabla f_i(\bar{x}_i^k)\right|^2 | \mathcal{F}^k\right] \leq \frac{1}{n} \sum_{i=1}^{n} \left( 2L^2\mathbb{E}\left[\left|x_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right] + 2L^2\mathbb{E}\left[\left|x_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right] \right)
\]
(34)
\[
\leq \frac{1}{n} \sum_{i=1}^{n} \left( 2L^2\mathbb{E}\left[\left|x_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right] + 2L^2\mathbb{E}\left[\left|x_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right] \right) + 4L^2\mathbb{E}\left[\left|\tau_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right].
\]
(35)
The proof follows by taking the total expectation on both sides of (35).
3) By Young’s inequality and the result in 2), \(\mathbb{E}[\|\bar{\beta}\|^2]\) satisfies
\[
\mathbb{E}[\|\bar{\beta}\|^2] \leq 2\mathbb{E}[\|\nabla f(\bar{x}_i^k)\|^2] + 2\mathbb{E}\left[\left\{\frac{1}{n} \sum_{i=1}^{n} \left( a_k^i - \nabla f_i(\bar{x}_i^k) \right)^2 \right\} \right] \leq 2\mathbb{E}[\|\nabla f(\bar{x}_i^k)\|^2] + \frac{2}{n} \sum_{i=1}^{n} \left( 6L^2\mathbb{E}\left[\left|x_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right] + 4L^2\mathbb{E}\left[\left|\tau_i^k - \bar{x}_i^k\right|^2 | \mathcal{F}^k\right] \right)
\]
where the last inequality follows from (11).
\(\square\)

B. Proof of Proposition 15
Proof:
1) Recall that \(\bar{x}_i^k = \bar{x}_i\). By (5) and (7)
\[
\left\|\bar{x}_i^{k+1} - \bar{x}_i^{k} \right\|^2 \leq \left\|\left(W \otimes I_d\right)[\bar{x}_i - \bar{x}_i - \eta(\bar{x}_i - \bar{x}_i)] \right\|^2 \leq 2\rho^2\left\|\bar{x}_i - \bar{x}_i\right\|^2 + 2\rho^2\eta^2\left\|\bar{x}_i - \bar{x}_i\right\|^2.
\]
(36)
We take the total expectation on both sides of (36) to obtain
\[
\mathbb{E}[\left\|\bar{x}_i^{k+1} - \bar{x}_i^{k} \right\|^2] \leq 2\rho^2\mathbb{E}[\left\|\bar{x}_i - \bar{x}_i\right\|^2] + 2\rho^2\eta^2\mathbb{E}[\left\|\bar{x}_i - \bar{x}_i\right\|^2].
\]
(37)
2) Recall the Bernoulli distribution in GT-VR that
\[
\mathbb{E}[\left\{\tilde{x}_i^{k+1} \right\} | \mathcal{F}^{k+1}] = P \mathbb{E}[\left\{\tilde{x}_i^{k+1} \right\} | \mathcal{F}^{k+1}] + (1 - P) \mathbb{E}[\left\{\tilde{x}_i^{k+1} \right\} | \mathcal{F}^{k+1}].
\]
Then
\[
\mathbb{E}[\left\|\bar{x}_i^{k+1} - x_i^{k+1} \right\|^2 | \mathcal{F}^{k+1}] = \mathbb{E}[\left\|\bar{x}_i^{k+1} - \left(\tilde{x}_i^{k+1} \right) \right\|^2 | \mathcal{F}^{k+1}] + \mathbb{E}[\left\{\tilde{x}_i^{k+1} \right\} | \mathcal{F}^{k+1}].
\]
(38)
Summing over \(i\) and taking the total expectation on both sides
\[
\mathbb{E}[\left\{\bar{x}_i^{k+1} - x_i^{k+1} \right\}] = P \mathbb{E}[\left\{\bar{x}_i^{k+1} - x_i^{k+1} \right\}] + (1 - P) \mathbb{E}[\left\{\bar{x}_i^{k+1} - x_i^{k+1} \right\}].
\]
For the second term
\[
\mathbb{E}[\left\{\bar{x}_i^{k+1} - x_i^{k+1} \right\}] = \mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^k \right\} + \mathbb{E}[\left\{\bar{x}_i^k \right\} | \mathcal{F}^{k+1}].
\]
\[
\mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^k \right\}] = E[\left\{\bar{x}_i^{k+1} - \bar{x}_i^k \right\}] + \mathbb{E}[\left\{\bar{x}_i^{k} \right\} | \mathcal{F}^{k+1}] - \mathbb{E}[\left\{\bar{x}_i^k \right\} | \mathcal{F}^{k+1}].
\]
where the last inequality holds by 3) in Proposition 14. By (37) and (39)
\[
\mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^{k+1} \right\}] \leq P\left(2\rho^2\mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^{k+1} \right\}] + 2\rho^2\eta^2\mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^{k+1} \right\}] \right) \times \left(\left(\eta^2 + \frac{\eta}{\beta}\right)\frac{2n}{\beta}\mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^{k+1} \right\}] + 12L^2\mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^{k+1} \right\}] + 8L^2\mathbb{E}[\left\{\bar{x}_i^{k+1} - \bar{x}_i^{k+1} \right\}] \right)
\]
where the last inequality holds by 3) in Proposition 14.
3) By (6), we have
\[ y^{k+1} - \bar{y}^{k+1} = (W \otimes I_d)(y^k - \bar{y}^k + \nu^{k+1} - \nu^k) = -\nu^{k+1} + \bar{y}^k. \quad (40) \]

For the term \( v^{k+1} - v^k - \nu^{k+1} + \nu^k \) in (40), it satisfies
\[
\|v^{k+1} - v^k - \nu^{k+1} + \nu^k\|^2 \\
= \|v^{k+1} - v^k\|^2 + \|\nu^{k+1} - \nu^k\|^2 \\
- 2 \sum_{i=1}^n (\nu_i^{k+1} - \nu_i^k, \bar{b}_i^{k+1} - \bar{b}_i^k) \\
= \|v^{k+1} - v^k\|^2 - \|\nu^{k+1} - \nu^k\|^2 \\
\leq \|v^{k+1} - v^k\|^2. \\
(41)
\]

Hence, it follows from Lemma 13, (40), and (41) that:
\[
\|y^{k+1} - \bar{y}^{k+1}\|^2 \leq 2\rho^2(\|y^k - \bar{y}^k\|^2 + \|v^{k+1} - v^k\|^2).
\]

By taking the total expectation on both sides
\[
\mathbb{E}\|y^{k+1} - \bar{y}^{k+1}\|^2 \leq 2\rho^2\mathbb{E}\|y^k - \bar{y}^k\|^2 + 2\rho^2\mathbb{E}\|v^{k+1} - v^k\|^2.
\]

The second term \( \mathbb{E}\|v^{k+1} - v^k\|^2 \) in the abovementioned inequality satisfies
\[
\mathbb{E}\|v^{k+1} - v^k\|^2 \\
\leq 2\mathbb{E}(\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2) \\
+ 2\mathbb{E}(\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2) \\
\leq 2L^2\mathbb{E}\|x^{k+1} - x^k\|^2 \\
+ 2\mathbb{E}(\|\nabla f(x^{k+1}) - \nabla f(x^k)\|^2) \\
\leq 2L^2\mathbb{E}\|x^{k+1} - x^k\|^2 + 4\mathbb{E}\|\nabla f(x^{k+1})\|^2 \\
+ 4\mathbb{E}\|\nabla f(x^k)\|^2 \\
\leq 2L^2\mathbb{E}\|x^{k+1} - x^k\|^2 + 8L^2\mathbb{E}\|x^{k+1} - \bar{x}^{k+1}\|^2 \\
+ 8L^2\mathbb{E}\|\nabla f(x^{k+1})\|^2 \\
+ 8L^2\mathbb{E}\|\nabla f(x^k)\|^2.
\]

where the last equality is from result 1) in Proposition 14. Then, for the first term \( \|x^{k+1} - x^k\|^2 \) in (42), it satisfies
\[
\|x^{k+1} - x^k\|^2 \\
= \|(W \otimes I_d)(y^k - \bar{y}^k) - \eta(W \otimes I_d)\|y^k\|^2 \\
= \|(W \otimes I_d - I_d)(y^k - \bar{y}^k) - \eta(W \otimes I_d)(y^k - \bar{y}^k) - \eta I_n \otimes \bar{b}^k\|^2 \\
\leq 2\|(W \otimes I_d - I_d)(y^k - \bar{y}^k) - \eta(W \otimes I_d)(y^k - \bar{y}^k)\|^2 \\
+ 2\eta^2 n\|\bar{b}^k\|^2 \\
\leq 4\|y^k - \bar{y}^k\|^2 + 4\eta^2 \|y^k - \bar{y}^k\|^2 + 2\eta^2 n\|\bar{b}^k\|^2.
\]

Taking the total expectation, we have
\[
\mathbb{E}\|x^{k+1} - x^k\|^2 \\
\leq 4\mathbb{E}\|y^k - \bar{y}^k\|^2 + 4\eta^2 \mathbb{E}\|y^k - \bar{y}^k\|^2 \\
+ 2\eta^2 n\left(2\mathbb{E}\|\nabla f(x^k)\|^2 + \frac{2}{n} \sum_{i=1}^n (6L^2\mathbb{E}\|x_i^k - \bar{x}_i^k\|^2 \\
+ 4L^2\mathbb{E}\|\nu_i^k - \bar{x}_i^k\|^2)\right) \\
\leq 4\mathbb{E}\|y^k - \bar{y}^k\|^2 + 4\eta^2 \mathbb{E}\|y^k - \bar{y}^k\|^2 \\
+ 4\eta^2 n\mathbb{E}\|\nabla f(x^k)\|^2 + 24\eta^2 L^2\mathbb{E}\|y^k - \bar{y}^k\|^2 + 24\eta^2 L^2\mathbb{E}\|\nu^k - \bar{\nu}^k\|^2 + 16\eta^2 L^2\mathbb{E}\|\nu^k - \bar{\nu}^k\|^2 \\
= (4 + 24\eta^2 L^2)\mathbb{E}\|x^k - \bar{x}^k\|^2 + 4\eta^2 \mathbb{E}\|y^k - \bar{y}^k\|^2 + 16\eta^2 L^2\mathbb{E}\|\nu^k - \bar{\nu}^k\|^2.
\]

Hence, \( \mathbb{E}\|y^{k+1} - \bar{y}^{k+1}\|^2 \) satisfies
\[
\mathbb{E}\|y^{k+1} - \bar{y}^{k+1}\|^2 \\
\leq 2\rho^2\mathbb{E}\|y^k - \bar{y}^k\|^2 + 2\rho^2\mathbb{E}\|v^{k+1} - v^k\|^2 \\
\leq 2\rho^2\mathbb{E}\|y^k - \bar{y}^k\|^2 + 2\rho^2 \left[ 2L^2\mathbb{E}\|x^{k+1} - x^k\|^2 \\
+ 8L^2\mathbb{E}\|x^{k+1} - \bar{x}^{k+1}\|^2 + 8L^2\mathbb{E}\|\nabla f(x^k)\|^2 + 16\eta^2 L^2\mathbb{E}\|\nu^k - \bar{\nu}^k\|^2 \right] \\
\leq 2\rho^2\mathbb{E}\|y^k - \bar{y}^k\|^2 + 2\rho^2 \left[ 48\eta^2 L^2\mathbb{E}\|x^k - \bar{x}^k\|^2 + 16\eta^2 L^2\mathbb{E}\|\nu^k - \bar{\nu}^k\|^2 \right] \\
\leq 2\rho^2\mathbb{E}\|y^k - \bar{y}^k\|^2 + 2\rho^2 \left[ 48\eta^2 L^2(1 + 6n\rho^2 + 4\eta^2 \mathbb{E}\|\nabla f(x^k)\|^2) \\
+ 16\rho^2 L^2(1 + 6n\rho^2 + 2\rho^2) \right] \left[ 1 + \eta \beta + \left( \eta + \frac{\eta}{\beta} \right)^2 \right] \\
\times \left( \frac{\eta^2 + \frac{\eta}{\beta}^2}{\left( \eta + \frac{\eta}{\beta} \right)^2} \right) \mathbb{E}\|\nabla f(x^k)\|^2. \\
\]

Now, with Assumption 5, we have proved all inequalities in Proposition 15.
Xia Jiang received the bachelor’s degree in control science and engineering from Shandong University, Jinan, China, in 2017. She is currently pursuing the Ph.D. degree in control science and engineering with the School of Automation, Beijing Institute of Technology, Beijing, China.

Her current research interests include distributed optimization, large-scale machine learning, and distributed computation of semidefinite problems.

Xianlin Zeng (Member, IEEE) received the B.S. and M.S. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2009 and 2011, respectively, and the Ph.D. degree in mechanical engineering from Texas Tech University, Lubbock, TX, USA, in 2015.

He is currently an Associate Professor with the Key Laboratory of Intelligent Control and Decision of Complex Systems, School of Automation, Beijing Institute of Technology, Beijing, China. His current research interests include distributed optimization, distributed control, and distributed computation of network systems.

Jian Sun (Senior Member, IEEE) received the bachelor’s degree from the Department of Automation and Electric Engineering, Jilin Institute of Technology, Changchun, China, in 2001, the master’s degree from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences (CAS), Beijing, China, in 2004, and the Ph.D. degree from the Institute of Automation, CAS, in 2007.

He was a Research Fellow with the Faculty of Advanced Technology, University of Glamorgan, Pontypridd, U.K., from 2008 to 2009. He was a Post-Doctoral Research Fellow with the Beijing Institute of Technology, Beijing, from 2007 to 2010. In 2010, he joined the School of Automation, Beijing Institute of Technology, where he has been a Professor since 2013. His current research interests include networked control systems, time-delay systems, and security of cyber-physical systems.

Dr. Sun is an Editorial Board Member of the Journal of Systems Science and Complexity and Acta Automatica Sinica.

Jie Chen (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from the Beijing Institute of Technology, Beijing, China, in 1986, 1996, and 2001, respectively, all in control theory and control engineering.

From 1989 to 1990, he was a Visiting Scholar with California State University, Long Beach, CA, USA. From 1996 to 1997, he was a Research Fellow with the School of Engineering, University of Birmingham, Birmingham, U.K. He is currently a Professor with the School of Automation, Beijing Institute of Technology, where he serves as the Director of the Key Laboratory of Intelligent Control and Decision of Complex Systems. He also serves as the President of Tongji University, Shanghai, China. His research interests include complex systems, multiagent systems, multiobjective optimization and decision, and constrained nonlinear control.

Prof. Chen is a fellow of the International Federation of Accountants and a member of the Chinese Academy of Engineering. He has served on the editorial boards for several journals, including the IEEE TRANSACTIONS ON CYBERNETICS, the International Journal of Robust and Nonlinear Control, and Science China Information Sciences. He is currently the Editor-in-Chief of Unmanned Systems and the Journal of Systems Science and Complexity.