Distributed Stochastic Optimization over Time-Varying Noisy Network*

Zhan Yu, Daniel W. C. Ho, Deming Yuan

Abstract

This paper is concerned with distributed stochastic multi-agent optimization problem over a class of time-varying network with slowly decreasing communication noise effects. This paper considers the problem in composite optimization setting which is more general in noisy network optimization. It is noteworthy that existing methods for noisy network optimization are Euclidean projection based. We present two related different classes of non-Euclidean methods and investigate their convergence behavior. One is distributed stochastic composite mirror descent type method (DSCMD-N) which provides a more general algorithm framework than former works in this literature. As a counterpart, we also consider a composite dual averaging type method (DSCDA-N) for noisy network optimization. Some main error bounds for DSCMD-N and DSCDA-N are obtained. The trade-off among stepsizes, noise decreasing rates, convergence rates of algorithm is analyzed in detail. 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 show that an optimal rate of $O(1/\sqrt{T})$ in nonsmooth convex optimization can be obtained for proposed methods under appropriate communication noise condition. Moreover, convergence rates in different orders are comprehensively derived in both expectation convergence and high probability convergence sense.

1 Introduction

In recent years, the problem of minimizing a sum of locally known convex objective functions that are distributed over a network is studied extensively (see [1, 2, 5, 8, 9, 10, 16, 17, 19, 20, 22, 23, 24, 29, 30, 31, 32, 33, 34, 35, 36]). Such problem arises in a variety of application domains, such as localization in sensor networks (see e.g. [18, 36]), smart grid (see e.g. [28]), utility maximization (see e.g. [9]), allocation of resources in microeconomics (see e.g. [12]). Earlier fundamental works focus on the unconstrained optimization of a smooth function known to several agents while distributing processing of components of the decision variables (see e.g. [1, 4, 15]). Very recent researches have turned attention to problems in which each agent has its own associated (perhaps nonsmooth) objective function (see e.g. [1, 5, 8, 10, 23]). For solving this kind of problem, a variety of methods have emerged recently. In these methods, 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 progress of distributed optimization in numerous aspects.

*This work was partially supported by Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11202819, CityU 11200717). Zhan Yu and Daniel W. C. Ho are with the Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (e-mail: zhanyu2-c@my.cityu.edu.hk, mathyuzhan@gmail.com; madaniel@cityu.edu.hk), Deming Yuan is with the School of Automation, Nanjing University of Science and Technology, Nanjing, China (e-mail: dmyuan1012@gmail.com)
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 unconstraint set of decision variable. Consequently, studies on distributed stochastic subgradient method appears \(^{(23)}\). 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 same period, the work \(^{(5)}\) provides a novel distributed method to better capture the direct structure of the network topology, the convergence relies on a core matrix analysis result in \(^{(6)}\). In a different way, the work in \(^{(16)}\) presents a (stochastic) dual averaging-based method, the method is based on maintaining and forming weighted averages of subgradients throughout the network. In what follows, several works which improve \(^{(16)}\) appear (e.g. \(^{(25, 26)}\)). Very recently, distributed gradient tracking method is developed as a powerful tool for solving smooth and strongly convex optimization problem (e.g. \(^{(27)}\)). On the other hand, several works also turn to investigate the case when local objective functions are nonconvex (e.g. \(^{(20, 29, 33)}\)). Mirror descent technique has been utilized in distributed optimization domain recently (e.g. \(^{(24)}\)), 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. \(^{(11, 13, 24, 31)}\)), 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. Since the realization of consensus is an essentially necessary condition for convergence of these methods.

In centralized and distributed optimization, convergence rate is an important object. Providing convergence rate for different classes of distributed algorithms is a hot topic. As we know, a rate of \(O(\frac{1}{\sqrt{T}})\) is best known in centralized nonsmooth convex optimization problem. The rate has been obtained in centralized setting in work like \(^{(3)}\). To design method which recovers to this rate is also meaningful for nonsmooth distributed convex optimization problem. The reason is that, with same convergence rate, distributed methods has much more advantages and flexibilities over centralized methods on handling network optimization issues. For example, overburdening nodes, node failures, too large data sets. Hence, in aforementioned distributed optimization circumstance, it would be good if a rate of \(O(\frac{1}{\sqrt{T}})\) can be obtained by designing distributed optimization method via information communication among nodes.

The main goal of this paper is to study distributed optimization problems by addressing following considerations: (i) Since uncertain stochastic disturbances always exist in real life environment, it is desirable to consider the topic of solving distributed optimization problem over network in which communication noises exist among nodes by designing stochastic gradient methods. (ii) The existing optimization methods over noisy network are Euclidean gradient projection based (see e.g. \(^{(8, 17)}\)), is it possible to consider some more general frameworks and provide general methods to solve them in some class of noisy network? (iii) 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 like \(^{(5, 17)}\). Explicit description of convergence rate is still absent in this literature. Is it possible to derive convergence rate results in this setting under relaxed stepsize condition? Also, the optimization methods are very poorly explored over noisy network, can we develop some methods in the literature? To this end, we consider multi-agent composite optimization problems over time-varying noisy network in this paper. Specifically, we analyze two related forms of problems that differ essentially on the status of the regularization term. One has the form

\[
\text{minimize}_{x \in \mathcal{X}} \quad F(x) = \sum_{i=1}^{N} F_i(x) := \sum_{i=1}^{N} [f_i(x) + \chi_i(x)],
\]  

(1.1)
where $\mathcal{X}$ is non-empty 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 $i$. The other related problem has the form

$$\min_{x \in \mathcal{X}} F(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) + \eta(x),$$

in which $\eta$ is a global regularization function. The objective of the agents is to cooperatively solve the above problems. In recent years, there are a few works on distributed methods treating with the aforementioned problems with composite framework (e.g. [11], [21], [30]). [11] mainly focuses on online optimization and develops an online two-point bandit feedback mirror decent based method. [21] considers the primal-dual type method in composite setting. [30] analyzes the decentralized proximal gradient type algorithm in composite setting. From a different perspective, this work considers the network with communication noise and attempts to develop distributed optimization methods that are suitable to noisy network. Meanwhile, we study the convergence of the distributed composite optimization methods in noisy circumstance.

In this paper, inspired by stochastic approximation theory in [8, 16, 17], we develop two related classes of stochastic optimization methods for solving above two types of composite optimization problems. The problems are considered over a class of time-varying network that has slowly decreasing communication noise effects among nodes in information transmission process. We propose distributed stochastic composite mirror descent (DSCMD-N method) for Problem (1.1), distributed stochastic composite dual averaging (composite DSCDA-N method) for Problem (1.2). The convergence results are analyzed in detail. Specifically, we are interested in the convergence behavior of the methods under different selections of stepizes, the expected convergence bound and high probability convergence bound are established respectively. In what follows, the discussion on selection of stepsizes and corresponding convergence rates are provided. Note that, by taking composite regularization function into consideration, this work also extends the former works in same literature to a more general setting. For the proposed DSCMD-N, by implementing Bregman divergence instead of former Euclidean distance in works like [8], [17], the DSCMD-N method extends the projection structure of these methods to more general setting. Explicit rate $O\left(\frac{1}{\sqrt{T}}\right)$ result is obtained for expected function error for DSCMD-N method under appropriate selection of stepsize. For dual averaging type method (known as “lazy” mirror decent method), we also propose a DSCDA-N method for noisy setting, the convergence behavior is described by convergence bound in terms of $\alpha_t$ (stepsize for stochastic gradient) and $r_t$ (decreasing rate for noise vector) and some cross terms of them. The error bound obtained in this work describes some intrinsic trade-off between the stepsizes $\{\alpha_t\}$ and the slowly decreasing rate $\{r_t\}$ of communication noise.

The technical contributions of this paper can be summarized as follows: (1) composite optimization is investigated in time-varying noisy network optimization. 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. In contrast to former work in distributed composite optimization literature (e.g. [30]), this work allows the objective functions to be nonsmooth. Also, the decision space $\mathcal{X}$ does not need to be the whole space (unconstrained set), a very special case for optimization issue. This fact makes the proposed methods more flexible to handle optimization problems when tough smoothness conditions are added on objective functions and constraint set. Also, the methods are convenient for a class of optimization over time-varying network, in contrast to static network.

(2) Two new methods are presented for optimization over a class of noisy network. Existing works in same literature like [8], [17] are all Euclidean projection based. By presenting DSCMD-N method, we extend these former works to a more general setting in the proposed network model. Since Bregman divergence is utilized, the underlying geometry structure of 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 a special case, when we take distance generating function as $\frac{1}{2}\| \cdot \|^2$ and consider the regularization term $\chi_i = 0$, the DSCMD-N method degenerates to the Euclidean projection-based method that former works have discussed on noisy network optimization. Also, in order to solve problem (1.2) when a global regularization term is considered, we provide the other related different class of method DSCDA-N to solve it. This method stands as a counterpart of DSCMD-N. Also, to the best of our knowledge, this is the first dual averaging type method that is fit for optimization over noisy network.

(3) The convergence behavior for DSCMD-N and DSCDA-N are investigated comprehensively. We obtain two types of convergence bounds for expected error: expected bound and high probability bound. The bounds are in terms of some cross terms consisting of stochastic gradient stepsizes $\{\alpha_t\}$ and communication noise decreasing rate $\{r_t\}$. The stepsize selection is comprehensively analyzed under effects of different network noise decreasing rates. The corresponding convergence rates are obtained. All these rates are first achieved in the setting of optimization over noisy network. We also show that the optimal expected rate and high probability of $O(\frac{1}{\sqrt{t}})$ can be obtained under some conditions on stepizes and noise decreasing rate. The almost sure convergence type results are derived for the local sequence.

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

2 Problem setting and preliminaries

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

The objective of the paper is to cooperatively solve the composite optimization Problem (1.1) and Problem (1.2) through communication among the agents of a multi-agent system described by graph $\mathcal{G}_t$. The decision space $\mathcal{X} \subseteq \mathbb{R}^n$ for the state variable $x$ is a convex and closed set. 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 paper is denoted by $\mathcal{X}^\star$ with optimal value $f(x^\star)$ for any $x^\star \in \mathcal{X}^\star$. The following standard assumption is made on the graph $\mathcal{G}_t$.

Assumption 1. The communication matrix $P_t$ is doubly stochastic. i.e., $\sum_{i=1}^N [P_t]_{ij} = 1$ and $\sum_{j=1}^N [P_t]_{ij} = 1$ for any $i, j \in \mathcal{V}$. There exists some positive integer $B$ such that the graph $(\mathcal{V}, \bigcup_{s=0}^{B} \mathcal{E}_s)$ is strongly connected for every $s \geq 0$. There exists a scalar $0 < \theta < 1$ such that $[P_t]_{ii} \geq \theta$ for all $i \in \mathcal{V}$ and $t$, and $[P_t]_{ij} \geq \theta$ if $(j, i) \in \mathcal{E}_t$.

In this paper, $P(t, s) = P_t P(t-1) \cdots P_s$ is used to denote 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 multi-agent time-varying network.
Lemma 1. Under Assumption 4, for all $i, j \in V$ and $t, s$ satisfying $t \geq s \geq 1$, we have $|P(t, s)_{ij} - \frac{1}{t^{s}}| \leq \omega \gamma^{t-s}$, in which $\omega = (1 - \frac{\phi}{4N^2})^{-2}$ and $\gamma = (1 - \frac{\phi}{4N^2})^{\frac{1}{2}}$.

The methods in this paper is first-order stochastic approximation based. We make some assumptions on subgradients of the objective functions. We assume that the nodes can only compute the noisy subgradients of its corresponding objective functions. In what follows, we use $\mathcal{F}_t$ to denote the $\sigma$-algebra of the history up to time $t$. In this paper, we assume that all random processes are adapted to the filtration $\mathcal{F}_t$.

Assumption 2. At any point $x \in X$, let the stochastic subgradient $\tilde{g}_i(x)$ be such that $\mathbb{E}[\tilde{g}_i(x)|\mathcal{F}_{t-1}] = g_i(x) \in \partial f_i(x)$ and $\mathbb{E}[\|\tilde{g}_i(x)\|^2|\mathcal{F}_{t-1}] \leq G_i^2$.

This paper focuses on the noisy network optimization. We assume that the time-varying noise exists over the network. The noise needs to be considered in information communication process of state variables of agents. In this model, denote the communication noise between node $i$ and node $j$ at instance $t$ by $\{r_t \xi_{ij}^t\}$ with a slowly decreasing communication noise rate $\{r_t\}$ that is not "faster" than $O(1/t)$. An introduction of $\{r_t\}$ is also motivated by some denoise engineering over network in real world. $\{r_t\}$ can be treated as a class of denoise mechanism on the links to face network distortion in some settings. In modern large-scale information transportation network, denoise technique has deep influence on the quality of the information among nodes. Some digital signal processing methods such as wavelet transform theory and digital filtering technique have been widely used for information communication denoising over network. On the other hand, in a different literature, slowly decreasing noise structure has been also considered in problems like global optimization of nonlinear stochastic systems (e.g. [35]) and minimization of computational energy function over Hopfield neural network (e.g. [33]). Hence it would be interesting to investigate some effects of decreasing communication noise on the convergence of distributed optimization methods. Now we present the following basic assumption on communication noise.

Assumption 3. At any time instance $t$, the noise on link $(i, j)$ is independent of the noise on link $(i', j')$ for $i \neq i', j \neq j'$. The communication noise $\{r_t \xi_{ij}^t\}$, $i, j \in V$ over the time-varying network is a random sequence with $\mathbb{E}[\|\xi_{ij}^t\|^2|\mathcal{F}_{t-1}] \leq \nu$. The slowly decreasing communication noise index $r_t$ has a decreasing order slower than $O(1/t)$ which will be discussed in detail later.

The following Azuma-Hoeffding lemma is needed to derive high probability bound and rate later.

Lemma 2. Let $\{X_t\}$ be a martingale difference sequence satisfying $|X_t| \leq \tau_t$, then for any $\epsilon > 0$, $\text{Prob}(\sum_{t=1}^{T} X_t \geq \epsilon) \leq \exp \left( - \frac{\epsilon^2}{2 \sum_{t=1}^{T} \tau_t^2} \right)$.

In 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 $\Phi : K \rightarrow \mathbb{R}$ to capture the geometric properties of the decision variables from some Banach space $K$. In this paper, we will consider $K = \mathbb{R}^n$ endowed with a norm $\| \cdot \|$ which may be a non-Euclidean norm, allowing us to capture 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) $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}$, 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 3, for the Bregman divergence, we need the following assumption.

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 $\{b_i\}_{i=1}^{N}$ in $\mathbb{R}^n$, the Bregman divergence satisfies the separate convexity in the following sense: $D_\Phi(a, \sum_{i=1}^{N} \nu_i b_i) \leq \sum_{i=1}^{N} \nu_i D_\Phi(a, b_i)$, $\nu_i \in [0, 1]$ and $\sum_{i=1}^{N} \nu_i = 1$. 
3 Algorithms: main convergence results

3.1 Distributed stochastic mirror descent for composite optimization over noisy network

In this section, we consider problem (1.1), minimizing $F(x) = \sum_{i=1}^{N} f_i(x) := \sum_{i=1}^{N} [f_i(x) + \chi_i(x)]$ over 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 $x_i^t$ evolves as follows

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

(3.1)

$$
x_i^{t+1} = \arg \min_{x \in \mathcal{X}} \{ g_i^t(x) + \frac{1}{\alpha_t} D_{\phi}(x, y_i^t) + \chi_i(x) \},
$$

(3.2)

where $[P^t]_{ij}$, $i, j \in \mathcal{V}$ denotes the elements of communication weight matrix $P^t$ satisfying assumptions in Assumption 2. 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 estimate (noisy observation). 3.1 describes the noisy information communication process between $i$ and its neighbors. Then, query the stochastic subgradient oracle at $y_i^t$ to get a stochastic subgradient $\tilde{g}_i^t := \tilde{g}_i(y_i^t)$, such that $E[\tilde{g}_i^t | F_{t-1}] = g_i(y_i^t) \in \partial f_i(y_i^t)$ is a subgradient of $f_i$ at $y_i^t$. In (3.2), we perform a Bregman projection for variable $y_i$ to decision space $\mathcal{X}$ to get variable $x_i^{t+1}$. A composite mirror descent scheme is considered in this Bregman projection with stepsize $\alpha_t$ and composite term $\chi_i(x)$. We remark that the 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 $g_{\chi_i}$. In this section $\partial \chi_i(x)$ is used to denote the subdifferential set of $\chi_i$ at $x \in \mathcal{X}$. denote $G_{\chi_i} = \sup_{\|g\| \leq 2 \sigma_{\Phi}} \sup_{x \in \mathcal{X}} \partial \chi_i(x) \|g\|$ and suppose $\sup_{i \in \mathcal{V}} G_{\chi_i} \leq G_{\chi}$ for a bound $G_{\chi}$. In this section, we assume $\sup_{x,y \in \mathcal{X}} D_{\phi}(x, y) = D_{\phi, X}$. Then, in fact, the strong convexity of $\Phi$ implies $D_{\chi} \leq \frac{2}{\sigma_{\Phi}} D_{\phi, X}$. To investigate the convergence behavior of DSCMD-N, we denote the Bregman projection error by $\xi_i^t = x_i^{t+1} - y_i^t$. We start with the following error estimate on $\xi_i^t$.

**Lemma 3.** The Bregman projection error satisfies $E[\|\xi_i^t\|^2] \leq \frac{G_{\chi} + G_{\chi}}{\sigma_{\Phi}} - \alpha_t$.

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

$$
(\alpha_t \tilde{g}_i^t + \nabla \Phi(x_i^{t+1}) - \nabla \Phi(y_i^t) + \alpha_t h_i^{t+1}, x - x_i^{t+1}) \geq 0, \forall x \in \mathcal{X}.
$$

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

$$
(\alpha_t \tilde{g}_i^t + \nabla \Phi(x_i^{t+1}) - \nabla \Phi(y_i^t) + \alpha_t h_i^{t+1}, y_i^t - x_i^{t+1}) \geq 0.
$$

The above inequality implies that

$$
(\alpha_t (\tilde{g}_i^t + h_i^{t+1}), y_i^t - x_i^{t+1}) \geq (\nabla \Phi(y_i^t) - \nabla \Phi(x_i^{t+1}), y_i^t - x_i^{t+1}).
$$

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

$$
\alpha_t (\|\tilde{g}_i^t\|^2 + G_{\chi}) \|y_i^t - x_i^{t+1}\| \geq \sigma_{\Phi} \|y_i^t - x_i^{t+1}\|^2.
$$
Eliminate same term $\|y_i^t - x_i^{t+1}\|^2$ on both sides and take conditional expectation on $\mathcal{F}_{t-1}$, we have

$$
\mathbb{E}\left[\|E_t^i\| | \mathcal{F}_{t-1}\right] \leq \frac{1}{\sigma_{\Phi}} \left( \mathbb{E}\left[\|G_t\| | \mathcal{F}_{t-1}\right] + G_{\chi}\right) \alpha_t.
$$

(3.3)

The desired result is obtained after taking total expectation of 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 $\bar{x}^t = \frac{1}{N} \sum_{i=1}^N x_i^t$.

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

$$
\sum_{t=1}^T \sum_{i=1}^N \mathbb{E}\left[\|x_i^t - x_j^t\| \right] \leq \frac{2N\omega}{1 - \gamma} \|x_j^0\| + \left( 4N + \frac{2N^2\omega}{1 - \gamma} \right) \sum_{t=0}^T \left[ \frac{G_f + G_\chi}{\sigma_{\Phi}} \alpha_t + N \sqrt{\nu_{\mathcal{F}_t}} \right]
$$

(3.4)

**Proof.** For $\forall i \in \mathcal{V}$, set $\xi^i_j = \sum_{j=1}^N [P_t^i]_{ij} \xi^i_j$, by iterating recursively, it can be obtained that

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

$$
\bar{x}^t = \frac{1}{N} \sum_{s=1}^t \sum_{j=1}^N (c_j^{s-1} + r_{s-1} \xi_j^{s-1}) + \frac{1}{N} \sum_{j=1}^N x_j^0.
$$

Then it follows that

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

$$
+ \sum_{s=1}^t \sum_{j=1}^N \|P(t-1,s)\|_{ij} - \frac{1}{N} \|c_j^{s-1} + r_{s-1} \xi_j^{s-1}\|
$$

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

$$
+ \frac{1}{N} \sum_{j=1}^N \|c_j^{t-1} + r_{t-1} \xi_j^{t-1}\| + \|c_i^{t-1} + r_{t-1} \xi_i^{t-1}\|.
$$

Since $\mathbb{E}[\|\xi_j^{t-1}\|] = \mathbb{E}[\| \sum_{j=1}^N [P^{t-1}]_{ij} c_j^{t-1} \|] \leq \sum_{j=1}^N \mathbb{E}[\|c_j^{t-1}\|] \leq N \sqrt{\nu_{\mathcal{F}_t}}$, then $\mathbb{E}[\|c_j^{t-1} + r_{t-1} \xi_j^{t-1}\|] \leq \mathbb{E}[\|c_j^{t-1}\| + r_{t-1} \mathbb{E}[\|\xi_j^{t-1}\|]] \leq G_{\ell_2} G_\chi \alpha_t + N \sqrt{\nu_{\mathcal{F}_t-1}}$. Combine these inequalities, it follows that, for any $i \in \mathcal{V}$,

$$
\mathbb{E}[\|x_i^t - \bar{x}^t\|] \leq \omega \gamma^{-t-1} \sum_{j=1}^N \|x_j^0\| + \sum_{s=1}^{t-1} \omega \gamma^{t-s-1} \left( \frac{G_f + G_\chi}{\sigma_{\Phi}} \alpha_s + N \sqrt{\nu_{\mathcal{F}_s}} \right)
$$

$$
+ 2 \left( \frac{G_f + G_\chi}{\sigma_{\Phi}} \alpha_t + N \sqrt{\nu_{\mathcal{F}_t-1}} \right).
$$

(3.5)

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

$$
\sum_{t=1}^T \sum_{i=1}^N \mathbb{E}[\|x_i^t - \bar{x}^t\|] \leq \frac{N\omega}{1 - \gamma} \sum_{j=1}^N \|x_j^0\| + \left( 2N + \frac{N^2\omega}{1 - \gamma} \right) \sum_{t=0}^T \left( \frac{G_f + G_\chi}{\sigma_{\Phi}} \alpha_t + N \sqrt{\nu_{\mathcal{F}_t}} \right).
$$

(3.6)
Note that the bound on right hand side of (3.5) 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 (3.3). Sum up from $t = 1$ to $T$ and $i = 1$ to $N$ to $\mathbb{E}[|x_j^T - \bar{x}^T|]$, use the triangle inequality $\mathbb{E}[|x_j^T - \bar{x}^T|] \leq \mathbb{E}[|x_j^T - \bar{x}^T|] + \mathbb{E}[|\bar{x}^T - \bar{x}^t|]$, and combine with (3.6), the result in theorem is obtained. \hfill \blacksquare

**Lemma 5.** Let $\{x_i^t\}, \{y_i^t\}$ be the sequences in DSCMD-N. Let $\{\alpha_t\}$ be a non-increasing stepsize. Then we have

$$
\chi_i(x_i^{t+1}) - \chi_i(x^*) + (\bar{g}_i^t, y_i^t - x^*) \leq \frac{1}{\alpha_t} [D_{\Phi}(x^*, y_i^t) - D_\Phi(x^*, x_i^{t+1})] + \frac{\alpha_t}{2\sigma_\Phi} \|\bar{g}_i^t\|^2. \tag{3.7}
$$

**Proof.** According to the first-order optimality of the DSCMD-N, there exists $h_i^{t+1} \in \partial \chi_i(x_i^{t+1})$,

$$
\langle \alpha_i \bar{g}_i^t + \nabla \Phi(x_i^{t+1}) - \nabla \Phi(y_i^t) + \alpha_i h_i^{t+1}, x - x_i^{t+1} \rangle \geq 0, \ \forall x \in \mathcal{X}.
$$

Set $x = x^*$ in above inequality, and rearrange terms, we have

$$
\langle \alpha_i \bar{g}_i^t, x_i^{t+1} - x^* \rangle \leq \langle \nabla \Phi(y_i^t) - \Phi(x_i^{t+1}), x_i^{t+1} - x^* \rangle + \alpha_i \langle h_i^{t+1}, x^* - x_i^{t+1} \rangle
$$

$$
\leq D_\Phi(x^*, y_i^t) - D_\Phi(x^*, x_i^{t+1}) - D_\Phi(x^*, y_i^t) + \chi_i(x^*) - \chi_i(x_i^{t+1})
$$

$$
\leq D_\Phi(x^*, y_i^t) - D_\Phi(x^*, x_i^{t+1}) - \frac{\sigma_\Phi}{2} \|x_i^{t+1} - y_i^t\|^2 + \chi_i(x^*) - \chi_i(x_i^{t+1}), \tag{3.8}
$$

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,

$$
\langle \alpha_i \bar{g}_i^t, x_i^{t+1} - x^* \rangle \geq \langle \alpha_i \bar{g}_i^t, x_i^{t+1} - y_i^t \rangle + \langle \alpha_i \bar{g}_i^t, y_i^t - x^* \rangle \tag{3.9}
$$

$$
\geq -\frac{\alpha_i^2}{2\sigma_\Phi} \|\bar{g}_i^t\|^2 - \frac{\sigma_\Phi}{2} \|x_i^{t+1} - y_i^t\|^2 + \alpha_i \chi_i(x^*) - \chi_i(x_i^{t+1}).
$$

Combine (3.8) and (3.9), it follows that

$$
\langle \alpha_i \bar{g}_i^t, y_i^t - x^* \rangle \leq D_\Phi(x^*, y_i^t) - D_\Phi(x^*, x_i^{t+1}) + \frac{\alpha_i^2}{2\sigma_\Phi} \|\bar{g}_i^t\|^2 + \alpha_i [\chi_i(x^*) - \chi_i(x_i^{t+1})].
$$

The proof is concluded after dividing both sides by $\alpha_i$ in above inequality. \hfill \blacksquare

**Lemma 6.** Let $\{x_i^t\}, \{y_i^t\}$ be the sequences in DSCMD-N, then there holds $\mathbb{E}[|y_i^t - x_i^t|] \leq \sum_{j=1}^N \mathbb{E}[|x_j^t - x_i^t|] + N \sqrt{\nu_{r_1}}$ for any $i, l \in \mathcal{V}$.

**Proof.** According to the structure of DSCMD-N and the fact that the matrix $P^t$ is doubly stochastic,

$$
\|y_i^t - x_i^t\| = \|\sum_{j=1}^N [P_i^{t}]_{ij} [x_j^t - x_i^t] + r_l \sum_{j=1}^N [P_i^{t}]_{ij} \xi_j \| \leq \sum_{j=1}^N [P_i^{t}]_{ij} \|x_j^t - x_i^t\| + r_l \sum_{j=1}^N \|\xi_j\|.
$$

Take expectation over $\mathcal{F}_{i-1}$, use Assumption $\mathcal{X}$ and the fact that $0 \leq [P_i^{t}]_{ij} < 1$, then take total expectation, the lemma is concluded. \hfill \blacksquare
Lemma 7. Let \( \{x_t^j\} \) be the sequences in DSCMD-N, the noise sequence \( \{r_t\xi_t^j\} \) is defined as before, we have \( \mathbb{E}[D_\Phi(x^*, x_t^j + r_t\xi_t^j)] \leq \mathbb{E}[D_\Phi(x^*, x_t^j)] + \frac{2}{\sigma_\Phi} D_\Phi x L_\Phi \sqrt{\nu r_t} + L_\Phi \nu r_t^2. \)

Proof. According to mean value formula, there exists a \( \zeta \in [0, 1] \) such that \( \Phi(x_t^j + r_t\xi_t^j) = \Phi(x_t^j) + \langle \nabla \Phi(x_t^j + r_t\xi_t^j), r_t\xi_t^j \rangle \), then it follows that

\[
\begin{align*}
D_\Phi(x^*, x_t^j + r_t\xi_t^j) &= \Phi(x^*) - \Phi(x_t^j + r_t\xi_t^j) - \langle \nabla \Phi(x_t^j + r_t\xi_t^j), x^* - x_t^j - r_t\xi_t^j \rangle \\
&= \Phi(x^*) - \Phi(x_t^j) - \langle \nabla \Phi(x_t^j + \zeta r_t\xi_t^j), r_t\xi_t^j \rangle - \langle \nabla \Phi(x_t^j + r_t\xi_t^j), x^* - x_t^j - r_t\xi_t^j \rangle \\
&= \Phi(x^*) - \Phi(x_t^j) - \langle \nabla \Phi(x_t^j), x^* - x_t^j \rangle - \langle \nabla \Phi(x_t^j + r_t\xi_t^j), r_t\xi_t^j \rangle \\
&\leq D_\Phi(x^*, x_t^j) + (1 - \zeta) L_\Phi r_t^2\|\xi_t^j\|^2 + L_\Phi D_\chi r_t\|\xi_t^j\| \\
&\leq D_\Phi(x^*, x_t^j) + D_\chi L_\Phi\|\xi_t^j\| r_t + L_\Phi\|\xi_t^j\|^2 r_t^2,
\end{align*}
\]

in which the first inequality follows from Cauchy inequality and gradient \( L_\Phi \)-Lipschitz condition of \( \Phi \), the second inequality follows from the fact \( 0 \leq 1 - \zeta \leq 1 \). Take conditional expectation on \( \mathcal{F}_{t-1} \) on both sides, use Assumption \( \square \) and note that \( D_\chi \leq \frac{1}{\sigma_\Phi} D_\Phi x \), the result is obtained after taking total expectation. \( \blacksquare \)

Now return to (3.7), take conditional expectation over \( \mathcal{F}_{t-1} \) on both sides of (3.7), we have

\[
\mathbb{E}[\chi_t(x_t^{i+1})|\mathcal{F}_{t-1}] - \chi_t(x^*) \leq \frac{1}{\alpha_t} \left[ D_\Phi(x^*, y_t^i) - \mathbb{E}[D_\Phi(x^*, x_t^{i+1})|\mathcal{F}_{t-1}] \right] + \frac{\alpha_t}{2\sigma_\Phi} \mathbb{E}[\|\xi_t^j\|^2|\mathcal{F}_{t-1}].
\]

Take total expectation on both sides of above inequality, we have

\[
\Delta_{1,t}^1 + \Delta_{2,t}^2 \leq \Delta_{3,t}^3, \tag{3.10}
\]

in which we denote \( \Delta_{1,t}^1 = \mathbb{E}[\langle g_i(y_t^i), y_t^i - x^*\rangle], \Delta_{1,t}^2 = \mathbb{E}[\chi_t(x_t^{i+1})] - \chi_t(x^*), \Delta_{3,t}^3 = \frac{1}{\alpha_t} \left[ \mathbb{E}[D_\Phi(x^*, y_t^i)] - \mathbb{E}[D_\Phi(x^*, x_t^{i+1})] \right] + \frac{2\sigma_\Phi}{\alpha_t} \mathbb{E}[\|\xi_t^j\|^2]. \)

Before coming to the main result, we need the following lemma for \( \Delta_{3,t}^3 \).

Lemma 8. Under Assumptions 1-4, if \( \{\alpha_t\}, \{r_t\} \) be non-increasing positive sequences, then the following bound result for \( \Delta_{3,t}^3 \) holds,

\[
\sum_{i=1}^T \sum_{t=1}^N |\Delta_{3,t}^3| \leq \frac{ND_\Phi x}{\alpha_T} + \frac{NG_\Phi}{\alpha_T} + \frac{2\sigma_\Phi}{\alpha_t} D_\Phi x L_\Phi \sqrt{\nu r_t} + N L_\Phi \nu r_t + N L_\Phi \nu, \tag{3.11}
\]
Proof. Since 

$$y_t^i = \sum_{j=1}^{N} [P^t]_{ij} (x^*_j + r_l \xi^i_j),$$

separate convexity of $D_\Phi(\cdot, \cdot)$ implies that

$$\sum_{t=1}^{T} \sum_{l=1}^{N} |\Delta^3_{t,l}|$$

$$\leq \sum_{t=1}^{T} \frac{1}{\alpha_t} \left[ \sum_{l=1}^{N} \left[ \sum_{j=1}^{N} [P^t]_{ij} \mathbb{E}[D_\Phi(x^*, x^*_j + r_l \xi^i_j)] - \sum_{i=1}^{N} \mathbb{E}[D_\Phi(x^*, x^*_j)] \right] + \frac{N G^2_\Phi}{2 \sigma_\Phi} \sum_{t=0}^{T} \alpha_t \right]$$

$$\leq \sum_{t=1}^{T} \frac{1}{\alpha_t} \left[ \sum_{l=1}^{N} \mathbb{E}[D_\Phi(x^*, x^*_j)] - \sum_{i=1}^{N} \mathbb{E}[D_\Phi(x^*, x^*_j)] \right] + N \left[ 2G_\Phi + \sqrt{\frac{2}{\sigma_\Phi}} D_{\Phi,X} \nu L_{\Phi} \right] \sqrt{\nu} \sum_{t=1}^{T} \frac{r_t}{\alpha_t}$$

$$+ NL_{\Phi} \nu \sum_{t=1}^{T} \frac{r_t^2}{\alpha_t} + \frac{N G^2_\Phi}{2 \sigma_\Phi} \sum_{t=0}^{T} \alpha_t,$$

in which the second inequality is obtained by double stochasticity of matrix $P^t$ and Lemma 3, the result is obtained after eliminating same terms in the summation in above equality. \[\Box\]

Now we are ready to give the main result of this section. Denote $\hat{x}_T^l = \frac{1}{T} \sum_{t=1}^{T} x^*_t, x^* = \arg\min_{x \in X} f(x)$. The following result describes the expected bound for DSCMD-N in terms of stepsizes $\{\alpha_t\}$, noise decreasing rates $\{r_t\}$.

**Theorem 1.** Let the Assumptions 1-4 hold. If $\{\alpha_t\}, \{r_t\}$ are positive non-increasing sequences, then for DSCMD-N method, for any $l \in \mathcal{V}$, we have

$$\mathbb{E}[F(\hat{x}_T^l)] - F(x^*) \leq \frac{C_1}{T} + \frac{C_2}{T \alpha_T} + \frac{C_3}{T} \sum_{t=0}^{T} \alpha_t + \frac{C_4}{T} \sum_{t=0}^{T} r_t + \frac{C_5}{T} \sum_{t=0}^{T} \frac{r_t^2}{\alpha_t} + \frac{C_6}{T} \sum_{t=0}^{T} \frac{r_t^2}{\alpha_t},$$

in which

$$C_1 = \frac{2N \omega}{1-\gamma} [(N + 1)G_f + NG_\lambda] \cdot \|x^0\|, \quad C_2 = ND_{\Phi,X}^2,$$

$$C_3 = \left( 4N + \frac{2N^2 \omega}{1-\gamma} \right)([(N + 1)G_f + NG_\lambda] + NG_\lambda) \cdot \frac{G_f + G_\lambda}{\sigma_\Phi} + \frac{NG^2_\Phi}{2 \sigma_\Phi},$$

$$C_4 = (4N + \frac{2N^2 \omega}{1-\gamma})[(N + 1)G_f + NG_\lambda] N \sqrt{\nu} + (G_f + G_\lambda)N^2 \sqrt{\nu},$$

$$C_5 = \sqrt{\frac{2}{\sigma_\Phi}} D_{\Phi,X} L_{\Phi} \sqrt{\nu}, \quad C_6 = NL\Phi \nu,$$

and $\omega = (1 - \frac{\theta}{\alpha_T})^{-2}, \; \gamma = (1 - \frac{\theta}{\alpha_T})^{1/3}$. 

10
Proof. We prove the result by estimating the terms in (3.10). For any index $l \in V$, 
\[
\langle g_l(y_t^l), y_t^l - x^* \rangle 
\geq f_t(y_t^l) - f_t(x^*) 
= f_t(y_t^l) - f_t(x_t^L) + f_t(x_t^L) - f_t(x^*) 
\geq -G_f \|y_t^l - x_t^L\|_2 + \langle f_t(x_t^L) - f_t(x^*) \rangle 
\geq -G_f \|y_t^l - x_t^L\|_2 - G_f \|x_t^L - x^*\|_2 + \langle f_t(x_t^L) - f_t(x^*) \rangle.
\]
After taking expectation and using Lemma 6, it follows that 
\[
\Delta^1_{i,t} \geq -N \sqrt{\nu G_f \sigma_t} - G_f \|x_t^L - x^*\|_2 + \langle f_t(x_t^L) - f_t(x^*) \rangle.
\]

Denote $f = \sum_{i=1}^N f_i$, $\chi = \sum_{i=1}^N \chi_i$, and denote the bound on the right hand side in Lemma 4 by $B_T$, sum up both sides and use Lemma 4 it follows that 
\[
\sum_{t=1}^T \sum_{i=1}^N \left[ \Delta^1_{i,t} \right] \geq -(N + 1) G_f B_T - N^2 \sqrt{\nu G_f} \sum_{t=1}^T r_t + \sum_{t=1}^T \mathbb{E} [f(x_t^L) - f(x^*)].
\]
(3.13)

On the other hand, for any index $l \in V$, 
\[
\chi_i(x_t^{l+1}) - \chi_i(x^*) 
= \left[ \chi_i(x_t^{l+1}) - \chi_i(y_t^l) \right] + \left[ \chi_i(y_t^l) - \chi_i(x_t^L) \right] + \left[ \chi_i(x_t^L) - \chi_i(x^*) \right] 
\geq -G_f \|x_t^{l+1} - y_t^l\|_2 - G_f \|y_t^l - x_t^L\|_2 + \langle \chi_i(x_t^L) - \chi_i(x^*) \rangle.
\]
After taking expectation on both sides, using Lemma 3 and Lemma 6, we have 
\[
\Delta^2_{i,t} \geq -N G_f \chi \|x_t^L - y_t^l\|_2 - G_f \|y_t^l - x_t^L\|_2 + \langle \chi_i(x_t^L) - \chi_i(x^*) \rangle.
\]
(3.14)

Sum up both sides of (3.10) from $i = 1$ to $N$ and $t = 1$ to $T$, combine it with (3.13), (3.14). Lemma 8 The desired result is obtained after substituting $B_T$, using Lemma 4 dividing both sides by $T$ and using the convexity of $f_i$, $i = 1, 2, ..., N$. ■

Under a boundedness assumption of stochastic gradient and network noise, the following high probability bound holds for DSCD-N.

**Theorem 2.** Under the assumptions of Theorem 7, if we assume in addition that $\|\tilde{y}_t^L\|_2 \leq G_f$ and $\|\xi_t^L\|_2^2 \leq \nu$, then for DSCD-N, for any $l \in V$, we have, for any $\delta \in (0,1)$, with probability of at least $1 - \delta$, 
\[
F(x^*_t) - F(x^*) \leq \frac{C_1}{T} + \frac{C_2}{T \log T} + \frac{C_3}{T} \sum_{t=0}^T \sigma_t + \frac{C_4}{T} \sum_{t=0}^T \sigma_t r_t + \frac{C_5}{T} \sum_{t=0}^T \alpha_t r_t + \frac{C_6}{T} \sum_{t=0}^T \alpha_t r_t + 2 \sqrt{G_f \nu} \sqrt{\frac{\log(1/\delta)}{\sqrt{T}}},
\]
in which $C_1 \sim C_6$ are defined as in Theorem 7.

Proof. For saving space, we just show the difference between the proof for this result and the above expected bound result. Come back to (5.7), if we denote \( \Delta^1_{i,t} = \langle g_i(x_t^l), y_t^l - x^* \rangle, \Delta^2_{i,t} = \chi_i(x_t^{l+1}) - \chi_i(x^*), \Delta_{i,t}^3 = \frac{1}{\alpha_t} \left[ D_f(x^L, y_t^l) - D_f(x^L, x_t^{l+1}) \right], X_{t,i} = \langle g_i(y_t^l) - \tilde{g}_i^L, y_t^l - x^* \rangle \), then (5.7) can be written in the form of \( \Delta^1_{i,t} + \Delta^2_{i,t} \leq \hat{\Delta}_{i,t}^3 + X_{i,t}. \hat{\Delta}_{i,t}^4, \Delta_{i,t}^5, \Delta_{i,t}^6 \) corresponds to \( \hat{\Delta}_{i,t}^1, \hat{\Delta}_{i,t}^2, \).
We investigate a distributed dual averaging type method (DS-SCDA-N) for solving it in next section. If we denote \(X_t = \sum_{i=1}^{N} X_{i,t}\) and sum up both sides from \(t = 1\) to \(T\) and \(i = 1\) to \(N\), it follows that

\[
\sum_{t=1}^{T} \sum_{i=1}^{N} |\Delta_{i,t}^1| + \sum_{t=1}^{T} \sum_{i=1}^{N} |\Delta_{i,t}^2| \leq \sum_{t=1}^{T} \sum_{i=1}^{N} |\bar{\Delta}_{i,t}^3| + \sum_{t=1}^{T} X_t.
\]

(3.15)

Note that \(\mathbb{E}[X_t | \mathcal{F}_{t-1}] = 0\), the bound condition \(|\bar{g}_i^2| \leq G_f\) and Cauchy inequality implies \(|X_t| \leq 2NG_fD_X\), then \(\{X_t\}\) is a bounded martingale difference sequence. Use Azuma-Hoeffding inequality to \(\{X_t\}\), we have for any \(\epsilon > 0\)

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

(3.16)

Setting the above probability to \(\delta\), we have, with probability at least \(1 - \delta\),

\[
\sum_{t=1}^{T} X_t \leq 2\sqrt{2G_fD_XN}\sqrt{\log \frac{1}{\delta}}.
\]

(3.17)

On the other hand, it is easy to see that, with bound assumptions \(|\bar{g}_i^2| \leq G_f\) and \(|\xi_i^1|^2 \leq \nu\) in hand, the estimate result of Lemma 4 and Lemma 7 holds without taking expectation. Therefore we know (3.11), (3.13), (3.14) hold with \(\Delta_{i,t}^1, \Delta_{i,t}^2, \tilde{\Delta}_{i,t}^3\) replaced by \(\Delta_{i,t}^1, \Delta_{i,t}^2, \tilde{\Delta}_{i,t}^3\). Combining these three estimates with (3.17) and (3.19), dividing both sides by \(T\) and using the convexity of \(F_t\), \(i = 1, 2, \ldots, N\), we obtain the desired result.

For relatively comprehensive study on distributed stochastic composite optimization over noisy network. We consider Problem (1.2) that has different regularization feature from Problem (1.1). We investigate a distributed dual averaging type method (DS-SCDA-N) for solving it in next section.

### 3.2 Distributed stochastic dual averaging method for composite optimization over noisy network

In this section, we consider problem (1.2), minimizing \(F(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x) + \eta(x)\) over noisy network. We solve the problem by providing a distributed stochastic composite dual averaging type method which we call it the DS-SCDA-N method. DS-SCDA-N deals with \(N\) pairs of vector variables \((x_i^t, z_i^t) \in \mathcal{X} \times \mathbb{R}^p\), with the \(i\)-th pair associated with node \(i \in \mathcal{V}\). In DS-SCDA-N, we query the stochastic subgradient oracle at variable \(x_i^t\) to get a stochastic subgradient \(\tilde{g}_i^t := \tilde{g}_i(x_i^t)\), such that \(\mathbb{E}[\tilde{g}_i^t | \mathcal{F}_{t-1}] = g_i(x_i^t) \in \partial f_t(x_i^t)\) is a subgradient of \(f_t\) at \(x_i^t\). Then the algorithm is presented as follows:

\[
\begin{align*}
z_i^{t+1} &= \sum_{j=1}^{N} [\mathcal{P}^t]_{ij}(z_j^t + r_t \xi_{ij}^t) + \tilde{g}_i^t, \\
x_i^{t+1} &= \arg\min_{x \in \mathcal{X}} \{z_i^{t+1} + \frac{1}{\alpha_t} \Psi(x) + \tilde{\eta}(x)\}.
\end{align*}
\]

(3.18), (3.19)

In above proposed dual-averaging scheme, the variable \(z_i^t\) of node \(i\) participates in the information communication process with its neighbours, hence, for agent \(i\), the information communication of the noisy corrupted version \(z_j^t + r_t \xi_{ij}^t\), \(j = 1, 2, \ldots, N\) is considered in (3.18). In composite dual averaging scheme (3.10), \(\alpha_t\) is the stepsize. Global composite function \(\eta\) is supposed to be some simple regularization function with supremum subgradient \(G_\eta\). The proximal function \(\Psi(x) : \mathcal{X} \to \mathbb{R}\)
\(\mathbb{R}\) is assumed to be \(\sigma_{\Psi}\)-strongly convex in the sense \(\Psi(x) \geq \Psi(y) + \langle \nabla \Psi(y), x - y \rangle + \frac{\sigma_{\Psi}^2}{2} \|x - y\|^2\) for any \(x, y \in \mathcal{X}\). To analyze the convergence behavior of DSCDA-N under slowly decreasing communication noise effects over network, this part starts with the following expected bound for variable \(\bar{z}_t^i\).

**Lemma 9.** For any \(i \in \mathcal{V}\) and \(t = 1, 2, ..., T\), the variables \(z_t^i\) and their average \(\bar{z}_t^i\) satisfy

\[
\mathbb{E}[\|z_t^i - \bar{z}_t^i\|] \leq \left( \frac{N\omega}{1 - \gamma} + 2 \right) G_f + \omega N^2 \sqrt{\nu} \sum_{s=1}^{T-2} r_s \gamma^{t-s-2} + 2N \sqrt{\nu} r_{t-1}. \tag{3.20}
\]

**Proof.** By iteration on sequence \(z_t^i\),

\[
z_t^i = \sum_{s=1}^{t-1} \sum_{j=1}^{N} [P(t-1, s)]_{ij} (\bar{g}_j^{s-1} + r_{s-1} \xi_j^{s-1}) + \bar{g}_i^{t-1} + r_{t-1} \xi_i^{t-1}, \tag{3.21}
\]

\[
\bar{z}_t^i = \frac{1}{N} \sum_{s=1}^{t-1} \sum_{j=1}^{N} (\bar{g}_j^{s-1} + r_{s-1} \xi_j^{s-1}) + \frac{1}{N} \sum_{j=1}^{N} (\bar{g}_j^{t-1} + r_{t-1} \xi_j^{t-1}). \tag{3.22}
\]

By taking substraction of above inequality and taking norm, we have

\[
\|z_t^i - \bar{z}_t^i\| \leq \sum_{s=1}^{t-1} \sum_{j=1}^{N} \| [P(t-1, s)]_{ij} \| \cdot (\|\bar{g}_j^{s-1}\| + r_{s-1} \|\xi_j^{s-1}\|)
\]

\[
+ \|\bar{g}_i^{t-1}\| + r_{t-1} \|\xi_i^{t-1}\| + \frac{1}{N} \sum_{j=1}^{N} (\|g_j^{t-1}\| + r_{t-1} \|\xi_j^{t-1}\|).
\]

Taking expectation on both sides and use Lemma 1 \text{ and Assumptions 2 and 3} we have

\[
\mathbb{E}[\|z_t^i - \bar{z}_t^i\|] \leq \sum_{s=1}^{t-1} \sum_{j=1}^{N} \omega \gamma^{t-s-1} \left[ G_f + N \sqrt{\nu} r_{s-1} \right] + 2 [G_f + N \sqrt{\nu} r_{t-1}]
\]

\[
\leq \left( \frac{N\omega}{1 - \gamma} + 2 \right) G_f + \omega N^2 \sqrt{\nu} \sum_{s=1}^{T-2} r_s \gamma^{t-s-1} + 2N \sqrt{\nu} r_{t-1},
\]

which finishes the proof. \(\square\)

Denote \(\Pi_{\Phi}(x) = \arg\min_{x \in \mathcal{X}} \{\langle x, y \rangle + \frac{1}{\alpha_{T}}\Psi(y) + t\eta(y)\}\). The following lemmas in \([16]\) are needed in following analysis.

**Lemma 10.** For any \(x, y \in \mathcal{X}\), we have \(\|\Pi_{\Phi}(x) - \Pi_{\Phi}(y)\| \leq \frac{\alpha_{T}}{\alpha_T} \|x - y\|\).

**Lemma 11.** Let \(\{\alpha_t\}\) be a nonincreasing sequence and \(G_t \in \mathbb{R}^n\) be an arbitrary sequence of vectors. For a sequence \(\{W_t\}\), if \(W_{t+1} = \Pi_{\Phi}(\sum_{s=1}^{t} G_s)\). Then for any \(x^* \in \mathcal{X}\), \(\sum_{t=1}^{T} \left[ (G_t, W_t - x^*) + \eta(W_t) - \eta(x^*) \right] \leq \frac{1}{\alpha_T} \Psi(x^*) + \frac{1}{2} \sum_{t=1}^{T} \alpha_{t-1} \|G_t\|^2\).

Now we are ready to present the first main result of this section. We still denote \(\tilde{x}_t^i = \frac{1}{T} \sum_{t=1}^{T} x_t^i\). The following theorem describes the expected error bound for approximating sequence \(\tilde{x}_t^i = \frac{1}{T} \sum_{t=1}^{T} x_t^i\) generated by DSCDA-N method.
Theorem 3. Under Assumptions 1-3. If \( \{\alpha_t\} \) and \( \{r_t\} \) are non-increasing sequences. Then for problem (1.2) over noisy network, the DSCDA-N method achieves the following expected error bound: for any \( i \in V \),

\[
E[F(\tilde{x}^i_t)] - F(x^*) 
\leq \left[ 3G_f + G_n \left( \frac{N\omega}{1 - \gamma} + 2 \right) G_f + G^2_f \right] \frac{1}{T} \sum_{t=0}^{T} \alpha_t + \frac{3G_f + G_n}{\sigma_f} \omega N^2 \sqrt{\nu} \cdot \frac{1}{T} \sum_{t=1}^{T} \alpha_t \sum_{s=1}^{t-1} r_{s-1} \gamma^{t-s-1} \\
+ \frac{3G_f + G_n}{\sigma_f} 2N\sqrt{\nu} \cdot \frac{1}{T} \sum_{t=1}^{T} \alpha_t r_{t-1} + \frac{N\nu}{T} \sum_{t=0}^{T} \alpha_t r^2_t + \frac{\sqrt{\nu} D^2}{T} \sum_{t=1}^{T} r_t.
\]

Proof. denote \( y_t = \Pi^*_y(\bar{z}^i) \), it follows that

\[
F(x^*_t) - F(x^*) \\
= F(x^*_1) - F(y^1) + F(y^1) - F(x^*) \\
\leq (G_f + G_n)\|\Pi^*_y(z^*_1) - \Pi^*_y(\bar{z}^i)\| + [F(y^1) - F(x^*)] := \Omega^1_1 + \Omega^1_2. \tag{3.23}
\]

Use Lemma [10] by taking total expectation, it is easy to see

\[
\sum_{t=1}^{T} E[\Omega^1_t] \leq \frac{G_f + G_n}{\sigma_f} \sum_{t=1}^{T} \alpha_t \mathbb{E}[\|z^*_t - \bar{z}^i\|]. \tag{3.24}
\]

In what follows, we consider the bound for \( \sum_{t=1}^{T} E[\Omega^2] \). Denote \( f = \sum_{t=1}^{T} f_t \),

\[
\sum_{t=1}^{T} [F(y^t) - F(x^*)] \\
= \sum_{t=1}^{T} \frac{1}{N} [f(y^t) - f(x^*)] + \sum_{t=1}^{T} [f(y^t) - f(x^*)] \\
= \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} [f_j(x^*_j) - f(x^*_j)] + \sum_{t=1}^{T} [f(y^t) - f(x^*_t)] + \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} [f_j(y^t) - f_j(x^*_j)] \\
\leq \left( \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} g_j(x^*_j), y^t - x^* \right) + \sum_{t=1}^{T} [f(y^t) - f(x^*_t)] + \left( \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} (g_j(x^*_j), x^*_j - y^t) \\
+ \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} [f_j(y^t) - f_j(x^*_j)] \right) := s_1^t + s_2^t. \tag{3.25}
\]

We treat with \( s_1^t \) first as follow,

\[
s_1^t = \sum_{t=1}^{T} \left[ \frac{1}{N} \sum_{j=1}^{N} (g_j^t + r_t \xi_j^t), y^t - x^* \right] + \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} (g_j(x^*_j), y^t - x^*) \\
+ \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} (\tilde{g}_j^t, y^t - x^*) - \sum_{t=1}^{T} \frac{1}{N} \sum_{j=1}^{N} (r_t \xi_j^t, y^t - x^*). \tag{3.26}
\]
By using Lemma 11 with \( W_t = y_t \) and \( G_t = \tilde{z}_t = \frac{1}{N} \sum_{j=1}^{N} (g_j^{t-1} + r_t^* \xi_j^{t-1}) \), we have

\[
\sum_{t=1}^{T} \left( \frac{1}{N} \sum_{j=1}^{N} (\tilde{g}_j^t + r_t^* \xi_j^t), y_t^t - x^* \right) + \eta(y_t^t) - \eta(x^*) \leq \frac{1}{\alpha T} \Psi(x^*) + \frac{1}{2} \sum_{t=1}^{T} \alpha_t \left( \frac{1}{N} \sum_{j=1}^{N} (g_j^t + r_t^* \xi_j^t) \right)^2 \\
\leq \frac{1}{\alpha T} \Psi(x^*) + \frac{1}{2} \sum_{t=0}^{T} \alpha_t \left( \frac{1}{N} \sum_{j=1}^{N} (\|\tilde{g}_j^t\| + r_t \|\xi_j^t\|) \right)^2 \\
\leq \frac{1}{\alpha T} \Psi(x^*) + \frac{1}{N} \sum_{t=0}^{T} \alpha_t \left( \|\tilde{g}_j^t\|^2 + r_t^2 \|\xi_j^t\|^2 \right).
\]

Since \( \|\xi_j^t\|^2 = \|\sum_{t=1}^{T} [P_t \xi_j^t]\|^2 \leq \sum_{t=1}^{T} \|P_t [\xi_j^t]\|^2 \), which follows from the convexity of \( \| \cdot \|^2 \).

Take total expectation, we have \( \mathbb{E}[\|\xi_j^t\|^2] \leq N \nu \). Then after taking total expectation, it follows that

\[
\sum_{t=1}^{T} \mathbb{E} \left( \left( \frac{1}{N} \sum_{j=1}^{N} (\tilde{g}_j^t + r_t^* \xi_j^t), y_t^t - x^* \right) + \eta(y_t^t) - \eta(x^*) \right) \leq \frac{1}{\alpha T} \Psi(x^*) + G_f^2 \sum_{t=0}^{T} \alpha_t + N \nu \sum_{t=0}^{T} \alpha_t r_t^2.
\]

For the second term of \( s_1^t \) in (3.26), we have

\[
\mathbb{E}[\langle g_j(x_j^t) - \tilde{g}_j^t, y_t^t - x^* \rangle] = \mathbb{E}[\mathbb{E}[\langle g_j(x_j^t) - \tilde{g}_j^t, y_t^t - x^* \rangle|F_{t-1}]] = \mathbb{E}[\langle \mathbb{E}[g_j(x_j^t) - \tilde{g}_j^t|F_{t-1}], y_t^t - x^* \rangle] = 0.
\]

Also, since \( \mathbb{E}[\langle r_t \xi_j^t, y_t^t - x^* \rangle] \leq r_t \mathbb{E}[\|\xi_j^t\| \cdot \|y_t^t - x^*\|] \leq r_tD_X \mathbb{E}[\|\xi_j^t\|] \leq \sqrt{\nu} D_X r_t \), we have

\[
\left| - \frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} \mathbb{E}[\langle r_t \xi_j^t, y_t^t - x^* \rangle] \right| \leq \sqrt{\nu} D_X \sum_{t=1}^{T} r_t.
\]

The above inequalities imply that

\[
\mathbb{E}[s_1^t] \leq \frac{1}{\alpha T} \Psi(x^*) + G_f^2 \sum_{t=0}^{T} \alpha_t + N \nu \sum_{t=0}^{T} \alpha_t r_t^2 + \sqrt{\nu} D_X \sum_{t=1}^{T} r_t.
\]

On the other hand, note that \( \langle g_j(x_j^t), x_j^t - y^t \rangle \leq \|g_j(x_j^t)\| \cdot \|x_j^t - y^t\| \leq G_f \cdot \frac{2\alpha}{\nu} \|z_j^t - \tilde{z}^t\| \) and \( f_j(y^t) - f_j(x_j^t) \leq G_f \|y^t - x_j^t\| \leq G_f \cdot \frac{2\alpha}{\nu} \|z_j^t - \tilde{z}^t\| \). After taking total expectation, we have

\[
\mathbb{E}[s_2^t] \leq \frac{2G_f}{N\sigma_\Psi} \sum_{j=1}^{N} \sum_{t=1}^{T} \alpha_t \mathbb{E}[\|z_j^t - \tilde{z}^t\|].
\]

Combine (3.25) with (3.29) and (3.30) and take total expectation, it follows that

\[
\mathbb{E}[\Omega_2^t] \leq \frac{1}{\alpha T} \Psi(x^*) + G_f \sum_{t=0}^{T} \alpha_t + N \nu \sum_{t=0}^{T} \alpha_t r_t^2 + \sqrt{\nu} D_X \sum_{t=1}^{T} r_t + \frac{2G_f}{N\sigma_\Psi} \sum_{j=1}^{N} \sum_{t=1}^{T} \alpha_t \mathbb{E}[\|z_j^t - \tilde{z}^t\|].
\]
take total expectation of (3.23) and sum up from $t = 1$ to $T$, and combine with (3.24) and (3.31), we have

$$
\sum_{t=1}^{T} \mathbb{E}[F(x^t_i) - F(x^*)] \leq \frac{G_f + G_n}{\sigma_f} \sum_{t=1}^{T} \alpha_t \mathbb{E}[\|z^t_i - z^t\|] + \frac{2G_f}{N\sigma_f} \sum_{t=1}^{T} \alpha_t \mathbb{E}[\|z^t_i - z^t\|]
$$

$$
+ \frac{1}{\alpha_T} \Psi(x^*) + G_f^2 \sum_{t=0}^{T} \alpha_t + N\nu \sum_{t=0}^{T} \alpha_t + \sqrt{T} \sum_{t=1}^{T} r_t.
$$

If we use $B_t$ to denote the bound on the right hand side of (3.20) in Lemma 9, note that $B_t$ does not depend on $i \in V$, then the above inequality

$$
\leq \frac{3G_f + G_n}{\sigma_f} \sum_{t=1}^{T} \alpha_t B_t + \frac{1}{\alpha_T} \Psi(x^*) + G_f^2 \sum_{t=0}^{T} \alpha_t + N\nu \sum_{t=0}^{T} \alpha_t + \sqrt{T} \sum_{t=1}^{T} r_t.
$$

(3.32)

After substituting $B_t$ and rearranging terms, we have

$$
\sum_{t=1}^{T} \mathbb{E}[F(x^t_i)] - F(x^*) \leq \left[ \frac{3G_f + G_n}{\sigma_f} \left( \frac{N\omega}{1 - \gamma} + 2 \right) G_f + G_f^2 \right] \sum_{t=0}^{T} \alpha_t + \frac{3G_f + G_n}{\sigma_f} \omega N^2 \sqrt{T} \sum_{t=1}^{T} \alpha_t \sum_{s=1}^{t-1} r_{s-1} \gamma^{t-s-1} + \frac{3G_f + G_n}{\sigma_f} 2N \sqrt{T} \sum_{t=1}^{T} \alpha_t \sum_{s=1}^{t-1} r_{s-1} \gamma^{t-s-1}
$$

$$
+ \frac{3G_f + G_n}{\sigma_f} \frac{2N \sqrt{T}}{\gamma^T} \sum_{t=1}^{T} \alpha_t r_{t-1} + \frac{\Psi(x^*)}{\alpha_T} + \frac{N\nu}{\alpha_T} \sum_{t=0}^{T} \alpha_t r_t
$$

$$
+ 2 \sqrt{2} G_f D_X \frac{\sqrt{\log(2/\delta)}}{\sqrt{T}} + \sqrt{2} N \sqrt{T} D_X \left( \sum_{t=1}^{T} r_t^2 \right)^{1/2} \frac{\sqrt{\log(2/\delta)}}{\sqrt{T}}.
$$

(3.33)

The desired result is obtained by dividing both sides by $T$ and noting the convexity of $F$. ■

In what follows, after assuming appropriate conditions for stochastic gradients $\{g_i^t\}$ and noise vectors $\{\xi_i^t\}$, we provide a high probability convergence bound of $F(\tilde{x}_i^T) - F(x^*)$ for DSCDA-N method. The main tool is the Azuma-Hoeffding inequality in Lemma 2.

**Theorem 4.** Under assumptions of Theorem 3, if we further assume $\|\tilde{g}_i^t\| \leq G_f$ and $\{\xi_i^t\}$ is a bounded martingale difference sequence with $\|\xi_i^t\|^2 \leq \nu$, then for any $i \in V$, for any $\delta \in (0, 1)$, we have, with probability of at least $1 - \delta$, the DSCDA-N method satisfies

$$
F(\tilde{x}_i^T) - F(x^*) \leq \left[ \frac{3G_f + G_n}{\sigma_f} \left( \frac{N\omega}{1 - \gamma} + 2 \right) G_f + G_f^2 \right] \sum_{t=0}^{T} \alpha_t + \frac{3G_f + G_n}{\sigma_f} \omega N^2 \sqrt{T} \sum_{t=1}^{T} \alpha_t \sum_{s=1}^{t-1} r_{s-1} \gamma^{t-s-1} + \frac{3G_f + G_n}{\sigma_f} \frac{2N \sqrt{T}}{\gamma^T} \sum_{t=1}^{T} \alpha_t r_{t-1} + \frac{\Psi(x^*)}{\alpha_T} + \frac{N\nu}{\alpha_T} \sum_{t=0}^{T} \alpha_t r_t
$$

$$
+ 2 \sqrt{2} G_f D_X \frac{\sqrt{\log(2/\delta)}}{\sqrt{T}} + \sqrt{2} N \sqrt{T} D_X \left( \sum_{t=1}^{T} r_t^2 \right)^{1/2} \frac{\sqrt{\log(2/\delta)}}{\sqrt{T}}.
$$

Proof. We turn to consider $s_i^t$ in (3.20) again. Set $X_t = \sum_{j=1}^{N} (g_j(x_j^t) - \tilde{g}_j^t, y_j^t - x^*)$, recall that (3.28) holds and $y^t$ is measurable w.r.t. $\mathcal{F}_{t-1}$. Also by using the bounded norm condition for stochastic gradient, we have $|X_t| \leq 2G_f D_X N$, it follows that $\{X_t\}$ is a bounded martingale difference sequence. Then by using Azuma-Hoeffding inequality to $\{X_t\}$, we have, for any $\epsilon_t$,

$$
\text{Prob}(\sum_{t=1}^{T} X_t \geq \epsilon_t) \leq \exp\left(-\frac{\epsilon_t^2}{2T(2G_f D_X N)^2}\right),
$$

(3.34)
On the other hand, set $Y_t = \sum_{j=1}^{N} (r_t \xi^j_t : x^* - y^j)$, then \( \{Y_t\} \) is a bounded martingale difference sequence with \( |Y_t| \leq N r_t \|\xi_t^j\| D_X \leq N^2 \sqrt{D_X} r_t \), by using Azuma-Hoeffding inequality to \( Y_t \), it follows that, for any \( \epsilon_2 > 0 \),

\[
\Pr\left( \sum_{t=1}^{T} Y_t \geq \epsilon_2 \right) \leq \exp\left(-\frac{\epsilon_2^2}{2N^4 \nu D_X^2 \sum_{t=1}^{T} r_t^2}\right). \tag{3.35}
\]

Now set \( \frac{\epsilon_2}{2} = \exp\left(-\frac{\epsilon_2^2}{2N^4 \nu D_X^2 \sum_{t=1}^{T} r_t^2}\right) \) and \( \frac{\epsilon_2}{2} = \exp\left(-\frac{\epsilon_2^2}{2N^4 \nu D_X^2 \sum_{t=1}^{T} r_t^2}\right) \), we obtain \( \epsilon_1 = 2\sqrt{2} G_f D_X N \sqrt{T} \sqrt{\log(2/\delta)} \) in (3.34) and \( \epsilon_2 = \sqrt{2} N^2 \sqrt{D_X} (\sum_{t=1}^{T} r_t^2)^{1/2} \sqrt{\log(2/\delta)} \) in (3.35). Then it follows that

\[
\Pr\left( \sum_{t=1}^{T} X_t + \sum_{t=1}^{T} Y_t \geq 2\sqrt{2} G_f D_X N \sqrt{T} \sqrt{\log\frac{2}{\delta}} + \sqrt{2} N^2 \sqrt{D_X} (\sum_{t=1}^{T} r_t^2)^{1/2} \sqrt{\log\frac{2}{\delta}} \right) \leq \delta. \tag{3.36}
\]

With the norm bound condition of \( \|\tilde{g}_t\|, \|\xi_t^j\| \), the corresponding previous estimates of Lemma 3.24, 3.27, 3.30 still hold without taking expectations. Combine these estimates with (3.34) and (3.35), we obtain the high probability bound version of (3.33), which reads that, with probability at least \( 1 - \delta \), there holds

\[
\sum_{t=1}^{T} F(x_t^*) - F(x^*) \leq \left[ \frac{3G_f + G_g}{\sigma_y} (\frac{N \omega}{1 - \gamma} + 2G_f + G_f^2) \sum_{t=0}^{T} \alpha_t + \frac{3G_f + G_g}{\sigma_y} N^2 \sqrt{T} \cdot \sum_{t=1}^{T} \alpha_t \sum_{s=1}^{r_{t-1}} t^{1-s-1} \right] + \frac{3G_f + G_g}{\sigma_y} 2N \sqrt{T} \cdot \sum_{t=1}^{T} \alpha_t r_{t-1} + \psi(x^*) \frac{1}{\alpha T} + N \nu \sum_{t=0}^{T} \alpha_t r_t^2 + 2\sqrt{2} G_f D_X \sqrt{T} \sqrt{\log\frac{2}{\delta}} + \sqrt{2} N \sqrt{D_X} (\sum_{t=1}^{T} r_t^2)^{1/2} \sqrt{\log\frac{2}{\delta}}
\]

The desired result is obtained by dividing both sides by \( T \) of above inequality and using the convexity of \( F \). \( \blacksquare \)

Compare the assumptions in Theorem 2 and Theorem 4, we note that, to obtain the high probability bound for DSCDA-N method, we need the condition that the communication noise sequence \( \{\xi_t^j\} \) is bounded martingale difference sequence. Meanwhile, in DSCMD-N method, only bounded noise sequence condition is needed for \( \{\xi_t^j\} \) to get the high probability bound. This fact shows a difference on the requirement of noise condition to obtain high probability bounds for the two methods in this paper.

4 Convergence rates

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

17
4.1 DSCMD-N

The following theorem provides a general expected bound for expected error \( E[F(\hat{x}_T^T)] - F(x^*) \) in terms of the total iteration step \( T \) with a general stepsize consideration in form of \( \alpha_t = \frac{1}{(t+1)^{1/2}} \) and noise decreasing rate in form of \( r_t = \frac{1}{(t+1)^{1/2}} \).

**Proposition 1.** Under conditions of Theorem 3 if the sequences \( \{\alpha_t\} \), \( \{r_t\} \) in the DSCMD-N method are \( \alpha_t = \frac{1}{(t+1)^{1/2}} \) and \( r_t = \frac{1}{(t+1)^{1/2}} \), \( t = 1,2,\ldots,T \). Suppose that \( 0 < \kappa_1 < \kappa_2 \leq 1 \) and \( 2\kappa_2 - \kappa_1 \neq 1 \), then we have

\[
\mathbb{E}[F(\hat{x}_T^T)] - F(x^*) \\
\leq \left( C_4 + \frac{1}{1 - (2\kappa_2 - \kappa_1)} \right) C_6 \left( C_7 + \frac{1}{1 - (2\kappa_2 - \kappa_1)} \right) \frac{1}{T} + 2^{\kappa_1} C_2 \frac{1}{T^{1-\kappa_1}} + 2^{1-\kappa_1} C_3 \frac{1}{T^{\kappa_1}} + 2^{1-\kappa_2} C_4 \frac{1}{T^{\kappa_2}} \\
+ \frac{2^{1-\kappa_2} C_5 \left( 1 - (1 - (2\kappa_2 - \kappa_1) \right) \frac{1}{T^{\kappa_2 - \kappa_1}} }{1 - (2\kappa_2 - \kappa_1)} + \frac{C_6}{1 - (2\kappa_2 - \kappa_1) \right) \frac{1}{T^{2\kappa_2 - \kappa_1}}}, \\
\text{if } \kappa_1 \in (0,1), \kappa_2 \in (0,1); \text{ and}
\]

\[
\leq \left( C_4 + \frac{2 - \kappa_1}{1 - \kappa_1} C_6 \right) \frac{1}{T} + 2^{\kappa_1} C_2 \frac{1}{T^{1-\kappa_1}} + \frac{2^{\kappa_1} C_3 \left( 1 - (2\kappa_2 - \kappa_1) \right) \frac{1}{T^{\kappa_1}} }{1 - (1 - (2\kappa_2 - \kappa_1) \right) \frac{1}{T^{\kappa_2}}}, \\
\text{if } \kappa_1 \in (0,1), \kappa_2 = 1,
\]

in which \( C_1 \sim C_6 \) are defined as in Theorem 3.

**Proof.** See Appendix.

The following corollary shows a selection of \( \alpha_t \) such that the DSCMD-N achieve the optimal rate in expectation under the case when the network has a communication noise decreasing rate \( r_t = \frac{1}{(t+1)^{1/2}} \).

**Corollary 1.** Under conditions of Theorem 3, suppose the sequences \( \{\alpha_t\} \), \( \{r_t\} \) in the DSCMD-N method are \( \alpha_t = \frac{1}{(t+1)^{1/2}} \) and \( r_t = \frac{1}{(t+1)^{1/2}} \), \( t = 1,2,\ldots,T \). 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 \( l \in V, T \geq 3 \), the DSCMD-N method achieves an expected rate of \( O(\frac{1}{T}) \) as follow:

\[
\mathbb{E}[F(\hat{x}_T^T)] - F(x^*) \leq 3C/\sqrt{T}.
\]

in which \( C_1 \sim C_6 \) are defined as in Theorem 3.

**Proof.** By using Proposition 1 to the case when \( \kappa_1 = 1/2 \) and \( \kappa_2 = 1 \), we have

\[
\mathbb{E}[F(\hat{x}_T^T)] - F(x^*) \leq \left( C_1 + 3C_6 \right) \frac{1}{T} + 4C_4 \frac{\ln T}{T} + \left( \sqrt{2}C_2 + 2\sqrt{2}C_3 + 2\sqrt{2}C_5 \right) \frac{1}{\sqrt{T}}.
\]

After taking the maximum of the coefficients \( C = \max\{C_1 + 3C_6, 4C_4, \sqrt{2}C_2 + 2\sqrt{2}C_3 + 2\sqrt{2}C_5\} \) and noting that \( \frac{1}{T} \leq \frac{\ln T}{T} \leq \frac{1}{T} \) when \( T \geq 3 \), the result is obtained.

**Remark 1.** In fact, for a general order pair \( (\kappa_1, \kappa_2) \) of \( \alpha_t = O(\frac{1}{(t+1)^{1/2}}) \), \( \kappa_1 \in (0,1) \) and \( \kappa_2 \in (0,1) \), by using similar idea with Corollary 1, we have a rate of \( O(\frac{1}{T}) \) with \( \kappa = \min\{1 - \kappa_1, 1/2 - \kappa_2, \kappa_1 - \kappa_2 - 1\} = \min\{1 - \kappa_1, 1/2 - \kappa_2, \kappa_2 - 1\} \). If \( 0 < \kappa < 1/2 \), then \( \kappa = \min\{1 - \kappa_1, 1/2 - \kappa_2, \kappa_2 - 1\} \leq 1/2 \) which presents a worse rate. If \( \kappa_1 = 1/2 \), then \( \kappa = \min\{1/2, \kappa_2 - 1/2\} = \kappa_2 - 1/2 < 1/2 \), for \( \kappa_1 < \kappa_2 < 1 \), which is also a worse rate than \( O(1/\sqrt{T}) \). Hence, the rate \( O(1/\sqrt{T}) \) can be obtained only when \( (\kappa_1, \kappa_2) = (1/2, 1) \).
Next, we consider the high probability convergence rate for DSCMD-N by presenting following results.

**Proposition 2.** Under conditions of Theorem\[3\] let the sequences \(\{\alpha_t\}, \{r_t\}\) in the DSCDA-N method be \(\alpha_t = \frac{1}{\sqrt{T}}\) and \(r_t = \frac{1}{t^2}, t = 1, 2, ..., T\). Then for any \(l \in \mathcal{V}, T \geq 3\), we have, for any \(\delta \in (0,1)\), with probability of at least \(1 - \delta\),
\[
F(\hat{x}_T^l) - F(x^*) \leq (C_1 + 3C_6) \frac{1}{T} + 4C_4 \ln T \left[ \sqrt{2C_2 + 2\sqrt{2}C_3 + 2\sqrt{2}C_5 + 2\sqrt{2}G_f X N \sqrt{\ln(1/\delta)} } \right] \frac{1}{\sqrt{T}},
\]
in which \(C_1 \sim C_6\) are defined as in Theorem\[3\].

**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 \(2\sqrt{2}G_f X N \sqrt{\ln(1/\delta)} / \sqrt{T}\) (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 2.** Under conditions of Proposition\[3\] for any \(\delta \in (0,1)\), set \(C_\delta = \max\{C_1 + 3C_6, 4C_4, \sqrt{2C_2 + 2\sqrt{2}C_3 + 2\sqrt{2}C_5 + 2\sqrt{2}G_f X N \sqrt{\ln(1/\delta)} }\}\). Then for any \(l \in \mathcal{V}, T \geq 3\), we have, for any \(\delta \in (0,1)\), with probability of at least \(1 - \delta\), the DSCMD-N method achieves the following rate
\[
F(\hat{x}_T) - F(x^*) \leq 3C_\delta / \sqrt{T}.
\]

**Proof.** The result follows directly from Proposition\[3\].

**Remark 2.** Now we make a comparison between the results on DSCMD-N in this work and main existing works in this literature\[3, 17\]. \[17\] is a seminal work on distributed optimization over noisy network. Both of the works\[3, 17\] consider standard distributed Euclidean projection-based algorithms to minimize the objective function \(\sum_{i=1}^N f_i(x)\) associated with local functions \(f_i, i \in \mathcal{V}\). Their approaches rely on a standard Robbins-Monro stepsize summability condition \(\sum_{t=0}^\infty \alpha_t = \infty\) and \(\sum_{t=0}^\infty \alpha_t^2 < \infty\) to ensure the almost sure convergence of \(\{x_t^l\}\) to the solution set \(X^*\). In this work, 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\[3, 17\], the geometric feature of the underlying space becomes easier to capture by selecting different types of mirror map (distance-generating function) \(\Phi\).

**Remark 3.** Here, we mention a very special case: when we consider regularizer \(\chi_1 = 0\) and mirror map \(\Phi = \frac{1}{2} \| \cdot \|^2\), then the algorithm degenerates to\[3\] if a zeroth-order gradient oracle is used. Moreover, we relax the aforementioned stepsize assumptions (hence the stepsize \(\alpha_t = \frac{1}{(t + 1)^2}\) with \(\kappa \in (1, 1/2]\) can be used, this stepsize can not be considered and used in\[3, 17\]) and derive the explicit convergence rate in expectation. On the way to the convergence in expectation, we also relax an assumption of noise \(\{\xi_{ij}\}\) in contrast to\[3\]. In fact, we do not require the martingale difference condition \(\mathbb{E}[\xi_{ij}^2] = 0\) to get 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 noisy network optimization setting.
4.2 DSCDA-N

To investigate some specific convergence rates of DSCDA-N. We start with the following result on expected error $E[F(x^*_T)] - F(x^*)$ for DSCDA-N. The result presents a general expected error bound in terms of total iteration steps $T$. The stepsizes sequence are in general form of $\alpha_t = \frac{1}{(t+1)^{cT}}$, the communication noise decreasing rates are in form of $r_t = \frac{1}{(t+1)^{cT}}$.

**Proposition 3.** Under conditions of Theorem £ Supposing the sequences $\{\alpha_t\}$ and $\{r_t\}$ in the DSCDA-N method satisfy $\alpha_t = \frac{1}{(t+1)^{cT}}$, $\alpha_t \in (0, 1)$, $r_t = \frac{1}{(t+1)^{cT}}$, $r_t \in (0, 1)$, $t = 1, 2, ..., T$. Denote $\kappa_m = \min\{\kappa_1, \kappa_2\}$. Suppose $\kappa_1 + \kappa_2 \neq 1$, $\kappa_1 + 2\kappa_2 \neq 1$. Then we have, for any $l \in \mathcal{V}$, $T \geq 2$, the following expected rate holds for the DSCDA-N method,

\[
E[F(x^*_T)] - F(x^*) \leq \left[\left(3G_f + G_\eta \left(\frac{N\omega}{1-\gamma} + 2\right)G_f + G_f^2\right)\right]^{1-\kappa_1} \frac{1}{1-\kappa_1} T^{\kappa_1} + \left[3G_f + G_\eta \frac{\omega N^2 \sqrt{\nu}}{1-\gamma} \frac{2^{1-\kappa_m}}{1-\kappa_m} T^{\kappa_1} \right]
\]

Proof. See Appendix.

We are ready to show the following important convergence rates in expectation for DSCDA-N.

**Corollary 3.** Under conditions of Proposition 3 for the DSCDA-N method, if $\kappa_1 \in (0, 1)$, $\kappa_2 \in (0, 1)$, and $\kappa_1 + \kappa_2 \neq 1$, $\kappa_1 + 2\kappa_2 \neq 1$, denote $\bar{\kappa} = \min\{\kappa_1, \kappa_2, 1-\kappa_1, 1\}$, set $c(\kappa_2)$ equals $\sqrt{\nu D}(2^\kappa) \kappa_1 = 1$ if we take

\[
\bar{C} = \max \left\{ \left[3G_f + G_\eta \left(\frac{N\omega}{1-\gamma} + 2\right)G_f + G_f^2\right]^{1-\kappa_1} \frac{1}{1-\kappa_1} T^{\kappa_1} + \left[3G_f + G_\eta \frac{\omega N^2 \sqrt{\nu}}{1-\gamma} \frac{2^{1-\kappa_m}}{1-\kappa_m} \right] \right\}
\]

we have the estimate: $E[F(x^*_T)] - F(x^*) \leq 7\bar{C}/T^\bar{\kappa}$, with $\bar{\kappa}$ satisfying $\bar{\kappa} \in (0, 1/2)$; Furthermore, when $\kappa_1 = 1/2$, $\kappa_2 \in (1/2, 1)$, if we take

\[
\bar{C}' = \max \left\{ \left[3G_f + G_\eta \left(\frac{N\omega}{1-\gamma} + 2\right)G_f + G_f^2\right]^{1-\kappa_1} \frac{1}{1-\kappa_1} T^{\kappa_1} + \left[3G_f + G_\eta \frac{\omega N^2 \sqrt{\nu}}{1-\gamma} \frac{2^{1-\kappa_m}}{1-\kappa_m} \right] \right\}
\]

we have $E[F(x^*_T)] - F(x^*) \leq 6\bar{C}'/\sqrt{T}$.
Proof. The result follows from the fact that \( \frac{1}{\sqrt{T+1}}, \frac{1}{T+1}, \frac{1}{T}, \frac{1}{T^2}, \frac{\ln T}{T}, \frac{1}{T^2} \leq \frac{1}{T} \) and \( \bar{r}_1 \leq \frac{1}{T^2} \). For the second argument, note that, when \( \kappa_1 = \frac{1}{T} < \kappa_2 \), we have \( \kappa_m = \frac{1}{T^2} \) and \( \frac{1}{T^{2+i}}, \frac{1}{T^{2+i+1}}, \frac{1}{T^{2+i+2}}, \frac{1}{T^{2+i+3}}, \frac{1}{T^{2+i+4}}, \frac{1}{T^2}, \frac{1}{T^3}, \frac{1}{T^4} \leq \frac{1}{T^2} \), then the result follows.

Next, we investigate the high probability convergence rate of DSCDA-N when \( \alpha_t = \frac{1}{\sqrt{T+1}} \), \( r_t = \frac{1}{(\frac{1}{T+1})^2} \), \( \kappa_2 \in (0, 1] \). The following result describes a detailed relation between the stepsize sequence, communication noise decreasing rate and the high probability rate of the DSCDA-N method. It can be seen that, when \( \kappa_2 \) (the order index for network noise decreasing rate) lies in three different ranges \( (1, 1/2), (1/2, 1), (0, 1/2) \), three totally different high probability \( T \)-rates are derived as follows.

**Corollary 4.** Under conditions of Theorem \[4\], if the stepsize \( \{\alpha_t\} \) and decreasing rate \( \{r_t\} \) of communication noise satisfy \( \alpha_t = \frac{1}{\sqrt{T+1}}, r_t = \frac{1}{(\frac{1}{T+1})^2}, \kappa_2 \in (0, 1] \). Then we have, for any \( \delta \in (0, 1), T \geq 3 \), with probability of at least \( 1 - \delta \),

\[
F(\bar{x}_1^T) - F(x^*) \leq \begin{cases} 
3\bar{C}_1(\delta)/T\kappa_2, & \kappa_2 \in (0, 1/2); \\
4\bar{C}_2(\delta)/\sqrt{\ln T}/\sqrt{T}, & \kappa_2 = 1/2; \\
2\bar{C}_3(\delta)/\sqrt{T}, & \kappa_2 \in (1/2, 1], 
\end{cases}
\]

in which

\[
\bar{C}_1(\delta) = \max\{\left[3\bar{G}_\gamma + \bar{G}_{\psi} \left( \frac{N\varphi}{1-\gamma} + 2 \right) \bar{G}_f + G_f^2 \right] 2\sqrt{2} + \Psi(x^*) \sqrt{2} + 2\sqrt{2} \bar{G}_f \bar{D}_X \sqrt{\log \frac{2}{\delta}} + \frac{3\bar{G}_\gamma + \bar{G}_{\psi} \nu D_X \bar{D}_X \sqrt{2 \kappa_2}}{\sqrt{2 \kappa_2}}, \frac{3\bar{G}_\gamma + \bar{G}_{\psi} \omega N^2 X \sqrt{2 \kappa_2}}{\sqrt{2 \kappa_2}}, \nu D_X \}
\]

\[
\bar{C}_2(\delta) = \max\{\left[3\bar{G}_\gamma + \bar{G}_{\psi} \left( \frac{N\varphi}{1-\gamma} + 2 \right) \bar{G}_f + \frac{\omega N^2 X \sqrt{2 \kappa_2}}{1-\gamma} \right] 2\sqrt{2} + \Psi(x^*) \sqrt{2} + 2\sqrt{2} \bar{G}_f \bar{D}_X \sqrt{\log \frac{2}{\delta}}, \frac{3\bar{G}_\gamma + \bar{G}_{\psi} \nu D_X \sqrt{2 \kappa_2}}{\sqrt{2 \kappa_2}}, \nu D_X \}
\]

\[
\bar{C}_3(\delta) = \max\{\left[3\bar{G}_\gamma + \bar{G}_{\psi} \left( \frac{N\varphi}{1-\gamma} + 2 \right) \bar{G}_f + \frac{\omega N^2 X \sqrt{2 \kappa_2}}{1-\gamma} \right] 2\sqrt{2} + \Psi(x^*) \sqrt{2} + 2\sqrt{2} \bar{G}_f \bar{D}_X \sqrt{\log \frac{2}{\delta}} + \frac{\sqrt{2} \nu D_X \bar{D}_X \sqrt{2 \kappa_2}}{\sqrt{2 \kappa_2}}, \frac{3\bar{G}_\gamma + \bar{G}_{\psi} \omega N^2 X \sqrt{2 \kappa_2}}{\sqrt{2 \kappa_2}}, \nu D_X \}
\]

Proof. See Appendix. ■

**Remark 4.** The above Corollary presents obvious effects of the decreasing rate of communication noise on the high probability rates of DSCDA-N. When \( \kappa_2 \) is sufficiently close to 0 (which means that, the communication noise has a very slow decreasing rate), the corresponding high probability convergence speed is very slow with \( 3\bar{C}_1(\delta)/T\kappa_2 \) with probability at least \( 1 - \delta \). Meanwhile, the corollary presents three different types of rates under different noise decreasing rates.

**Remark 5.** Based on aforementioned results, we make an observation on stepsize \( \alpha_t = O\left(\frac{1}{\sqrt{T+1}}\right) \), \( \kappa_1 \in (0, 1) \) and noise decreasing rate \( r_t = O\left(\frac{1}{(\frac{1}{T+1})^2}\right) \), \( \kappa_2 \in (0, 1] \) in DSCMD-N and DSCDA-N methods. If we consider the setting when composite terms \( \chi_i = 0, \ i = 1, 2, ..., N \) in Problem \[1.4] \) and term \( \eta = 0 \) in Problem \[1.2] \), it can be seen that, to achieve a rate of \( O(1/\sqrt{T}) \), for both DSCMD-N and DSCDA-N, order index \( \kappa_1 \) should be coercively set to be 1/2. However, for index \( \kappa_2 \), for DSCMD-N, the only case for \( \kappa_2 \) for a rate of \( O(1/\sqrt{T}) \) is \( \kappa_2 = 1 \), other values \( \kappa_2 \) would make the rate worse than \( O(1/\sqrt{T}) \). This shows a very tough condition on the behavior of communication noise decreasing for DSCMD-N to obtain optimal convergence rate. Meanwhile, for DSCDA-N, the value of \( \kappa_2 \) in \( (1/2, 1] \) all can make the rate be \( O(1/\sqrt{T}) \). This fact shows that DSCDA-N has a wider range of communication noise decreasing index \( \kappa_2 \) for an optimal rate of \( O(1/\sqrt{T}) \) in the literature. This fact reflects some potential advantages of DSCDA-N over DSCMD-N in some noisier network optimization settings.
Remark 6. In contrast to existing works on noisy network optimization, the paper also considers composite terms that serve as regularization terms for two types of composite optimization problem (local regularizer $\chi_i$, $i = 1, 2, ..., N$ in Problem (1.1) and global regularizer term $\eta$ in Problem (1.2)). The utilization of the regularization terms makes the methods more flexible to present some structure types of the solution of optimization problem. Meanwhile, the structure of Problem (1.1) and DSCMD-N method allow the regularization term $\chi$ to be independent of each other. There are several choices of $\chi_i$, $\eta$ that are often considered to promote different structure types of solutions of optimization problem. For example, the indicator function of $X_i$, $I_X(x)$; The $l^p$-norm squared function $\frac{1}{p}\|x\|^p$, $p \in (1, 2]$; Sparsity inducing regularizer $\lambda \|x\|_1$, $\lambda > 0$; $l^\infty$-norm $\lambda \|x\|_\infty$, $\lambda > 0$; entropy function $\sum_{i=1}^m[x_i]\log[x_i]$; mixed regularizer $\frac{1}{2}\|x\|^2 + \lambda_2\|x\|_1$, $\lambda_1, \lambda_2 > 0$.

Till now, we observe that all the approximating sequences of convergence results in this paper are in weighted average form $\hat{x}_t = \frac{1}{t}\sum_{i=1}^t x_i^t$, $i = 1, 2, ..., N$. 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 $\{x_i^t\}$ or $\{x_i^t\}$ in distributed composite optimization setting? To this end, we provide following almost sure convergence results for DSCMD-N. The counterpart result for DSCDA-N can be obtained in a similar way. 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 5. Under conditions of Theorem 4 in the DSCMD-N method, suppose the stepsize sequences $\{\alpha_t\}$, noise decreasing rate $\{\eta_t\}$ are $\alpha_t = \frac{1}{\sqrt{t+1}}$ and $\eta_t = \frac{1}{t+1}$, $t = 1, 2, ..., T$. If we take $C = \max\{C_1 + 3C_0, 4C_1, \sqrt{2}C_2 + 2\sqrt{2}C_3 + 2\sqrt{2}C_5\}$, in which $C_1 \sim C_0$ is as in Theorem 7. Then for any $t \in V$, $T \geq 3$, for sequence $\{x_i^t\}$ and their average version $\{\hat{x}_i^t\}$ generated from DSCMD-N method, we have almost surely,

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

and

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

Proof. See Appendix. $\blacksquare$

5 Conclusion

This paper studies the noisy network optimization problems. Two typical related distributed stochastic composite optimization problems over noisy network are considered. Two new methods DSCMD-N and DSCDA-N are presented to solve them respectively. Convergence of the methods are systematically studied. Novel convergence rates are obtained in several different situations under different detailed discussions on stepsize $\{\alpha_t\}$ and communication noise decreasing rate $\{\eta_t\}$. These 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 are obvious in stochastic circumstances.

6 Appendix

Proof of Proposition 1

22
Proof. Note that, under the conditions that $0 < \kappa_1 < \kappa_2 < 1$ and $2\kappa_2 - \kappa_1 \neq 1$, 
\[
\frac{1}{\kappa_2^2} = \frac{(T+1)^{\kappa_1}}{1+T}, \quad T \geq 1, \quad \frac{1}{\kappa_1^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1, \quad \frac{1}{\kappa_1^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1.
\]

Theorem 1, we obtain the first argument. For the second argument, we substitute these bounds into Theorem 3, the desired estimate is obtained after substituting these estimates into Theorem 1 again.

Proof of Proposition 3.

Proof. Substitute the stepsize sequences \(\{\alpha_t\}, \{r_t\}\) into the bounds in Theorem 3

\[
\frac{1}{\kappa_1^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad \frac{1}{\kappa_1^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad \frac{1}{\kappa_1^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1.
\]

Note that, under the conditions that $0 < \kappa_1 < \kappa_2 < 1$ and $2\kappa_2 - \kappa_1 \neq 1$, 
\[
\frac{1}{\kappa_2^2} = \frac{(T+1)^{\kappa_1}}{1+T}, \quad T \geq 1, \quad \frac{1}{\kappa_1^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1, \quad \frac{1}{\kappa_1^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1.
\]

Theorem 1, we obtain the first argument. For the second argument, we substitute these bounds into Theorem 3, the desired estimate is obtained after substituting these estimates into Theorem 1 again.

Proof of Corollary 4.

Proof. When $\kappa_2 \in (0, 1/2)$, \(\sum_{t=1}^T r_t^2 = \sum_{t=1}^T (2^{-1/2} + 2^{-1/2})^2 \leq \int_0^T \frac{dt}{(1+T)^{\kappa_2}} \leq \frac{(T+1)^{2-\kappa_2}}{1-2\kappa_2}, \) then we have 
\[\int \left(\sum_{t=1}^T r_t^2\right)^{1/2} \leq \frac{(T+1)^{2-\kappa_2}}{1-2\kappa_2} \leq \frac{2}{1-2\kappa_2} \frac{T^{2-\kappa_2}}{1-2\kappa_2}.\]

Combine this estimate with similar estimates in the proof of Proposition 3 with $\kappa_1 = 1/2 > \kappa_2$, we have, with probability of at least $1 - \delta$, 
\[F(\hat{x}_T^n) - F(x^*)\]
\[\leq \left[3G_f + G_n \left(\frac{N\omega}{1-\gamma} + 2G_f + G_f^2\right)\right] \frac{2\sqrt{2}}{\sigma_\phi} \frac{1}{\sqrt{T}} + \left[3G_f + G_n \frac{\omega^2 N^2 \sqrt{\nu}}{1+\kappa_2} \frac{1}{1-\gamma} \frac{1}{1-\kappa_2} \frac{T^{2-\kappa_2}}{1-2\kappa_2}\right] \frac{1}{\sqrt{T}} + \left[3G_f + G_n \frac{N\omega}{1-\gamma} + 2G_f + G_f^2\right] \frac{2\sqrt{2}}{\sigma_\phi} \frac{1}{\sqrt{T}} + \left[\frac{2\sqrt{2}}{\sigma_\phi} \frac{1}{\sqrt{T}} + \frac{N\nu}{\kappa_2} \frac{1}{\sqrt{T}} \frac{2}{1-2\kappa_2} \frac{T^{2-\kappa_2}}{1-2\kappa_2} \right] \frac{1}{\sqrt{T}}
\]

After taking $\tilde{C}_1(\delta)$ as defined in Corollary 3 and noting that 
\[\frac{1}{1-\delta} < \frac{1}{\sqrt{T}} < \frac{1}{\sqrt{1-2\kappa_2}}
\]
we obtain the first argument. When $\kappa_2 = 1/2$, note that, 
\[
\frac{1}{\kappa_2^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1, \quad \frac{1}{\kappa_2^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1, \quad \frac{1}{\kappa_2^2} = \frac{(T+1)^{\kappa_2}}{1+T}, \quad T \geq 1.
\]
when $T \geq 3$ (same procedure implies $\sum_{t=1}^{T} r_{t}^{2} = \sum_{t=1}^{T} \frac{1}{T} \leq \frac{2 \ln(T)}{T}$ and $\frac{1}{T} \sum_{t=0}^{T} \frac{\sqrt{t}}{(t+1)^{3/2}} \leq \frac{1}{T} (1 + \int_{0}^{\infty} \frac{dt}{(t+1)^{3/2}}) \leq \frac{3}{4}$. Substitute these estimates into Theorem 4, we have, for $T \geq 3$, with probability of at least $1 - \delta$,

\[
F(\hat{x}_{1}^{T}) - F(x^{*}) \\
\leq \left[ \frac{3G_{f} + G_{g}}{\sigma_{\psi}} \left( \frac{N}{1 - \gamma} + 2 \right) G_{f} + \frac{\omega N^{2} \sqrt{\nu}}{1 - \gamma} + G_{f}^{2} \right] \frac{2 \sqrt{2}}{\sqrt{T}} + \left[ \sqrt{2} + \frac{3G_{f} + G_{g}}{\sigma_{\psi}} 4N \sqrt{\nu} \ln(T) + \Psi(x^{*}) \sqrt{2} \right] \frac{1}{\sqrt{T}} \\
+ \frac{3N\nu}{T} + \frac{2 \sqrt{2} G_{f} D_{x}}{\sqrt{T}} \frac{\log(2/\delta)}{\sqrt{T}} + 2N \sqrt{D_{x}} \sqrt{\log(2/\delta)} \frac{1}{\sqrt{T}}.
\]

After taking $\bar{C}_{2}(\delta)$ as defined in Corollary 4 and noting that $\frac{1}{T} \leq \frac{\ln(T)}{T} \leq \frac{\sqrt{10}}{\sqrt{T}}$ when $T \geq 3$, the second argument is obtained. For $\kappa_{2} \in (1/2, 1]$, note that $\frac{1}{T} \sum_{t=1}^{T} \alpha_{t} r_{t-1} \leq \sum_{t=1}^{\infty} \frac{1}{(t+1)^{3/2}} \leq \frac{2\kappa_{2} + 1}{2\kappa_{2} - 1}$, $\sum_{t=1}^{T} r_{t}^{2} \leq \int_{0}^{\infty} \frac{dt}{(t+1)^{3/2}} \leq \frac{1}{2\kappa_{2} - 1}$. Substitute these estimates into Theorem 4, we have

\[
F(\hat{x}_{1}^{T}) - F(x^{*}) \\
\leq \left[ \frac{3G_{f} + G_{g}}{\sigma_{\psi}} \left( \frac{N}{1 - \gamma} + 2 \right) G_{f} + \frac{\omega N^{2} \sqrt{\nu}}{1 - \gamma} + G_{f}^{2} \right] \frac{2 \sqrt{2}}{\sqrt{T}} + \left[ \sqrt{2} + \frac{3G_{f} + G_{g}}{\sigma_{\psi}} 4N \sqrt{\nu} \ln(T) + \Psi(x^{*}) \sqrt{2} \right] \frac{1}{\sqrt{T}} \\
+ \frac{3N\nu}{T} + \frac{2 \sqrt{2} G_{f} D_{x}}{\sqrt{T}} \frac{\log(2/\delta)}{\sqrt{T}} + 2N \sqrt{D_{x}} \sqrt{\log(2/\delta)} \frac{1}{\sqrt{T}}.
\]

The result follows after a selection of $\bar{C}_{2}(\delta)$ as defined in Corollary 4.

Proof of Corollary 4

\textbf{Proof.} From Corollary 4, we have $\mathbb{E}[F(\hat{x}_{1}^{T})] - F(x^{*}) \leq 3C/\sqrt{T}, \ i \in V$. This implies $\lim_{T \to \infty} \left( \mathbb{E}[F(\hat{x}_{1}^{T})] - F(x^{*}) \right) = 0$. Since $F(\hat{x}_{1}^{T}) - F(x^{*})$ is nonnegative for all $T \geq 0$, by applying Fatou lemma we arrive at $\lim_{T \to \infty} F(\hat{x}_{1}^{T}) = F(x^{*})$, a.s. On the other hand, we have already assumed that $X$ is bounded and $F$ is continuous (since $F_{1}$ is continuous for all $i \in V$). Weierstrass Theorem implies that the accumulation point set of $\{\hat{x}_{1}^{T}\}$ exists. The above inequality and the continuity of $F$ implies at least one of the accumulation points minimizes the summation of Problem (1), which means

$$
\lim_{T \to \infty} \text{dist}(\hat{x}_{1}^{T}, x^{*}) = 0, \ i \in V.
$$

Due to the convexity of $F$ and the fact that $\hat{x}_{1}^{T}$ is a convex combination of $x_{1}^{T}, x_{2}^{T}, ..., x_{n}^{T}$. We have $\min_{1 \leq i \leq T} F(x_{i}^{T}) \leq F(\hat{x}_{1}^{T}), \ i \in V$. Then it follows that $0 \leq \min_{1 \leq i \leq T} F(x_{i}^{T}) - F(x^{*}) \leq F(\hat{x}_{1}^{T}) - F(x^{*}), \ i \in V$. After taking expectation on both sides, we have

$$
0 \leq \mathbb{E} \left[ \min_{1 \leq i \leq T} F(x_{i}^{T}) - F(x^{*}) \right] \leq \mathbb{E} \left[ F(\hat{x}_{1}^{T}) - F(x^{*}) \right], \ i \in V.
$$

(6.2)
Take limit on both sides of above inequality and use Squeeze theorem, we have \( \lim_{T \to \infty} \mathbb{E}\left[ \min_{1 \leq t \leq T} F(x^i_t) - F(x^*) \right] \). Use Fatou lemma again, we have \( \lim_{T \to \infty} \mathbb{E}\left[ \min_{1 \leq t \leq T} F(x^i_t) - F(x^*) \right] = 0 \). Since \( \min_{1 \leq t \leq T} F(x^i_t) \) is a lower bounded non-increasing sequence in \( T \), hence \( \lim_{T \to \infty} \min_{1 \leq t \leq T} F(x^i_t) \) exists and \( \lim_{T \to \infty} \min_{1 \leq t \leq T} F(x^i_t) = \lim_{T \to \infty} \min_{1 \leq t \leq T} F(x^i_t) = F(x^*), \quad i \in V \). Then, using similar argument of getting (6.1), we arrive at \( \lim_{T \to \infty} \left( \min_{1 \leq t \leq T} \text{dist}(x^i_t, X^*) \right) = 0, \quad i \in V \).

References

[1] A. Nedic and A. Ozdaglar, Distributed subgradient methods for multi-agent optimization, IEEE Trans. on Automat. Control, 54(2009), pp. 48-61.

[2] A. Nedic and A. Olshevsky, Stochastic gradient-push for strongly convex functions on time-varying directed graphs, IEEE Trans. on Automat. Control, 61(2016), pp. 3936-3947.

[3] A. Nedic and S. Lee, On stochastic subgradient mirror-descent algorithm with weighted averaging, SIAM J. Optim., 24(2014), pp. 84-107.

[4] B. Touri and A. Nedic, Distributed consensus over network with noisy links, 12th International Conference on Information Fusion, IEEE, (2009), pp. 146-154.

[5] C. Xi and U. A. Khan, Distributed subgradient projection algorithm over directed graphs, IEEE Trans. on Automat. Control, 62(2016), pp. 3986-3992.

[6] K. Cai and H. Ishii, Average consensus on general strongly connected digraphs, Automatica, 48(2012), pp. 2750-2761.

[7] D. P. Bertsekas, J. N. Tsitsiklis, Parallel and distributed computation: numerical methods, 23(1889), Englewood Cliffs, NJ: Prentice hall.

[8] D. Wang, J. Zhou, Z. Wang, and W. Wang, Random gradient-free optimization for multiagent systems with communication noises under a time-varying weight balanced digraph, IEEE Trans. Systems Cybernet: Systems, accepted, (2018).

[9] D. P. Palomar, M. Chiang, (2007). Alternative distributed algorithms for network utility maximization: Framework and applications, IEEE Trans. on Automat. Control, 52(2017), pp. 2254-2269.

[10] 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., 54(2016), pp. 2872-2892.

[11] D. Yuan, Y. Hong, D. W. C Ho, and S. Xu, Distributed mirror descent for online composite optimization, IEEE Trans. on Automat. Control, DOI: 10.1109/TAC.2020.2987379.

[12] D. G. Luenberger, Microeconomic theory, Mcgraw-Hill College, (1995).

[13] H. J. Kushner, Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo, SIAM J. Applied Math., 47(1987), pp. 169-185.

[14] J. N. Tsitsiklis, D. Bertsekas, and M. Athans, Distributed asynchronous deterministic and stochastic gradient optimization algorithms, IEEE Trans. on Automat. Control, 31(1986), pp. 803-812.
[15] J. N. Tsitsiklis, *Problems in decentralized decision making and computation*, Ph.D. dissertation, Massachusetts Institute of Technology, (1984).

[16] J. C. Duchi, A. Alekh, and J. W. Martin, *Dual averaging for distributed optimization: Convergence analysis and network scaling*, IEEE Trans. on Automat. Control, 2011 [Online] version.

[17] K. Srivastava, A. Nedic, D. M. Stipanovic, *Distributed constrained optimization over noisy networks* In 49th IEEE Conference on Decision and Control (CDC) IEEE, (2010), pp. 1945-1950.

[18] M. G. Rabbat, R. D. Nowak, *Decentralized source localization and tracking [wireless sensor networks]*. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 3). IEEE.

[19] M. Akbari, B. Gharesifard, and T. Linder, *Distributed online convex optimization on time-varying directed graphs*, IEEE Trans. Control of Network Systems, 4(2015), pp. 417-428.

[20] P. Di Lorenzo and G. Scutari, *Distributed nonconvex optimization over time-varying networks*, In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2016), pp. 4124-4128.

[21] P. Bianchi, W. Hachem, and F. Iutzeler, *A stochastic primal-dual algorithm for distributed asynchronous composite optimization*, In 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), (2014), pp. 732-736.

[22] P. Yi, Y. Hong, and F. Liu, *Distributed gradient algorithm for constrained optimization with application to load sharing in power systems*. Systems Control Lett., 83(2015), pp. 45-52.

[23] S. S. Ram, A. Nedic, and V. V Veeravalli, *Distributed stochastic subgradient projection algorithms for convex optimization*, Journal of Optim. Theory and Applications, 147(2010), pp. 516-545.

[24] S. Shahrampour, A. Jadbabaie, *Distributed online optimization in dynamic environments using mirror descent*, IEEE Trans. on Automat. Control, 63(2017), pp. 714-725.

[25] S. Lee, S. A. Nedic, M. Raginsky, *Stochastic dual averaging for decentralized online optimization on time-varying communication graphs*, IEEE Trans. on Automat. Control, 62(2017), pp. 6407-6414.

[26] S. Liu, P. Y. Chen, and A. O. Hero, *Accelerated distributed dual averaging over evolving networks of growing connectivity*, IEEE Trans. Signal Process., 66(2018), 1845-1859.

[27] S. Pu and A. Nedic, *Distributed Stochastic Gradient Tracking Methods*, Math. Program., (2020) pp. 1-49.

[28] T. H. Chang, and A. Nedic, *Distributed constrained optimization by consensus-based primal-dual perturbation method*. IEEE Trans. on Automat. Control, 59(2014), pp. 1524-1538.

[29] T. Tatarenko and B. Touri, *Non-convex distributed optimization*. IEEE Trans. on Automat. Control, 62(2017), 3744-3757.

[30] W. Shi, Q. Ling, G. Wu, and W. Yin, *A proximal gradient algorithm for decentralized composite optimization*, IEEE Trans. on Signal Process., 63(2015), pp. 6013-6023.
[31] W. Zhang, P. Zhao, W. Zhu, S. C. Hoi, and T. Zhang, *Projection-free distributed online learning in networks*, Proceedings of the 34th International Conference on Machine Learning, 70(2017), pp. 4054-4062.

[32] X. Zeng, P. Yi, Y. Hong, and L. Xie, *Distributed continuous-time algorithms for non-smooth extended monotropic optimization problems*, SIAM J. Control Optim., 56(2018), pp. 3973-3993.

[33] Y. Tang, N. Li, *Distributed zero-order algorithms for nonconvex multi-agent optimization*, 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), (2019), pp. 781-786.

[34] Y. Wang, W. Zhao, Y. Hong, and M. Zamani, *Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks*, SIAM J. Control Optim., 57(2019), pp. 2821-2842.

[35] Y. He, L. Wang, and D. Xu, *Chaotic simulated annealing with decaying chaotic noise*, IEEE Trans. on Neural Networks, 13(2002), pp. 1526-1531.

[36] Y. Zhang, Y. Lou, Y. Hong and L. Xie, *Distributed projection-based algorithms for source localization in wireless sensor networks*, IEEE Trans. Wireless Communications, 14(2015), pp. 3131-3142.