Distributed Stochastic Constrained Composite Optimization Over Time-Varying Network With a Class of Communication Noise
Zhan Yu, Daniel W. C. Ho, Fellow, IEEE, Deming Yuan, Member, IEEE, and Jie Liu

Abstract—This article is concerned with the distributed stochastic multiagent-constrained optimization problem over a time-varying network with a class of communication noise. This article considers the problem in composite optimization setting, which is more general in the literature of noisy network optimization. It is noteworthy that the mainstream existing methods for noisy network optimization are Euclidean projection based. Based on the Bregman projection-based mirror descent scheme, we present a non-Euclidean method and investigate their convergence behavior. This method is the distributed stochastic composite mirror descent type method (DSCMD-N), which provides a more general algorithm framework. Some new error bounds for DSCMD-N are obtained. To the best of our knowledge, this is the first work to analyze and derive convergence rates of optimization algorithm in noisy network optimization. We also show that an optimal rate of $O(1/\sqrt{T})$ in nonsmooth convex optimization can be obtained for the proposed method under appropriate communication noise condition. Moreover, novel convergence results are comprehensively derived in expectation convergence, high probability convergence, and almost surely sense.

Index Terms—Communication noise, composite optimization, distributed optimization, mirror descent, multiagent network.

I. INTRODUCTION

In recent years, distributed consensus control and optimization problems over networked system are studied extensively. These problems arise in a variety of application domains, such as localization in sensor networks (e.g., [26] and [33]), smart grid (e.g., [33]), utility maximization (e.g., [12]), and allocation of resources in microeconomics (e.g., [17]). On the other hand, there always exists noise in a realistic scenario, and intrinsic disturbances with different types of noises often appear in many multiagent networked systems. For networked systems with active communication channels, besides the inherent disturbance, probably one of the most important issues on disturbance is the communication noise. The communication noise is unavoidable in signal transmission and information communication process. Recently, there have been many works circling around the effects of noise among nodes of the networks on control or optimization methods (e.g., [5], [8], [11], [18], [22]–[25], [27], [32], [38], and [41]). These works focus on different classes of noises, including some common classes, such as bounded noise, decay noise, and mean-zero noise.

In this article, we mainly consider the minimization of a sum of locally known convex functions that are distributed over a network with a class of communication noise. In this model, each agent has its own associated (perhaps nonsmooth) objective function (e.g., [1], [9], [11], [13], and [29]). For solving this kind of problem, a variety of methods has emerged recently. In these methods, the distributed optimization method has been shown to be one of the most powerful methods for its advantage of saving energy and reducing unnecessary waste of resources. Recent years have witnessed the progress of distributed optimization in numerous aspects. Modern studies on distributed optimization start from the classical distributed subgradient method [1]. The seminal research [1] is inspired by a deterministic gradient descent model over network system. Also, their work treats with an unconstrained decision variable. Consequently, studies on the distributed stochastic subgradient method appears [29]. Disturbance on subgradient is considered to capture the dynamical environment in real world. In [2], a distributed gradient-push method is established without requiring information of either the number of agents or the graph sequence. In the same period, the work [9] provides a novel distributed method to better capture the direct structure of the network topology, and the convergence relies on a core matrix analysis result in [10]. In a different way, the work in [20] presents a (stochastic) dual averaging-based method, and the method is based on maintaining and forming weighted averages of subgradients throughout the network. In what follows, several works that improve [20] appear (e.g., [30] and [31]). On the other hand, several works also turn to investigate the case when local objective functions are nonconvex (e.g., [39]). The mirror descent technique has
been utilized in the distributed optimization domain recently (e.g., [21]), and one of the main features of mirror descent is that it can better reflect the geometry of underlying space. Moreover, online distributed optimization has also become a new direction recently (e.g., [14], [19], and [35]), and online distributed methods are often investigated to handle the dynamical environment of local objective functions. Based on the sum structure of the global objective function, a great deal of works aiming at solving the problem are consensus based. The realization of consensus is an essentially necessary condition for the convergence of these methods.

The main goal of this article is to study distributed optimization problems by addressing the following considerations: 1) since uncertain stochastic disturbances always exist in a real-life environment, it is desirable to consider the topic of solving a distributed optimization problem over the network in which some class of communication noises exist among nodes; 2) the existing optimization methods over noisy network are Euclidean gradient projection based (see, [11] and [25]), is it possible to consider some more general frameworks and provide general methods to solve them in some class of noisy network? and 3) although under suitable conditions, in the setting of distributed optimization when communication noise exists over network, almost surely existence result of the optimal solution is proven for gradient descent-based methods in main existing works, such as [11] and [25]. The explicit description of convergence rate is still absent in this literature. Is it possible to derive convergence rate results in settings when some class of communication noise exists? Also, the optimization methods are very poorly explored over the noisy network, it is desirable to develop some algorithms for such optimization problem. To this end, we consider a multiagent composite optimization problem over the time-varying noisy network in this article. Specifically, we analyze the following problem:

\[
\min_{x \in \mathcal{X}} F(x) = \sum_{i=1}^{N} F_i(x) = \sum_{i=1}^{N} [f_i(x) + \chi_i(x)]
\]

where \( \mathcal{X} \) is a nonempty convex constraint set and each local cost function \( f_i \) (only known to node \( i \)) is convex and maybe nonsmooth. \( \chi_i \) is a simple convex regularization function associated with node \( i \). In recent years, there are a few works on distributed methods treating with the aforementioned problems with composite framework (e.g., [14]). Yuan et al. [14] mainly focused on online optimization and developed an online two-point bandit feedback mirror decent based method. Shi and Yang [34] analyzed the distributed optimization problem over relay-assisted networks. From a different perspective, the work in this article considers the network with communication noise and attempts to develop distributed optimization methods that are suitable to the noisy network. Meanwhile, we study the convergence of the distributed composite optimization methods in noisy circumstance.

In this article, inspired by the stochastic approximation theory in [11] and [25], we develop a class of stochastic optimization method for solving the above composite optimization problem. The problems are considered over the time-varying network that has a class of communication noise effects among nodes in the information transmission process. We propose distributed stochastic composite mirror descent (DSCMD-N) method for problem (1). The convergence results are analyzed in detail. Specifically, we are interested in the convergence behavior of the methods under different selections of stepsizes. The expected convergence bound and high probability convergence bound are established, respectively. In what follows, the discussion on the selection of stepsizes and corresponding convergence rates is provided. Note that by taking the composite regularization function into consideration, this work also extends the former works in the same literature to a more general setting. For the proposed DSCMD-N, by implementing the Bregman divergence instead of the former Euclidean distance in works, such as [11] and [25], the DSCMD-N method extends the projection structure of these methods to a more general setting. The explicit rate \( O(1/\sqrt{T}) \) result is obtained for an expected function error for the DSCMD-N method under the appropriate selection of stepsize. The convergence behavior is described by the convergence bound in terms of \( \alpha_t \) (stepsize for stochastic gradient) and \( r_t \) (decaying rate for noise vector) and some cross terms of them. The error bound obtained in this work describes some intrinsic tradeoff between \( \{\alpha_t\} \) and \( \{r_t\} \).

The technical contributions of this article can be summarized as follows.

1) New method DSCMD-N is presented for distributed optimization over a class of noisy network. Existing works in the same literature, such as [11] and [25], are all Euclidean projection based. By presenting the DSCMD-N method, we extend these former works to a more general setting in the proposed network model. In contrast to the previous work in the same literature (e.g., [25]), since the Bregman divergence is utilized, the underlying geometry structure of the distributed optimization problem is better reflected. The flexible selection of mirror map (distance generating function) can enable us to generate efficient updates to face the noisy network optimization. As special cases, when we take the distance generating function as the norm squared function \( (1/2)\|\|^{2} \), the entropy induced function \( \sum_{i=1}^{N} \ln |x_i| \) or \( l_p \) norm squared function \( (1/2)\|\|^{2} \), and take \( \chi_i \) as some specific regularizers, the DSCMD-N method can include a wide range of algorithm class that previous works on noisy network optimization does not consider. In addition, the intrinsic results among the network noise, non-Euclidean structure, and the composite terms are provided (Theorem 1).

2) The convergence behaviors for DSCMD-N are comprehensively investigated. We obtain two types of error bounds for expected error: a) expected bound and b) high probability bound. The bounds are in terms of some cross terms consisting of stepsizes \( \{\alpha_t\} \) and communication noise decaying rate \( \{r_t\} \). The stepsize selection is comprehensively conducted under the effects of \( \{r_t\} \) in different orders. The corresponding convergence rates are obtained. To the best of our knowledge, all these rates are first achieved in the setting of optimization over the noisy network. We also show that the optimal
expected rate and high probability of $O(1/\sqrt{T})$ can be obtained under some conditions on $[\alpha_t]$ and $[r_t]$. Furthermore, a new almost sure convergence type result is derived for the local sequence. This result is new in the literature of distributed optimization.

3) Composite optimization is investigated in a class of time-varying noisy network. Hence, for different purposes on some concrete distributed optimization problems, flexible selection of regularization terms becomes possible. By taking appropriate regularization terms into consideration, the proposed methods are potentially flexible to reflect certain structure features of the solution of distributed optimization problem. This work allows the objective functions to be nonsmooth. This fact makes the proposed methods more flexible to handle optimization problems when tough smoothness conditions are added on objective functions. Also, the methods are convenient for optimization over a class of time-varying network, in contrast to the static network.

Notation and Terminology: Denote the n-dimension Euclidean space by $\mathbb{R}^n$, and the set of positive real numbers by $\mathbb{R}^+$. For a vector $v \in \mathbb{R}^n$, use $||v||$ and $[v]_i$ to denote its Euclidean 2-norm and its i-th entry. Use $||v||_1$ to denote 1-norm $||v||_1 = ||v_1|| + ||v_2|| + \cdots + ||v_n||$. The inner product of two vectors $v_1$ and $v_2$ is denoted by $(v_1, v_2)$. For a matrix $M \in \mathbb{R}^{n \times n}$, denote the element in the i-th row and j-th column by $[M]_{ij}$. Use $I_n$ to denote the identity matrix. A function $f$ is $\sigma_i$-strongly convex over domain $\mathcal{X}$ if for any $x, y \in \mathcal{X}$ and $\zeta \in [0, 1]$, $f(\zeta x + (1-\zeta) y) \leq \sigma_i(f(x)+f(y)) - (\frac{\sigma_i}{2}) \|x-y\|^2$. Denote the gradient operator by $\nabla$, when $f$ is differentiable, the $\sigma_i$-strongly convex inequality above is equivalent to $f(x) \geq f(y) + (\nabla f(y), x-y) + \frac{\sigma_i}{2} \|x-y\|^2$. For two functions $f$ and $g$, write $f(n) = O(g(n))$ if there exist $N < \infty$ and positive constant $C < \infty$ such that $f(n) \leq C g(n)$ for $n \geq N$. For a random variable $X$, use $\mathbb{E}[X]$ to denote its expected value.

II. PROBLEM SETTING AND PRELIMINARIES

Let $G_t = (\mathcal{V}, E_t, P_t)$ be a directed graph, which denotes the information communication among the nodes at time $t$. $\mathcal{V} = \{1, 2, \ldots, N\}$ is the node set. $E_t = \{(i, j) | [P]_{ij} > 0, i, j \in \mathcal{V}\}$ is the set of active links with $P_t$ being the weight matrix at time $t$. $(i, j) \in E_t$ corresponds to the case when agent $i$ and agent $j$ have information communication at time $t$.

The objective of this article is to cooperatively solve the composite optimization problem (1) through communication among the agents of a multiagent system described by graph $G_t$ in a constrained setting. The decision space $\mathcal{X} \subseteq \mathbb{R}^n$ for the state variable $x$ is a convex and compact set. Recall that a compact constrained condition on $\mathcal{X}$ is standard and commonly considered in works on mirror descent type methods (e.g., [3] and [21]). For agent $i \in \mathcal{V}$, we assume that there is a corresponding local cost function $f_i$, $f_i$ is assumed to be convex and perhaps nonsmooth. We assume that the set of nonempty optimal solution of the problems considered in this article is denoted by $\mathcal{X}^*$ with the optimal value $f(x^*)$ for any $x^* \in \mathcal{X}^*$. The following standard assumption is made on the graph $G_t$.

Assumption 1: The communication matrix $P_t$ is doubly stochastic. That is, $\sum_{i=1}^{N} [P]_{ij} = 1$ and $\sum_{j=1}^{N} [P]_{ij} = 1$ for any $i, j \in \mathcal{V}$. There exists some positive integer $\beta$ such that the graph $(\mathcal{V}, \cup_{t=0}^{\infty} E_t)$ is strongly connected for every $s \geq 0$. There exists a scalar $0 < \theta < 1$ such that $[P]_{ij} \geq \theta$ for all $i \in \mathcal{V}$ and $t$, and $[P]_{ij} \geq \theta$ if $(j, i) \in E_t$.

The network model in Assumption 1 is widely used in distributed multiagent optimization community (e.g., [1] and [13]). For realizing the doubly stochasticity, in many settings, we can construct a doubly stochastic matrix according to the rules in [44]. In this article, $P(t, s) = P^t P^{t-1} \cdots P^s$ is used to denote the transition matrix when $t \geq s$; the notation $P(t, t+1) = I_n$ is also used. The following consequence in [1] is basic for the analysis over the multiagent time-varying network.

Lemma 1: Under Assumption 1, for all $i, j \in \mathcal{V}$ and $t, s$ satisfying $t \geq s \geq 1$, we have $[P(t, s)]_{ij} = \frac{1}{N} \leq \omega\gamma^{t-s}$ in which

$$\omega = \left(1 - \frac{\theta}{4N^2}\right)^2$$

$$\gamma = \left(1 - \frac{\theta}{4N^2}\right)^{\frac{1}{N}}.$$
magnitude of the communication noises decays to zero
\[ \|n'_t\| \to 0, \text{ a.s.} \]

In this article, in order to investigate some novel explicit convergence rates under the theoretical framework of rate analysis, we assume that the noise \( \{n'_t\} \) satisfies Assumption 3 with magnitude decaying rate \( r_t \). In fact, the noise shares a similar noise decaying feature with [18] and [42]

\[ \mathbb{E}\left[ \|n'_t\| \right] \to 0. \]

Assumptions in this article can include some types of noises that the assumption \( \sum_{t=1}^{\infty} \mathbb{E}\left[\|n'_t\|^2\right]F_{t-1} < \infty \) does not cover. For example, the noise \( n'_t = r_t \xi'_t \) with \( r_t \geq O(1/\sqrt{t}) \) and all \( \xi'_t \) have equal nonzero \( \mathbb{E}\left[\|\xi'_t\|^2\right] > 0 \) satisfies assumptions in this article, but does not satisfy assumption \( \sum_{t=1}^{\infty} \mathbb{E}\left[\|n'_t\|^2\right]F_{t-1} < \infty \) as in [18], since \( \sum_{t=1}^{\infty} r_t = \infty \) when \( r_t \geq O(1/\sqrt{t}) \). The fact shows that our work can cover some noise types that the work under \( \sum_{t=1}^{\infty} \mathbb{E}\left[\|n'_t\|^2\right]F_{t-1} < \infty \) assumption cannot handle. In practice, as stated in [4], [5], and [18] and simulation examples of [27], there exists such type of noise model in common network engineering, for example, some injected false data and noise signals decay with time in some network adversarial attack, or the communication noise decays when a multiagent system moves away gradually from noisy source, or the noise is caused by a dropped external noise source.

The following Azuma–Hoeffding lemma [6], [7] is needed to derive high probability bound and rate of DSCMD-N later.

**Lemma 2:** Let \( \{X_t\} \) be a martingale difference sequence satisfying \( |X_t| \leq \tau_t \), then for any \( \epsilon > 0 \)

\[ \mathbb{P}\left( \sum_{t=1}^{T} X_t \geq \epsilon \right) \leq \exp\left( -\frac{\epsilon^2}{2 \sum_{t=1}^{T} \tau_t^2} \right). \]

In the optimization literature, mirror descent is a powerful extension of classical gradient descent. Generally, in contrast to gradient descent, for a given decision space defined on a Hilbert space, the mirror descent can relax the Hilbert space structure and employ a mirror map \( \Phi : \mathcal{X} \to \mathbb{R}^n \) to better reflect the geometric properties of the decision variables from some Banach space \( \mathcal{K} \). In this article, we will consider \( \mathcal{K} = \mathbb{R}^n \) endowed with a norm \( \|\cdot\| \), which may be a non-Euclidean norm, that can better reflect the non-Euclidean geometric structures of decision variable from \( \mathbb{R}^n \). To introduce the basic distributed mirror descent scheme, we consider a continuously differentiable \( \sigma_{\Phi} \)-strongly convex mirror map (distance generating function) \( \Phi : \mathbb{R}^n \to \mathbb{R}^n \), and define the Bregman divergence associated with \( \Phi \) as

\[ D_{\Phi}(x, y) = \Phi(x) - \Phi(y) - \langle \nabla \Phi(y), x - y \rangle. \]

In Section III, for the Bregman divergence, we need the following assumption. The assumption is standard in investigations of mirror descent type methods [3], [21].

**Assumption 4:** We assume that the mirror map \( \Phi \) is chosen such that \( \|\nabla \Phi(x) - \nabla \Phi(y)\| \leq L_{\Phi} \|x - y\| \) for any \( x, y \in \Omega \) for some \( L_{\Phi} \). For any vectors \( a \) and \( \{b_t\}_{t=1}^{N} \) in \( \mathbb{R}^n \), the Bregman divergence satisfies the separate convexity in the following sense: \( D_{\Phi}(a, \sum_{i=1}^{N} v_i b_i) \leq \sum_{i=1}^{N} v_i D_{\Phi}(a, b_i), \) \( v_i \in [0, 1] \) and \( \sum_{i=1}^{N} v_i = 1 \).

III. DSCMD-N Algorithm: Main Convergence Results

In this section, we consider problem (1), minimizing \( F(x) = \sum_{i=1}^{N} F_i(x) := \sum_{i=1}^{N} [f_i(x) + \chi_i(x)] \) over the noisy network. We solve the problem by providing a distributed stochastic composite mirror descent method, which we call it the DSCMD-N method. In the algorithm, for each \( i \in \mathcal{V} \), the local variable \( x'_i \) evolves as follows:

\[ y'_i = \sum_{j=1}^{n} [P^s]_{ij} (\delta_{ij} + r_t \xi'_j) \]

\[ x'_{i+1} = \arg \min_{x \in \mathcal{X}} \left\{ \mathbb{E}\left[ g_{t}^{ij} (x, y'_i) + \frac{1}{a_t} D_{\Phi}(x, y'_i) + \chi_i (x) \right] \right\} \]

where \( [P^s]_{ij}, i, j \in \mathcal{V} \) denotes the elements of the communication weight matrix \( P^s \) satisfying the conditions in Assumption 1. It denotes the weight assigned by node \( i \) to the estimate coming from node \( j \). In the algorithm, we are concerned with the case when communication links are noisy with noise assumptions in Assumption 3. Therefore, the node \( i \) has only access to a noise corrupted value of its neighbor’s local decision variable (noisy observation). The algorithm structure in (2) and (3) is mainly inspired by the classical distributed gradient descent in [1]. The basic idea is that, “first communication, then projection.” Equation (2) describes the noisy information communication process between \( i \) and its neighbors. After such information communication, we obtain an intermediate variable \( y'_i \). Then, query the stochastic subgradient oracle at \( y'_i \) to obtain a stochastic subgradient \( \tilde{g}_i(y'_i) = \tilde{g}_i(y'_i) \), such that \( \mathbb{E}[\tilde{g}_i(y'_i)] = g_i(y'_i) \) is a subgradient of \( f_i \) at \( y'_i \). In (3), we perform a stochastic Bregman projection for variable \( y'_i \) to decision space \( \mathcal{X} \) to obtain variable \( x'_{i+1} \). After this non-Euclidean projection, we finish an updating procedure to obtain the variable \( x'_{i+1} \). A composite mirror descent scheme is considered in this Bregman projection with stepsize \( a_t \) and composite term \( \chi_i(x) \) (see, [45]). In this article, for technical consideration, we treat with the settings that noisy state variable lies in the decision manifold. We remark that the simple composite function \( \chi_i \) associated with node \( i \) can be different from each other. Here, \( \chi_i(x), i \in \mathcal{V} \) are supposed to be some simple convex regularization function with supremum subgradient \( G_{\chi} \). In this section, \( \tilde{g}_i(x) \) is used to denote the subdifferential set of \( \chi_i \) at \( x \in \mathcal{X} \). For the subgradients of \( \chi_i \), we denote

\[ G_{\chi} = \sup_{g \in \tilde{g}_i(x)} \|g\| \]

and

\[ G_{\chi} = \sup_{i \in \mathcal{V}} G_{\chi}. \]

In this section, we also denote

\[ D_{\Phi, \mathcal{X}} = \sup_{x, y \in \mathcal{X}} D_{\Phi}(x, y). \]

In fact, the finiteness of \( G_{\chi} \) and \( D_{\Phi, \mathcal{X}} \) follows from the compactness of \( \mathcal{X} \), and the strong convexity of \( \Phi \) implies \( \sup_{x, y \in \mathcal{X}} \|x - y\| \leq \sqrt{2/\sigma_{\Phi} D_{\Phi, \mathcal{X}}} \). In this article, it is unnecessary to know the concrete value of \( G_{\chi} \) and \( D_{\Phi, \mathcal{X}} \). The
finiteness of them is enough to provide the rigorous convergence analysis of the algorithm. To investigate the convergence behavior of DSCMD-N, we denote the Bregman projection error by

$$
ep_i^t = x_i^{t+1} - y_i^t.
$$

We start with the following error estimate on $\ep_i^t$.

**Lemma 3:** The Bregman projection error satisfies

$$
\mathbb{E}[\|\ep_i^t\|^2] \leq \frac{G_i + G_x}{\sigma_\Phi} \alpha_t.
$$

**Proof:** According to the first-order optimality condition, there exists $\bar{h}_i^{t+1} \in \partial \chi_i(x_i^{t+1})$ such that

$$
\left(\alpha \bar{g}_i^t + \nabla \Phi(x_i^{t+1})\right) - \nabla \Phi(y_i^t) + \alpha \bar{h}_i^{t+1}, x - x_i^{t+1}\right) \geq 0 \quad \forall x \in \mathcal{X}.
$$

Setting $x = y_i^t$ in the above inequality, we obtain that

$$
\left(\alpha \bar{g}_i^t + \nabla \Phi(x_i^{t+1})\right) - \nabla \Phi(y_i^t) + \alpha \bar{h}_i^{t+1}, y_i^t - x_i^{t+1}\right) \geq 0.
$$

The above inequality implies that

$$
\left(\alpha \bar{g}_i^t + \bar{h}_i^{t+1}, y_i^t - x_i^{t+1}\right) \geq \nabla \Phi(y_i^t) - \nabla \Phi(x_i^{t+1}), y_i^t - x_i^{t+1}\right).
$$

Use the Cauchy inequality to the left-hand side and $\sigma_\Phi$-strong convexity of $\Phi$ to the right-hand side of the above inequality, it can be obtained that

$$
\alpha_t\left(\|\bar{g}_i^t\|^2 + G_x\right) \|y_i^t - x_i^{t+1}\| \geq \sigma_\Phi \|y_i^t - x_i^{t+1}\|^2.
$$

Eliminating the same term $\|y_i^t - x_i^{t+1}\|^2$ on both sides and taking conditional expectation on $\mathcal{F}_{t-1}$, we have

$$
\mathbb{E}[\|\ep_i^t\| | \mathcal{F}_{t-1}] \leq \frac{1}{\sigma_\Phi} \left(\mathbb{E}[\|\bar{g}_i^t\|^2 | \mathcal{F}_{t-1}] + G_x\right) \alpha_t.
$$

The desired result is obtained after taking total expectation of the above inequality on both sides.

We are ready to give the following disagreement result, which is necessary to establish the main convergence result of this section. In what follows, for nodes with estimates $x_i^t$, $i \in \mathcal{V}$, we denote the average estimate of them at time $t$ by:

$$
\bar{x}^t = \frac{1}{N} \sum_{i=1}^N x_i^t.
$$

The average $\bar{x}^t$ will play an important bridge role in the following proofs. In distributed convex optimization, the consensus among agent is often crucial to the convergence of the global objective function. The next result describes such crucial relation on the disagreement among agents, which is important for us to derive main results of the article.

**Lemma 4:** Under Assumptions 1–3, let $\{\xi_i^t\}$ be the sequences in DSCMD-N. Then, for any $i \in \mathcal{V}$

$$
\sum_{t=1}^T \sum_{i=1}^N \mathbb{E}[\|x_i^t - x_i^{t+1}\|^2] \leq \frac{2N \omega}{1 - \gamma} \sum_{j=1}^N \|x_j^0\|^2
$$

$$
+ \left(4N + \frac{2N^2 \omega}{1 - \gamma} \right) \sum_{t=0}^{T} \left[\frac{G_i + G_x}{\sigma_\Phi} \alpha_t + N \sqrt{\nu r_t}\right].
$$

**Proof:** For all $i \in \mathcal{V}$, set $\xi_i^t = \sum_{j=1}^N [P^t]_{ij} \xi_j^{t-1}$, by iterating recursively, it can be obtained that

$$
x_i^t = \sum_{s=1}^t \sum_{j=1}^N [P(t - 1, s)]_{ij} \left(\ep_j^{s-1} + r_s - 1 \xi_j^{s-1}\right)
$$

$$
+ \sum_{j=1}^N [P(t - 1, 0)]_{ij} x_j^0
$$

$$
\bar{x}^t = \frac{1}{N} \sum_{s=1}^t \sum_{j=1}^N \left(\ep_j^{s-1} + r_s - 1 \xi_j^{s-1}\right) + \frac{1}{N} \sum_{j=1}^N x_j^0.
$$

Then, it follows that

$$
\|x_i^t - \bar{x}^t\| \leq \sum_{j=1}^N \|P(t - 1, j)\| \cdot \|x_j^0\|
$$

$$
+ \sum_{s=1}^t \sum_{j=1}^N [P(t - 1, s)]_{ij} \left(\ep_j^{s-1} + r_s - 1 \xi_j^{s-1}\right)
$$

$$
\leq \omega \gamma^{t-1} \sum_{i=1}^N \|x_j^0\| + \sum_{s=1}^{t-1} \omega \gamma^{t-s-1} \sum_{j=1}^N \|\ep_j^{s-1} + r_s - 1 \xi_j^{s-1}\|
$$

$$
+ \frac{1}{N} \sum_{j=1}^N \|\ep_j^{t-1} + r_t - 1 \xi_j^{t-1}\| + \|\ep_i^t + r_t - 1 \xi_i^{t-1}\|.
$$

Since $\mathbb{E}[\|\xi_j^{t-1}\|] = \mathbb{E}[\|\sum_{j=1}^N [P^{t-1}]_{ij} \xi_j^{t-1}\|] \leq \sum_{j=1}^N \mathbb{E}[\|\xi_j^{t-1}\|] \leq N \sqrt{\nu}$, then $\mathbb{E}[\|\ep_j^{t-1} + r_t - 1 \xi_j^{t-1}\|] \leq \frac{G_i + G_x}{\sigma_\Phi} \alpha_t + N \sqrt{\nu r_{t-1}}$. Combining these inequalities, it follows that for any $i \in \mathcal{V}$:

$$
\mathbb{E}[\|x_i^t - \bar{x}^t\|] \leq \omega \gamma^{t-1} \sum_{j=1}^N \|x_j^0\|
$$

$$
+ \sum_{s=1}^{t-1} \omega \gamma^{t-s-1} \left(\frac{G_i + G_x}{\sigma_\Phi} \alpha_t + N \sqrt{\nu r_{t-1}}\right)
$$

$$
+ \frac{1}{N} \sum_{j=1}^N \left(\frac{G_i + G_x}{\sigma_\Phi} \alpha_t + N \sqrt{\nu r_{t-1}}\right)\|x_j^0\|
$$

$$
+ \left(4N + \frac{2N^2 \omega}{1 - \gamma} \right) \sum_{t=0}^{T} \left[\frac{G_i + G_x}{\sigma_\Phi} \alpha_t + N \sqrt{\nu r_t}\right].
$$

Summing up both sides of the above inequality from $t = 1$ to $T$ and $i = 1$ to $N$, it follows that:

$$
\sum_{t=1}^T \sum_{i=1}^N \mathbb{E}[\|x_i^t - \bar{x}^t\|] \leq \frac{N \omega}{1 - \gamma} \sum_{j=1}^N \|x_j^0\|
$$

$$
+ \left(2N + \frac{N^2 \omega}{1 - \gamma} \right) \sum_{t=0}^{T} \left[\frac{G_i + G_x}{\sigma_\Phi} \alpha_t + N \sqrt{\nu r_t}\right].
$$

Note that the bound on the right-hand side of (6) does not depend on the index $i$. For any index $j \in \mathcal{V}$, $\mathbb{E}[\|x_j^t - \bar{x}^t\|]$ also satisfies the bound in (6). Summing up from $t = 1$ to $T$ and $i = 1$ to $N$ to $\mathbb{E}[\|x_j^t - \bar{x}^t\|]$, using the triangle inequality $\mathbb{E}[\|x_j^t - x_j^0\|] \leq \mathbb{E}[\|x_j^t - \bar{x}^t\|] + \mathbb{E}[\|x_j^0 - x_j^0\|]$, and combining with (7), the result in theorem is obtained.
Lemma 5: Let \( \{x_i^j\} \) and \( \{y_i^j\} \) be the sequences in DSCMD-N. Let \( \{\alpha_i\} \) be a nonincreasing stepsize. Then, we have
\[
\chi_i(x_i^{j+1}) - \chi_i(x_i) + \langle g_i^j, x_i^j - x_i \rangle 
\leq \frac{1}{\alpha_i} \left[ D_\Phi(x_i, y_i) - D_\Phi(x_i^{j+1}) \right] + \frac{\alpha_i}{2\sigma_\Phi} \|g_i^j\|^2. 
\]
Proof: According to the first-order optimality of the DSCMD-N, there exists \( h_i^{j+1} \in \partial \chi_i(x_i^{j+1}) \)
\[
\langle \alpha_i g_i^j + \nabla \Phi(x_i^{j+1}) - \nabla \Phi(y_i^j) + \alpha_i h_i^{j+1}, x_i - x_i^{j+1} \rangle \geq 0 \forall x_i \in \mathcal{X}.
\]
Setting \( x = x_i^* \) in the above inequality, and rearranging, we have
\[
\begin{align*}
\alpha_i g_i^j, x_i^{j+1} - x_i^* & \leq \left( \nabla \Phi(y_i^j) - \Phi(x_i^{j+1}), x_i^{j+1} - x_i^* \right) \\
& \leq D_\Phi(x_i, y_i^j) - D_\Phi(x_i^{j+1}) - D_\Phi(x_i^*, y_i^j) \\
& \quad + \chi_i(x_i^*) - \chi_i(x_i^{j+1}) \\
& \leq D_\Phi(x_i^*, y_i^j) - D_\Phi(x_i, y_i^j) - \frac{\sigma_\Phi}{2} \|x_i^{j+1} - y_i^j\|^2 \\
& \quad + \chi_i(x_i^*) - \chi_i(x_i^{j+1}).
\end{align*}
\]
(9)

in which the second inequality follows from the three point inequality and the second inequality follows from the definition of \( D_\Phi(\cdot, \cdot) \) and \( \sigma_\Phi \)-strong convexity of \( \Phi \). Also
\[
\begin{align*}
\alpha_i g_i^j, x_i^{j+1} - x_i^* & \leq \alpha_i g_i^j, x_i^{j+1} - y_i^j + \alpha_i g_i^j, y_i^j - x_i^* \\
& \leq -\frac{\alpha_i^2}{2\sigma_\Phi} \|g_i^j\|^2 - \frac{\sigma_\Phi}{2} \|x_i^{j+1} - y_i^j\|^2 \\
& \quad + \alpha_i g_i^j, y_i^j - x_i^*.
\end{align*}
\]
(10)

Combining (9) and (10), it follows that:
\[
\begin{align*}
\alpha_i g_i^j, x_i^{j+1} - x_i^* & \leq D_\Phi(x_i^*, y_i^j) - D_\Phi(x_i, y_i^j) \\
& \quad + \frac{\alpha_i^2}{2\sigma_\Phi} \|g_i^j\|^2 + \alpha_i \left[ \chi_i(x_i^*) - \chi_i(x_i^{j+1}) \right].
\end{align*}
\]

The proof is concluded after dividing both sides by \( \alpha_i \) in the above inequality.

Lemma 6: Let \( \{x_i^j\} \) and \( \{y_i^j\} \) be the sequences in DSCMD-N. Then, there holds
\[
\mathbb{E}[\|y_i^j - x_i^j\|^2] \leq \sum_{j=1}^N \mathbb{E}[\|x_i^j - x_i^j\|^2] + N \sqrt{\nu} r_i
\]
for any \( i, l \in \mathcal{V} \).

Proof: According to the structure of DSCMD-N and the fact that the matrix \( P^i \) is doubly stochastic
\[
\|y_i^j - x_i^j\|^2 = \sum_{j=1}^N \langle P^i \rangle_{lj} [x_i^j - x_i^j] + r_i \sum_{j=1}^N \|P^i \|_{lj} \|x_i^j - x_i^j\| \\
\leq \sum_{j=1}^N \langle P^i \rangle_{lj} [x_i^j - x_i^j] + r_i \sum_{j=1}^N \|x_i^j\|.
\]

Taking expectation over \( \mathcal{F}_{l-1} \), using Assumption 3 and the fact that \( 0 \leq \langle P^i \rangle_{lj} < 1 \), and then taking total expectation, the lemma is concluded.

Lemma 7: Let \( \{x_i^j\} \) be the sequences in DSCMD-N, the noise sequence \( \{r_i^j\}\) is defined as before, we have
\[
\mathbb{E}[D_\Phi(x_i^*, x_i^j + r_i^j)] = \mathbb{E}[D_\Phi(x_i^*, x_i^j)] + \sqrt{(2)/(\sigma_\Phi)} D_\Phi \chi L_\Phi \sqrt{r_i} + L_\Phi L_\Phi^2.
\]

Proof: According to mean value formula, there exists \( \zeta \in [0, 1] \) such that \( \Phi(x_i^j + r_i^j) = \Phi(x_i^j) + (\nabla \Phi(x_i^j + \zeta r_i^j), r_i^j) \)
then it follows that:
\[
\begin{align*}
D_\Phi(x_i^*, x_i^j + r_i^j) & = \Phi(x_i^*) - \Phi(x_i^j + r_i^j) \\
& = \langle \nabla \Phi(x_i^j + r_i^j), x_i^* - x_i^j - r_i^j \rangle \\
& = \langle \nabla \Phi(x_i^j + r_i^j), x_i^* - x_i^j \rangle \rangle \\
& = \langle \nabla \Phi(x_i^j + r_i^j), x_i^* - x_i^j \rangle \\
& = \Phi(x_i^*) - \Phi(x_i^j) - \langle \nabla \Phi(x_i^j + r_i^j), x_i^* - x_i^j \rangle \\
& = \langle \nabla \Phi(x_i^j + r_i^j), x_i^* - x_i^j \rangle \\
& \leq D_\Phi(x_i^*, x_i^j) + (1 - \zeta) L_\Phi r_i^j \|x_i^* - x_i^j\|^2 + L_\Phi L_\Phi r_i^j \|x_i^* - x_i^j\|^2 \\
& \leq D_\Phi(x_i^*, x_i^j) + D_\Phi(x_i^*, x_i^j) \|x_i^* - x_i^j\|^2
\end{align*}
\]
in which the first inequality follows from the Cauchy inequality and gradient \( L_\Phi \)-Lipschitz condition of \( \Phi \), and the second inequality follows from the fact \( 0 \leq 1 - \zeta \leq 1 \). Taking conditional expectation on \( \mathcal{F}_{l-1} \) on both sides, using Assumption 3, and noting that \( D_\chi \leq \sqrt{(2)/(\sigma_\Phi)} D_\Phi \chi \), the result is obtained after taking the total expectation.

Now, returning to (8), taking the conditional expectation over \( \mathcal{F}_{l-1} \) on both sides of (8), we have
\[
\begin{align*}
\mathbb{E}[g_i(y_i^j, y_i^j - x_i^*)] + \mathbb{E}[\chi_i(x_i^{j+1}) - \chi_i(x_i^*)] \\
\leq \frac{1}{\alpha_i} \left[ D_\Phi(x_i^*, y_i^j) - D_\Phi(x_i^*, x_i^{j+1}) \right] \\
& + \frac{\alpha_i}{2\sigma_\Phi} \mathbb{E}[\|g_i^j\|^2].
\end{align*}
\]

Take the total expectation on both sides of the above inequality, we have
\[
\Delta_i^1 + \Delta_i^2 \leq \Delta_i^3
\]
(11)
in which we denote
\[
\begin{align*}
\Delta_i^1 & = \mathbb{E}[g_i(y_i^j, y_i^j - x_i^*)] \\
\Delta_i^2 & = \mathbb{E}[\chi_i(x_i^{j+1}) - \chi_i(x_i^*)] \\
\Delta_i^3 & = \frac{1}{\alpha_i} \left[ D_\Phi(x_i^*, y_i^j) - D_\Phi(x_i^*, x_i^{j+1}) \right] + \frac{\alpha_i}{2\sigma_\Phi} \mathbb{E}[\|g_i^j\|^2].
\end{align*}
\]

Before coming to the main result, we need the following lemma for \( \Delta_i^3 \).

Lemma 8: Under Assumptions 1–4, if \( \{\alpha_i\} \) and \( \{r_i\} \) be non-increasing positive sequences, then the following bound result for \( \Delta_i^3 \) holds:
\[
\begin{align*}
\sum_{i=1}^T \sum_{l=1}^N \Delta_i^3 & \leq \frac{N^2 D_\Phi \chi L_\Phi \sqrt{v}}{\alpha_i} + \frac{N^2 \sigma_\Phi}{2\alpha_i} \sum_{i=1}^T \alpha_i \\
& + N \sum_{i=1}^T \sum_{l=1}^N \sum_{i=1}^N \sum_{l=1}^N \frac{1}{\alpha_i} \left[ D_\Phi(x_i^*, x_i^j) \right] + \mathbb{E}[\|g_i^j\|^2].
\end{align*}
\]
Proof: Since \( y_i^t = \sum_{j=1}^{N} P_j^t (y_i^j + r_j^t \xi_i^j) \), separate convexity of \( D_{\phi}(\cdot, \cdot) \) implies that

\[
\sum_{t=1}^{T} \sum_{i=1}^{N} \left[ \Delta_{i,t}^2 \right] \leq \sum_{t=1}^{T} \frac{1}{\alpha_t} \left[ \sum_{i=1}^{N} \sum_{j=1}^{N} \left[ P_j^t y_i^j E \left[ D_{\phi} \left( x^*, x_j^t + r_j^t \xi_i^j \right) \right] \right] \right] - \sum_{t=1}^{T} \frac{N}{\alpha_{t-1}} \left[ \sum_{i=1}^{N} \sum_{j=1}^{N} \left[ P_j^t y_i^j E \left[ D_{\phi} \left( x^*, x_j^{t+1} \right) \right] \right] \right] + 2G_{\phi} \sqrt{\sum_{i=0}^{T-1} \frac{r_i}{\alpha_t}} + NL_{\phi} \sqrt{\sum_{i=0}^{T-1} \frac{r_i^2}{\alpha_t}}
\]

in which the second inequality is obtained by double stochasticity of matrix \( P^t \) and Lemma 7, and the result is obtained after eliminating same terms in the summation in the above equality.

Now, we are ready to give the main result of this section. Denote

\[
\tilde{x}^t_i = \frac{1}{T} \sum_{i=1}^{T} x_i^t
\]

and

\[
x^* = \arg \min_{x \in X} F(x).
\]

The following result describes the expected bound for DSCMD-N in terms of stepsizes \( \alpha_t \) and noise decaying rates \( \alpha_t \).

Theorem 1: Let the Assumptions 1-4 hold. If \( \{\alpha_t\} \) and \( \{\gamma_t\} \) are positive nonincreasing sequences, then for DSCMD-N method, for any \( l \in V \), we have

\[
\mathbb{E}[F(\tilde{x}^t_i)] - F(x^*) \leq \frac{C_1}{T} + \frac{C_2}{T} + \frac{C_3}{T} \sum_{t=0}^{T} \alpha_t + \frac{C_4}{T} \sum_{t=0}^{T} r_t + \frac{C_5}{T} \sum_{t=0}^{T} \alpha_t + \frac{C_6}{T} \sum_{t=0}^{T} \alpha_t
\]

in which

\[
C_1 = \frac{2N\omega}{1 - \gamma} [N + 1] G_f + NG_x \sum_{j=1}^{N} \|x_j^0\|
\]

\[
C_2 = N\Delta_{\phi,X}^2
\]

\[
C_3 = \left( \frac{4N + 2N^2 \omega}{1 - \gamma} \right) [(N + 1) G_f + NG_x]
\]

\[
+ NG_x \left( \frac{G_f + G_x}{\sigma_{\phi}} + \frac{NG_x^2}{2\sigma_{\phi}} \right)
\]

\[
C_4 = \left( \frac{4N + 2N^2 \omega}{1 - \gamma} \right) [(N + 1) G_f + NG_x] N\sqrt{\nu}
\]

\[
+ (G_f + G_x) N^2 \sqrt{\nu}
\]

\[
C_5 = \sqrt{\frac{2}{\sigma_{\phi}}} D_{\phi} C L_{\phi} N \sqrt{\nu} \quad C_6 = NL_{\phi} \nu
\]

and \( \omega = (1 - [\theta/4N^2])^{-2} \), \( \gamma = (1 - [\theta/4N^2])^{(1/2)} \).

Proof: We prove the result by estimating the terms in (11). For any index \( l \in V \)

\[
\mathbb{E}[g_i(y_i^t), y_i^t - x^*] \geq f_i(y_i^t) - f_i(x^*)
\]

\[
= f_i(y_i^t) - f_i(x^*) + f_i(x^*) - f_i(x^*)
\]

\[
\geq -G_f \|y_i^t - x_i^t\| - G_f \|x_i^t - x_i^t\| + f_i(x^*) - f_i(x^*)
\]

\[
\geq -G_x \|x_i^t - x_i^t\| - G_x \|x_i^t - x_i^t\| + f_i(x^*) - f_i(x^*)
\]

In which the second inequality follows from \( g_i(x) = \mathbb{E}[g_i(x)|f_{i-1}] \leq \mathbb{E}[\|g_i(x)|f_{i-1}] \leq G_f \). After taking expectation and using Lemma 6, it follows that \( \Delta_{i,t}^2 \geq -G_f \sum_{i=1}^{N} \mathbb{E}[\|x_i^t - x_i^t\|] + \mathbb{E}[\|x_i^t - x_i^t\|] + \mathbb{E}[f_i(x^*) - f_i(x^*)] \). Denoting \( f = \sum_{i=1}^{N} f_i \) and \( \chi = \sum_{i=1}^{N} \chi_i \) and denoting the bound on the right-hand side in Lemma 4 by \( B_T \), summing up both sides and using Lemma 4, it follows that:

\[
\sum_{t=1}^{T} \sum_{i=1}^{N} \left[ \Delta_{i,t}^2 \right] \geq -(N + 1) G_f B_T - N^2 \sqrt{\nu} G_f \sum_{t=1}^{T} r_t
\]

\[
+ \sum_{t=1}^{T} \mathbb{E}[f(x_i^t) - f(x^*)].
\]

On the other hand, for any index \( l \in V \)

\[
\chi_i(\tilde{x}_i^t) - \chi_i(x^*)
\]

\[
= \left[ \chi_i(\tilde{x}_i^{t+1}) - \chi_i(x^*) \right] + \left[ \chi_i(x^*) - \chi_i(x^*) \right]
\]

\[
\geq -G_x \|\tilde{x}_i^{t+1} - x_i^{t+1}\| - G_x \|x_i^{t+1} - x_i^{t+1}\| + \left[ \chi_i(x^*) - \chi_i(x^*) \right].
\]

After taking expectation on both sides, using Lemmas 3 and 6, we have \( \Delta_{i,t}^2 \geq -G_x \|\tilde{x}_i^{t+1} - x_i^{t+1}\| - G_x \mathbb{E}[\|\tilde{x}_i^{t+1} - x_i^{t+1}\|] \). Summing up from \( i = 1 \) to \( N \) and \( t = 1 \) to \( T \) on both sides, we obtain

\[
\sum_{t=1}^{T} \sum_{i=1}^{N} \left[ \Delta_{i,t}^2 \right] \geq -N G_x \sum_{i=1}^{N} \mathbb{E}[\|x_i^t - x_i^t\|] - G_x \mathbb{E}[\|\tilde{x}_i^{t+1} - x_i^{t+1}\|] - \sum_{t=1}^{T} \sum_{i=1}^{N} \left[ \chi_i(x_i^t) - \chi_i(x^*) \right].
\]

(15)

Sum up both sides of (11) from \( i = 1 \) to \( N \) and \( t = 1 \) to \( T \), and combine it with (14), (15), and Lemma 8. The desired result is obtained after substituting \( B_T \), using Lemma 4,
dividing both sides by $T$ and using the convexity of $F_i$, $i = 1, 2, \ldots, N$.

It can be seen that the last three terms involve the noise-related parameters $\{r_i\}$, which play an important role on bounding the expected error as in Theorem 1. Under a boundedness assumption of stochastic gradient and network noise, the following high probability bound holds for DSCMD-N.

**Theorem 2**: Under the assumptions of Theorem 1, if we assume in addition that $\|\hat{g}_i\| \leq G$ and $\|\hat{g}_i^j\| \leq v$, then for DSCMD-N, for any $l \in \mathcal{V}$, we have, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, there holds

$$F(\hat{x}_T^l) - F(x^*) \leq \frac{C_1}{T} + \frac{C_2}{T^{1/2}} + \frac{C_3}{T} \sum_{t=0}^{T} \alpha_t + \frac{C_4}{T} \sum_{t=0}^{T} r_t$$

$$+ \frac{C_5}{T} \sum_{t=0}^{T} r_t^2 + \frac{C_6}{T} \sum_{t=0}^{T} \alpha_t^2 + 2\sqrt{2}G_{d}D_{X}N \sqrt{\log(1/\delta)}/\sqrt{T}$$

in which $C_1$-$C_6$ are defined as in Theorem 1.

**Proof**: For saving space, we just show the difference between the proof for this result and the above expected bound result. Coming back to (8), if we denote $\hat{\Delta}_{i,t} = (g_i(x_i^t), y_i^t - x^i)$, $\hat{\Delta}_{2,i} = \chi_i(x_i^{t+1}) - \chi_i(x_i)$, $\hat{\Delta}_{3,i} = (1/\alpha_t)[D_{\phi}(x_i, y_i^t) - D_{\phi}(x_i, x_i^{t+1})]$, $X_i = (g_i^t, y_i, y_i^t - x_i^t)$, then (8) can be written in the form of $\Delta_{i,t}^1 + \Delta_{i,t}^2 \leq \Delta_{i,t}^3 + X_i$, $\hat{\Delta}_{i,t}$, $\hat{\Delta}_{2,i}$, and $\hat{\Delta}_{3,i}$ corresponds to $\Delta_{i,t}^1$, $\Delta_{i,t}^2$, and $\Delta_{i,t}^3$ in (11) only up to a procedure of taking expectation. If we denote $X_i = \sum_{t=1}^{N} X_i$, and sum both sides from $t = 1$ to $T$ and $i = 1$ to $N$, it follows that:

$$\sum_{t=1}^{T} \sum_{i=1}^{N} \Delta_{i,t}^1 + \sum_{t=1}^{T} \sum_{i=1}^{N} \Delta_{i,t}^2 \leq \sum_{t=1}^{T} \sum_{i=1}^{N} \Delta_{i,t}^3 + \sum_{t=1}^{T} X_i.$$  \hfill (16)

Note that $\mathbb{E}[X_i|F_{T-1}] = 0$, the bound condition $\|\hat{x}_i\| \leq G_f$ and Cauchy inequality imply $\|X_i\| \leq 2NG_fD_X$, then $X_i$ is a bounded martingale difference sequence. Using the Azuma–Hoeffding inequality (Lemma 2) to $X_i$, we have for any $\epsilon > 0$

$$\text{Prob} \left( \sum_{t=1}^{T} X_i \geq \epsilon \right) \leq \exp \left( - \frac{\epsilon^2}{2T(2G_fD_X\sqrt{N})^2} \right).$$ \hfill (17)

Setting the above probability upper bound to $\delta$, we have, with the probability at least $1 - \delta$

$$\sum_{t=1}^{T} X_i \leq 2\sqrt{2}G_fD_XN \sqrt{T} \log(1/\delta).$$ \hfill (18)

On the other hand, it is easy to see that with bound assumptions $\|\hat{g}_i\| \leq G$ and $\|\hat{g}_i^j\| \leq v$ in hand, the estimate result of Lemmas 4 and 7 holds without taking expectation. Therefore, we know (12), (14), and (15) hold with $\Delta_{i,t}^1$, $\Delta_{i,t}^2$, and $\Delta_{i,t}^3$ replaced by $\hat{\Delta}_{i,t}^1$, $\hat{\Delta}_{i,t}^2$, and $\hat{\Delta}_{i,t}^3$. Combining these three estimates with (18) and (16), dividing both sides by $T$ and using the convexity of $F_i$, $i = 1, 2, \ldots, N$, we obtain the desired result.

In this section, we have rigorously proved the error bound results in both expectation sense and high probability sense. Based on these theoretical bounds, we are ready to provide the convergence rate analysis in the next section.

**IV. CONVERGENCE RATES OF DSCMD-N**

In this section, we provide a general framework for convergence rate analysis by selecting different step sizes under different effects of noise decaying rates $\{r_i\}$. We also show that in some situations of $\{r_i\}$, by selecting some step sizes of $\{\alpha_t\}$, the best achievable rate of $O(1/T)$ for the centralized subgradient method for nonsmooth convex optimization, can be obtained for DSCMD-N. We present the results on expected rate and high probability rate in the following section.

The following proposition provides a general expected bound for expected error $\mathbb{E}[F(\hat{x}_T^l)] - F(x^*)$ in terms of the total iteration step $T$ with a general stepsize consideration in the form of $\alpha_t = [(1)/(t+1)^{\kappa}]$ and the noise decaying rate in the form of $r_t = [(1)/(t+1)^{\kappa}]$.

**Proposition 1**: Under conditions of Theorem 1, if the sequences $\{\alpha_t\}$ and $\{r_t\}$ in the DSCMD-N method are $\alpha_t = [(1)/(t+1)^{\kappa}]$ and $r_t = [(1)/(t+1)^{\kappa}]$, $t = 1, 2, \ldots, T$. Suppose that $0 < \kappa_1 < \kappa_2 \leq 1$ and $2\kappa_2 - \kappa_1 \neq 1$, then for any $l \in \mathcal{V}$, we have

$$\mathbb{E}[F(\hat{x}_T^l)] - F(x^*) \leq \left( C_1 + \frac{\|1 - 2(2\kappa_2 - \kappa_1)\|C_6}{1 - (2\kappa_2 - \kappa_1)} \right) \frac{1}{T} + 2\kappa_2 C_2 \frac{1}{T^{1-\kappa_2}}$$

$$+ \frac{2^{1-\kappa_2}C_4}{1 - (2\kappa_2 - \kappa_1)} \frac{1}{T^{1-\kappa_2}} + \frac{C_6}{1 - (2\kappa_2 - \kappa_1)} \frac{1}{T^{2\kappa_2 - \kappa_1}}$$

if $\kappa_1 \in (0, 1)$, $\kappa_2 \in (0, 1)$; and

$$\leq \left( C_1 + \frac{\|1 - 2(2\kappa_2 - \kappa_1)\|C_6}{1 - (2\kappa_2 - \kappa_1)} \right) \frac{1}{T} + \frac{2^{1-\kappa_2}C_4}{1 - (2\kappa_2 - \kappa_1)} \frac{1}{T^{1-\kappa_2}}$$

$$+ \frac{2^{1-\kappa_2}C_3}{1 - (2\kappa_2 - \kappa_1)} \frac{1}{T^{1-\kappa_2}} + 4C_4 \frac{\ln T}{T^{\kappa_2 - 1}}, \text{ if } \kappa_1 \in (0, 1), \kappa_2 = 1$$

in which $C_1$-$C_6$ are defined as in Theorem 1.

**Proof**: See Appendix A.

The following result provides a class of novel convergence rates for a general class of noise decaying rate.

**Corollary 1**: Under conditions of Theorem 1, suppose the sequences $\{\alpha_t\}$ and $\{r_t\}$ in the DSCMD-N method are $\alpha_t = [(1)/(t+1)^{\kappa}]$ and $r_t = [(1)/(t+1)^{\kappa}]$, with $\kappa \in ((1/2), 1)$ and $\kappa \neq (3/4)$. If the constant $C_6$ is taken as

$$C_6 = \max \left\{ C_1 + \frac{\|1 - 2(2\kappa - \frac{1}{2})\|C_6, \sqrt{2}C_2 + 2\sqrt{2}C_3 \right.$$

$$\frac{2^{1-\kappa}C_4}{1 - \kappa} \frac{2^{1-2\kappa}C_3}{2 - \kappa} \frac{C_6}{\sqrt{2} - 2\kappa} \left. \right\}$$

then for any $l \in \mathcal{V}$, $T \geq 3$, the DSCMD-N method achieves an expected rate of $O((1)/(T^{\kappa-1/2}))$ in the following sense:

$$\mathbb{E}[F(\hat{x}_T^l)] - F(x^*) \leq 5C_6/T^{\kappa-1/2}$$ \hfill (19)

in which $C_1$-$C_6$ are defined as in Theorem 1.
Proof: By using Proposition 1 to \( \kappa_1 = 1/2 \) and \( \kappa_2 = \kappa \), it follows that:
\[
\mathbb{E}[F(\tilde{x}_t^2)] - F(x^*) \leq \left( C_1 + \frac{1 - 2(2\kappa - \frac{1}{2})}{1 - (2\kappa - \frac{1}{2})} \right) \frac{1}{T} + \left( \sqrt{2}C_2 + 2\sqrt{2}C_3 \right) \frac{1}{\sqrt{T}} + \frac{2^{1 - \kappa}C_4}{1 - \kappa} \frac{1}{T^x} + \frac{2^{1 - \kappa}C_5}{3 - \kappa} \frac{1}{T^{x - \frac{1}{2}}} + \frac{C_6}{\sqrt{2} - 2\kappa \sqrt{T^{2x - 1}}},
\]
\[
\text{Note that when } \kappa \in ([1/2], 1), \text{ there holds } \frac{1}{T} < \frac{1}{T^x} < \frac{1}{\sqrt{T}} < \frac{1}{T^{x - \frac{1}{2}}}
\]
and
\[
\frac{1}{T^{2x - 1}} < \frac{1}{T^x}.
\]
Therefore, after taking the maximum coefficient \( C_\kappa \) as above, the desired result holds.

The following corollary shows a selection of \( \alpha_t \) such that DSCMD-N achieves the optimal rate in expectation under the case when the network has a communication noise decaying rate \( r_t = [(1)/(t + 1)] \).

Corollary 2: Under conditions of Theorem 1, suppose the sequences \( \{\alpha_t\} \) and \( \{r_t\} \) in the DSCMD-N method are \( \alpha_t = [(1)/(\sqrt{t + 1})] \) and \( \alpha_t = [(1)/(t + 1)] \), \( t = 1, 2, \ldots, T \). If we take \( C = \max\{C_1 + 3C_6, 4C_4, \sqrt{2}C_2 + 2\sqrt{2}C_3, 2\sqrt{2}C_5\} \), then for any \( \delta \leq \kappa \), \( \kappa \leq 3 \), the DSCMD-N method achieves an expected rate of \( O(1/(\sqrt{T}) \) as follows:
\[
\mathbb{E}[F(\tilde{x}_t^2)] - F(x^*) \leq 3C/\sqrt{T}.
\]

Proof: By using Proposition 1 to the case when \( \kappa_1 = 1/2 \) and \( \kappa_2 = 1 \), we have
\[
\mathbb{E}[F(\tilde{x}_t^2)] - F(x^*) \leq \left( C_1 + 3C_6 \right) \frac{1}{T} + 4C_4 \frac{\ln T}{T} + \left( \sqrt{2}C_2 + 2\sqrt{2}C_3 + 2\sqrt{2}C_5 \right) \frac{1}{\sqrt{T}}.
\]

Proposition 2: Under conditions of Theorem 2, let the sequences \( \{\alpha_t\} \) and \( \{r_t\} \) in the DSCMD-N method be \( \alpha_t = [(1)/(\sqrt{t + 1})] \) and \( r_t = [(1)/(t + 1)] \), \( t = 1, 2, \ldots, T \). Then, for any \( \delta < \kappa \), \( \kappa \leq 3 \), we have, for any \( \delta < 0 \), the probability of at least \( 1 - \delta \)
\[
F(\tilde{x}_t^2) - F(x^*) \leq \left( C_1 + 3C_6 \right) \frac{1}{T} + 4C_4 \frac{\ln T}{T} + \left( \sqrt{2}C_2 + 2\sqrt{2}C_3 + 2\sqrt{2}C_5 + 2\sqrt{2}G_D \alpha_{\log(1/\delta)} \right) \frac{1}{\sqrt{T}},
\]
in which \( C_1 \leq C \) are defined as in Theorem 1.

Proof: The proof has the similar procedure with Corollary 1 by using the general bounds for terms of \( \alpha_t \) and \( r_t \). The result is obtained by combining an additional term of \( 2\sqrt{2}G_D \alpha_{\log(1/\delta)} \sqrt{T} \) (this term appears since we consider high probability bound this time).

The high probability optimal rate of \( O(1/\sqrt{T}) \) for DSCMD-N is obtained in the following corollary.

Corollary 3: Under conditions of Proposition 2, for any \( \delta < 0 \), \( \kappa \leq 3 \), we have, for any \( \delta < 0 \), the probability of at least \( 1 - \delta \), and the DSCMD-N method achieves the following rate:
\[
F(\tilde{x}_t^2) - F(x^*) \leq 3C_{\delta}/\sqrt{T}.
\]

Proof: The result follows directly from Proposition 2.

Remark 3: Now, we make a comparison between the results on DSCMD-N in this work and some main existing works in this literature [11, 25]. Reference [25] is a seminal work on distributed optimization over the noisy network. Both of the works [11, 25] consider standard distributed Euclidean projection-based algorithms to minimize the objective function \( \sum_{i=1}^N f_i(x) \) associated with local functions \( f_i, i \in V \). Their approaches rely on a standard Robbins–Monro stepsize summability condition \( \sum_{i=0}^\infty \alpha_t = \infty \) and \( \sum_{i=0}^\infty \alpha_t^2 < \infty \) to ensure the almost sure convergence of \( \{x_t\} \) to the solution set \( \mathcal{X}^* \). In this work, the DSCMD-N method is introduced in a more general setting (composite optimization) when regularization terms are considered. Hence, we are able to handle the optimization problem from different angles by selecting different types of regularizers. Also, the Bregman divergence is utilized instead of the Euclidean projection in [11] and [25]; therefore, the proposed algorithm can better reflect the geometric feature of the underlying decision space when selecting different types of mirror map (distance-generating function) \( \Phi \).

Remark 4: Here, we mention a special case: when we consider regularizer \( \chi_i = 0 \) and mirror map \( \Phi = (1/2)\|\cdot\|^2 \), the algorithm degenerates to [11] if a zeroth-order gradient oracle is used. Moreover, we relax the aforementioned stepsize assumptions (hence, the stepsize \( \alpha_t = [(1)/(t + 1)^{\kappa}] \) with \( \kappa \in (1, 2] \) can be used, this stepsize cannot be considered and used in [11] and [25]) and derive the explicit convergence rate in expectation. On the way to the convergence in expectation, we also relax an assumption of noise \( \{\xi_t\} \) in contrast to [11]. In fact, we do not require the martingale difference condition \( \mathbb{E}[\xi_t^2] = 0 \) to obtain expectation...
convergence results. As an important counterpart of convergence in expectation, high probability bound and rate are also obtained via Azuma–Hoeffding inequality, which enriches the convergence class of distributed optimization methods in this literature. These convergence rates and bounds are new in the noisy network optimization setting.

Remark 5: In contrast to existing works on noisy network optimization, this article also considers composite terms that serves as regularization terms for the composite optimization problem [local regularizer $\chi_i$, $i = 1, 2, \ldots, N$ in problem (1)]. The utilization of the regularization terms makes the method more flexible to present some structure types of the solution of optimization problem. Meanwhile, the structure of problem (1) and DSCMD-N method allows the regularization term $\chi_i$ associated with agent $i$ to be independent of each other. There are several choices of $\chi_i$ and $\eta$ that are often considered to promote different structure types of solutions of the optimization problem. For example, the indicator function of $\mathcal{X}$, the $p$-norm squared function $(1/2)\|x\|_p^2$, $p \in (1, 2]$; sparsity inducing regularizer $\lambda\|x\|_1$, $\lambda > 0$; Lipschitz–product $\lambda\|x\|_\infty$, $\lambda > 0$; entropy function $\sum_{i=1}^n \|x_i\| \log \|x_i\|$; mixed regularizer $[(\lambda_1)/(2)]\|x\|_2^2 + (\lambda_2)|x|_1$, $\lambda_1, \lambda_2 > 0$.

Till now, we observe that all the approximating sequences of convergence results in this article are in weighted average form $\hat{x}_t = (1/T) \sum_{i=1}^T x^*_t$, $i = 1, 2, \ldots, N$. The expectation convergence and high probability convergence result are derived. A question rises that can we present some almost sure convergence results for local sequence $\{\alpha_t\}$ or $\{\chi_t\}$ in distributed composite optimization setting? To this end, we provide following almost sure convergence results for DSCMD-N. In the following, we use $\text{dist}(x, M)$ to denote the distance from a point $x$ to the closed set $M$. Namely, $\text{dist}(x, M) = \inf\{\|x - m\| : m \in M\}$.

Corollary 4: Under conditions of Theorem 1, in the DSCMD-N method, suppose the stepsize sequences $\{\alpha_i\}$ and noise decreasing rate $\{r_{t}\}$ are $\alpha_i = [(1)/(\sqrt{t+1})]$ and $r_{t} = [(1)/(t+1)]$, $t = 1, 2, \ldots, T$. If we take $C = \max\{C_1 + 3C_6, 4C_4, \sqrt{2}/C_2 + 2\sqrt{2}/C_3 + 2\sqrt{2}/C_5\}$, in which $C_1 - C_6$ is as in Theorem 1, then for any $\epsilon \in \mathcal{V}$, $T \geq 3$, for sequence $\{\chi_t\}$ and their average version $\{\chi_t\}$ generated from DSCMD-N method, we have almost surely

$$\lim_{T \to \infty} F(x^*_T) = F(x^*), \quad \lim_{T \to \infty} \text{dist}(x^*_T, x^*) = 0 \ \forall i \in \mathcal{V}$$

and

$$\lim_{T \to \infty} \min_{1 \leq t \leq T} F(x^*_t) = F(x^*)$$

$$\lim_{T \to \infty} \left( \min_{1 \leq t \leq T} \text{dist}(x^*_t, x^*) \right) = 0 \ \forall i \in \mathcal{V}.$$

Proof: See Appendix B.

V. EXPERIMENT PERFORMANCE

In this section, we consider the following regularized distributed estimation problem (e.g., [2]):

$$\min_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N (a_i \|x - b_i\|^2 + \chi_i(x))$$

over a class of noisy sensor network. We consider the domain $\mathcal{X} = \{x \in \mathbb{R}^n : -100 \leq |x_i| \leq 100, i = 1, 2, \ldots, N\}$. The real number $a_i$ and $n$-dimension vector $b$ are randomly generated according to uniform distribution in $[-30, 30]$ and $[-30, 30]^n$ ($n$-product of interval of $[-30, 30]$). We first consider the case when the regularization term is chosen as $\chi_i(x) = \lambda_i \|x\|_1$, $i = 1, 2, \ldots, N$ and also conduct the experiment with the well-known 1-norm sparsity inducing regularizer $\chi_i(x) = \lambda \|x\|_1$, $i = 1, 2, \ldots, N$. We apply the DSCMD-N to solve the above problem and show their performance under different selection of regularization parameters $\lambda$. In DSCMD-N, we take the Bregman divergence as $D_{\Phi}(x, y) = (1/2)\|x - y\|^2$. We consider the situation that the gradient noise and communication noise are all generated independent and identically distributed from the normal distribution $\mathcal{N}(0, I_{1 \times n})$, and the noise decaying rate of the noise source satisfies $r_t = [(1)/(t+1)]$. The network size is $N = 30$, and the dimension of the decision variable $x$ is $n = 10$. Then, for $\lambda = 0.2, 0.2$, if we use the stepsize $\alpha_i = [(1)/(\sqrt{t+1})]$ in DSCMD-N, when $\chi_i(x) = \lambda_i \|x\|_1$, $i = 1, 2, \ldots, N$, the experiments on relative error $\|\text{dist}(x^*_T, F(x^*))\|/\|F(x^*)\|$ for DSCMD-N are performed as in Figs. 1 and 2. When $\chi_i(x) = \lambda \|x\|_1$, $i = 1, 2, \ldots, N$, the experiments on the relative error $\|\text{dist}(x^*_T, F(x^*))\|/\|F(x^*)\|$ for DSCMD-N are performed as in Figs. 3 and 4. In the experiments, for DSCMD-N, the convergence behaviors of three randomly chosen nodes from the network are presented. In each figure, there is also a green line reflecting the average convergence performance of the algorithm, and the average result is based on an average 30 times of tests. In all of the first four figures, satisfactory convergence behaviors of DSCMD-N are witnessed from the experiments. The accuracies of the relative errors at iteration step $T = 2000$ are all around $10^{-3}$. Also, when the experiment is conducted by using 1-norm regularizer, the global performance in Figs. 3 and 4 is shown to be better than Figs. 1 and 2 where the 2-norm square regularizer is used.

Next, we conduct experiments to verify some noise influence on the convergence of DSCMD-N. In these experiments,
we provide comparisons in different aspects. These include: 1) the comparison between noise-involved setting and noise-free setting and 2) the comparison on convergence with different noise decaying rates. To verify the theoretical results, we conduct two groups of experiments. The first group of experiment is conducted with noise variables $\xi$ generated from $\mathcal{N}(20, 10I_{10\times 10})$, and the noise variables $\xi_j^t$ in the second group are generated from $\mathcal{N}(25, 20I_{10\times 10})$. The regularization term $\chi(x) = \|x\|_1$ and stepsize $\alpha_0 = [(1)/(\sqrt{t+1})]$ are used in both experiments. In both experiments, the green line denotes the experiments in which the noise is not considered. The other three lines denote the experiments with different noise decaying rates $r_1 = [(1)/(t+1)^{2}]$, $r_2 = 1, 0.95, 0.9$. It can be obviously witnessed from both experiments that the convergence in the noise-involved case is always slower than the noise-free case (green line). Although the noises are generated from different means and variances ($\mathcal{N}(20, 10I_{10\times 10})$ and $\mathcal{N}(25, 20I_{10\times 10})$) in our experiments, this performance has been observed obviously in both Figs. 5 and 6. This phenomenon coincides with the practical setting and reflects the influence of communication noise on convergence. Meanwhile, for different decaying order $\kappa$, it can be seen that when $\kappa$ is smaller, the convergence is slower. This fact directly reflects the influence of noise decaying rates on the convergence.

This nice experiment performance verifies the main theoretical results of Corollaries 1 and 2 in this article.

VI. CONCLUSION

This article has studied a class of noisy network optimization problems. One distributed stochastic composite optimization problem over the noisy network is considered. Based on the Bregman non-Euclidean projection scheme, a new method DSCMD-N is presented to solve them, respectively. Convergence of the methods is systematically studied. New convergence rates are obtained in several different situations under different detailed discussions on stepsize $[\alpha_t]$ and communication noise decreasing rate $[r_t]$. These new convergence results include expectation convergence, high probability convergence, and almost sure convergence. These results enrich the exploration in noisy network optimization. The rates for expectation convergence and high probability convergence are first derived in the literature. Since we have considered randomness on both network links and gradients, the potential value of the methods is obvious in stochastic circumstances. The experiments verify the theoretical results in this article.

There are some possible extensions of the studies in this articles. For example, the current results may be extended to online distributed optimization settings. It would be interesting and challenging to establish the same type of results in this article after relaxing several conditions on the network. It would also be interesting to consider distributed nonconvex optimization over a noisy network. It would be meaningful to develop the results in this article to the optimization problem with nonuniform constraints and nonuniform stepizes.

APPENDIX A

PROOF OF PROPOSITION 1

Proof: Note that, under the conditions that $0 < \kappa_1 < \kappa_2 < 1$ and $2\kappa_2 - \kappa_1 \neq 1$, $[1/(T\alpha_T)] = [(T+1)^{\kappa_1}]/(T) \leq [(2T)^{\kappa_1}]/(T) = [(2^{\kappa_1})/(T^{1-\kappa_1})]$, $T \geq 1$. The regularizer term $\chi(x) = \|x\|_1$ and stepsize $\alpha_0 = [(1)/(\sqrt{t+1})]$ are used in both experiments. In both experiments, the green line denotes the experiments in which the noise is not considered. The other three lines denote the experiments with different noise decaying rates $r_1 = [(1)/(t+1)^{2}]$, $r_2 = 1, 0.95, 0.9$. It can be obviously witnessed from both experiments that the convergence in the noise-involved case is always slower than the noise-free case (green line). Although the noises are generated from different means and variances ($\mathcal{N}(20, 10I_{10\times 10})$ and $\mathcal{N}(25, 20I_{10\times 10})$) in our experiments, this performance has been observed obviously in both Figs. 5 and 6. This phenomenon coincides with the practical setting and reflects the influence of communication noise on convergence. Meanwhile, for different decaying order $\kappa$, it can be seen that when $\kappa$ is smaller, the convergence is slower. This fact directly reflects the influence of noise decaying rates on the convergence.

This nice experiment performance verifies the main theoretical results of Corollaries 1 and 2 in this article.

VI. CONCLUSION

This article has studied a class of noisy network optimization problems. One distributed stochastic composite optimization problem over the noisy network is considered. Based on the Bregman non-Euclidean projection scheme, a new method DSCMD-N is presented to solve them, respectively. Convergence of the methods is systematically studied. New convergence rates are obtained in several different situations under different detailed discussions on stepsize $[\alpha_t]$ and communication noise decreasing rate $[r_t]$. These new convergence results include expectation convergence, high probability convergence, and almost sure convergence. These results enrich the exploration in noisy network optimization. The rates for expectation convergence and high probability convergence are first derived in the literature. Since we have considered randomness on both network links and gradients, the potential value of the methods is obvious in stochastic circumstances. The experiments verify the theoretical results in this article.

There are some possible extensions of the studies in this articles. For example, the current results may be extended to online distributed optimization settings. It would be interesting and challenging to establish the same type of results in this article after relaxing several conditions on the network. It would also be interesting to consider distributed nonconvex optimization over a noisy network. It would be meaningful to develop the results in this article to the optimization problem with nonuniform constraints and nonuniform stepizes.
(1/T)(2 + \int_0^T \int_0^t [(dt)/(\sqrt{t^2 - x^2})]) = (1/T)(2 + \int_0^T \int_0^t 1) = (1/T)(2 + \sum_{i=1}^{N}(1/(t+1) ≤ 1 + \int_0^t 1) = 1 + \ln(T+1) ≤ 1 + \ln 2T ≤ 2 \ln 2 = 4 \ln T, T ≥ 3. (1/T) \sum_{i=1}^{N} (1/(t+1)^{1/2}) = \int_0^T \int_0^t 1) = (1/T) \sum_{i=1}^{N} (1/(t+1)^{1/2}) ≥ 1 + \int_0^T \int_0^t (1/(t+1)^{1/2}) = (2 - \kappa (1/k)) (1/T), and the second argument is obtained after substituting these estimates into Theorem 1 again.

APPENDIX B
PROOF OF COROLLARY 4

Proof: According to Corollary 2, we have E[∥x^T - \hat{x}^T∥^2] ≤ C^3 / \sqrt{T}, i ∈ V. This implies lim_{T→∞} E[∥x^T - \hat{x}^T∥^2] = 0. Since F(\hat{x}^T) - F(x^T) is non-negative for all T ≥ 0, by applying the Fatou lemma, we arrive at lim_{T→∞} E[∥x^T - \hat{x}^T∥^2] = 0, a.s.

Due to the convexity of F and the fact that \hat{x}^T is a convex combination of x_1, x_2, ..., x_T, we have min_{1≤T} F(x_1) ≤ F(\hat{x}^T), i ∈ V. Then, it follows that 0 ≤ min_{1≤T} F(x_1) ≤ F(x^T) ≤ F(\hat{x}^T) - F(x^T), i ∈ V. After taking expectation on both sides, we have

0 ≤ E[\min_{1≤T} F(x_1) - F(x^T)] ≤ E[∥x^T - \hat{x}^T∥^2] = 0, i ∈ V.

Taking limit on both sides of the above inequality and using the Squeeze theorem, we have lim_{T→∞} E[∥x^T - \hat{x}^T∥^2] = 0. Using the Fatou lemma again, we have

lim_{T→∞} E[∥x^T - \hat{x}^T∥^2] = 0. Since min_{1≤T} F(x_1) is a lower bounded nonincreasing sequence in T, hence, min_{1≤T} F(x_1) exists and lim_{T→∞} min_{1≤T} F(x_1) = lim_{T→∞} min_{1≤T} F(x_1) = min_{1≤T} F(x_1) = F(x^T), i ∈ V. Then, using a similar argument of obtaining (21), we arrive at lim_{T→∞} min_{1≤T} dist(x_1, x^T) = 0, i ∈ V.

REFERENCES

[1] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," IEEE Trans. Autom. Control, vol. 54, no. 1, pp. 48–61, Jan. 2009.
[2] A. Nedic and A. Olshewsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," IEEE Trans. Autom. Control, vol. 61, no. 12, pp. 3936–3947, Dec. 2016.
[3] A. Nedic and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," SIAM J. Optim., vol. 24, no. 1, pp. 84–107, 2014.
[4] A. Khatib, B. Touri, and T. Basar, "Consensus in the presence of an adversary," in Proc. 3rd IFAC Workshop Distrib. Estimation Control Netw. Syst., vol. 45, 2012, pp. 276–281.
[5] A. Akhavan, M. Pontil, and A. B. Tsybakov, "Distributed zero-order optimization under adversarial noise," 2021, arXiv:2102.01121.
[6] B. Bercu, B. Delyon, and E. Rio, Concentration Inequalities for Sums and Martingales. Cham, Switzerland: Springer, 2015.
[7] S. P. Lalley, "Concentration inequalities," Lecture Notes, Univ. Chicago, Chicago, IL, USA, 2013.
[8] B. Touri and A. Nedic, "Distributed consensus over network with noisy links," in Proc. 12th Inf. Conf. Inf. Fusion, 2009, pp. 146–154.
[9] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," IEEE Trans. Autom. Control, vol. 62, no. 8, pp. 3986–3992, Aug. 2017.
[10] K. Cai and H. Ishii, "Average consensus on general strongly connected digraphs," Automatica, vol. 48, no. 11, pp. 2750–2761, 2012.
[11] D. Wang, J. Zhou, Z. Wang, and W. Wang, "Random gradient-free optimization for multiautonomous systems with communication noises under a time-varying weight balanced digraph," IEEE Trans. Syst. Man., Cybern., Syst., vol. 50, no. 1, pp. 281–289, Jan. 2020.
[12] D. P. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," IEEE Trans. Autom. Control, vol. 52, no. 12, pp. 2254–2269, Dec. 2017.
[13] D. Yuan, D. W. C. Ho, and Y. Hong, "On convergence rate of distributed stochastic gradient algorithm for convex optimization with inequality constraints," SIAM J. Control Optim., vol. 54, no. 5, pp. 2872–2892, 2016.
[14] D. Yuan, Y. Hong, D. W. C. Ho, and S. Xu, "Distributed mirror descent for online composite optimization," IEEE Trans. Autom. Control, vol. 66, no. 2, pp. 714–729, Feb. 2021.
[15] D. Yuan, D. W. C. Ho, and G.-P. Jiang, "An adaptive primal-dual subgradient algorithm for online distributed constrained optimization," IEEE Trans. Cybern., vol. 48, no. 11, pp. 3045–3055, Nov. 2018.
[16] D. Yuan, D. W. C. Ho, and S. Xu, "Regularized primal-dual subgradient method for distributed constrained optimization," IEEE Trans. Cybern., vol. 46, no. 9, pp. 2100–2118, Sep. 2016.
[17] D. G. Luenberger, Microeconomic Theory. Singapore: McGraw-Hill, 1995.
[18] H. Li, B. Jin, and W. Yan, "Distributed model predictive control for linear systems under communication noise: Algorithm, theory and implementation," Automatica, vol. 125, Mar. 2021, Art. no. 109422.
[19] J. Li, C. Gu, Z. Wu, and T. Huang, "Online learning algorithm for distributed convex optimization with time-varying coupled constraints and bandit feedback," IEEE Trans. Cybern., early access, May 22, 2020, doi: 10.1109/TCYB.2020.2990796.
[20] J. C. Duchi, A. Alekh, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," IEEE Trans. Autom. Control, vol. 57, no. 3, pp. 592–606, Mar. 2012.
[21] J. Li, G. Chen, Z. Dong, and Z. Wu, "Distributed mirror descent method for multi-agent optimization with delay," Neurocomputing, vol. 177, pp. 643–650, Feb. 2016.
[22] J. Lu and D. Ho, "Stabilization of complex dynamical networks with noise disturbance under performance constraint," Nonlinear Anal. Real World Appl., vol. 12, no. 4, pp. 1974–1984, 2011.
[23] J. Hu, Z. Wang, and G.-P. Liu, "Delay compensation-based state estimation for time-varying complex networks with incomplete observations and dynamical bias," IEEE Trans. Cybern., early access, Jan. 15, 2021, doi: 10.1109/TCYB.2020.3034323.
[24] L. Feng, J. Cao, and L. Liu, "Robust analysis of discrete time noises for stochastic systems and application in neural networks," Int. J. Control, vol. 93, no. 12, pp. 2908–2921, 2020.
[25] K. Srivastava, A. Nedic, and D. M. Stipanovic, "Distributed constrained optimization over noisy networks," in Proc. 49th IEEE Conf. Decis. Control (CDC), 2010, pp. 1945–1950.
[26] M. G. Rabbat and R. D. Nowak, "Decentralized source localization and tracking [wireless sensor networks]," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., vol. 5, 2004, p. 921.
[27] J. He, M. Zhou, P. Cheng, L. Shi, and J. Chen, "Consensus under directed graphs," in Proc. 54th IEEE Conf. Decision Control (CDC), 2015, pp. 3045–3052, Sep. 2015.
[28] P. Yi, Y. Hong, and F. Liu, "Distributed gradient algorithm for constrained optimization with application to load sharing in power systems," Int. J. Control, vol. 83, pp. 55–72, Sep. 2010.
[29] S. S. Ram, A. Nedic, and V. V. Vreevall, "Distributed stochastic subgradient projection algorithms for convex optimization," J. Optim. Theory Appl., vol. 147, no. 3, pp. 516–545, 2010.
[30] S. Lee, S. A. Nedić, and M. Raginsky, “Stochastic dual averaging for decentralized online optimization on time-varying communication graphs,” IEEE Trans. Autom. Control, vol. 62, no. 12, pp. 6407–6414, Dec. 2017.

[31] S. Liu, P.-Y. Chen, and A. O. Hero, “Accelerated distributed dual averaging over evolving networks of growing connectivity,” IEEE Trans. Signal Process., vol. 66, no. 7, pp. 1845–1859, Apr. 2018.

[32] S. Liu, Z. Wang, B. Shen, and G. Wei, “Partial-neurons-based state estimation for delayed neural networks with state-dependent noises under redundant channels,” Inf. Sci., vol. 547, pp. 931–944, Feb. 2021.

[33] T.-H. Chang, A. Nedić, and A. Scaglione, “Distributed constrained optimization by consensus-based primal-dual perturbation method,” IEEE Trans. Autom. Control, vol. 59, no. 6, pp. 1524–1538, Jun. 2014.

[34] C.-X. Shi and G.-H. Yang, “Distributed composite optimization over relay-assisted networks,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 51, no. 10, pp. 6587–6598, Oct. 2021, doi: 10.1109/TSMC.2019.2963452.

[35] W. Zhang, P. Zhao, W. Zhu, S. C. Hoi, and T. Zhang, “Projection-free distributed online learning in networks,” in Proc. 34th Int. Conf. Mach. Learn., vol. 70, 2017, pp. 4054–4062.

[36] J. Gao, P. Zhu, W. Xiong, J. Cao, and L. Zhang, “Asymptotic synchronization for stochastic memristor-based neural networks with noise disturbance,” J. Franklin Inst., vol. 353, no. 13, pp. 3271–3289, 2016.

[37] X. Zeng, P. Yi, Y. Hong, and L. Xie, “Distributed continuous-time algorithms for nonsmooth extended monotropic optimization problems,” SIAM J. Control Optim., vol. 56, no. 6, pp. 3973–3993, 2018.

[38] X. Wang, Z. Wang, M. Wu, and X. Liu, “State estimation for discrete time-delayed genetic regulatory networks with stochastic noises under the round-robin protocols,” IEEE Trans. Nanobiosci., vol. 17, no. 2, pp. 145–154, Apr. 2018.

[39] Y. Tang and N. Li, “Distributed zero-order algorithms for nonconvex multi-agent optimization,” in Proc. 57th Ann. Allerton Conf. Commun. Control Comput. (Allerton), 2019, pp. 781–786.

[40] Y. Wang, W. Zhao, Y. Hong, and M. Zaman, “Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks,” SIAM J. Control Optim., vol. 57, no. 4, pp. 2821–2842, 2019.

[41] Y. Yuan, Z. Wang, P. Zhang, and H. Dong, “Nonfragile near-optimal control of stochastic time-varying multiagent systems with control-and state-dependent noises,” IEEE Trans. Cybern., vol. 49, no. 7, pp. 2605–2617, Jul. 2019.

[42] N. Chatzizisis and M. M. Zavlanos, “A distributed algorithm for convex constrained optimization under noise,” IEEE Trans. Autom. Control, vol. 61, no. 9, pp. 2496–2511, Sep. 2016.

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

[44] B. Gharesifard and J. Cortes, “Distributed strategies for generating weight-balanced and doubly stochastic digraphs,” Eur. J. Control, vol. 18, no. 6, pp. 539–557, 2012.

[45] S. Ghadimi, G. Lan, and H. Zhang, “Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization,” Math. Program., vol. 155, nos. 1–2, pp. 267–305, 2016.

Zhan Yu received the Ph.D. degree in applied mathematics from the City University of Hong Kong, Hong Kong, in 2021.

He is currently a Postdoctoral Fellow with the School of Data Science, City University of Hong Kong. His current research interests include learning theory, optimization theory, and applied harmonic analysis.

Jie Liu received the B.S. degree from the Hefei University of Technology, Hefei, China, in 2020. He is currently pursuing the Ph.D. degree with the Department of Mathematics, City University of Hong Kong, Hong Kong.

His current research interests include distributed optimization, machine learning, and multiagent systems.

Daniel W. C. Ho (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees in mathematics from the University of Salford, Salford, U.K., in 1980, 1982, and 1986, respectively.

From 1985 to 1988, he was a Research Fellow with the Industrial Control Unit, University of Strathclyde, Glasgow, U.K. In 1989, he joined the City University of Hong Kong, Hong Kong, where he is currently the Chair Professor of Applied Mathematics, and an Associate Dean with the College of Science. He has over 250 publications in scientific journals. His current research interests include control and estimation theory, complex dynamical distributed networks, multiagent systems, and stochastic systems.

Dr. Ho was a recipient of the Chang Jiang Chair Professor Awarded by the Ministry of Education, China, in 2012 and the ISI Highly Cited Researchers Award in Engineering by Clarivate Analytics from 2014 to 2021. He has been on the Editorial Board of a number of journals, including the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IET Control Theory and Its Applications, the Journal of the Franklin Institute, and the Asian Journal of Control.

Deming Yuan (Member, IEEE) received the B.Sc. degree in electrical engineering and automation and the Ph.D. degree in control science and engineering from the Nanjing University of Science and Technology, Nanjing, China, in 2007 and 2012, respectively.

He is currently a Professor with the School of Automation, Nanjing University of Science and Technology. His current research interests include algorithms for distributed optimization, control, and machine learning.

Prof. Yuan received the Australia Award Endeavour Research Fellowship to further his research on distributed optimization, and was hosted by the Australian National University in 2018. He has been on the Editorial Board of Transactions of the Institute of Measurement and Control.