Distributed Stochastic Subgradient Optimization Algorithms
Over Random and Noisy Networks

Tao Li †, Keli Fu ‡ and Xiaozheng Fu ‡

August 21, 2020

Abstract: We study distributed stochastic optimization by networked nodes to cooperatively minimize a sum of convex cost functions. The network is modeled by a sequence of time-varying random digraphs with each node representing a local optimizer and each edge representing a communication link. We consider the distributed subgradient optimization algorithm with noisy measurements of local cost functions’ subgradients, additive and multiplicative noises among information exchanging between each pair of nodes. By stochastic Lyapunov method, convex analysis, algebraic graph theory and martingale convergence theory, it is proved that if the local subgradient functions grow linearly and the sequence of digraphs is conditionally balanced and uniformly conditionally jointly connected, then proper algorithm step sizes can be designed so that all nodes’ states converge to the global optimal solution almost surely.

Keywords: Distributed stochastic convex optimization, Additive and multiplicative communication noise, Random graph, Subgradient.

AMS subject classifications. 05C80, 65K10, 68W15, 68W20.

1 Introduction

In recent years, distributed cooperative optimization over networks has attracted extensive attentions, such as the economic dispatch in power grids \[1,2\], the traffic flow control in intelligent transportation networks \[3,4\], and the cooperative source localization by sensor networks \[5,6\], et al. For these networked systems, each node is a local optimizer with certain capabilities of data collection, storage, calculation and communication. In distributed optimization algorithms, each node only needs information related to its local cost function, and exchanges information with neighboring nodes to iteratively update its state such that the global optimal solution is asymptotically achieved. Distributed optimization algorithms have great advantages in solving large-scale optimization problems which are difficult to deal with by centralized algorithms.

Considering the various uncertainties in practical network environments, distributed stochastic optimization algorithms have been widely studied by scholars. The (sub)gradients of local cost functions are used in many distributed optimization algorithms. However, it is difficult to get accurate (sub)gradients in many practical applications. For example, in distributed statistical machine learning \[7,8\], the local loss functions are the mathematical expectations of random functions so that the local optimizers can only obtain the measurement of the (sub)gradients with random noises. The influence of (sub)gradient measurement noises have been considered for distributed optimization algorithms. In
the local (sub)gradient noises are required to be independent, with zero mean and bounded second-order moments. In [19, 20, 21, 22], the local (sub)gradient noise process is a martingale difference sequence. Besides measurement noises, the information exchange among nodes is often affected by communication noises [23, 24], and the structure of the network often changes randomly due to packet dropouts, link/node failures and recreations, which are especially serious in wireless networks. The case with i.i.d. random graphs is studied in [25, 26, 27, 28]. In [29, 30], the graph is randomly selected at each time instant from a family of digraphs which are jointly strongly connected. Especially, the random graphs at different time instants are supposed to be independent with each other in [29]. The case with Markovian switching graphs is discussed in [31, 32].

Most of the above works consider the effects of random switching of network structure, (sub)gradient measurement and communication link noises on distributed optimization algorithms separately. However, a variety of random factors may co-exist in practical environment. In distributed statistical machine learning algorithms, the (sub)gradients of local loss functions cannot be obtained accurately, the graphs may change randomly and the communication links may be noisy. Many scholars have studied distributed optimization with multiple uncertain factors, and have obtained excellent results. Both (sub)gradient noises and random graphs are considered in [33, 34, 35]. In [33], the local gradient noises are independent with bounded second-order moments. They defined a random activation graph sequence to describe the network. Only a single node is activated and exchanges information with its adjacent nodes at each time instant, and the graph sequence is i.i.d. In [34, 35, 36], the (sub)gradient measurement noises are martingale difference sequences and their second-order conditional moments depend on the states of the local optimizers. The random graph sequences in [34, 35, 36, 37] are i.i.d. with connected and undirected mean graphs. In addition, additive communication noises are considered in [36, 37].

In addition to uncertainties in information exchange, different assumptions on the cost functions have been discussed. In [9, 21, 22, 38, 39, 40, 41, 42], the convex local cost functions are required to be differentiable and their gradients are Lipschitz continuous. The gradients of the local cost functions are bounded in [38, 39, 40, 41], and the local cost functions are twice continuously differentiable in [42]. In [43, 25, 24, 37, 44, 45, 46, 47], the case with convex non-differentiable local cost functions is studied and the local subgradients are required to be bounded. In [15, 33], the local cost function is decomposed into a sum of differentiable and non-differentiable parts with the gradient of the differentiable part being Lipschitz continuous and the subgradient of the non-differentiable part being bounded.

Though the above works have made a deep research on distributed stochastic optimization, the practical cases may be more complex. In the regression problem of LASSO (Least absolute shrinkage and selection operator), the local cost functions are not differentiable and their subgradients are not bounded functions. What’s more, the local optimizers can only obtain noisy measurements of local subgradients, and the graphs may change randomly without spatial and temporal dependency. Besides, additive and multiplicative communication noises may co-exist in communication links. In summary, the conditions of such problems are much weaker than those required in the existing works.

Motivated by distributed statistical learning over uncertain communication networks, we study the distributed stochastic convex optimization by networked local optimizers to cooperatively minimize a sum of local convex cost functions. The network is modeled by a sequence of time-varying random digraphs which may be spatially and temporally dependent. The local cost functions are not required to be differentiable, nor do their subgradients need to be bounded. The local optimizers can only obtain measurement information of the local subgradients with random noises. And the additive and multiplicative communication noises co-exist in communication links. We consider the distributed stochastic subgradient optimization algorithm. By algebraic graph theory, convex analysis, and non-negative supermartingale convergence theorem, we prove that if the sequence of random digraphs is conditionally balanced and uniformly conditionally jointly connected, then the states of all local optimizers converge to
the same global optimal solution almost surely. Compared to the existing works, the main contributions of our paper are listed as follows.

I. The structure of the networks among optimizers is modeled by a more general sequence of random digraphs. The weighted adjacency matrices are not required to have special statistical properties such as independency with identical distribution, Markovian switching, or stationarity, etc. The edge weights are also not required to be nonnegative at every time instants. By introducing the concept of conditional digraphs and nonnegative supermartingale convergence theory, uniformly conditionally joint connectivity condition is established to ensure the convergence of the distributed stochastic optimization algorithms. The joint connectivity condition for Markovian and deterministic switching graphs, and the connectivity condition on the mean graph for i.i.d. graphs are all special cases of our condition.

II. The co-existence of random graphs, subgradient measurement noises, additive and multiplicative communication noises are considered. Compared to the case with only a single random factor, the coupling terms of different random factors will inevitably appear on the right hand side of the difference inequality of the square of the difference between optimizers’ states and any given vector. What’s more, multiplicative noises relying on the relative states between adjacent local optimizers make states, graphs and noises coupled together. Therefore, it becomes more complex to estimate the mean square upper bound of the local optimizers’ states. We first introduce the property of conditional independence to deal with the coupling term of different random factors. Then, we prove that the mean square upper bound of the coupling term between states, network graphs and noises depends on the second-order moment of the difference between optimizers’ states and the given vector. Finally, we get an estimate of the mean square divergence rate of the local optimizers’ states in terms of the step sizes of the algorithm.

III. We do not require bounded subgradients of local cost functions. Compared to [9, 21, 22, 38, 39, 36, 40, 41] and [25, 44, 45, 46], we only assume that the subgradients of local cost functions are linearly growth functions. As a result, the existing methods for estimating the terms which couple the subgradients with the errors between local optimizers’ states and the global optimal solution are no longer applicable. To this end, we substitute the mean square divergence rate of the local optimizers’ states into the Lyapunov function difference inequality of the state consensus error, and obtain mean square average consensus and the convergence rate, based on which we prove that the states of all local optimizers converge to the same global optimal solution almost surely by the non-negative supermartingale convergence theorem.

The paper is organized as follows. In Section 2, the problem formulation and an example of distributed statistical machine learning for the described problem is given. In Section 3, the main results are presented. In Section 4 conclusion remarks and future research topics are given.

Notation and symbols: \( \mathbf{1}_N \): \( N \)-dimensional vector with all ones; \( \mathbf{0}_N \): \( N \)-dimensional vector with all zeros; \( I_N \): \( N \)-dimensional identity matrix; \( O_{m \times n} \): \( m \times n \) dimensional zero matrix; \( \mathbb{R} \): the set of real numbers; \( A \geq B \): matrix \( A - B \) is positive semi-definite; \( A \geq B \): matrix \( A - B \) is a nonnegative matrix; \( A \otimes B \): the Kronecker product of matrices \( A \) and \( B \); \( A^T \): the transpose of matrix \( A \); \( Tr(A) \): the trace of matrix \( A \); \( \lambda_2(A) \): the second minimum eigenvalue of a real symmetric matrix \( A \); \( \text{diag} \left( B_1, \ldots, B_n \right) \): the block diagonal matrix with entries being \( B_1, \ldots, B_n \); \( \| A \| \): the Frobenius-norm of matrix \( A \); \( E[\mathcal{G}] \): the mathematical expectation of random variable \( \mathcal{G} \); \( | S | \): the cardinal number of set \( S \); \( \lceil x \rceil \): the minimal integer greater than or equal to real number \( x \); \( b_n = O ( r_n ) \): \( \lim \sup_{n \to \infty} \frac{b_n}{r_n} < \infty \), where \( \{ b_n, n \geq 0 \} \) is a real sequence and \( \{ r_n, n \geq 0 \} \) is a positive real sequence; \( b_n = o ( r_n ) \): \( \lim_{n \to \infty} \frac{b_n}{r_n} = 0 \); \( T_{\gamma}(k) = \sigma(\eta(j), 0 \leq j \leq k), k \geq 0, T_{\gamma}(-1) = \{ \Omega, \emptyset \} \), where \( \{ \eta(k), k \geq 0 \} \) is a sequence of random vectors or matrices, and \( \Omega \) is the sample space; \( \text{dom}(f) \): the domain of function \( f \); \( d_f(\bar{x}) \): a subgradient of the convex function \( f \) at \( \bar{x} \), which is a vector satisfying

\[
\frac{d_f(\bar{x})}{d_f(\bar{x})} \leq f(x), \forall x \in \text{dom}(f); \tag{1.1}
\]
\( d_f(\bar{x}) \): the sub-differential set of the convex function \( f \) at \( \bar{x} \), which is a nonempty set denoting the set of
all subgradients of \( f \) at \( \bar{x} \) (**2**).

### 2 Problem formulation

Consider a network with \( N \) nodes. Each node represents a local optimizer. The objective of the network is to solve the optimization problem:

\[
\min_{x \in \mathbb{R}^n} \ f(x) = \sum_{i=1}^{N} f_i(x), \tag{2.1}
\]

where each \( f_i(\cdot) : \mathbb{R}^n \to \mathbb{R} \) is a convex function, representing the local cost function, which is only known to optimizer \( i \). For the problem (**2.1**), denote the optimal value by \( f^* = \min_{x \in \mathbb{R}^n} \ f(x) \) and the set of optimal solutions by \( X^* = \{ x \in \mathbb{R}^n : f(x) = f^* \} \).

The information structure of the network is described by a sequence of random digraphs \( \{ \mathcal{G}(k) = (\mathcal{V}, \mathcal{E}_{\mathcal{G}(k)}, \mathcal{A}_{\mathcal{G}(k)}), k \geq 0 \} \), where \( \mathcal{V} = \{ 1, \ldots, N \} \) is the set of nodes, \( \mathcal{E}_{\mathcal{G}(k)} \) is the set of edges at time instant \( k \), and \( (j, i) \in \mathcal{E}_{\mathcal{G}(k)} \) if and only if the \( j \)th optimizer can send information to the \( i \)th optimizer directly. The neighborhood of the \( i \)th optimizer at time instant \( k \) is denoted by \( \mathcal{N}_i(k) = \{ j \in \mathcal{V} | (j, i) \in \mathcal{E}_{\mathcal{G}(k)} \} \). \( \mathcal{A}_{\mathcal{G}(k)} = [a_{ij}(k)]_{i,j=1}^{N} \) is the generalized weighted adjacency matrix at time instant \( k \), where \( a_{ii}(k) = 0 \), and \( a_{ij}(k) \neq 0 \Leftrightarrow j \in \mathcal{N}_i(k) \), representing the weight on channel \((j, i)\) at time instant \( k \). The generalized Laplacian matrix of the digraph \( \mathcal{G}(k) \) is denoted by \( \mathcal{L}_{\mathcal{G}(k)} = [L_{ij}(k)]_{i,j=1}^{N} \). Let \( \hat{\mathcal{G}}(k) = (\mathcal{V}, \mathcal{E}_{\mathcal{G}(k)} \cup \mathcal{E}_{\hat{\mathcal{G}}(k)}), \mathcal{A}_{\hat{\mathcal{G}}(k)} = \mathcal{L}_{\mathcal{G}(k)} + \mathcal{L}_{\hat{\mathcal{G}}(k)} \) be the symmetrized graph of \( \mathcal{G}(k) \), where \((i, j) \in \mathcal{E}_{\hat{\mathcal{G}}(k)} \) if and only if \((j, i) \in \mathcal{E}_{\mathcal{G}(k)} \). Denoted \( \hat{\mathcal{L}}_{\hat{\mathcal{G}}(k)} = \frac{\mathcal{L}_{\mathcal{G}(k)} + \mathcal{L}_{\hat{\mathcal{G}}(k)}}{2} \). If \( a_{ij}(k) \geq 0, \forall, i, j \in \mathcal{V} \), then the generalized weighted adjacency matrix \( \mathcal{A}_{\mathcal{G}(k)} \) and the generalized Laplacian matrix \( \hat{\mathcal{L}}_{\hat{\mathcal{G}}(k)} \) degenerate to the weighted adjacency matrix and Laplacian matrix in usual sense, respectively. And \( \hat{\mathcal{L}}_{\hat{\mathcal{G}}(k)} \) is the Laplacian matrix of \( \hat{\mathcal{G}}(k) \) if and only if \( \mathcal{G}(k) \) is balanced (**2.3**).

We consider the following distributed stochastic subgradient algorithm

\[
x_i(k + 1) = x_i(k) + c(k) \sum_{j \in \mathcal{N}_i(k)} a_{ij}(k)(y_{ji}(k) - x_i(k)) - a(k) \bar{d}_{ji}(x_i(k)), \quad k \geq 0, \quad i \in \mathcal{V}, \tag{2.2}
\]

where \( x_i(k) \in \mathbb{R}^n \) is the state of the \( i \)th optimizer at time instant \( k \), representing its local estimate of the global optimal solution to the problem (**2.1**); \( x_i(0) \in \mathbb{R}^n, i = 1, 2, \ldots, N \) are the initial values; \( c(k) \) and \( a(k) \) are the time-varying step sizes; \( y_{ji}(k) \in \mathbb{R}^n \) denotes the measurement of the neighbouring optimizer \( j \)’s state by optimizer \( i \) at time instant \( k \), which is given by

\[
y_{ji}(k) = x_j(k) + \psi_{ji}(x_j(k) - x_i(k))\xi_{ji}(k), \quad j \in \mathcal{N}_i(k), \quad i \in \mathcal{V}, \tag{2.3}
\]

where \( \{\xi_{ji}(k), k \geq 0\} \) is the sequence of communication noises in channel \((j, i)\), \( \psi_{ji}(\cdot) : \mathbb{R}^n \to \mathbb{R} \) is the noise intensity function; and \( \bar{d}_{ji}(x_i(k)) \) denotes the noisy measurement of the subgradient \( d_{ji}(x_i(k)) \) by optimizer \( i \), i.e.

\[
\bar{d}_{ji}(x_i(k)) = d_{ji}(x_i(k)) + \zeta_{ji}(k), \tag{2.4}
\]

where \( \{\zeta_{ji}(k), k \geq 0\} \) is the measurement noise sequence.

Denote \( X(k) = [x_1^T(k), \ldots, x_N^T(k)]^T, \xi(k) = [\xi_{11}^T(k), \ldots, \xi_{N1}^T(k); \ldots; \xi_{1N}^T(k), \ldots, \xi_{NN}^T(k)]^T, \) where \( \xi_{ji}(k) \equiv 0, \) if \( j \notin \mathcal{N}_i(k) \) for all \( k \geq 0, \) and \( \xi(k) = [\xi_{i1}^T(k), \ldots, \xi_{iN}^T(k)]^T \). For the optimization model (**2.1**), the measurement model (**2.3**) and (**2.4**), we have the following assumptions.

**Assumption 1** (Linear growth condition) There exist nonnegative constants \( \sigma_{di} \) and \( C_{di} \) such that \( \|d_{ji}(x)\| \leq \sigma_{di}\|x\| + C_{di}, x \in \mathbb{R}^n \) for all \( d_{ji}(x) \in \partial f_i(x), i = 1, \ldots, N. \)
Assumption 2 There exists a σ-algebra flow \( \{ F(k), k \geq 0 \} \), such that \( \{ \xi(k), F(k), k \geq 0 \} \) and \( \{ A_{G(k)}, F(k), k \geq 0 \} \) are adaptive processes. The communication noise process \( \{ \xi(k), F(k), k \geq 0 \} \) is a vector-valued martingale difference and there exists a positive constant \( C_\xi \) such that \( \sup_{k \geq 0} E[\|\xi(k)\|^2 | F(k-1)] \leq C_\xi \) a.s. For any given time instant \( k \), \( \sigma(\xi(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

Assumption 3 For the σ-algebra flow given by Assumption 2, the subgradient measurement noise process \( \{ \zeta(k), F(k), k \geq 0 \} \) is a vector-valued martingale difference. There exist nonnegative constants \( \sigma_\zeta \) and \( C_\zeta \) such that \( E[\|\zeta(k)\|^2 | F(k-1)] \leq \sigma_\zeta \|X(k)\|^2 + C_\zeta \) a.s. For any given time instant \( k \), \( \sigma(\zeta(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

Assumption 4 There exist nonnegative constants \( \sigma_{ji} \) and \( b_{ji}, i, j \in V \), such that \( |\psi_{ji}(x)| \leq \sigma_{ji} \|x\| + b_{ji}, \forall x \in \mathbb{R}^n \).

Assumption 5 The set \( X^* \) of optimal solutions is a non-empty countable set.

We give an example of distributed statistical machine learning satisfying the model assumptions above. The local cost function \( f_i \) is the risk function associated with the \( i \)th optimizer’s local data, i.e.

\[
f_i(x) = E[\ell_i(x; \mu_i)] + R_i(x),
\]

where \( \ell_i(\cdot; \cdot) \) is a loss function which is convex with respect to its first argument, \( \mu_i \) is the data sample of optimizer \( i \), and \( R_i : \mathbb{R}^n \rightarrow \mathbb{R} \) is a convex regularization term (\(|49|\)). It is known that \( L_1 \)-regularization and \( L_2 \)-regularization are two common regularization methods in machine learning. An example of \( L_2 \)-regularization is given in [21], for which Assumption 1 naturally holds. If the quadratic loss is considered with \( L_1 \)-regularization, then it is called the LASSO regression problem:

\[
\min_{x \in \mathbb{R}^n} \sum_{i=1}^{N} (E[\ell_i(x; u_i(k), p_i(k))] + \kappa \|x\|_1),
\]

where

\[
\ell_i(x; u_i(k), p_i(k)) = \frac{1}{2} \|p_i(k) - u_i(k)x\|^2,
\]

with

\[
p_i(k) = u_i^T(k)x_0 + v_i(k),
\]

in which \( x_0 \in \mathbb{R}^n \) is an unknown parameter, \( u_i(k) \in \mathbb{R}^n \) is the regression vector of the \( i \)th optimizer, and \( v_i(k) \) is the local measurement noise. Random sequences \( \{u_i(k), k \geq 0\} \) and \( \{v_i(k), k \geq 0\} \) are mutually independent i.i.d. Gaussian sequences with distributions \( \mathcal{N}(0, R_{a,ij}) \) and \( \mathcal{N}(0, \sigma_{v,ij}^2) \), respectively. For this case, \( f_i(x) = E[\ell_i(x; u_i(k), p_i(k))] + \kappa \|x\|_1 \).

If we apply the algorithm (\ref{eq:1}) to the problem (\ref{eq:15})-\ref{eq:19}, then it can be verified that Assumptions 1-3 hold. See appendix B for details.

We consider the following conditions of algorithm step sizes:

\[
\sup_{k \geq 0} E[\|\xi(k)\|^2 | F(k-1)] \leq C_\xi \) a.s. For any given time instant \( k \), \( \sigma(\xi(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

\]

\[
\sup_{k \geq 0} E[\|\zeta(k)\|^2 | F(k-1)] \leq \sigma_\zeta \|X(k)\|^2 + C_\zeta \) a.s. For any given time instant \( k \), \( \sigma(\zeta(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

\]

\[
\sup_{k \geq 0} E[\|\zeta(k)\|^2 | F(k-1)] \leq \sigma_\zeta \|X(k)\|^2 + C_\zeta \) a.s. For any given time instant \( k \), \( \sigma(\zeta(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

\]

\[
\sup_{k \geq 0} E[\|\zeta(k)\|^2 | F(k-1)] \leq \sigma_\zeta \|X(k)\|^2 + C_\zeta \) a.s. For any given time instant \( k \), \( \sigma(\zeta(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

\]

\[
\sup_{k \geq 0} E[\|\zeta(k)\|^2 | F(k-1)] \leq \sigma_\zeta \|X(k)\|^2 + C_\zeta \) a.s. For any given time instant \( k \), \( \sigma(\zeta(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

\]

\[
\sup_{k \geq 0} E[\|\zeta(k)\|^2 | F(k-1)] \leq \sigma_\zeta \|X(k)\|^2 + C_\zeta \) a.s. For any given time instant \( k \), \( \sigma(\zeta(k)) \) and \( \sigma(A_{G(k)}, A_{G(k+1)}, \ldots) \) are conditionally independent given \( F(k-1) \).

\]
(C1) $c(k) \downarrow 0$, $\alpha(k) \downarrow 0$, $\sum_{k=0}^\infty \alpha(k) = \infty$, $\sum_{k=0}^\infty \alpha^2(k) < \infty$, $\sum_{k=0}^\infty c(k) = \infty$, $\sum_{k=0}^\infty c^2(k) < \infty$, $c(k) = O(c(k+1))$, $k \to \infty$;

(C2) $\lim_{k \to \infty} \frac{c^2(k)}{\alpha(k)} = 0$;

(C3) For any given positive constant $C$, $\sum_{k=0}^\infty \alpha(k) \exp(-C \sum_{t=0}^k \alpha(t)) < \infty$;

(C4) For any given positive constant $C$, $\lim_{k \to \infty} \frac{\alpha(k) \exp(C \sum_{t=0}^k \alpha(t))}{\alpha(k)} = 0$;

(C5) For any given positive constant $C$, the sequence $\{\alpha(k) \exp(C \sum_{t=0}^k \alpha(t)), k \geq 0\}$ decreases monotonically for sufficiently large $k$ and $\alpha(k) \exp(C \sum_{t=0}^k \alpha(t)) - \alpha(k+1) \exp(C \sum_{t=0}^{k+1} \alpha(t)) = O(\alpha^2(k) \exp(2C \sum_{t=0}^k \alpha(t)))$.

Remark 2.1 There exist step sizes satisfying Conditions (C1)-(C5). For example,

\[
\begin{align*}
\alpha(k) &= \frac{\alpha_1}{(k + 3) \ln(\tau_1(k + 3))}, \quad \tau_1 \in (0, 1], \\
c(k) &= \frac{\alpha_2}{(k + 3)^{\tau_2} \ln^{\tau_2}(k + 3)}, \quad \tau_2 \in (0.5, 1), \quad \tau_3 \in (-\infty, 1],
\end{align*}
\]

where $\alpha_1, \alpha_2$ are given positive constants.

3 Main results

Let $D(k) = \text{diag}(a_{11}^T(k), \ldots, a_{N1}^T(k)) \otimes I_n$, where $a_{ij}^T(k)$ is the $i$th row of $A_{\mathcal{G}(k)}$, $i = 1, \ldots, N$, $\psi_i(k) = \text{diag}(\psi_{i1}(x_1(k) - x_i(k)), \ldots, \psi_{iN}(x_N(k) - x_i(k)))$, $i = 1, \ldots, N$, $\Psi(k) = \text{diag}(\psi_1(k), \ldots, \psi_N(k)) \otimes I_n$, and $d(k) = [d_{11}^T(x_1(k)), \ldots, d_{N1}^T(x_N(k))]^T$.

Rewrite the algorithm (2.2)-(2.4) in a compact form as

\[
X(k + 1) = ((I_N - c(k)P_{\mathcal{G}(k)} \otimes I_n)X(k) + c(k)D(k)\Psi(k)\xi(k) - \alpha(k)d(k)) + \zeta(k).
\]

(3.1)

For any non-negative integer $k$ and positive integer $h$, denote

\[
\lambda^h_k = \lambda_2 \left( \sum_{i=k}^{k+h-1} E[\hat{L}_{\mathcal{G}(i)}|\mathcal{F}(k-1)] \right).
\]

(3.2)

Note that if $\mathcal{G}(k) \in \Gamma$, then $E[\hat{L}_{\mathcal{G}(i)}|\mathcal{F}(k-1)]$ is a symmetric matrix a.s., and $\lambda^h_k$ is well defined.

Denote the consensus error vector $\delta(k) = (P \otimes I_n)X(k)$ and the Lyapunov function $V(k) = ||\delta(k)||^2$, where $P = I_N - \frac{1}{N}1_N1_N^T$. By det$(I_N - P)$ = det$((\lambda - 1)I_N + \frac{1}{N}1_N1_N^T) = (\lambda - 1 + \frac{1}{N})(\lambda - 1)^{N-1}$, we have $||P|| = \sqrt{4\max(P^TP)} = 1$. By $(P \otimes I_n)(1_N1_N^T \otimes I_n) = 0_{nN \times nN}$, we have $(P \otimes I_n)\delta(k) = (P \otimes I_n)\delta(k)$. Therefore,

\[
\begin{align*}
(P \otimes I_n)((I_N - c(k)P_{\mathcal{G}(k)} \otimes I_n)X(k) &= (P \otimes I_n)X(k) - c(k)(P \otimes I_n)(P_{\mathcal{G}(k)} \otimes I_n)X(k) \\
= \delta(k) - c(k)(P \otimes I_n)\delta(k) \\
= ((I_N - c(k)P_{\mathcal{G}(k)} \otimes I_n)) \delta(k),
\end{align*}
\]

(3.3)

which together with (3.1) gives

\[
\delta(k + 1) = ((I_N - c(k)P_{\mathcal{G}(k)} \otimes I_n)\delta(k) + c(k)(P \otimes I_n)D(k)\Psi(k)\xi(k)
\]

(3.4)
For the problem (2.1), the algorithm (2.2)-(2.4) and the associated random graph sequence \( \{G(k), k \geq 0\} \in \Gamma_1 \), assume that

(a) Assumptions 7-15 and Conditions (C1)-(C5) hold;

(b) there exists a deterministic positive integer \( h \), positive constants \( \theta \) and \( \rho_0 \), such that

\[
\inf_{|m| \geq 0} \lambda^{|m|} \geq \theta \quad \text{a.s.};
\]

\[
\sup_{k \geq 0} \left[ E \|L_{G(k)}\|^2_{\max(k-1)} |F(k-1)| \right]^{\sup(k-1)} \leq \rho_0 \quad \text{a.s.}
\]

Then, there exists a random vector \( z^* \) taking values in \( X^* \), such that \( \lim_{k \to \infty} x_i(k) = z^* \ a.s., \ i = 1, \ldots, N \).

**Remark 3.1** Condition (b.1) is called the uniformly conditionally joint connectivity condition (2.3), i.e. the conditional digraphs over the intervals \([mh, (m+1)h - 1], m \geq 0\) are jointly connected, and the average algebraic connectivity is uniformly bounded away from zero.

Next, we consider two special classes of random graph sequences, i.e. \( \{G(k), k \geq 0\} \) is a Markov chain with countable state space and \( \{G(k), k \geq 0\} \) is an independent process with uncountable state space. For these two special cases, Condition (b.1) of Theorem 3.1 becomes more intuitive and Condition (b.2) is weakened.

Denote \( S_1 = \{A_j, j = 1, 2, \ldots, \} \), which is a countable set of generalized weighted adjacency matrices and denote the associated generalized Laplacian matrix of \( A_j \) by \( L_j \). Let \( \hat{L}_j = \frac{L_j + L_j^T}{2} \). We consider the following random graph sequences

\[
\Gamma_2 = \{\{G(k), k \geq 0\} | \{A_{G(k)}, k \geq 0\} \subseteq S_1 \ is \ a \ homogeneous \ and \ uniformly \ ergodic \ Markov \ chain \ with \ unique \ stationary \ distribution \ \pi; E [A_{G(k)}|A_{G(k-1)}] \geq O_{N \times N} \ a.s., \ \text{and the associated digraph of} \ E [A_{G(k)}|A_{G(k-1)}] \ is \ balanced \ a.s., \ k \geq 0\}.
\]

Here, \( \pi = [\pi_1, \pi_2, \ldots]^{T}, \ \pi_j \geq 0, \sum_{j=1}^{N} \pi_j = 1, \) where \( \pi_j \) denotes the stationary probability at \( A_j \). For the concept and properties of uniformly ergodic Markov chains, the readers may refer to [50]. For Markovian switching graph sequences, we have the following corollary.

**Corollary 3.1** For the problem (2.1), the algorithm (2.2)-(2.4) and the associated random graph sequence \( \{G(k), k \geq 0\} \in \Gamma_2 \), assume that

(i) Assumptions 7-15 and Conditions (C1)-(C5) hold;

(ii) the associated graph of the Laplacian matrix \( \sum_{j=1}^{N} \pi_j L_j \) contains a spanning tree;

(iii) \( \sup_{|k| \geq 1} \|\hat{L}_j\| < \infty \).

Then, there exists a random vector \( z^* \) taking values in \( X^* \), such that \( \lim_{k \to \infty} x_i(k) = z^* \ a.s., i = 1, \ldots, N \).

**Proof.** From the definition of \( \Gamma_2 \), we know that \( \Gamma_2 \subseteq \Gamma_1 \). Then, similar to the proof of Theorem 2 in [23], we get that Condition (b.1) of Theorem 3.1 holds by Condition (ii). And from Condition (iii), we know that Condition (b.2) of Theorem 3.1 holds. Finally, the conclusion of Corollary 3.1 is obtained by Theorem 3.1.
Consider the independent graph sequences
\[ \Gamma_3 = \{ (\mathcal{G}(k), k \geq 0) \mid (\mathcal{G}(k), k \geq 0) \} \text{ is an independent process, } E [\mathcal{A}_{\mathcal{G}(k)}] \geq O_{N \times N}, \text{ a.s.,} \]
and the associated digraph of \( E [\mathcal{A}_{\mathcal{G}(k)}] \) is balanced a.s., \( k \geq 0 \).

For independent graph sequences, we have the following corollary, whose proof is omitted.

**Corollary 3.2** For the problem (2.1), the algorithm (2.2)-(2.4) and the associated random graph sequence \( \{ \mathcal{G}(k), k \geq 0 \} \in \Gamma_3 \), assume that
(i) Assumptions 1 and Conditions (C1)-(C5) hold;
(ii) there exists a positive integer \( h \) such that
\[ \inf_{m \geq 0} \left\{ A_2 \left( \sum_{i=0}^{(m+1)h-1} E \left[ L_{\mathcal{G}(i)} \right] \right) \right\} > 0; \]
(iii) \( \sup_{k \geq 0} E \left[ \| L_{\mathcal{G}(k)} \|^2 \right] < \infty. \)

Then, there exists a random vector \( z^* \) taking values in \( X^* \), such that \( \lim_{k \to \infty} x_i(k) = z^* \) a.s., \( i = 1, \cdots, N. \)

---

**4 Conclusions**

We have studied the distributed stochastic optimization by networked nodes to cooperatively minimize a sum of convex cost functions. The distributed stochastic subgradient optimization algorithm with two different time-varying diminishing step sizes has been considered. Compared to the existing literature, our model is more widely applicable in the sense that i) the graphs are not required to be spatially and temporally independent, and their edge weights are not necessarily nonnegative almost surely; ii) the measurement covers both additive and multiplicative communication noises; iii) the local cost functions do not need to be differentiable, nor do their subgradients need to be bounded. By stochastic Lyapunov method, convex analysis, algebraic graph theory and martingale convergence theory, it has been proved that if the local subgradient functions grow linearly and the sequence of digraphs is conditionally balanced and uniformly conditionally jointly connected, then proper algorithm step sizes can be designed so that all nodes’ states converge to the global optimal solution almost surely. Constrained stochastic optimization is an important research area for its widespread applications. The existence of constraints may destroy the linearity in the information evolution which leads to the failure of analyzing the piecewise consensus of nodes’ states by using the piecewise binomial expansion of the product of random matrices. We leave this topic for future investigation.

---

**A Proofs of Theorem 3.1**

The proof of Theorem 3.1 needs the following Lemmas A.1-A.6.

**Lemma A.1** For the problem (2.1), consider the algorithm (2.2)-(2.4). If Assumption [1] holds, then \( \| d(k) \|^2 \leq 2\sigma_d^2 \| X(k) \|^2 + 2NC_d^2 \), where \( \sigma_d = \max_{1 \leq i \leq N} \{ \sigma_{di} \} \) and \( C_d = \max_{1 \leq i \leq N} \{ C_{di} \}. \)

**Proof.** By Assumption [1] we have \( \| d(k) \|^2 = \sum_{i=1}^{N} \| d_i(k) \|^2 \leq \sum_{i=1}^{N} (\sigma_{di} \| x_i(k) \| + C_{di})^2 \leq 2\sigma_d^2 \| X(k) \|^2 + 2NC_d^2. \)

**Lemma A.2** For the problem (2.1), consider the algorithm (2.2)-(2.4). If Assumption [3] holds, then \( E[\| z(k) \|^2] \leq \sigma_z E[\| X(k) \|^2] + C_z. \)

---

8
Proof. By Assumption $\exists$ we have $E[\|\zeta(k)\|^2] = E[E[\|\zeta(k)\|^2|\mathcal{F}(k-1)]] \leq \sigma \varepsilon E[\|X(k)\|^2] + C_{\varepsilon}$. \hfill $\Box$

**Lemma A.3** For the problem (2.1), consider the algorithm (2.2)–(2.4). If Assumption $\exists$ holds, then for any given $x \in \mathbb{R}^n$, we have

$$||\Psi(k)||^2 \leq 4\sigma^2 ||X(k) - 1_N \otimes x||^2 + 2b^2.$$ 

(A.1)

Especially,

$$||\Psi(k)||^2 \leq 4\sigma^2 V(k) + 2b^2.$$ 

(A.2)

where $\sigma = \max_{1 \leq i, j \leq N} \{ \sigma_{ji} \}, b = \max_{1 \leq i, j \leq N} \{ b_{ji} \}$.

Proof. By the definition of $\Psi(k)$ and Assumption $\exists$ we have

$$||\Psi(k)||^2 = \max_{1 \leq i, j \leq N} (\psi_{ji}(x_j(k) - x_i(k))^2$$

$$\leq \max_{1 \leq i, j \leq N} [2\sigma^2 ||x_j(k) - x_i(k)||^2 + 2b^2]$$

$$\leq 4\sigma^2 \max_{1 \leq i, j \leq N} \left( ||x_j(k) - x_i||^2 + ||x_i(k) - x_i||^2 \right) + 2b^2$$

$$\leq 4\sigma^2 \sum_{i=1}^{N} ||x_j(k) - x_i||^2 + 2b^2$$

$$= 4\sigma^2 ||X(k) - 1_N \otimes x||^2 + 2b^2.$$ 

Therefore, (A.1) holds. Then replacing $x$ by $\frac{1}{N} \sum_{i=1}^{N} x_i(k)$ gives (A.2). \hfill $\Box$

**Lemma A.4** For the problem (2.1), consider the algorithm (2.2)–(2.4). Suppose that $\{ G(k), k \geq 0 \} \in \Gamma_1$, Assumptions $\exists$ hold, and there exists a positive constant $\rho_0$, such that $\sup_{k \geq 0} \left[ E[||\mathcal{L}_{G(k)}||^2|\mathcal{F}(k-1)]\right]^{\frac{1}{2}} \leq \rho_0$ a.s.. Then

$$E[V(k+1)] \leq (1 + 2c^2(k)(\rho_0^2 + 8\sigma^2 C_{\varepsilon} p_1)) E[V(k)] + 8b^2 C_{\varepsilon} p_1 c^2(k)$$

$$+ 2a^2(k)(2\sigma \varepsilon + 3\sigma_D^2) E[||X(k)||^2] + 2a^2(k)(2\sigma \varepsilon + 3NC_D^2)$$

$$- 2E[\alpha(k) d^2(k) (P \otimes I_n) \delta(k)], \quad \forall k \geq 0.$$ 

(A.3)

And for any given $x \in \mathbb{R}^n$,

$$E[||X(k+1) - 1_N \otimes x||^2]$$

$$\leq (1 + 2c^2(k)(\rho_0^2 + 8\sigma^2 C_{\varepsilon} p_1) + 4a^2(k)(2\sigma \varepsilon + 3\sigma_D^2)) E[||X(k) - 1_N \otimes x||^2]$$

$$+ 8b^2 C_{\varepsilon} p_1 c^2(k) + 2a^2(k)(2\sigma \varepsilon + 3NC_D^2 + 2(3\sigma_D^2 + 2\sigma \varepsilon)) N ||x||^2$$

$$- 2\alpha(k) E[d^2(k) (X(k) - 1_N \otimes x)], \quad \forall k \geq 0.$$ 

(A.4)

where $p_1$ is a positive constant satisfying $\sup_{k \geq 0} E[|\mathcal{E}_{G(k)}| \max_{1 \leq i, j \leq N} a_{ij}^2(k)|\mathcal{F}(k-1)|] \leq \rho_1$ a.s., $\sigma_D$ and $C_D$ are defined in Lemma [A.7].

Proof. By the definition of $V(k), \{ A.3 \}$, $||p - q||^2 \leq 2||p||^2 + 2||q||^2$ and $2p^T q \leq ||p||^2 + ||q||^2, \forall p, q \in \mathbb{R}^n$, we have

$$V(k+1) \leq V(k) - 2c(k) \delta^T(k) \left( \frac{L_{G(k)}^T P + P L_{G(k)}}{2} \right) \otimes I_n$$

$$+ 2c^2(k)||P||^2 \| L_{G(k)} \otimes I_n \|^2 ||\delta(k)||^2$$

9
\[ +2\left(c(k)D(k)\Psi(k)\xi(k) - \alpha(k)\xi(k)\right)^T(P \otimes I_n)((I_N - c(k)P)L_{\hat{G}(k)} \otimes I_n)\delta(k) \]
\[ +4c^2(k)\left(\|P \otimes I_n)D(k)\Psi(k)\xi(k)\|^2 + 4\alpha^2(k)\|P \otimes I_n)\xi(k)\|^2\]
\[ -2\alpha(k)\delta^T(k)(P \otimes I_n)\delta(k) + 3\|\alpha(k)(P \otimes I_n)d(k)\|^2. \] (A.5)

Now, we consider the mathematical expectation of each term on the right side of (A.5). For the 2nd term, noting that \( \hat{G}(k|k-1) \) is balanced a.s., we have \( E[P L_{\hat{G}(k)}|F(k-1)] = E[L_{\hat{G}(k)}|F(k-1)] \) a.s. Thus,
\[ E \left[ \frac{(L_{\hat{G}(k)}^T P^T + P L_{\hat{G}(k)}) \otimes I_n}{2} \right] F(k-1) \right] = E[L_{\hat{G}(k)} \otimes I_n|F(k-1)] \geq c_n \] a.s., and then, by \( \delta(k) \in F(k-1) \), we have
\[ E \left[ \frac{(L_{\hat{G}(k)}^T P^T + P L_{\hat{G}(k)}) \otimes I_n}{2} \right] \delta(k) \]
\[ = E \left[ \frac{(L_{\hat{G}(k)}^T P^T + P L_{\hat{G}(k)}) \otimes I_n}{2} \right] \delta(k) | F(k-1) \right] \]
\[ = E \left[ \frac{(L_{\hat{G}(k)}^T P^T + P L_{\hat{G}(k)}) \otimes I_n}{2} \right] | F(k-1) \delta(k) \]
\[ \geq 0. \] (A.6)

For the 3rd term, by \( \sup_{k \geq 0} \left[ E[\|L_{\hat{G}(k)} \otimes I_n\|^2|F(k-1)] \right] \leq \rho_0 a.s. \) and \( \delta(k) \in F(k-1) \), we get
\[ E \left[ \|L_{\hat{G}(k)} \otimes I_n\|^2 | F(k-1) \right] \leq \rho_0 \] a.s. and \( \delta(k) \in F(k-1) \), and Assumption 2 we have
\[ E[\xi^T(k)\Psi^T(k)D^T(k)(P \otimes I_n)((I_N - c(k)P)L_{\hat{G}(k)} \otimes I_n)\delta(k)] \]
\[ = E \left[ E[\xi^T(k)\Psi^T(k)D^T(k)(P \otimes I_n)((I_N - c(k)P)L_{\hat{G}(k)} \otimes I_n)\delta(k)] \right] \]
\[ = E \left[ E[\xi^T(k)|F(k-1)]\Psi^T(k)(P \otimes I_n)E[D^T(k)((I_N - c(k)P)L_{\hat{G}(k)} \otimes I_n)|F(k-1)]\delta(k) \right] \]
\[ = 0. \] (A.7)

where the third “=” is obtained from \( \|L_{\hat{G}(k)} \otimes I_n\| = \|L_{\hat{G}(k)}\| \). By \( \delta(k) \in F(k-1) \), and Assumption \( 2 \) we have
\[ E[\xi^T(k)\Psi^T(k)D^T(k)(P \otimes I_n)((I_N - c(k)P)L_{\hat{G}(k)} \otimes I_n)\delta(k)] = 0. \] Thus, for the 4th term, combining the above two equations gives
\[ E \left[ (c(k)D(k)\Psi(k)\xi(k) - \alpha(k)\xi(k))^T(P \otimes I_n)((I_N - c(k)P)L_{\hat{G}(k)} \otimes I_n)\delta(k) \right] = 0. \] (A.8)

Noting that \( E_{\hat{G}(k)} \) \( \max_{1 \leq i, j \leq N} \alpha_{i,j}^2(k) \leq N(N-1) \) \( \max_{1 \leq i, j \leq N} \alpha_{i,j}^2(k) \leq N(N-1) \|L_{\hat{G}(k)}\|_F^2 \), by \( \|L_{\hat{G}(k)}\|_F^2 \leq nN \|L_{\hat{G}(k)}\|^2 \)
\[ \text{and } \sup_{k \geq 0} \left[ E[\|L_{\hat{G}(k)}\|^2|F(k-1)] \right] \leq \rho_0 a.s., \text{ we know that } \rho_1 \text{ exists. Thus, for the 5th term, by (A.2) in Lemma A.3, we have} \]
\[ E[\xi^T(k)\Psi^T(k)D^T(k)(P \otimes I_n)(P \otimes I_n)D(k)\Psi(k)\xi(k)] \]
\[ \leq E[\|\Psi(k)\|^2 \|D^T(k)D(k)\| \|P \otimes I_n\|^2 \|P \otimes I_n\| \|\xi(k)\| \|F(k-1)\|] \]
\[ = E[\|\Psi(k)\|^2 \|D^T(k)D(k)\| \|\xi(k)\| \|F(k-1)\|] \]

10
we get $X$ term, by (X) = $\| \parallel X \parallel k \parallel I k (\parallel n k) \parallel 1, we have 

$$E[\xi^T(k)(P \otimes I_n)\xi(k)] \leq \|P \otimes I_n\|^2 E[\|\xi(k)\|^2] \leq \sigma^2 \|X(k)\|^2 + C_\epsilon.$$ (A.10)

For the last term on the right side of (A.5), from Lemma (A.1), we have

$$E[\alpha(k)d^T(k)(P \otimes I_n)^T \alpha(k)(P \otimes I_n)d(k)] \leq 2\sigma^2(k)(\|P \otimes I_n\|^2 E[\|d(k)\|^2]) + N C_n^2. (A.11)$$

Taking mathematical expectations on both sides of (A.5), by (A.6) + (A.11), we get (A.3).

In the following, we prove (A.4). Noting that $L_{\theta(k)}1_N = 0_{nN}$, we have $(L_{\theta(k)} \otimes I_n)(1_N \otimes x) = (L_{\theta(k)}1_N) \otimes (I_n \otimes x) = 0_{nN \times nN}$. Thus, by (A.1), $\|p - q\|^2 \leq 2\|p\|^2 + 2\|q\|^2$ and $2p^Tq \leq \|p\|^2 + \|q\|^2$, $\forall p, q \in \mathbb{R}^n$, we get

$$\|X(k + 1) - 1_N \otimes x\|^2 \leq \|X(k) - 1_N \otimes x\|^2 - 2c(k)(X(k) - 1_N \otimes x)^T (L_{\theta(k)}^T + L_{\theta(k)}) \otimes I_n (X(k) - 1_N \otimes x) + 2c(k)D(k)\Psi(k)\xi(k) + 4\sigma^2(k)(\|D(k)\|\Psi(k)\xi(k))^2 + 3\sigma^2(k)(\|d(k)\|d(k))^2.$$ (A.12)

Now, we consider the mathematical expectation of each term on the right side of (A.12). For the 2nd term, by $(X(k) - 1_N \otimes x) \in F(k - 1)$, we get

$$E[(X(k) - 1_N \otimes x)^T (L_{\theta(k)}^T + L_{\theta(k)}) \otimes I_n (X(k) - 1_N \otimes x)] = E[E[(X(k) - 1_N \otimes x)^T (L_{\theta(k)}^T \otimes I_n) (X(k) - 1_N \otimes x) | F(k - 1)] = E[(X(k) - 1_N \otimes x)^T (\overline{L}_{\theta(k)} \otimes I_n) (X(k) - 1_N \otimes x)] \geq 0.$$ (A.13)

For the 3rd term, by $\sup_{k \geq 0} E[\|L_{\theta(k)}\|^2 | F(k - 1)] \leq \rho_0$ a.s. and $(X(k) - 1_N \otimes x) \in F(k - 1)$, we have

$$E[\|L_{\theta(k)} \otimes I_n\|^2 (X(k) - 1_N \otimes x)^2] = E[E[\|L_{\theta(k)} \otimes I_n\|^2 | F(k - 1)] (X(k) - 1_N \otimes x)^2] \leq \rho_0 E[\|X(k) - 1_N \otimes x\|^2].$$ (A.14)

By $(X(k) - 1_N \otimes x) \in F(k - 1)$, Assumption 2, we have

$$E[\xi^T(k)\Psi(k)D^T(k)((I_N - c(k)\theta(k)) \otimes I_n)(X(k) - 1_N \otimes x)] = E[E[\xi^T(k)\Psi(k)D^T(k)((I_N - c(k)\theta(k)) \otimes I_n)(X(k) - 1_N \otimes x) | F(k - 1)]$$
Similarly, by Assumption 3, we have $E[\xi^T(k)(I_N - c(k)\mathcal{L}_{G(k)}) \otimes I_n](X(k) - 1_N \otimes x) = 0$. Thus, for the 4th term, by the above two equations, we get

$$E[(c(k)\mathcal{D}(k)\Psi(k)\xi(k) - \alpha(k)\xi(k))^T((I_N - c(k)\mathcal{L}_{G(k)}) \otimes I_n)(X(k) - 1_N \otimes x)] = 0.$$  \hspace{1cm} (A.15)

For the 5th term, by (A.1) in Lemma A.3, Assumption 2, we have

$$E[|\xi(k)|^2] \leq \max_{1 \leq i,j \leq N} a^2_{ij}(k) |\mathcal{F}(k - 1)\rangle.$$  \hspace{1cm} (A.16)

By $|\langle x|X\rangle|^2 \leq 2(|\langle x|X\rangle|^2 + |\langle -X|X\rangle|^2)$, Lemma A.1 and Lemma A.2, we derive $E[|d(k)|^2] \leq 4\sigma_d^2 E[|\langle x|X\rangle|^2] + 4\sigma_d^2 E[|\langle -X|X\rangle|^2] + 2NC_d^2$ and $E[|\xi(k)|^2] \leq 2\sigma_\xi E[|\langle x|X\rangle|^2] + 2\sigma_\xi^2 E[|\langle x|X\rangle|^2]^2 + C_\xi$. Then taking mathematical expectations on both sides of (A.12), by (A.13)-(A.16) and the above two inequalities, we get (A.4).

The following lemma gives a mean square upper bound of the divergence rate of the local optimizers’ states.

**Lemma A.5** For the problem (2.1), the algorithm (2.2)-(2.4) and the associated random graph sequence $\{G(k), k \geq 0\} \in \Gamma_1$, if Assumptions 1-4 and Conditions (C1)-(C3) hold, and there exists a positive constant $\rho_0$, such that $\sup_{k \geq 0} \left| E[|\mathcal{L}_{G(k)}|^2|\mathcal{F}(k - 1)\rangle\right|^2 \leq \rho_0$ a.s., then

$$E[|\langle x|X\rangle|^2] = O(\beta(k)),$$  \hspace{1cm} (A.17)

where $\beta(k) = \exp(C_0 \sum_{i=0}^k \alpha(t))$, and $C_0 = 1 + 2\rho_0^2 + 16\sigma_d^2 C_\xi \rho_1 + 8\sigma_\xi + 16\sigma_d^2$.

**Proof.** By $2p^Tq \leq \|p\|^2 + |q|^2$, $\forall p, q \in \mathbb{R}^n$ and Lemma A.1 we have

$$\begin{align*}
-2\alpha(k)d^T(k)(X(k) - 1_N \otimes x) & \leq \alpha(k)\left(\|d(k)\|^2 + \|X(k) - 1_N \otimes x\|^2\right) \\
& \leq \alpha(k)\left(2\sigma_d^2 |\langle x|X\rangle|^2 + 2NC_d^2 + \|X(k) - 1_N \otimes x\|^2\right) \\
& \leq \alpha(k)\left(2\sigma_d^2 |\langle x|X\rangle|^2 + 2N|\langle x|X\rangle|^2 + 2NC_d^2\right) \\
& = \alpha(k)(4\sigma_d^2 + 1)|\langle x|X\rangle|^2 + \alpha(k)(4\sigma_d^2 N|\langle x|X\rangle|^2 + 2NC_d^2), \hspace{1cm} (A.18)
\end{align*}$$

which together with (A.1) in Lemma A.4 leads to

$$\begin{align*}
E[|\langle x|X\rangle|^2 - 1_N \otimes x|^2] & \leq \left(1 + 2\sigma_d^2(\rho_0^2 + 8\sigma_\xi + 8\sigma_d^2 C_\xi \rho_1) + 4\alpha^2(2\sigma_\xi + 3\sigma_d^2)\right)E[|\langle x|X\rangle|^2 - 1_N \otimes x|^2] \\
& + 8\beta^2 C_\xi \rho_1 c^2(k) + 2\sigma_\xi^2(2\sigma_\xi + 3\sigma_d^2 + 2\sigma_\xi + 2\sigma_d^2 N|\langle x|X\rangle|^2)^2 \\
& + \alpha(k)(4\sigma_d^2 + 1)E[|\langle x|X\rangle|^2 - 1_N \otimes x|^2] + \alpha(k)(4\sigma_d^2 N|\langle x|X\rangle|^2 + 2NC_d^2) \\
& \leq \left(1 + 2\sigma_d^2(\rho_0^2 + 1) + 2\sigma_\xi^2(\rho_0^2 + 8\sigma_\xi + 8\sigma_d^2 C_\xi \rho_1) + 4\alpha^2(2\sigma_\xi + 3\sigma_d^2)\right)E[|\langle x|X\rangle|^2 - 1_N \otimes x|^2] \\
& + (4\sigma_d^2 N|\langle x|X\rangle|^2 + 2NC_d^2)\alpha(k) + 8\beta^2 C_\xi \rho_1 c^2(k).
\end{align*}$$

12
By Conditions (C1) and (C2), it is known that there exists a positive integer $k_0$, such that $\alpha(k) \geq \alpha^2(k)$ and $\alpha(k) \geq \epsilon^2(k)$, $\forall k \geq k_0$. Thus, from (A.19) with $x = 0_n$, we have

$$E[||X(k + 1)||^2] \leq \left(1 + \alpha(k)(4\sigma_0^2 + 1) + 2\alpha(k)(\rho_0^2 + 8\sigma_0^2C_{\epsilon}p_1) + 4\alpha(k)(2\sigma_\epsilon + 3\sigma_0^2)\right)E[||X(k)||^2] + 2NC_\alpha^2(\alpha + 8b^2C_{\epsilon}p_1\alpha(k) + 2\alpha(k)(2\sigma_\epsilon + 3NC_\alpha^2)$$

$$= \left(1 + C_0\alpha(k)\right)E[||X(k)||^2] + \tilde{C_0}\alpha(k), \quad k \geq k_0,$$

where $\tilde{C_0} = 8NC_\alpha^2 + 8b^2C_{\epsilon}p_1 + 4C_\epsilon$. This gives

$$E[||X(k + 1)||^2] \leq \prod_{i=k_0}^k \left[1 + C_0\alpha(i)\right]E[||X(k_0)||^2] + \tilde{C_0}\sum_{i=k_0}^k (1 + C_0\alpha(j))\alpha(i)$$

$$\leq \exp(C_0\sum_{i=k_0}^k \alpha(i))E[||X(k_0)||^2] + \tilde{C_0}\sum_{i=k_0}^k \alpha(i)\exp(C_0\sum_{j=0}^i \alpha(j))$$

$$\leq \beta(k)\left(E[||X(k_0)||^2] + \tilde{C_0}\sum_{i=k_0}^k \alpha(i)\exp(-C_0\sum_{j=0}^i \alpha(j))\right), \quad k \geq k_0.$$

Then by Condition (C3), we get

$$\limsup_{k \to \infty} \frac{E[||X(k + 1)||^2]}{\beta(k)} \leq E[||X(k_0)||^2] + \tilde{C_0}\sum_{i=k_0}^\infty \alpha(i)\exp(-C_0\sum_{j=0}^i \alpha(j)) < \infty,$$

i.e. $E[||X(k)||^2] = O(\beta(k - 1)) = O(\beta(k))$. \hfill \Box

The following lemma shows some important properties of the consensus error, i.e. the consensus error vanishes in mean square and almost surely, which will be used in the proof of Theorem 3.1

**Lemma A.6** For the problem (2.1), the algorithm (2.2)-(2.4) and the associated random graph sequence $\{\mathcal{G}(k), k \geq 0\} \in \Gamma_1$, assume that

(a) Assumptions [12] and Conditions (C1)-(C5) hold;

(b) there exists a positive integer $h$ and positive constants $\theta$ and $\rho_0$, such that

(b.1) $\inf_{m \geq 0} \lambda_{mh} \geq \theta$ a.s.;

(b.2) $\sup_{k \geq 0} \left[E[||L_G(k)||^2]^{\text{max}(k,2)}(\overline{F}(k - 1))\right]^{\text{max}(k,2)} \leq \rho_0$ a.s. then

$$E[V(k)] = O\left(\frac{\alpha(k)\beta(k)}{\epsilon(k)}\right),$$

(A.20)

and

$$V(k) \to 0, \quad k \to \infty \text{ a.s.},$$

(A.21)

where $\beta(k)$ is given in Lemma A.3.

**Proof.** Let $\Phi(m, s) = (I_N - c(m - 1)P_L\mathcal{G}(m-1)) \cdots (I_N - c(s)P_L\mathcal{G}(s)), m > s \geq 0, \Phi(s, s) = I_N, s \geq 0$. By (3.4) and some iterative calculations, we have

$$\delta(m + 1)h = (\Phi((m + 1)h, mh) \otimes I_n)\delta(mh) + \delta_{m_0}^{mh} - a_{m_0}^{mh},$$

(A.22)
where

\[
\tilde{\Lambda}_m^{mh} = \sum_{j=m}^{(m+1)h-1} (\Phi((m+1)h, j+1)P \otimes I_n)(c(j)D(j)\Psi(j)\xi(j) - \alpha(j)\zeta(j)), \tag{A.23}
\]

\[
\tilde{d}_m^{mh} = \sum_{j=m}^{(m+1)h-1} \alpha(j)(\Phi((m+1)h, j+1)P \otimes I_n)d(j). \tag{A.24}
\]

By the definition of \(V(k)\), (A.22) and \(-2(\tilde{\Lambda}_m^{mh})^T(\tilde{d}_m^{mh}) \leq (\tilde{\Lambda}_m^{mh})^T(\tilde{\Lambda}_m^{mh}) + (\tilde{d}_m^{mh})^T(\tilde{d}_m^{mh})\), we get

\[
V((m+1)h) \leq \delta^T(mh)(\Phi^T((m+1)h, mh)\Phi((m+1)h, mh) \otimes I_n)\delta(mh) + 2(\tilde{\Lambda}_m^{mh})^T(\tilde{\Lambda}_m^{mh}) + 2\delta^T(mh)(\Phi^T((m+1)h, mh) \otimes I_n)\tilde{\Lambda}_m^{mh} + 2(\tilde{d}_m^{mh})^T(\tilde{d}_m^{mh}) - 2\delta^T(mh)(\Phi^T((m+1)h, mh) \otimes I_n)\tilde{d}_m^{mh}. \tag{A.25}
\]

We now consider the mathematical expectation of each term on the right side of (A.25). For the first term, by Condition (C1), we know that there exists a positive integer \(m_0\) and a positive constant \(C_2\), such that \(c^2(mh) \leq C_2c^2((m+1)h)\), \(\forall \ m \geq m_0\), and \(c(k) \leq 1\), \(\forall \ k \geq m_0h\). From Condition (b.2) and the conditional Lyapunov inequality, we have

\[
\sup_{k \geq 0} E[\|L_{G(k)}\|^2|\mathcal{F}(k-1)] \leq \sup_{k \geq 0} [E[\|L_{G(k)}\|^2]^{1/2}|\mathcal{F}(k-1)]^{1/2} \leq \rho_0 \ a.s., \ \forall \ 2 \leq i \leq 2^h. \tag{A.26}
\]

By multiplying term by term, applying conditional Hölder inequality, noting that \(c(mh)\) decreases monotonously as \(m\) increases, and from (A.26), we have

\[
E\left[\left\|\Phi^T((m+1)h, mh)\Phi((m+1)h, mh) - I_N + \sum_{i=m}^{(m+1)h-1} c(i)(L_{G(i)}^TP^T + P L_{G(i)})(\mathcal{F}(mh-1)\right]\right] \leq (C_2 \sum_{i=2}^{2^h} M_{2i}^j \rho_0^j)c^2((m+1)h)
= C_1c^2((m+1)h), \ m \geq m_0,
\]

where \(C_1 = C_2[1 + \rho_0^{2h} - 1 - 2h\rho_0]\). From \(\delta(mh) \in \mathcal{F}(mh-1)\), we have \(V(mh) \in \mathcal{F}(mh-1)\). Noting that \(\|A \otimes I_n\| = \|A\|\), we have

\[
E\left[\left\|\delta^T(mh)\left[\Phi^T((m+1)h, mh)\Phi((m+1)h, mh) - I_N + \sum_{i=m}^{(m+1)h-1} c(i)(L_{G(i)}^TP^T + P L_{G(i)})(I_m)(\delta(mh)\right]\right]\right] \leq E\left[\left\|\Phi^T((m+1)h, mh)\Phi((m+1)h, mh) - I_N + \sum_{i=m}^{(m+1)h-1} c(i)(L_{G(i)}^TP^T + P L_{G(i)})(I_m)(\mathcal{F}(mh)\right]\right]V(mh) = E\left[\left\|\Phi^T((m+1)h, mh)\Phi((m+1)h, mh) - I_N + \sum_{i=m}^{(m+1)h-1} c(i)(L_{G(i)}^TP^T + P L_{G(i)})(I_m)(\mathcal{F}(mh)\right]\right]V(mh) \leq C_1c^2((m+1)h)E[V(mh)], \ m \geq m_0. \tag{A.27}
\]
Noting that \( G(i|i-1) \) is balanced a.s., it is known that \( G(i | mh - 1) \) is balanced a.s., \( mh \leq i \leq (m+1)h - 1 \). Then by Condition (b.1), we get

\[
E \left[ \delta^T(mh) \left[ \sum_{i=mh}^{(m+1)h-1} c(i)(P L_{G(i)} + L_{G(i)}^T P) \otimes I_n \right] \delta(mh) \right] = E \left[ \delta^T(mh) \left[ \sum_{i=mh}^{(m+1)h-1} c(i)(P L_{G(i)} + L_{G(i)}^T P) \otimes I_n \right] \delta(mh) \right] \cdot |T(mh - 1)|
\]

\[
= 2E \left[ \delta^T(mh) \left[ \sum_{i=mh}^{(m+1)h-1} c(i)E[|L_{G(i)} \otimes I_n](T(mh - 1)|] \delta(mh) \right] \right] \geq 2c((m+1)h)E \left[ \delta^T(mh) \left[ \sum_{i=mh}^{(m+1)h-1} E[|L_{G(i)} \otimes I_n](T(mh - 1)|] \delta(mh) \right] \right] \geq 2c((m+1)h)E[V(mh)],
\]

which together with (A.27) gives

\[
E[\tilde{\delta}^T(mh)(\Phi^T((m+1)h,mh)\Phi((m+1)h,mh) \otimes I_n)\delta(mh)] = E \left[ \delta^T(mh) \left[ \sum_{i=mh}^{(m+1)h-1} c(i)[L_{G(i)}^TP + PL_{G(i)}] \otimes I_n \right] \delta(mh) \right] + E[V(mh)] \leq 1 - 2\theta c((m+1)h) + C_1c^2((m+1)h)E[V(mh)], \quad m \geq m_0.
\]

For the second term on the right side of (A.25), by (A.23), we have

\[
(\tilde{\lambda}_m^{mh})^T(\tilde{\lambda}_m^{mh}) = (\tilde{\xi}_m^{mh} - \tilde{\zeta}_m^{mh})^T(\tilde{\xi}_m^{mh} - \tilde{\zeta}_m^{mh}) \leq 2(\tilde{\xi}_m^{mh})^T(\tilde{\xi}_m^{mh}) + 2(\tilde{\zeta}_m^{mh})^T(\tilde{\zeta}_m^{mh}),
\]

where \( \tilde{\xi}_m^{mh} = \sum_{j=mh}^{(m+1)h-1} c(j)\Phi((m+1)h,j+1)P \otimes I_n)D(j)\Psi(j)\xi(j) \) and \( \tilde{\zeta}_m^{mh} = \sum_{j=mh}^{(m+1)h-1} \alpha(j)\Phi((m+1)h,j+1)P \otimes I_n)\zeta(j) \). Then by Cr-inequality, we get

\[
E[|\tilde{\lambda}_m^{mh})^T(\tilde{\lambda}_m^{mh})] \leq h \sum_{j=mh}^{(m+1)h-1} c^2(j)E[|\xi(j)^T(j)^T D(j)((P \Phi^T((m+1)h,j+1)) \Phi((m+1)h,j+1)P \otimes I_n)D(j)\Psi(j)\xi(j)] \leq h \sum_{j=mh}^{(m+1)h-1} c^2(j)E[|\Phi^T((m+1)h,j+1)\Phi((m+1)h,j+1)|] \times|D(j)|^2|\Psi(j)|^2|\xi(j)|^2.
\]

From Assumption 2, it follows that

\[
E[(\Phi^T((m+1)h,j+1)\Phi((m+1)h,j+1)]|D(j)|^2|\Psi(j)|^2|\xi(j)|^2
\]
From conditional H"older inequality, Assumption 3, Lemma A.2 and (A.34), it follows that

\[ E[||\Psi(j)||^2 E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||
\times||D(j)||^2||\xi(j)||^2|\mathcal{F}(j-1)|] \]

\[ = E[||\Psi(j)||^2 E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||
\times||D(j)||^2|\mathcal{F}(j-1)| \cdot E[||\xi(j)||^2|\mathcal{F}(j-1)|] \]

\[ \leq C_\xi E[||\Psi(j)||^2 E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||
\times||D(j)||^2|\mathcal{F}(j-1)|], \quad mh \leq j \leq (m+1)h-1. \]  
\( (A.32) \)

From Condition (b.2), it is known that there exists a positive constant \( \rho_1 \), such that

\[ \sup_{k \geq 0} [E[||D(k)||^4|\mathcal{F}(k-1)|]]^{1/2} \leq \rho_1 \quad a.s. \]  
\( (A.33) \)

By Condition (b.2) and (A.26), we have

\[ |E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||^2|\mathcal{F}(j-1)|]|^2 \leq C_3 \quad a.s., \]  
\( (A.34) \)

where \( C_3 = \left\{ 2^{2(h-1)} \right\} \sum_{l=0}^{2(h-1)} M_{l,2(h-1)-l} \left( \begin{array}{c} h-1 \end{array} \right) \}, \quad M_{l,2(h-1)-l} \) is the combinatorial number of choosing \( l \) elements from \( 2(h-1) \). By (A.33), (A.34), and conditional H"older inequality, we get

\[ E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||^2|\mathcal{F}(j-1)|] \]

\[ \leq \{ E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||^2|\mathcal{F}(j-1)|] \}^{1/2} \]

\[ \times \{ E[||D(j)||^4|\mathcal{F}(j-1)|] \}^{1/2} \]

\[ \leq \rho_1 C_3 \quad a.s. \]  
\( (A.35) \)

From Lemma [A.3] we have \( ||\Psi(j)||^2 \leq 4\sigma^2 ||X(j)||^2 + 2b^2 \). Then, by (A.31), (A.32) and (A.35), we get

\[ E[\tilde{\zeta}_{mh}^T \tilde{\zeta}_{mh}] \leq hp_1 C_3 C_\xi \sum_{j=mh}^{(m+1)h-1} c^2(j)(4\sigma^2 E[||X(j)||^2] + 2b^2). \]  
\( (A.36) \)

From conditional H"older inequality, Assumption 3, Lemma A.2 and (A.34), it follows that

\[ E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||^2|\mathcal{F}(j-1)|] \]

\[ = E \left[ \left( E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||^2|\mathcal{F}(j-1)|] \right)^{1/2} \right] \]

\[ \times \phi((m+1)h, j+1)||^2|\mathcal{F}(j-1)| \]

\[ \leq C_3(\sigma_\xi E[||X(j)||^2] + C_\xi), \quad mh \leq j \leq (m+1)h-1, \]

which leads to

\[ E[\tilde{\zeta}_{mh}^T \tilde{\zeta}_{mh}] \]

\[ \leq hE \left[ \sum_{j=mh}^{(m+1)h-1} \alpha^2(j)\zeta^T(j)P\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)P\zeta(j) \right] \]

\[ \leq h \sum_{j=mh}^{(m+1)h-1} \alpha^2(j)E[||\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||^2] \]

\[ \leq hC_3 \sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(\sigma_\xi E[||X(j)||^2] + C_\xi). \]  
\( (A.37) \)
Thus, by (A.30), (A.36) and (A.37), we get
\[
E[\bar{\Lambda}_m^{nm} (\bar{\Lambda}_m^{nm})^T] \leq 2h\rho_1 C_3 C_\xi \sum_{j=mn}^{(m+1)h-1} c^2 \left(j(A\cdot\sigma_2 E[||X(j)||^2] + 2h^2) + 2hC_3 \sum_{j=mn}^{(m+1)h-1} \alpha^2(j)(\sigma_\xi E[||X(j)||^2] + C_\xi). \right. (A.38)
\]

For the third term on the right side of (A.25), by \(\delta(mh) \in F(j - 1), j \geq mh\), Assumption 2, we have
\[
E[\delta^T(mh)(\Phi^T((m+1)h, mh)\Phi((m+1)h, j+1)P \otimes I_n)] = 0, \quad mh \leq j \leq (m+1)h - 1, \quad m \geq 0. \quad (A.39)
\]
Similarly, from Assumption 3, we have \(E[\delta^T(mh)(\Phi^T((m+1)h, mh)\Phi((m+1)h, j+1)P \otimes I_n)\zeta(j)] = 0, \quad mh \leq j \leq (m+1)h - 1, \quad m \geq 0.\) This together with (A.23) and (A.39) gives
\[
E[\delta^T(mh)(\Phi((m+1)h, mh) \otimes I_n)^T \bar{\Lambda}_m^{nm}] = 0. \quad (A.40)
\]
By Lemma 1 and conditional Hölder inequality, we have
\[
E[\|\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||^2|d(j)||^2] \leq E[\|\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)||F(j - 1)||^2 + 2NC_\delta^2 j] \leq C(\sigma_\delta^2 E[||X(j)||^2] + 2NC_\delta^2). \quad j \geq mh,
\]
where the first “=” is derived by \(X(j) \in F(j - 1)\). This together with (A.24) and the Cauchy inequality gives
\[
E[\delta^T(mh)(\bar{\Lambda}_m^{nm})^T] \leq h \alpha^2(j) E[d^T(j)(P^T\Phi^T((m+1)h, j+1)\Phi((m+1)h, j+1)P \otimes I_n)d(j)] \leq \frac{1}{\alpha(mh)} E[\delta^T(mh)(\Phi^T((m+1)h, mh) \otimes I_n)\delta(mh)] + \frac{1}{\alpha(mh)} E[(\bar{\Lambda}_m^{nm})^T(\bar{\Lambda}_m^{nm})]
\]
\[
\quad -2E[\delta^T(mh)(\Phi^T((m+1)h, mh) \otimes I_n)\bar{\Lambda}_m^{nm}] \leq \alpha(mh)E[\delta^T(mh)(\Phi^T((m+1)h, mh) \otimes I_n)\delta(mh)] + \frac{1}{\alpha(mh)} E[(\bar{\Lambda}_m^{nm})^T(\bar{\Lambda}_m^{nm})]
\]
\[
17
\]
\[ \leq \alpha(mh)\left[1 - 2\beta c((m + 1)h) + C_1c^2((m + 1)h)\right]E[V(mh)] \\
+ \frac{hc_3}{\alpha(mh)} \sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(2\sigma_d^2E[||X(j)||^2] + 2NC_d^2) \]  

(A.42)

Taking the mathematical expectations on both sides of (A.25), by (A.29), (A.38), (A.40), (A.41) and (A.42), we get

\[ E[V((m + 1)h)] \leq \left(1 + \alpha(mh)\right)\left[1 - 2\beta c((m + 1)h) + C_1c^2((m + 1)h)\right]E[V(mh)] \\
+ \left(\frac{1}{\alpha(mh)} + 2\right)\left(\sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(2\sigma_d^2C_4\beta((m+1)h) + 2NC_d^2)\right) \\
+ 4\left(\sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(4\sigma_d^2C_4\beta((m+1)h) + 2b^2)\right) \\
+ hC_3 \sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(\sigma_d^2E[||X(j)||^2] + C_d^2), m \geq m_0. \]  

(A.43)

By Lemma[A.3] we know that there exists a constant \( C_4 > 0 \), such that

\[ E[||X(k)||^2] \leq C_4\beta(k), \forall k \geq 0. \]  

(A.44)

which implies

\[ E[||X(j)||^2] \leq C_4\beta(j) \leq C_4\beta((m + 1)h), mh \leq j \leq (m + 1)h, \]  

(A.45)

Noting that \( c(k) \) and \( \alpha(k) \) are monotonically decreasing, we have \( \alpha(j) \leq \alpha(mh), c(j) \leq c(mh), mh \leq j \leq (m + 1)h \). Thus, by (A.43) and (A.45), noting that \( \beta((m + 1)h) > 1 \), we get

\[ E[V((m + 1)h)] \leq \left(1 + \alpha(mh)\right)\left[1 - 2\beta c((m + 1)h) + C_1c^2((m + 1)h)\right]E[V(mh)] \\
+ \left(\frac{1}{\alpha(mh)} + 2\right)\left(\sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(2\sigma_d^2C_4\beta((m+1)h) + 2NC_d^2)\right) \\
+ 4\left(\sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(4\sigma_d^2C_4\beta((m+1)h) + 2b^2)\right) \\
+ hC_3 \sum_{j=mh}^{(m+1)h-1} \alpha^2(j)(\sigma_d^2C_4\beta((m+1)h) + C_d^2) \]  

\[ \leq \left[1 - 2\beta c((m + 1)h) + q_0(mh)\right]E[V(mh)] + p(mh), m \geq m_0. \]  

(A.46)

where \( q_0(mh) = C_1c^2((m+1)h) + \alpha(mh)(1 - 2\beta c((m + 1)h) + C_1c^2((m + 1)h)), p(mh) = C_5\alpha(mh)\beta((m+1)h) + (2C_3 + C_6)\alpha^2(mh)\beta((m + 1)h) + C_7c^2((m + 1)h)C_5 = 2h^2C_5(C_4^2C_4 + NC_d^2)), C_6 = 4h^2C_3(C_4 + C_d), C_7 = 4h^2C_3C_6(C_4^2C_4 + 2b^2). \) From Conditions (C1) and (C4), we obtain \( q_0(mh) = o(c((m + 1)h)) \), thus, there exists a positive integer \( m_1 \), such that

\[ 0 < 2\beta c((m + 1)h) - q_0(mh) \leq 1, \forall m \geq m_1. \]  

(A.47)

Let \( \Pi(k) = \frac{c(k)V(k)}{\alpha(k)\beta(k)}. \) By (A.46), (A.47) and the monotonically decreasing property of \( c(k) \), we get

\[ E[\Pi((m + 1)h)] = E\left[\frac{c((m + 1)h)V((m + 1)h)}{\alpha((m + 1)h)\beta((m + 1)h)}\right] \]
This implies that there exists a constant 

By Condition (C5), we have

which together with (A.50) leads to

This implies that there exists a constant \( C_8 > 0 \), such that

From Condition (C5), we know that there exists a positive integer \( k_1 \) and a positive constant \( C_9 \), such that \( \{\alpha(k)\beta(k), k \geq k_1\} \) is monotonically decreasing and

which together with (A.50) leads to

This together with (A.47) gives

where

Hence, from (A.48) and (A.51), we obtain

\[
E[\Pi((m+1)h)] \leq \left( 1 - q_1(mh) \right) E[\Pi(mh)] + \frac{c((m+1)h)p(mh)}{\alpha((m+1)h)\beta((m+1)h)}, \quad m \geq \max\{m_0, m_1\}. \tag{A.53}
\]
From \(c(k) = O(c(k+1))\) and Condition (C4), we have
\[
\lim_{m \to \infty} \frac{\alpha(mh)\beta(mh)}{c(m+1)h} = 0. \tag{A.54}
\]

By Conditions (C1) and (C4), we have
\[
q_0(mh) + hC_5\alpha(mh)\beta(mh)(1 + hC_5\alpha(mh)\beta(mh)) = o(c((m + 1)h)).
\]
Thus, by Condition (C1) and (A.54), we know that there exists a positive integer \(m_2\), such that
\[
0 < q_1(mh) \leq 1, \forall m \geq m_2, \tag{A.55}
\]
and
\[
\sum_{m=0}^{\infty} q_1(mh) = \infty. \tag{A.56}
\]
Noting that \(1 \leq \frac{\beta(k + h)}{\beta(k)} \leq \exp(hC_0\alpha(k))\) and \(\alpha(k) \downarrow 0\), we have \(\lim_{k \to \infty} \frac{\beta(k + h)}{\beta(k)} = 1\). Thus, from (49), we obtain
\[
\lim_{k \to \infty} \frac{\alpha(k)}{\alpha(k + h)} = \lim_{k \to \infty} \frac{\alpha(k)\beta(k)}{\alpha(k + h)\beta(k + h)} \lim_{k \to \infty} \frac{\beta(k + h)}{\beta(k)} = 1.
\]
This together with Conditions (C1), (C2) and the definition of \(p(mh)\) leads to
\[
\frac{p(mh)}{\alpha((m + 1)h)\beta((m + 1)h)} = \lim_{m \to \infty} \frac{C_5\alpha(mh)}{\alpha((m + 1)h)} + \lim_{m \to \infty} \frac{(2C_5 + C_6)\alpha^2(mh)}{\alpha((m + 1)h)} + \lim_{m \to \infty} \frac{C_7c^2(mh)}{\alpha((m + 1)h)} \tag{A.57}
\]
From the definition of \(q_1(mh), q_0(mh) = o(c((m + 1)h))\) and Condition (C4), we have
\[
\lim_{m \to \infty} \frac{q_1(mh)}{c((m + 1)h)} = 2\theta - \lim_{m \to \infty} \frac{q_0(mh)}{c((m + 1)h)} = \lim_{m \to \infty} \frac{hC_5C_0\alpha(mh)\beta(mh)(1 + hC_5C_0\alpha(mh)\beta(mh))}{c((m + 1)h)} = 2\theta. \tag{A.58}
\]
Thus, from (A.57) and (A.58), we have
\[
\lim_{m \to \infty} \frac{c((m + 1)h)p(mh)}{\alpha((m + 1)h)\beta((m + 1)h)q_1(mh)} = \lim_{m \to \infty} \frac{p(mh)}{q_1(mh)} = \frac{C_5}{2\theta}. \tag{A.59}
\]
This together with (A.53), (A.55), (A.56) and Theorem 1.2.22 in [51] leads to
\[
\limsup_{m \to \infty} E[\Pi(mh)] \leq \lim_{m \to \infty} \frac{c((m + 1)h)p(mh)}{\alpha((m + 1)h)\beta((m + 1)h)q_1(mh)} = \frac{C_5}{2\theta}. \tag{A.59}
\]
By \(2\sigma^2 q \leq \|p\|^2 + \|q\|^2, \forall p, q \in \mathbb{R}^n\) and Lemma [A.1] we get
\[
\begin{align*}
-2E[\alpha(k)d^T(k)(P \otimes I_n)\delta(k)] & \leq E[\|\alpha(k)d(k)\|^2] + \|P \otimes I_n\|^2 E[\|\delta(k)\|^2] \\
& \leq a^2(k)E[2\sigma^2 a^2\|X(k)\|^2 + 2NC^2 a^2] + E[V(k)].
\end{align*}
\]
which together with (A.3) in Lemma A.4 and (A.44) leads to
\[
E[V(k+1)] \leq 2(1 + c^2(k)(\rho_0^2 + 8\sigma^2C_\ell p_1))E[V(k)] + 2\alpha^2(k)C_4(2\sigma_\zeta + 4\sigma_\zeta^2)\beta(k) + 8b^2C_\ell p_1c^2(k) + 2\alpha^2(k)(2C_\zeta + 4NC_\zeta^2), \quad k \geq 0.
\]  
(A.60)

Let \( m_k = \lceil \frac{k}{h} \rceil \), then \( 0 \leq k - m_k h \leq h \), \( \forall \ k \geq 0 \). By (A.44), (A.60) and \( \beta(k) > 1, \forall \ k \geq 0 \), we have
\[
E[V(k+1)] \leq 2(1 + c^2(k)(\rho_0^2 + 8\sigma^2C_\ell p_1))E[V(k)] + 2\alpha^2(k)\beta(k)(2C_4\sigma_\zeta + 4C_4\sigma_\zeta^2 + 2C_\zeta + 4NC_\zeta^2) + 8b^2C_\ell p_1c^2(k)
\leq \sum_{i=m_k h}^{k} E[V(m_k h)] + \sum_{i=m_k h}^{k} \left[ 2(1 + c^2(j)(\rho_0^2 + 8\sigma^2C_\ell p_1))(2\alpha^2(j)\beta(j)(2C_4\sigma_\zeta + 4C_4\sigma_\zeta^2 + 2C_\zeta + 4NC_\zeta^2) + 8b^2C_\ell p_1c^2(j)) \right].
\]  
(A.61)

From Condition (C1), we have
\[
2(1 + c^2(i)(\rho_0^2 + 8\sigma^2C_\ell p_1)) \leq 2(1 + c^2(0)(\rho_0^2 + 8\sigma^2C_\ell p_1)) = \eta \beta, \quad \forall \ i \geq 0.
\]  
(A.62)

By Condition (C1), we know that \( \alpha^2(i) \leq \alpha^2(m_k h), \beta(i) \leq \beta(m_k h) \exp(C_0 \sum_{i=m_k h}^{(m_k+1)h} \alpha(i)) \leq \beta(m_k h) \exp(hC_0\alpha(0)) \), hence, by (A.61) and (A.62), we have
\[
E[V(k+1)] \leq \eta^h E[V(m_k h)] + \sum_{i=m_k h}^{k} \eta^h(2\alpha^2(i)\beta(i))(2C_4\sigma_\zeta + 4C_4\sigma_\zeta^2 + 2\alpha^2(i)\beta(i)(2C_4\sigma_\zeta + 4C_4\sigma_\zeta^2 + 2C_\zeta + 4NC_\zeta^2) + 8b^2C_\ell p_1c^2(i))
\leq \eta^h E[V(m_k h)] + 2h\eta^h(2C_4\sigma_\zeta + 4C_4\sigma_\zeta^2 + 2C_\zeta + 4NC_\zeta^2)\exp(C_0\alpha(m_k h)c(m_k h)) + 8h^2\eta^h b^2C_\ell p_1c^2(m_k h).
\]  
(A.63)

By Condition (C1) and (A.63), we get
\[
c(k+1)E[V(k+1)] \leq \frac{c(k+1)E[V(k+1)]}{\alpha(k+1)\beta(k+1)} \leq \frac{c(m_k h)E[V(k+1)]}{\alpha(m_k h)\beta(m_k h)} \leq \left[ \eta^h E[V(m_k h)] + 2h\eta^h(2C_4\sigma_\zeta + 4C_4\sigma_\zeta^2 + 2C_\zeta + 4NC_\zeta^2)\exp(C_0\alpha(m_k h)c(m_k h)) + 8h^2\eta^h b^2C_\ell p_1c^2(m_k h) \right] \frac{\alpha(m_k h)}{\alpha(m_k + 1)h}\frac{\alpha(k)}{\alpha(k+1)}
\]  
(A.64)

where the second “\( \leq \)” is obtained from \( 0 \leq k - m_k h \leq h \) and the monotonically decreasing property of \( \alpha(k) \) and \( c(k) \). Finally, from (A.57), (A.59), (A.64) and Condition (C1), we obtain
\[
\limsup_{k \to \infty} \frac{c(k+1)E[V(k+1)]}{\alpha(k+1)\beta(k+1)} \leq \frac{\eta^h C_5}{2\theta} < \infty,
\]  
(A.65)

i.e. (A.20) holds. From Condition (C4) with \( C = 4C_0 \), it follows that there exists a positive integer \( k_2 \) such that
\[
\frac{\alpha(k)\exp(4C_0 \sum_{i=0}^{k} \alpha(i))}{c(k)} \leq 1, \quad \forall \ k \geq k_2, \quad \text{which gives} \quad \frac{\alpha(k)}{c(k)} \leq \beta^{-4}(k), \quad \forall \ k \geq k_2.
\]  
By (A.20), we know that there exists a nonnegative constant \( C_{10} \) and a positive integer \( k_3 \), such that
\[
E[V(k)] \leq C_{10}\eta^h\beta^{-4}(k), \quad \forall \ k \geq k_3.
\]  
(A.66)

Take \( k_4 = \max(k_2, k_3) \), then we get
\[
E[V(k)] \leq C_{10}\eta^h\beta^{-3}(k), \quad \forall \ k \geq k_4.
\]  
(A.66)
By Hölder inequality, Lemma \ref{lem:A.1}, \ref{lem:A.44} and \ref{lem:A.66}, we have
\[
-2E[\alpha(k)d^T(k)(P \otimes I_n)\delta(k)] \leq 2E[\alpha(k)||d^T(k)||(P \otimes I_n)||\delta(k)||] \\
\leq 2\alpha(k)\left[E[||d^T(k)||^2]\right]^{\frac{1}{2}}\left[E[||\delta(k)||^2]\right]^{\frac{1}{2}} \\
\leq 2\alpha(k)\left[2\sigma^2_C E[||X(k)||^2] + 2NC^2_d\right]^{\frac{1}{2}}\left[E[|V(k)|]\right]^{\frac{1}{2}} \\
\leq 2\alpha(k)\left(2\sigma^2_C C_4 + 2NC^2_d\right)\beta(k)\left[|C_0\beta^{-3}(k)|\right]^{\frac{1}{2}} \\
\leq 2\sqrt{2(\sigma^2_C C_4 + NC^2_d)C_10\alpha(k)\beta^{-1}(k)}, \quad k \geq k_4.
\]
From the above, \ref{eq:A.3} in Lemma \ref{lem:A.4} and \ref{eq:A.44}, we get
\[
E[|V(k + 1)|] \leq (1 + 2c^2(k)\rho^2_0 + 8\sigma^2_C C_1\rho_1)E[|V(k)|] + 2(2\sigma_{\xi} + 3\sigma^2_d)C_4\alpha^2(k)\beta(k) + 8b^2C_\xi\rho_1c^2(k) \\
+ 2\sqrt{2(\sigma^2_d C_4 + NC^2_d)C_10\alpha(k)\beta^{-1}(k)}, \quad k \geq k_4.
\]
(A.67)
Taking conditional expectations on both sides of (A.67) gives
\[
E[|V(k + 1)||F(k - 1)|] \leq (1 + 2c^2(k)\rho^2_0 + 8\sigma^2_C C_1\rho_1)E[|V(k)|] + 2(2\sigma_{\xi} + 3\sigma^2_d)C_4\alpha^2(k)\beta(k) + 8b^2C_\xi\rho_1c^2(k) \\
+ 2\sqrt{2(\sigma^2_d C_4 + NC^2_d)C_10\alpha(k)\beta^{-1}(k)}, \quad k \geq k_4 a.s.
\]
(A.68)
From Conditions (C4) with \( C = C_0 \), it follows that there exists a positive integer \( k_5 \) such that \( \frac{\alpha(k)\beta(k)}{\epsilon(k)} \leq 1, \forall k \geq k_5 \), which means \( \alpha(k)\beta(k) \leq \epsilon(k), \forall k \geq k_5 \). This together with \( \beta(k) > 1 \) and \( \sum_{k=0}^{\infty} c^2(k) < \infty \) leads to \( \sum_{k=0}^{\infty} \alpha^2(k)\beta^2(k) < \sum_{k=0}^{k_5} \alpha^2(k)\beta^2(k) + \sum_{k=k_5}^{\infty} \alpha^2(k)\beta^2(k) \leq \sum_{k=0}^{k_5} \alpha^2(k)\beta^2(k) + \sum_{k=k_5}^{\infty} c^2(k) < \infty \). Then, by \ref{eq:A.68}, Conditions (C1), (C3) and Theorem 1 in \[52\], we obtain \( V(k) \rightarrow a \) random variable, \( k \rightarrow \infty \) a.s., which together with \ref{eq:A.20} and Conditions (C4) gives \ref{eq:A.21}. \qed

**Proof of Theorem 3.1:** Let \( \bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j(k) \). By \( d_j^T(k)(x_j(k) - x^*) < f_j(x^*) - f_j(x_j(k)) = f_j(x^*) - f_j(\bar{x}(k)) + f_j(\bar{x}(k)) - f_j(x_j(k)) \leq f_j(x^*) - f_j(\bar{x}(k)) + d_j^T(\bar{x}(k))(\bar{x}(k) - x_j(k)) \leq f_j(x^*) - f_j(\bar{x}(k)) + (\sigma_d||\bar{x}(k)|| + C_d)||\bar{x}(k) - x_j(k)|| , \) which gives
\[
-2\alpha(k)E[d^T(k)(X(k) - \mathbf{1}_N \otimes x^*)] \\
= -2\alpha(k)E\left[\sum_{i=1}^{N} d_j^T(k)(x_i(k) - x^*)\right] \\
\leq 2\alpha(k)E\left[\sum_{i=1}^{N} (f_j(x^*) - f_j(\bar{x}(k))) + \sum_{i=1}^{N} (\sigma_d||\bar{x}(k)|| + C_d)||\bar{x}(k) - x_i(k)||\right] \\
\leq 2\alpha(k)\sqrt{E\left[(f_j(x^*) - f_j(\bar{x}(k)))^2\right] + \sum_{i=1}^{N} (\sigma_d||\bar{x}(k)|| + C_d^2)\frac{1}{N}\left[\sum_{i=1}^{N} ||\bar{x}(k) - x_i(k)||^2\right]^{\frac{1}{2}}} \\
= 2\alpha(k)E[f_j(x^*) - f_j(\bar{x}(k))] + 2\alpha(k)\sqrt{E[(\sigma_d||\bar{x}(k)|| + C_d)||\delta(k)||^2]}.
\]
(A.69)
From \( ||p + q||^2 \leq 2||p||^2 + 2||q||^2, \quad p, q \in \mathbb{R}^n \) and \ref{eq:A.44}, we get
\[
E[\sigma_d||\bar{x}(k)|| + C_d^2] \leq 2E[\sigma^2_d||\bar{x}(k)||^2 + C_d^2] \\
\leq 2E\left[\frac{1}{N}\sigma^2_d\sum_{i=1}^{N} ||x_i(k)||^2 + C_d^2\right]
\]
\[22\]
\[
E[(\sigma_d||\bar{x}(k)|| + C_d)||\delta(k)||] \leq \left[ E[(\sigma_d||\bar{x}(k)|| + C_d^2)^2] \right]^{1/2} \left[ E[||\delta(k)||^2] \right]^{1/2} \\
\leq \sqrt{2(\frac{1}{N}\sigma_d^2 C_d + C_d^2)\beta(k)} \sqrt{C_1 \beta^2(k)} \\
= \sqrt{2(\frac{1}{N}\sigma_d^2 C_d + C_d^2)C_1 \beta^{-1}(k)}, \quad k \geq 0.
\]

This together with (A.4) in Lemma A.4 and (A.69) gives
\[
E[||X(k+1) - 1_N \otimes x^*||^2] \\
\leq \left( 1 + 2c^2(k)(\rho_0^2 + 8\sigma^2 C_{d\rho_1}) + 4\alpha^2(k)(2\sigma_c + 3\sigma_d^2) \right) E[||X(k) - 1_N \otimes x^*||^2] \\
+ 8b^2 C_{d\rho_1} c^2(k) + 2\alpha^2(k)(2C_c + 3NC_d^2) + 2(3\sigma_d^2 + 2\sigma_c) N ||x^*||^2 \\
- 2\alpha(k)E[f(\bar{x}(k)) - f^*] + 2 \sqrt{2(\sigma_d^2 C_d + C_d^2)C_1 \alpha(k) \beta^{-1}(k)}, \quad k \geq k_0.
\]

Since \( f(\cdot) \) is a convex function, from Jensen inequality we obtain \( E[f(\bar{x}(k))|F(k-1)] \geq f(E[\bar{x}(k)|F(k-1)]) = f(\bar{x}(k)). \) Taking conditional expectations on both sides of (A.70) gives
\[
E[||X(k+1) - 1_N \otimes x^*||^2|F(k-1)] \\
\leq \left( 1 + 2c^2(k)(\rho_0^2 + 8\sigma^2 C_{d\rho_1}) + 4\alpha^2(k)(2\sigma_c + 3\sigma_d^2) \right) ||X(k) - 1_N \otimes x^*||^2 \\
+ 8b^2 C_{d\rho_1} c^2(k) + 2\alpha^2(k)(2C_c + 3NC_d^2) + 2(3\sigma_d^2 + 2\sigma_c) N ||x^*||^2 \\
+ 2 \sqrt{2(\sigma_d^2 C_d + C_d^2)C_1 \alpha(k) \beta^{-1}(k)} - 2\alpha(k)(f(\bar{x}(k)) - f^*), \quad k \geq k_0 \ a.s.
\]

Since \( f(\bar{x}(k)) - f^* \geq 0 \), by Theorem 1 in [52] and Conditions (C1) and (C3), we obtain that the sequence \( \{||X(k) - 1_N \otimes x^*||^2, k \geq 0\} \) converges a.s. for any given \( x^* \in X^* \), and \( \sum_{k=0}^{\infty} \alpha(k)f(\bar{x}(k)) = f^* < \infty \) a.s., which together with \( f(\bar{x}(k)) \geq f^* \) and \( \sum_{k=0}^{\infty} \alpha(k) = \infty \) gives
\[
\lim_{k \to \infty} f(\bar{x}(k)) = f^* \ a.s.
\]  

Since for any given \( x^* \in X^* \), \( \{||X(k) - 1_N \otimes x^*||^2, k \geq 0\} \) converges, a.s., we know that for any given \( x^* \in X^* \), there is a measurable set \( \Omega_{x^*} \) with \( P(\Omega_{x^*}) = 1 \), such that for any given \( \omega \in \Omega_{x^*} \), \( ||X(k, \omega) - 1_N \otimes x^*||, k \geq 0 \) converges, which implies that \( \sup_{k \geq 0} ||X(k, \omega)|| < \infty \). Denote \( \Omega_1 = \{\omega|\lim_{k \to \infty} ||x_i(k, \omega) - \bar{x}(k, \omega)|| = 0, \ i = 1, \ldots, N\} \). From (A.21) in Lemma A.6 we know that \( P(\Omega_1) = 1 \). Denote \( \Omega_2 = \{\omega|\lim_{k \to \infty} f(\bar{x}(k, \omega)) = f^*\} \). From (A.71), it follows that \( P(\Omega_2) = 1 \). Denote \( \Omega = (\bigcap_{x^* \in X^*} \Omega_{x^*}) \cap \Omega_1 \cap \Omega_2 \).

From Assumption 5 it follows that \( P(\Omega) = 1 \). For any given \( \omega \in \Omega \), we know that there is a subsequence \( \{\bar{x}(k_1, \omega), l \geq 0\} \) of \( \{\bar{x}(k, \omega), k \geq 0\} \) such that \( \lim_{k \to \infty} f(\bar{x}(k_1, \omega)) = f^* \). By \( \omega \in \Omega \), it follows that \( \sup_{k \geq 0} ||X(k, \omega)|| < \infty \), which implies that \( \{\bar{x}(k, \omega), k \geq 0\} \) is bounded. By the continuity of \( f \) (due to convexity of \( f \) over \( \mathbb{R}^n \)) and the boundedness of \( \{\bar{x}(k, \omega), l \geq 0\} \), we know that there is a subsequence \( \{\bar{x}(k_f, \omega), l \geq 0\} \) of \( \{\bar{x}(k, \omega), l \geq 0\} \), converges to a point \( z^*(\omega) \) in \( X^* \), i.e. \( \lim_{l \to \infty} \bar{x}(k_f, \omega) = z^*(\omega) \), which gives \( \lim_{l \to \infty} ||x_i(k_f, \omega) - z^*(\omega)|| = 0, \ i = 1, 2, \ldots, N \). Then we get \( \lim_{l \to \infty} ||X(k_f, \omega) - 1_N \otimes z^*(\omega)|| = 0 \). This together with the convergence of \( \{||X(k, \omega) - 1_N \otimes z^*(\omega)||, k \geq 0\} \) leads to \( \lim_{k \to \infty} ||x_i(k, \omega) - z^*(\omega)|| = 0, \ i = 1, 2, \ldots, N \). Then by the arbitrariness of \( \omega \) and \( P(\Omega) = 1 \), we get that \( \lim_{k \to \infty} x_i(k) = z^* \) a.s., \( i = 1, \cdots, N \).
B Verification for the example in Section 2

Denote \( u(k) = (u_1^T(k), \ldots, u_M^T(k))^T \), \( v(k) = (v_1(k), \ldots, v_N(k))^T \). Suppose that \( \{\xi(k), k \geq 0\}, \{u(k), k \geq 0\}, \{v(k), k \geq 0\} \) and \( \{A_{ij}(k)\}, k \geq 0 \) are mutually independent.

Firstly, we will verify that Assumption 1 holds. Substituting (2.8) into (2.7) and taking the mathematical expectation, we have

\[
E[\ell_i(x; u_i(k), p_i(k))] = \frac{1}{2} E[\|u_i^T(k)x_0 + v_i(k) - u_i^T(k)x\|^2] = \frac{1}{2} E[\|u_i^T(k)(x_0 - x) + v_i(k)\|^2] = \frac{1}{2} E[(x_0 - x)^T u_i(k)u_i^T(k)(x_0 - x) + 2v_i(k)u_i^T(k)(x_0 - x) + v_i(k)v_i(k)] = \frac{1}{2} \left[ (x_0 - x)^T R_{u_i}(x_0 - x) + \sigma_{v_i} \right].
\]

then, \( \nabla E[\ell_i(x; \mu_i(k))] = R_{u_i}(x - x_0) \), and the subgradient of the local risk function \( f_i(x) \) is given by

\[
d_f(x) = R_{u_i}(x - x_0) + d_R(x), \quad \forall \, d_R(x) \in \partial R_i(x).
\]

By the definition of \( d_R(x) \), it is known that \( ||d_R(x)|| \leq \kappa, \forall \, d_R(x) \in \partial R_i(x) \). Hence, it can be obtained from (B.2) that

\[
||d_f(x)|| = ||R_{u_i}(x - x_0) + d_R(x)|| \leq ||R_{u_i}|| ||x|| + ||R_{u_i}|| ||x_0|| + \kappa,
\]

thus, Assumption 4 holds. The subgradients of the local cost functions are required to be bounded in [43, 44, 37, 24] which can not cover the case above, while our assumption covers both \( L_2 \)-regularization and \( L_1 \)-regularization.

Secondly, we will verify that Assumption 2 and Assumption 3 hold. For (2.6), the subgradients of local risk functions are measured with noises, i.e.

\[
\tilde{d}_f(x_i(k)) = d_f(x_i(k)) + \zeta_i(k),
\]

where

\[
\zeta_i(k) = (u_i(k)u_i^T(k) - R_{u_i})(x_i(k) - x_0) - u_i(k)v_i(k)
\]

is the subgradient measurement noise of the \( i \)th optimizer.

Let \( \mathcal{F}(k) = \sigma[\xi_i(t), u_i(t), v_i(t), A_{ij}(t), 0 \leq t \leq k, 1 \leq i, j \leq N], k \geq 0, \mathcal{F}(-1) = \{\emptyset, \Omega\} \). It can be derived from the algorithms (2.2), (2.3) that \( (x_i(k) - x_0) \in \mathcal{F}(k - 1) \subseteq \mathcal{F}(k), i = 1, \ldots, N \), and by (2.4), we obtain \( \zeta_i(k) \in \mathcal{F}(k) \), so \( \{\zeta_i(k), \mathcal{F}(k), k \geq 0\} \) is an adaptive process. Note that \( \{u_i(k), k \geq 0\} \) is i.i.d., \( \{u_i(k), k \geq 0\}, \{\xi_i(k), k \geq 0\}, \{A_{ij}(k), k \geq 0\} \) and \( \{v_i(k), k \geq 0\} \) are mutually independent. Then, \( \sigma(u_i(k)) \) and \( \mathcal{F}(k - 1) \) are mutually independent. Similarly, \( \sigma(v_i(k)) \) and \( \mathcal{F}(k - 1) \) are also mutually independent. Hence, from (B.4), we have

\[
E[\zeta_i(k)|\mathcal{F}(k - 1)] = E[(u_i(k)u_i^T(k) - R_{u_i})(x_i(k) - x_0) - u_i(k)v_i(k)|\mathcal{F}(k - 1)] = E[(u_i(k)u_i^T(k) - R_{u_i})(x_i(k) - x_0) - E[u_i(k)v_i(k)|\mathcal{F}(k - 1)]] = (E[u_i(k)u_i^T(k) - R_{u_i})(x_i(k) - x_0) - E[u_i(k)]E[v_i(k)] = 0 \quad a.s., \, \forall \, k \geq 0, \, i = 1, \ldots, N.
\]

Thus, \( \{\zeta_i(k), \mathcal{F}(k), k \geq 0\} \) is a martingale difference sequence. By (B.4), we have

\[
E[\zeta_i^T(k)\zeta_i(k)|\mathcal{F}(k - 1)]
\]
Substituting (B.6), (B.7) and (B.8) into (B.5) gives

\[
E \left[ \left( u_i(k)u_i^T(k) - R_{u,i} \right)(x_i(k) - x_0) - u_i(v_i(k)) \right]^T \times \left( u_i(k)u_i^T(k) - R_{u,i} \right) \left( x_i(k) - x_0 \right) - u_i(v_i(k)) \right] F(k - 1) \right] \\
= E \left[ (x_i(k) - x_0)^T (u_i(k)u_i^T(k) - R_{u,i})^T (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] \\
- 2 v_i(k)u_i^T(k) (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) + (v_i(k))^2 u_i^T(k) u_i(k) \right] F(k - 1) \right] a.s. \quad (B.5)
\]

Noting that \( \sigma(u_i(k)) \) and \( F(k - 1) \) are mutually independent, by \( x_i(k) \in F(k - 1) \), we have

\[
E \left[ (x_i(k) - x_0)^T (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] \\
= E \left[ (x_i(k) - x_0)^T (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] \\
= (x_i(k) - x_0)^T E \left[ (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] a.s. \quad (B.6)
\]

Noting that \( u_i(k) \) and \( v_i(k) \) are mutually independent, by \( x_i(k) \in F(k - 1) \) and \( E[v_i(k)] = 0 \), we have

\[
E \left[ -2 v_i(k)u_i^T(k) (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] \\
= -2 E \left[ v_i(k) \right] F(k - 1) \right] E \left[ (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] \\
= -2 E \left[ v_i(k) \right] F(k - 1) \right] E \left[ u_i^T(k) (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] a.s. \quad (B.7)
\]

It follows from the definitions of \( u_i(k) \) and \( v_i(k) \) that

\[
E \left[ (v_i(k))^2 u_i^T(k) u_i(k) \right] F(k - 1) \right] = E \left[ (v_i(k))^2 u_i^T(k) u_i(k) \right] \\
= E \left[ (v_i(k))^2 \right] E \left[ u_i^T(k) u_i(k) \right] \\
= \sigma_i^2 \text{Tr}(R_{u,i}) a.s. \quad (B.8)
\]

Substituting (B.6), (B.7) and (B.8) into (B.5) gives

\[
E \left[ \xi_i^T(k) \xi_i(k) \right] F(k - 1) \right] \\
= (x_i(k) - x_0)^T E \left[ (u_i(k)u_i^T(k) - R_{u,i}) (x_i(k) - x_0) \right] F(k - 1) \right] + \sigma_i^2 \text{Tr}(R_{u,i}) \\
\leq 2 E \left[ \|u_i(k)u_i^T(k) - R_{u,i}\|^2 \right] \|x_i(k)\|^2 + 2 E \left[ \|u_i(k)u_i^T(k) - R_{u,i}\|^2 \right] \|x_0\|^2 + \sigma_i^2 \text{Tr}(R_{u,i}) a.s.
\]

Denote \( \sigma_\zeta = \max_{1 \leq i \leq N} \left\{ 2 E \left[ \|u_i(k)u_i^T(k) - R_{u,i}\|^2 \right] \|x_0\|^2 + \sigma_i^2 \text{Tr}(R_{u,i}) \right\} \) and \( C_\zeta = N \max_{1 \leq i \leq N} \left\{ 2 E \left[ \|u_i(k)u_i^T(k) - R_{u,i}\|^2 \right] \|x_0\|^2 + \sigma_i^2 \text{Tr}(R_{u,i}) \right\} \).

Then we have

\[
E \left[ \zeta_i^T(k) \zeta_i(k) \right] F(k - 1) \right] = \sum_{i=1}^{N} E \left[ \zeta_i^T(k) \zeta_i(k) \right] F(k - 1) \right] \leq \sigma_\zeta \|X(k)\|^2 + C_\zeta a.s.
\]

Noting that \( \{\xi(k), k \geq 0\}, \{u(k), k \geq 0\}, \{v(k), k \geq 0\} \) and \( \{\mathcal{A}_{G_{(k)}}, k \geq 0\} \) are mutually independent, by Lemma A.1 in [23], we obtain that \( \sigma(\xi(k), \xi(k + 1), \ldots) \) and \( \sigma(\mathcal{A}_{G_{(k)}}, \mathcal{A}_{G_{(k+1)}}, \ldots) \) are conditionally independent given \( F(k - 1), \forall k \geq 0 \), which means that \( \sigma(\xi(k), \mathcal{A}_{G_{(k+1)}}, \ldots) \) and \( \sigma(\xi(k)) \) are conditionally independent given \( F(k - 1) \), i.e. \( \{\xi(k), k \geq 0\} \) satisfies Assumption 2.

By (B.4), we get \( \sigma(\zeta(k)) \subseteq \sigma[u_i(k), v_i(k), x_i(k), 1 \leq i \leq N] \). Then, by \( \sigma(x_i(k), 1 \leq i \leq N) \subseteq F(k - 1) \), we have \( \sigma(\zeta(k)) \subseteq \sigma[\sigma[u_i(k), v_i(k), 1 \leq i \leq N] \cup F(k - 1)] \). Therefore,

\[
\sigma(\sigma(\zeta(k)) \cup F(k - 1)) \subseteq \sigma[\sigma[u_i(k), v_i(k), 1 \leq i \leq N] \cup F(k - 1)]. \quad (B.10)
\]

Noting that \( \{\xi(k), k \geq 0\}, \{u(k), k \geq 0\}, \{v(k), k \geq 0\} \) and \( \{\mathcal{A}_{G_{(k)}}, k \geq 0\} \) are mutually independent, \( \{u(k), k \geq 0\} \) and \( \{v(k), k \geq 0\} \) are i.i.d., we have \( \sigma[u_i(k), v_i(k), 1 \leq i \leq N] \) is independent of \( \sigma[\mathcal{A}_{G_{(k)}}, \mathcal{A}_{G_{(k+1)}}, \ldots] \) and \( \sigma[\mathcal{A}_{G_{(k)}}, \mathcal{A}_{G_{(k+1)}}, \ldots] \) is independent of \( \sigma(\xi(k), \mathcal{A}_{G_{(k+1)}}, \ldots) \).
By Corollary 3 of Section 7.3 in [53], we have \( \sigma[\mathcal{A}_G(k), \mathcal{A}_G(k+1), \ldots] \) and \( \sigma[u_i(k), v_i(k), 1 \leq i \leq N] \) are conditionally independent given \( \mathcal{F}(k-1) \). Then, by Theorem 1(i) of Section 7.3 in [53], we obtain that for all \( A \in \sigma[\mathcal{A}_G(k), \mathcal{A}_G(k+1), \ldots] \),

\[
P(A \mid \sigma[u_i(k), v_i(k), 1 \leq i \leq N] \cup \mathcal{F}(k-1)) = P(A \mid \mathcal{F}(k-1)). \tag{B.11}
\]

By (B.10) and (B.11), we have

\[
P(A \mid \sigma[\zeta(k)] \cup \mathcal{F}(k-1)) = \mathbb{E}[\mathbb{1}_A \mid \sigma[\zeta(k)] \cup \mathcal{F}(k-1)] = \mathbb{E}[\mathbb{E}[\mathbb{1}_A \mid \mathcal{F}(k-1)] \mid \sigma[\zeta(k)] \cup \mathcal{F}(k-1)] = P(A \mid \mathcal{F}(k-1)).
\]

Furthermore, by Theorem 1(i) of Section 7.3 in [53], we obtain that \( \sigma[\zeta(k)] \) and \( \sigma[\mathcal{A}_G(k), \mathcal{A}_G(k+1), \ldots] \) are conditionally independent given \( \mathcal{F}(k-1) \), which together with (B.9) gives that \( \{\zeta(k), k \geq 0\} \) satisfies Assumption 3.

References

[1] P. Yi, Y. Hong, and F. Liu, “Distributed gradient algorithm for constrained optimization with application to load sharing in power systems,” *Systems Control Lett.*, vol. 83, no. 711, pp. 45–52, 2015.

[2] T. Yang, J. Lu, D. Wu, J. Wu, G. Shi, Z. Meng, and K. H. Johansson, “A distributed algorithm for economic dispatch over time-varying directed networks with delays,” *IEEE Trans. Ind. Electron.*, vol. 64, no. 6, pp. 5095–5106, 2017.

[3] R. Mohebifard and A. Hajbabaie, “Distributed optimization and coordination algorithms for dynamic traffic metering in urban street networks,” *IEEE Trans. Intel. Transp.*, vol. 20, no. 5, pp. 1930–1941, 2019.

[4] S. M. A. B. A. Islam and A. Hajbabaie, “Distributed coordinated signal timing optimization in connected transportation networks,” *Transport. Res. C-ener.*, vol. 80, pp. 272–285, 2017.

[5] Y. Zhang, Y. Lou, Y. Hong, and L. Xie, “Distributed projection-based algorithms for source localization in wireless sensor networks,” *IEEE Trans. Wirel. Commun.*, vol. 14, no. 6, pp. 3131–3142, 2015.

[6] S. Salari, I. Kim, and F. Chan, “Distributed cooperative localization for mobile wireless sensor networks,” *IEEE Wirel. Commun. Lett.*, vol. 7, no. 1, pp. 18–21, 2018.

[7] L. Xiao, “Dual averaging methods for regularized stochastic learning and online optimization,” *J. Mach. Learn. Res.*, vol. 11, no. 1, pp. 2543–2596, 2010.

[8] C. Yu, H. Tang, C. Renggli, S. Kassing, A. Singla, D. Alistarh, C. Zhang, and J. Liu, “Distributed learning over unreliable networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, (Long Beach, California, United States), pp. 7202–7212, 2019.
[9] T. Doan, J. Lubars, C. Beck, and R. Srikant, “Convergence rate of distributed random projections,” IFAC-Papers On Line, vol. 51, no. 23, pp. 373–378, 2018.

[10] S. Pu and A. Nedić, “A distributed stochastic gradient tracking method,” Math. Program., 2020.

[11] S. Pu and A. Garcia, “A flocking-based approach for distributed stochastic optimization,” Oper. Res., vol. 66, no. 1, pp. 267–281, 2017.

[12] X. Lian, W. Zhang, J. Liu, and C. Zhang, “Asynchronous decentralized parallel stochastic gradient descent,” in Proceedings of the 35th International Conference on Machine Learning (ICML), (Stockholm, Sweden), pp. 3043–3052, 2018.

[13] M. O. Sayin, N. D. Vanli, S. S. Kozat, and T. Basar, “Stochastic subgradient algorithms for strongly convex optimization over distributed networks,” IEEE Trans. Netw. Sci. Eng., vol. 4, no. 4, pp. 248–260, 2017.

[14] D. Yuan, Y. Hong, D. W. C. Ho, and G. Jiang, “Optimal distributed stochastic mirror descent for strongly convex optimization,” Automatica, vol. 90, no. 90, pp. 196–203, 2018.

[15] M. Nokleby and W. U. Bajwa, “Stochastic optimization from distributed streaming data in rate-limited networks,” IEEE Trans. Signal Inf. Process, vol. 5, no. 1, pp. 152–167, 2019.

[16] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, “Stochastic gradient push for distributed deep learning,” in Proceedings of the 36th International Conference on Machine Learning (ICML), (Long Beach, California, United States), pp. 344–353, 2019.

[17] H. Hendriks, F. Bach, and L. Massoulie, “Asynchronous accelerated proximal stochastic gradient for strongly convex distributed finite sums,” arXiv: 1901.09865, 2019.

[18] A. Nedić and A. Olshevsky, “Stochastic gradient-push for strongly convex functions on time-varying directed graphs,” IEEE Trans. Automat. Control, vol. 61, no. 12, pp. 3936–3947, 2016.

[19] S. S. Ram, A. Nedić, and V. V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” J. Optim. Theory Appl., vol. 147, no. 3, pp. 516–545, 2010.

[20] A. K. Sahu, D. Jakovetić, D. Bajovic, and S. Kar, “Communication-efficient distributed strongly convex stochastic optimization: Non-asymptotic rates,” arXiv:1809.02920, 2018.

[21] D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar, “Convergence rates for distributed stochastic optimization over random networks,” in Proceedings of the 57th IEEE Conference on Decision and Control (CDC), (Miami Beach, FL, United States), pp. 4238–4245, 2018.

[22] S. A. Alghunaim and A. H. Sayed, “Distributed coupled multiagent stochastic optimization,” IEEE Trans. Automat. Control, vol. 65, no. 1, pp. 175–190, 2020.

[23] T. Li and J. Wang, “Distributed averaging with random network graphs and noises,” IEEE Trans. Inform. Theory, vol. 64, no. 11, pp. 7063–7080, 2018.

[24] K. Srivastava, A. Nedić, and D. M. Stipanović, “Distributed constrained optimization over noisy networks,” in Proceedings of the 49th IEEE Conference on Decision and Control (CDC), (Atlanta, GA, United States), pp. 1945–1950, 2010.

[25] I. Lobel and A. E. Ozdaglar, “Distributed subgradient methods for convex optimization over random networks,” IEEE Trans. Automat. Control, vol. 56, no. 6, pp. 1291–1306, 2011.
[26] T. Wada, I. Masubuchi, K. Hanada, T. Asai, and Y. Fujisaki, “Distributed multi-objective optimization over randomly varying unbalanced networks,” IFAC-Papers On Line, vol. 50, no. 1, pp. 2403–2408, 2017.

[27] N. Bastianello, M. Todescato, R. Carli, and L. Schenato, “Distributed optimization over lossy networks via relaxed peaceman-rachford splitting: a robust admm approach,” in Proceedings of the 17th European Control Conference (ECC), Limassol, Cyprus, pp. 477–482, 2018.

[28] N. Bastianello, R. Carli, L. Schenato, and M. Todescato, “A partition-based implementation of the relaxed ADMM for distributed convex optimization over lossy networks,” in Proceedings of the 57th IEEE Conference on Decision and Control (CDC), Miami Beach, FL, United States, pp. 3379–3384, 2018.

[29] P. Rezaienia, B. Gharesifard, T. Linder, and B. Touri, “Distributed optimization on random graphs,” IFAC-Papers On Line, vol. 51, no. 23, pp. 391–395, 2018.

[30] S. S. Alaviani, Applications of fixed point theory to distributed optimization, robust convex optimization, and stability of stochastic systems. PhD thesis, Iowa State University, 2018.

[31] D. Wang, J. L. Wang, and W. Wang, “Discrete-time distributed optimization for multi-agent systems under markovian switching topologies,” in Proceedings of the IEEE International Conference on Control and Automation, (Chiang Mai), pp. 747–752, 2017.

[32] D. Wang, D. Wang, W. Wang, Y. Liu, and F. E. Alsaadi, “Distributed optimisation for multi-agent systems with the first-order integrals under markovian switching topologies,” Int. J. systems Sci., vol. 48, no. 9, pp. 1787–1795, 2017.

[33] M. Hong and T. H. Chang, “Stochastic proximal gradient consensus over random networks,” IEEE Trans. Signal Process., vol. 65, no. 11, pp. 2933–2948, 2017.

[34] P. Yi, J. Lei, and Y. Hong, “Distributed resource allocation over random networks based on stochastic approximation,” Systems Control Lett., vol. 114, pp. 44–51, 2018.

[35] A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, “Distributed zeroth order optimization over random networks: a Kiefer-Wolfowitz stochastic approximation approach,” in Proceedings of the 57th IEEE Conference on Decision and Control (CDC), Miami Beach, FL, United States, pp. 4951–4958, 2018.

[36] J. Lei, H. Chen, and H. Fang, “Asymptotic properties of primal-dual algorithm for distributed stochastic optimization over random networks with imperfect communications,” SIAM J. Control Optim., vol. 56, no. 3, pp. 2159–2188, 2018.

[37] K. Srivastava and A. Nedić, “Distributed asynchronous constrained stochastic optimization,” IEEE J. Sel. Top. Signal Process., vol. 5, no. 4, pp. 772–790, 2011.

[38] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, “On the linear convergence of the admm in decentralized consensus optimization,” IEEE Trans. Signal Process., vol. 62, no. 7, pp. 1750–1761, 2014.

[39] J. Wang and N. Elia, “Control approach to distributed optimization,” in Proceedings of the 48th Annual Allerton Conference on Communication, Control, and Computing, (Allerton, IL, USA), pp. 557–561, 2010.
[40] S. Lee and A. Nedić, “Distributed random projection algorithm for convex optimization,” *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 2, pp. 221–229, 2013.

[41] D. Jakovetic, J. Xavier, and J. M. F. Moura, “Fast distributed gradient methods,” *IEEE Trans. Automat. Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[42] A. Mokhtari, Q. Ling, and A. Ribeiro, “Network newton distributed optimization methods,” *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 146–161, 2017.

[43] A. Nedić and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Trans. Automat. Control*, vol. 54, no. 1, pp. 48–61, 2009.

[44] A. Nedić, A. Ozdaglar, and P. A. Parrilo, “Constrained consensus and optimization in multi-agent networks,” *IEEE Trans. Automat. Control*, vol. 55, no. 4, pp. 922–938, 2010.

[45] A. Nedić and A. Olshevsky, “Distributed optimization over time-varying directed graphs,” *IEEE Trans. Automat. Control*, vol. 60, no. 3, pp. 601–615, 2015.

[46] H. Li, S. Liu, Y. C. Soh, and L. Xie, “Event-triggered communication and data rate constraint for distributed optimization of multiagent systems,” *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 48, no. 11, pp. 1908–1919, 2018.

[47] S. Liu, Z. Qiu, and L. Xie, “Convergence rate analysis of distributed optimization with projected subgradient algorithm,” *Automatica*, vol. 83, no. 83, pp. 162–169, 2017.

[48] B. Dimitri, A. Nedić, and E. O. Asuman, *Convex Analysis and Optimization*. Boston: Athena Scientific, 2003.

[49] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2003.

[50] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*. London: Springer-Verlag, 1993.

[51] L. Guo, *Time-varying Stochastic System: Stability, Estimation and Control*. Changchun: Jilin Science and Technology Press, 1993.

[52] H. Robbins and D. Siegmund, “A convergence theorem for non negative almost supermartingales and some applications,” in *Herbert Robbins Selected Papers Ed. T. L. Lai, and D. Siegmund*, pp. 233–257, New York: Springer-Verlag, 1985.

[53] Y. S. Chow and T. Henry, *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer-Verlag, 1997.