Analysis of Decentralized Stochastic Successive Convex Approximation for composite non-convex problems

Basil M. Idrees*, Shivangi Dubey Sharma*, and Ketan Rajawat

Abstract

Successive Convex approximation (SCA) methods have shown to improve the empirical convergence of non-convex optimization problems over proximal gradient-based methods. SCA uses a strongly convex surrogate and offers a more flexible framework to solve such optimization problems. Further, in decentralized optimization, which aims to optimize a global function using only local information, the SCA framework has been successfully applied to achieve improved convergence. Still, the stochastic first order (SFO) complexity decentralized SCA algorithms have remained under-studied. While non-asymptotic convergence analysis has been studied for decentralized deterministic settings, its stochastic counterpart has only been shown to converge asymptotically.

We have analyzed a novel accelerated variant of the decentralized stochastic SCA that minimizes the sum of non-convex (possibly smooth) and convex (possibly non-smooth) cost functions. The algorithm viz. **D**ecentralized **M**omentum-based **S**tochastic **SCA** (**D-MSSCA**), iteratively solves a series of strongly convex subproblems at each node using one sample at each iteration. The key step in non-asymptotic analysis involves proving that the average output state vector moves in the descent direction of the global function. This descent allows us to obtain a bound on average *iterate progress* and *mean-squared stationary gap*. The recursive momentum-based updates at each node contribute to achieving stochastic first order (SFO) complexity of $\mathcal{O}(\epsilon^{-3/2})$ provided that the step sizes are smaller than the given upper bounds. Even with one sample used at each iteration and a non-adaptive step size, the rate is at par with the SFO complexity of decentralized state-of-the-art gradient-based algorithms. The rate also matches the lower bound for the centralized, unconstrained optimization problems. Through a synthetic example, the applicability of D-MSSCA is demonstrated.

Keywords

Decentralized, consensus, stochastic, non-convex optimization.

I. Introduction

We consider the following decentralized stochastic non-convex composite optimization problem:

$$U^\star = \min_{\mathbf{x}\in\mathbb{R}^d} \quad U(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} u_i(\mathbf{x}) + h(\mathbf{x}) \tag{$\mathcal{P}$}$$

$$\text{s.t.} \quad g(\mathbf{x}) \leq 0$$

* both the authors have equal contribution

where $u_i(\mathbf{x}) = \mathbb{E}\left[f_i(\mathbf{x}, \boldsymbol{\xi}_i)\right]$. Local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$ is assumed to be non-convex, smooth and known only to agent i. Regulariser $h : \mathbb{R}^d \to \mathbb{R}$ is a convex and possibly non-smooth function and constraint $g : \mathbb{R}^d \to \mathbb{R}$ is a convex function. Also, both g, h are publicly known. The form of problem ($\mathcal{P}$) arises in multiple areas, such as statistical inference, decision-making in sensor networks, and machine learning problems [1].

Existing methods to solve ($\mathcal{P}$) include projected and proximal stochastic gradient methods [2]–[7] and Successive Convex Approximation (SCA) [8], [9] methods. The performance of these algorithms is measured in terms of the number of stochastic first order (SFO) oracle calls required to reach a ϵ- Karush-Kuhn Tucher (KKT) point. While [2] proposes a projected DSGD-type algorithm for problems with a compact constraint set, [3] goes ahead and establishes the asymptotic convergence of DSGD for a family of non-convex, non-smooth functions. Further, in [4], a decentralized stochastic proximal primal-dual method called SPPDM is proposed, assuming that the epigraph of h is a polyhedral set. Only three works address non-asymptotic iteration complexity analysis for stochastic non-convex composite problems with a general convex non-differentiable regularizer h. While DProxSGT [6] achieves sub-optimal rate of $\mathcal{O}(\epsilon^{-2})$ without mean-squared smoothness assumption, ProxGT-SR-O/E [5] and DEEPSTORM [7] achieve an optimal convergence rate of $\mathcal{O}(\epsilon^{-3/2})$. However, the problem with ProxGT-SR-O/E [5] is that it uses large batches and more communication rounds at each iteration. Even though DEEPSTORM [7] overcomes the use of more communications rounds, it still uses small batches to achieve the optimal rate.

Unlike the above methods, SCA methods offer a more flexible framework to solve non-convex optimization problems since the first work done by [10]. At each iteration, SCA solves a convexified sub-problem formed by approximating the non-convex functions using convex functions called *surrogates*. Different from other competitive algorithms like Expectation-Minimization (EM) and Majorization-Minimization (MM) SCA offers more freedom in the choice of surrogates which can be tailored to a specific problem at hand [11], [12]. Even though there is a rich body of work on SCA [13]–[23], their non-asymptotic analysis has largely remained understudied. Under stochastic centralized settings, through non-asymptotic convergence analysis of SCA it was shown that AsySCA [22] archives a rate of $\mathcal{O}(\epsilon^{-2})$. Further, combining accelerated momentum-based updates with SCA has improved the rate to $\mathcal{O}(\epsilon^{-3/2})$ in [23].

Under decentralized settings, there are only a handful of SCA algorithms [8], [9], [24] including both deterministic and stochastic cases. In [24], the authors proposed an SCA-based decentralized algorithm, NEXT, . A stochastic variant, S-NEXT, was proposed in [9]; however, they did not apply the momentum to the update steps. In both works, only asymptotic convergence has been proven. Recently, [8] introduced a decentralized momentum-based algorithm that employs Nesterov-like momentum, providing the first non-asymptotic analysis of decentralized SCA methods for deterministic case. However, their analysis introduced a new metric tailored to SCA, and the detailed proofs were only provided for the case where the functions u_i-s are convex. To the best of our knowledge, there is no comprehensive non-asymptotic convergence analysis for decentralized consensus stochastic non-convex problems within the SCA rubric in the literature.

In this work, we have analyzed a novel Decentralized Momentum-based Stochastic SCA (D-MSSCA) algorithm to solve decentralized stochastic non-convex composite optimization problems. The D-MSSCA hinges on SCA techniques and iteratively solves a convexified subproblem at each node using recursive momentum [25] type local

TABLE I

COMPARISON OF ORACLE COMPLEXITIES OF DECENTRALIZED CONSENSUS STOCHASTIC NON-CONVEX COMPOSITE OPTIMIZATION ALGORITHMS FOR EXPECTATION (POPULATION RISK) PROBLEMS. (TO MAKE COMPARISONS FAIR, WE HAVE CONVERTED THE SFO-COMPLEXITIES OF ALL THE ALGORITHMS TO MATCH OUR DEFINITION OF (6))

Algorithm	**SFO complexity**	**Asymptotic/ Non-Asymptotic**	**Remarks**
projected DSGD [2]	-	Asymptotic	compact constraint set
[3]	-	Asymptotic	family of non-convex nonsmooth functions
SPPDM [4]	$\mathcal{O}(\epsilon^{-2})$	Non-Asymptotic	epigraph of h is polyhedral
ProxGT-SR-O/E [5]	$\mathcal{O}(\epsilon^{-3/2})$	Non-Asymptotic	Multiple communication, larger batches per iteration
S NEXT [9]	-	Asymptotic	SCA based
DProxSGT [6]	$\mathcal{O}(\epsilon^{-2})$	Non-Asymptotic	without MSS assumption
DEEPSTORM [7]	$\mathcal{O}(\epsilon^{-3/2})$	Non-Asymptotic	gradient-based, small batches per iteration
D-MSSCA (This work)	$\mathcal{O}(\epsilon^{-3/2})$	Non-Asymptotic	SCA based

gradient updates to reach the $\epsilon-$KKT point with an optimal convergence rate of $\mathcal{O}(\epsilon^{-3/2})$. Our analysis extends the methods used in gradient-based approaches [5], [7] to SCA framework. The key challenge in the convergence analysis was to form a descent inequality for the global function U at the average state-vector of D-MSSCA in terms of *iterate progress*, which we have overcome by utilizing the strong convexity of the surrogate. This descent helped us in bounding the mean-squared stationary gap, a more general metric than that used in [8] for analyzing non-convex functions. Unlike the Nesterov updates used in [8], our use of recursive momentum-type updates significantly contributes to this advancement. Finally, simulations on a synthetic problem empirically validate the theoretical findings.

A comparative performance of various state-of-the-art algorithms that can be used to solve ($\mathcal{P}$) has been provided in Table I. It can be observed that even with one sample per iteration, the proposed D-MSSCA algorithm is able to achieve the optimal convergence rate. Additional remarks are also provided in the table.

A. Notations

We denote vectors (matrices) using lowercase (uppercase) bold font letters. For a vector $\mathbf{x}$, we denote its transpose by $\mathbf{x}^\mathsf{T}$ and its i-th element by $[\mathbf{x}]_i$. Likewise, the (i,j)-th component of $\mathbf{A}$ is given by A_{ij}. An n-dimensional identity matrix is denoted by $\mathbf{I}_n$. The n-dimensional all-one vector is denoted by $\mathbf{1}_n$. The Kronecker product is denoted using $\otimes$. The d-dimensional average of any $nd-$dimensional vector $\mathbf{a}\in\mathbb{R}^{nd}$, is represented by $\bar{\mathbf{a}}=\frac{1}{n}\left(\mathbf{1}_n^\mathsf{T}\otimes\mathbf{I}_d\right)\mathbf{a}\in\mathbb{R}^d$. For a vector $\mathbf{y}$, $\|\mathbf{y}\|$ denotes its ℓ_2 (Euclidean) norm, Maximum eigenvalue of $\mathbf{A}$ is represented by $\lambda_{\max}(\mathbf{A})$. The set of neighbours of node i is denoted by $\mathcal{N}_i$. For a non-smooth function $h+\mathbf{1}_\mathcal{X}$, $\partial\left(h+\mathbf{1}_\mathcal{X}\right)|_{\mathbf{x}=\mathbf{a}}$ represents the set of sub-gradients of $h(x)+\mathbf{1}_\mathcal{X}(x)$ at $\mathbf{x}=\mathbf{a}$. For the ease of writing we have defined $\nabla_x\tilde{f}(\mathbf{x},\mathbf{x}_i^t,\boldsymbol{\xi}_i^t)\,|_{\mathbf{x}=\mathbf{a}}:=\boldsymbol{\nabla}\tilde{f}(\mathbf{a},\mathbf{x}_i^t,\boldsymbol{\xi}_i^t)$.

The rest of the paper is organized as follows. Section II discusses the proposed algorithm, and the assumptions on the problem are discussed. In Section III, the convergence proof of the proposed algorithm is presented. Section IV briefly describes the proposed algorithm's applicability to a synthetic problem.

II. Proposed method

A. Problem

Consider a network of n agents or nodes communicating over a fixed undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is a set of edges or links. An edge $(i, j) \in \mathcal{E}$ represents a communication link between nodes i and j. Now we re-write the decentralized problem ($\mathcal{P}$) as,

$$\min_{\mathbf{x} \in \mathcal{X}} \quad \frac{1}{n} \sum_{i=1}^{n} u_i(\mathbf{x}) + h(\mathbf{x}) \tag{$\mathcal{P}_c$}$$

where $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d \mid g(\mathbf{x}) < 0\}$.

B. Proposed Algorithm

We will now state the proposed algorithm. Each node is initialized at an arbitrary feasible point $\mathbf{x}_i^1 \in \mathcal{X}$, which satisfies $g(\mathbf{x}_i^1) \leq 0$. Each node i constructs a strong convex surrogate $\hat{f}_i$ and solves the following optimization problem based upon the private knowledge of f_i and public knowledge of h and g;

$$\hat{\mathbf{x}}_i^t = \underset{\mathbf{x}_i \in \mathcal{X}}{\arg\min} \quad \tilde{f}_i\left(\mathbf{x}_i, \mathbf{x}_i^t, \xi_i^t\right) + \pi_i^t\left(\mathbf{x}_i - \mathbf{x}_i^t\right) + h(\mathbf{x}_i) \tag{1}$$

where $\hat{f}$ is strongly convex and

$$\tilde{f}_i\left(\mathbf{x}_i, \mathbf{x}_i^t, \xi_i^t\right) = \hat{f}_i\left(\mathbf{x}_i, \mathbf{x}_i^t, \xi_i^t\right) + (1-\beta)\left(\mathbf{z}_i^{t-1} - \nabla f_i(\mathbf{x}_i^{t-1}, \xi_i^t)\right)\left(\mathbf{x}_i - \mathbf{x}_i^t\right) \tag{2}$$

with $\pi_i^t = \mathbf{y}_i^t - \mathbf{z}_i^t$ and β is the step size. One choice of $\hat{f}$ can be $\hat{f}_i\left(\mathbf{x}_i, \mathbf{x}_i^t, \xi_i^t\right) = f_i\left(\mathbf{x}_i^t, \xi_i^t\right) + \nabla f_i\left(\mathbf{x}_i^t, \xi_i^t\right)\left(\mathbf{x}_i - \mathbf{x}_i^t\right) + \frac{\mu}{2}\left\|\mathbf{x}_i - \mathbf{x}_i^t\right\|^2$. Each node then performs the following two updates

$$\mathbf{x}_i^{t+1} = \sum_{j=1}^{n} W_{i,j} \mathbf{v}_j^t = \sum_{j=1}^{n} W_{i,j}\left(\mathbf{x}_j^t + \alpha\left(\hat{\mathbf{x}}_j^t - \mathbf{x}_j^t\right)\right). \tag{3}$$

$$\mathbf{z}_i^{t+1} = \nabla f_i(\mathbf{x}_i^{t+1}, \xi_i^{t+1}) + (1-\beta)\left(\mathbf{z}_i^t - \nabla f_i(\mathbf{x}_i^t, \xi_i^{t+1})\right) \tag{4}$$

This update is inspired from [26]. It should be noted that in (4) we have used *local* momentum-based gradient estimator $\mathbf{z}_i^t$ [25], [27]. Also, it is noteworthy that the update (4) can also be seen as a convex combination of vanilla SGD and SARAH-type gradient estimator [28]. Finally, we perform a global gradient update:

$$\mathbf{y}_i^{t+1} = \sum_{j=1}^{n} W_{i,j}\left(\mathbf{y}_j^t + \mathbf{z}_j^{t+1} - \mathbf{z}_j^t\right). \tag{5}$$

The D-MSSCA algorithm is summarised in Algorithm 1:

C. Assumptions

We will now state the assumptions required for the proposed algorithm. The assumptions are divided among the following 3 heads, viz assumption on ($\mathcal{P}$), those on the surrogate, and those on the network,

Algorithm 1 Decentralized -**M**omentum based **S**tochastic **SCA** (**D-MSSCA**) at each node i

Require $\mathbf{x}_1^1 = \mathbf{x}_2^1 = \cdots = \mathbf{x}_n^1, \alpha, \beta > 0, \quad \tau_i, \{w_{ij}\}_{j=1}^n, \quad$ Sample $\xi_i^1, \mathbf{z}_i^0 = \nabla f_i(\mathbf{x}_i^0, \xi_i^1) = 0, \mathbf{y}_i^1 = \mathbf{z}_i^1 = \nabla f_i(\mathbf{x}_i^1, \xi_i^1)$,

for $t = 1$ **to** T **do**

  Minimize local surrogate as per (1)

  Obtain local update of the solution as per (3)

  Sample ξ_i^{t+1} and update the local gradient estimates as per (4)

  Update the global gradient estimates as per (5)

end for

Output $\tilde{\mathbf{x}}_T$ selected uniformly at random from $\{\hat{\mathbf{x}}_i^t\}_{0\le t\le T}^{i\in\mathcal{V}}$

1) Assumptions on $(\mathcal{P})$*:*

A1. U is bounded below, i.e., $\inf_{\mathbf{x}\in\mathbb{R}^d} U(x) > -\infty$

A2. Let $\mathcal{H}^t$ represent the history of the system generated by $\{\xi_i^\tau\}_{i=\{1,2,..,n\}}^{\tau\le t-1}$, then

$$\mathbb{E}\left[\nabla f_i(\mathbf{x}^t, \xi_i^t \mid \mathcal{H}^t)\right] = \nabla u_i(\mathbf{x}^t);$$

A3. Bounded Variance: $\mathbb{E}\left[\|\nabla f_i(\mathbf{x}, \xi_i^t) - \nabla u_i(\mathbf{x})\|^2\right] \le \sigma_i^2 \quad \forall \mathbf{x} \in \mathbb{R}^d, \bar{\sigma}^2 = \sum_{i=1}^n \sigma_i^2$

A4. Each local function f_i is L_i- smooth;

$$\mathbb{E}\left\|\nabla f_i(\mathbf{x}^t, \xi^t) - \nabla f_i(\mathbf{y}^t, \xi^t)\right\| = L_i \mathbb{E}\left\|\mathbf{x}^t - \mathbf{y}^t\right\|;$$

Global function u is $L-$smooth, and $L_{\max} = \max_i\{L_i\}$, where $\sum_{i=1}^n L_i \le L \le nL_{\max}$

Assumptions **A1**- **A4** are standard in the context of distributed optimization [29]. A direct consequence of **A1** is that for an initial point $\mathbf{x}_i^1 \in \mathcal{X}$ we have $U(\bar{\mathbf{x}}^1) - U^\star \le B_1$. Assumption **A4** implies that U is also L-smooth. is introduced to simplify the analysis.

2) Assumptions on the surrogate: Two assumptions on the surrogate choice are:

A5. Tangent matching: $\nabla \hat{f}_i(\mathbf{x}^t, \mathbf{x}^t, \xi_i^t) = \nabla f_i(\mathbf{x}^t, \xi_i^t)$;

A6. Each surrogate $\hat{f}_i$ of local function f_i is μ_i- strongly convex.

Assumptions **A5** and **A6** are standard in the context of SCA and restrict the choice of surrogates. A consequence of **A5** is that $\nabla \tilde{f}_i\left(\mathbf{x}_i^t, \mathbf{x}_i^t, \xi_i^t\right) = \mathbf{z}_i^t$.

3) Assumption on Network:

A7. Graph $\mathcal{G}\left(\mathcal{V}, \mathcal{E}\right)$ is undirected, connected and communication matrix $\mathbf{W}$ is doubly stochastic. $W_{i,i} > 0$ for all i in $\mathcal{N}$ and for $i \neq j$, $w_{ij} > 0$ wherever $(i,j) \in \mathcal{E}$, $W_{i,j} = 0$ otherwise.

D. Approximate optimality

The performance of the proposed algorithm is studied in terms of its SFO complexity. We define the following metric, viz *mean-squared stationary gap* [30], that provides the number of calls to the SFO oracle to achieve an ϵ-KKT point in expectation.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t)+\hat{\mathbf{w}}_i^t\right\|^2\right]\leq\epsilon \tag{6}$$

where $\hat{\mathbf{w}}_i^t\in\partial(h+\mathbf{1}_{\mathcal{X}})(\hat{\mathbf{x}}_i^t)$. If (6) holds then the output $\tilde{\mathbf{x}}$ of D-MSSCA is chosen uniformly at random from the set $\{\hat{\mathbf{x}}_i^t\}_{0\leq t\leq T}^{i\in\mathcal{V}}$ then we have $\mathbb{E}\left[\left\|\nabla u(\tilde{\mathbf{x}})+\hat{\mathbf{w}}_{\tilde{\mathbf{x}}}\right\|^2\right]\leq\epsilon$.

III. Convergence Analysis

In this section, we will provide a detailed convergence analysis of the D-MSSCA algorithm and compare its rate with other state-of-the-art algorithms. For the analysis, we have defined the concatenated vectors $\mathbf{x},\mathbf{y},\mathbf{z},\hat{\mathbf{x}}\in\mathbb{R}^{nd}$ by concatenating corresponding local vectors $\{\mathbf{x}_i\in\mathbb{R}^d\}_{i\in\mathcal{V}},\{\mathbf{y}_i\in\mathbb{R}^d\}_{i\in\mathcal{V}},\{\mathbf{z}_i\in\mathbb{R}^d\}_{i\in\mathcal{V}},\{\hat{\mathbf{x}}_i\in\mathbb{R}^d\}_{i\in\mathcal{V}}$ of all the nodes. Utilizing these concatenated vectors, we can write the update equation (3) and (5) in a more compact form as below,

$$\mathbf{x}^{t+1}=\underline{\mathbf{W}}\left(\mathbf{x}^t+\alpha\left(\hat{\mathbf{x}}^t-\mathbf{x}^t\right)\right), \tag{7}$$

$$\mathbf{y}^{t+1}=\underline{\mathbf{W}}\left(\mathbf{y}^t+\mathbf{z}^{t+1}-\mathbf{z}^t\right), \tag{8}$$

where, $\underline{\mathbf{W}}=\mathbf{W}\otimes\mathbf{I}_d\in\mathbb{R}^{nd\times nd}$. Furthermore, we have defined the concatenated local gradient vector $\nabla u(\mathbf{x}^t):=[\nabla u_1(\mathbf{x}_1^t)^\mathsf{T},\nabla u_2(\mathbf{x}_1^t)^\mathsf{T},\cdots,\nabla u_n(\mathbf{x}_n^t)^\mathsf{T}]^\mathsf{T}\in\mathbb{R}^{nd}$, where $\nabla u_i(\mathbf{x}_i^t)\in\mathbb{R}^d$ for all $i\in\mathcal{V}$. For the sake of brevity, we define for all t

$$\theta^t=\left\|\mathbf{x}^t-\frac{1}{n}\left(\mathbf{1}_n\otimes\mathbf{I}_d\right)\bar{\mathbf{x}}^t\right\| \qquad \text{(consensus error)}, \tag{9}$$

$$\boldsymbol{\delta}^t=\hat{\mathbf{x}}^t-\mathbf{x}^t \qquad \text{(iterate progress)}, \tag{10}$$

$$\phi^t=\mathbb{E}\left[\left\|\bar{\mathbf{z}}^t-\bar{\nabla}u(\mathbf{x}^t)\right\|^2\right] \qquad \text{(global gradient variance)}, \tag{11}$$

$$\upsilon^t=\mathbb{E}\left[\left\|\mathbf{z}^t-\nabla u(\mathbf{x}^t)\right\|^2\right] \qquad \text{(network gradient variance)}, \tag{12}$$

$$\varepsilon^t=\mathbb{E}\left[\left\|\mathbf{y}^t-\frac{1}{n}\left(\mathbf{1}_n\otimes\mathbf{I}_d\right)\bar{\mathbf{y}}^t\right\|^2\right] \qquad \text{(gradient tracking error)}. \tag{13}$$

We began our analysis by stating some standard results used in decentralized optimization in Lemma 1 [5], [26], [31]. Next, Lemmm 2, Lemma 3 and Lemma 4 establish the contraction relations for $\theta^t,\phi^t,\upsilon^t$ and ε^t. Proceeding further, using these contraction results, we bounded the cumulative error accumulation $\sum_{t=1}^T\mathbb{E}\left[\theta^t\right],\sum_{t=1}^T\phi^t,\sum_{t=1}^T\varepsilon^t$ and $\sum_{t=1}^T\upsilon^t$ in Lemma 6, Lemma 7, and Lemma 8. Lemma 7, then extends these findings to provide bounds on the accumulated average iterate progress $\sum_{t=1}^T\mathbb{E}\left[\delta^t\right]$ under certain step-size conditions. Finally, Theorem 1 uses all these results to quantify the SFO complexity of D-MSSCA.

The results presented in Lemma 2-8 are can be obtained by applying the findings of [26] to SCA. However, our bounds are different as we have defined the quantities in terms of $\boldsymbol{\delta}$ rather than υ, and the variations in the intermediate steps are detailed in the proof. The results in Lemma 9 and Lemma 10 are different from [26] due to our focus on average progress $\Delta^T = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]$. Results similar to those in Lemma 10 are presented in [8]; however, in that work, the bounds are achieved for a deterministic case. The results of Lemma 10 and Theorem 1 are entirely novel to the considered class of problem.

Lemma 1. *Under Assumptions* ***A4*** *and* ***A7****, we have the following results for all* $t \geq 1$*, where* $\mathbf{x}^t, \mathbf{y}^t, \mathbf{z}^t$ *are variables of D-MSSCA at iterate* t*.*

$$\left\|\underline{\mathbf{W}}\mathbf{x} - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\otimes\mathbf{I}_d\right)\mathbf{x}\right\| \leq \lambda_{\mathbf{W}}\left\|\mathbf{x} - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\otimes\mathbf{I}_d\right)\mathbf{x}\right\| \quad \forall\mathbf{x}\in\mathbb{R}^{nd}, \tag{14a}$$

$$\left\|\sum_{i=1}^{n}\nabla u_i(\bar{\mathbf{x}}^t) - \frac{1}{n}\left(\mathbf{1}_n^\mathsf{T}\otimes\mathbf{I}_d\right)\nabla u(\mathbf{x}^t)\right\|^2 \leq \frac{L^2}{n}\left\|\mathbf{x}^t - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\otimes\mathbf{I}_d\right)\mathbf{x}^t\right\|^2, \tag{14b}$$

$$\bar{\mathbf{y}}^t = \bar{\mathbf{z}}^t, \tag{14c}$$

$$\left\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\right\|^2 \leq \frac{1}{n}\left\|\mathbf{x} - \mathbf{y}\right\|^2 \quad \textit{for any} \quad \mathbf{x},\mathbf{y}\in\mathbb{R}^{nd}, \tag{14d}$$

where $\lambda_{\mathbf{W}} := \lambda_{\max}(\mathbf{W} - \frac{1}{n}\mathbf{1}\mathbf{1}^\mathsf{T})$. The proofs of the above results are straightforward and can be found in [12], [32]. In the proof of (14a), the contraction property of doubly stochastic symmetric matrices is used; in (14b), the smoothness of the functions u_i $(i\in\mathcal{V})$; in (14c), the special initialization condition of v^t and the doubly stochastic property of $\mathbf{W}$; and in (14d), the property of norm with the Cauchy-Schwarz inequality. The next Lemma bounds the consensus errors in the $\mathbf{x}^t-$updates (7) of the D-MSSCA algorithm.

Lemma 2. *Under Assumption* (**A7**)*, for the* $\mathbf{x}^t-$ *updates of D-MSSCA algorithm, the following inequality holds for all* $t \geq 2$ *and* $\eta_1 > 0$

$$(\theta^t)^2 \leq (1+\eta_1)\,{\lambda_{\mathbf{W}}}^2\left(\theta^{t-1}\right)^2 + \left(1+\frac{1}{\eta_1}\right)\alpha^2{\lambda_{\mathbf{W}}}^2\left\|\boldsymbol{\delta}^{t-1}\right\|^2.$$

The proof of Lemma 2 is provided in Appendix A and follows by applying update step (7), then separating the terms using Young's inequality. Finally, by utilizing the properties of the communication matrix $\mathbf{W}$ (14a); we obtain the desired results. Similar contraction bounds on θ^t have been achieved in various gradient-tracking based decentralized optimization algorithms [26], [32]. However, our bound is slightly different because we define it in terms of θ^{t-1} and $\left\|\boldsymbol{\delta}^{t-1}\right\|^2$ instead of θ^{t-1} and v^{t-1}, as seen in literature. Next, we will bound the gradient variances.

Lemma 3. *Under Assumptions* ***A2****-****A4****, and Assumption* ***A7****, The following inequalities hold for the iterates produced by D-MSSCA algorithm, where* $t \geq 2, 0 < \alpha < 1, \eta_1, \eta_2, \eta_3 > 0$

$$\begin{aligned}\phi^t \leq{}& (1-\beta)^2\phi^{t-1} + \frac{3L^2(1-\beta)^2}{n^2}\left(1+\frac{1}{\eta_2}\right)\mathbb{E}\left[\left(\theta^t\right)^2 + n\alpha^2\left\|\left(\frac{1}{n}\mathbf{1}^\mathsf{T}\otimes\mathbf{I}_d\right)\boldsymbol{\delta}^{t-1}\right\|^2 + \left(\theta^{t-1}\right)^2\right]\\ &+ \frac{(1+\eta_2)\beta^2\bar{\sigma}^2}{n^2},\end{aligned} \tag{15}$$

and

$$v^t \leq (1-\beta)^2 v^{t-1} + 3L^2(1-\beta)^2\left(1+\frac{1}{\eta_3}\right)\left[\left(\theta^t\right)^2 + n\alpha^2\left\|\left(\frac{1}{n}\mathbf{1}^\mathsf{T}\otimes\mathbf{I}_d\right)\boldsymbol{\delta}^{t-1}\right\|^2 + \left(\theta^{t-1}\right)^2\right]$$

$$+ (1+\eta_3)\beta^2\bar{\sigma}^2. \tag{16}$$

The proof of Lemma 3 proceeds along the similar lines as the proof in [26, Lemma 3]. However, our final bounds are in terms of $\bar{\boldsymbol{\delta}}^{t-1}$ rather than $\bar{\mathbf{z}}^{t-1}$ in [26]. This difference is due to the $\mathbf{x}-$update of D-MSSCA (3), which differs from GT-HSGD algorithm proposed in [26]. The proof uses the unbiased nature of the local gradient estimate $\mathbf{z}_i^t$. Furthermore, as the gradient estimate at each node is independent of those at other nodes given the history sequence $\mathcal{H}^t$, we can omit the cross terms of inner products appearing in the intermediate steps to obtain simplified expressions. Finally, by applying Assumption **A4** alongside $\mathbf{x}_i^t$ updates, we get the desired result.

Lemma 4. *Under Assumptions **A2**-**A4** and **A7**, the following inequality holds for* $\beta \in (0,1)$, $\forall t \geq 2$,

$$\varepsilon^t \leq \frac{1+\lambda_{\mathbf{W}}{}^2}{2}\varepsilon^{t-1} + \frac{4\beta^2\lambda_{\mathbf{W}}{}^2}{1-\lambda_{\mathbf{W}}{}^2}\upsilon^{t-1} + 3\lambda_{\mathbf{W}}{}^2\beta^2\bar{\sigma}^2 + \frac{36\lambda_{\mathbf{W}}{}^2L^2}{1-\lambda_{\mathbf{W}}{}^2}\mathbb{E}\left[(\theta^{t-1})^2\right] + \frac{36\alpha^2\lambda_{\mathbf{W}}{}^2L^2}{1-\lambda_{\mathbf{W}}{}^2}\mathbb{E}\left[\left\|\boldsymbol{\delta}^{t-1}\right\|^2\right].$$

The above lemma bounds the error in local estimation of the global gradient. The proof uses the $\mathbf{y}^t-$update (8), applies conditional expectation, and simplifies intermediate steps using Assumptions **A2**,**A3**, and **A4**. Finally, by utilizing the consensus error bound in Lemma 2 and few mathematical simplification we achieved the desired result. The elaborated proof can be found in Appendix B. The next Lemma provides the basic results of non-negative sequences, which are necessary to bound the error accumulation in subsequent lemmas.

Lemma 5. *The recursions of well-defined sequences can be bound as below:*

1) *Let* $\{V^t\}_{t\geq 0}$, $\{Q^t\}_{t\geq 0}$ *be non-negative sequences and* $C > 0$ *be some constant such that* $V^t \leq qV^{t-1} + Q^{t-1} + C$ *for some* $q \in (0,1)$ *and for all* $t \geq 1$. *Then the following inequality holds* $\forall T \geq 1$
$$\sum_{t=0}^{T} V^t \leq \frac{V^0}{1-q} + \frac{\sum_{t=0}^{T-1} Q^t}{1-q} + \frac{CT}{1-q}. \tag{17}$$
2) *Let* $\{V^t\}_{t\geq 1}$, $\{Q^t\}_{t\geq 1}$ *be non-negative sequences and* $C > 0$ *be some constant such that* $V^t \leq qV^{t-1} + Q^{t-1} + C$ *for some* $q \in (0,1)$ *and for all* $t \geq 2$. *Then the following inequality holds* $\forall T \geq 1$
$$\sum_{t=1}^{T} V^t \leq \frac{V_1}{1-q} + \frac{\sum_{t=2}^{T} Q^{t-1}}{1-q} + \frac{CT}{1-q}. \tag{18}$$

The above mentioned recursion results align with [26, Lemma 6], and the proof follows a similar approach. For the complete proof, refer to [26]. Using the stated lemmas, we can now establish upper bounds for the cumulative errors up to iteration T, as detailed in the following lemmas.

Lemma 6. *For the proposed D-MSSCA algorithm, following inequality holds:* $\forall T > 1$, $\alpha \in (0,1)$,

$$\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right] \leq \frac{4\alpha^2\lambda_{\mathbf{W}}{}^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right].$$

The above result can be obtained by summing both sides of the bound obtained in Lemma 2, for $1 \leq t \leq T$ and applying (18). The detailed proof is provided in Appendix C.

Lemma 7. *For the proposed D-MSSCA algorithm, following inequality holds:* $\forall T > 1$, $\beta, \alpha \in (0,1)$,

$$\sum_{t=1}^{T}\phi^t \leq \frac{\bar{\sigma}^2}{n^2 b_0 \beta} + \frac{2\beta\bar{\sigma}^2 T}{n^2} + \frac{6L^2\alpha^2}{n^2\beta}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right] + \frac{12L^2}{\beta n^2}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right],$$

$$\sum_{t=1}^{T} \upsilon^t \leq \frac{\bar{\sigma}^2}{b_0\beta} + 2\beta\bar{\sigma}^2 T + \frac{6L^2\alpha^2}{\beta}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right] + \frac{12L^2}{\beta}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right].$$

To prove Lemma 7, we have applied the results of Lemma 3, Lemma 5 and (**A3**). The proof of Lemma 7 is provided in Appendix D.

Lemma 8. *The following inequality holds for all* $t > 0$

$$\sum_{t=1}^{T} \varepsilon^t \leq \frac{24\alpha^2{\lambda_{\mathbf{W}}}^2 L^2}{\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}(3+2\beta)\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right] + \frac{24{\lambda_{\mathbf{W}}}^2 L^2}{\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}(3+4\beta)\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]$$
$$+ \frac{2{\lambda_{\mathbf{W}}}^2\beta^2\bar{\sigma}^2}{1-{\lambda_{\mathbf{W}}}^2}\left(\frac{4}{\left(1-{\lambda_{\mathbf{W}}}^2\right)b_0\beta} + \frac{8\beta T}{\left(1-{\lambda_{\mathbf{W}}}^2\right)} + 3T\right) + \frac{2}{1-{\lambda_{\mathbf{W}}}^2}\varepsilon^1.$$

To prove Lemma 8, we have applied the results of Lemma 4,5 and 7. The proof of Lemma 8 is provided in Appendix E.

The next Lemma is a key result in the analysis of D-MSSCA algorithm, offering a descent inequality for average (over the network) D-MSSCA updates with respect to the global function U.

Lemma 9. *Under Assumptions **A4** and **A7**, the following inequality holds for all* $T > 1$, $\gamma_1 > 0$, $\mu > 0$ *and* $0 < \alpha < 1$, *where* $\mathbf{x}^t, \mathbf{y}^t, \mathbf{z}^t$ *are iterate variables of D-MSSCA at iterate* t,

$$U(\bar{\mathbf{x}}^{T+1}) - U(\bar{\mathbf{x}}^1) \leq \frac{3L^2\alpha\gamma_1}{2n}\sum_{t=1}^{T}\mathbb{E}\left[(\theta^t)^2\right] + \frac{3\alpha\gamma_1}{2}\sum_{t=1}^{T}\phi^t + \frac{\alpha}{n}\left(-\mu + \frac{1}{2\gamma_1} + \frac{\alpha L}{2}\right)\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]$$
$$+ \frac{3\alpha\gamma_1}{2n}\sum_{t=1}^{T}\varepsilon^t.$$

The proof of Lemma 9 begins by defining the optimality condition of (1), and utilizing it with the properties of surrogate to get the descent direction. Further by using the convexity of $\mathbf{h}$ and the properties of the communication matrix $\mathbf{W}$, along with some mathematical simplifications, we achieve the desired result. The complete proof of Lemma 9 is detailed in Appendix F.

Now, we will use Lemma 6-9 to upper bound the average progress $\Delta^T = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]$.

Lemma 10. *Under considered assumption **A1**-**A7**, if* $0 < \beta = \alpha^2 < 1, \mu \geq \frac{6\sqrt{3}L}{n}\left(1 + \frac{8{\lambda_{\mathbf{W}}}^2}{(1-{\lambda_{\mathbf{W}}}^2)}\right)$ *and* $0 < \alpha \leq \min\left\{\frac{1}{114}, \frac{\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}{432{\lambda_{\mathbf{W}}}^2}, \frac{\left(1-{\lambda_{\mathbf{W}}}^2\right)^{2/3}}{8{\lambda_{\mathbf{W}}}^{2/3}}, \frac{\mu}{6L}, \frac{\mu^2\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}{48L^2{\lambda_{\mathbf{W}}}^2}\right\}$ *then the average progress of D-MSSCA algorithm is upper bounded for all* $T \geq 2$ *as below*

$$\Delta^T \leq \frac{4n}{\alpha T\mu}U(\bar{\mathbf{x}}^1) - \frac{4n}{\alpha T\mu}U^\star + \frac{24}{T\mu^2\left(1-{\lambda_{\mathbf{W}}}^2\right)}\left\|\nabla u(\mathbf{x}^1)\right\|^2 + \frac{12\bar{\sigma}^2}{2T\mu^2}\Bigg(\frac{1}{b_0\alpha^2 n} + \frac{2\alpha^2 T}{n} + \frac{4}{b_0^2\left(1-{\lambda_{\mathbf{W}}}^2\right)}$$
$$+ \frac{2{\lambda_{\mathbf{W}}}^2\alpha^4}{1-{\lambda_{\mathbf{W}}}^2}\left(\frac{4}{\left(1-{\lambda_{\mathbf{W}}}^2\right)b_0\alpha^2} + \frac{8\alpha^2 T}{\left(1-{\lambda_{\mathbf{W}}}^2\right)} + 3T\right)\Bigg).$$

Proof: We begin by substituting the bound of $\sum_{t=1}^{T}\phi^t$ from Lemma 7 into Lemma 9 and obtain,

$$U(\bar{\mathbf{x}}^{T+1}) - U(\bar{\mathbf{x}}^1) \leq \frac{3L^2\alpha\gamma_1}{2n}\sum_{t=1}^{T}\mathbb{E}\left[(\theta^t)^2\right] + \frac{\alpha}{n}\left(-\mu + \frac{1}{2\gamma_1} + \frac{\alpha L}{2}\right)\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]$$

$$+\frac{3\alpha\gamma_1}{2n}\sum_{t=1}^{T}\varepsilon^t+\frac{3\alpha\gamma_1}{2}\left(\frac{\bar{\sigma}^2}{n^2b_0\beta}+\frac{2\beta\bar{\sigma}^2T}{n^2}+\frac{6L^2\alpha^2}{n^2\beta}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{12L^2}{\beta n^2}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]\right).$$

On combining the common terms and substituting the bound of $\sum_{t=1}^{T}\varepsilon^t$ from Lemma 8, we get

$$\begin{aligned}U(\bar{\mathbf{x}}^{T+1})-U(\bar{\mathbf{x}}^1)\leq&\frac{3L^2\alpha\gamma_1}{2n}\left(1+\frac{12}{\beta n}\right)\sum_{t=1}^{T}\mathbb{E}\left[(\theta^t)^2\right]+\frac{\alpha}{n}\left(-\mu+\frac{1}{2\gamma_1}+\frac{\alpha L}{2}+\frac{9L^2\alpha^2\gamma_1}{n\beta}\right)\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]\\&+\frac{3\alpha\gamma_1\bar{\sigma}^2}{2n^2}\left(\frac{1}{b_0\beta}+2\beta T\right)+\frac{3\alpha\gamma_1}{2n}\left[\frac{24\alpha^2{\lambda_{\mathbf{W}}}^2L^2}{\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}(3+2\beta)\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{2}{1-{\lambda_{\mathbf{W}}}^2}\varepsilon^1\right.\\&\left.+\frac{24\left(3+4\beta\right){\lambda_{\mathbf{W}}}^2L^2}{\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]+\frac{2{\lambda_{\mathbf{W}}}^2\beta^2\bar{\sigma}^2}{1-{\lambda_{\mathbf{W}}}^2}\left(\frac{4}{\left(1-{\lambda_{\mathbf{W}}}^2\right)b_0\beta}+\frac{8\beta T}{\left(1-{\lambda_{\mathbf{W}}}^2\right)}+3T\right)\right].\end{aligned}$$

By combining the common terms and substituting the bound of $\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]$ from Lemma 6, we obtain

$$\begin{aligned}U(\bar{\mathbf{x}}^{T+1})-U(\bar{\mathbf{x}}^1)\leq&\frac{\alpha}{n}\left[\frac{6\alpha^2{\lambda_{\mathbf{W}}}^2\gamma_1L^2}{\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}\left(19+12\beta+\frac{12}{\beta n}+\frac{24{\lambda_{\mathbf{W}}}^2}{\left(1-{\lambda_{\mathbf{W}}}^2\right)^2}(3+4\beta)\right)-\mu+\frac{1}{2\gamma_1}+\frac{\alpha L}{2}\right.\\&\left.+\frac{9L^2\alpha^2\gamma_1}{n\beta}\right]\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{3\alpha\gamma_1}{n\left(1-{\lambda_{\mathbf{W}}}^2\right)}\varepsilon^1+\frac{3\alpha\gamma_1\bar{\sigma}^2}{2n}\left(\frac{1}{b_0\beta n}+\frac{2\beta T}{n}\right.\\&\left.+\frac{2{\lambda_{\mathbf{W}}}^2\beta^2}{1-{\lambda_{\mathbf{W}}}^2}\left(\frac{4}{\left(1-{\lambda_{\mathbf{W}}}^2\right)b_0\beta}+\frac{8\beta T}{\left(1-{\lambda_{\mathbf{W}}}^2\right)}+3T\right)\right).\end{aligned}$$

On defining $C_\mu=\left[\frac{6\alpha^2{\lambda_{\mathbf{W}}}^2\gamma_1L^2}{(1-{\lambda_{\mathbf{W}}}^2)^2}\left(19+12\beta+\frac{12}{\beta n}+\frac{24\lambda_{\mathbf{W}}^2}{(1-{\lambda_{\mathbf{W}}}^2)^2}(3+4\beta)\right)\mu+\frac{1}{2\gamma_1}+\frac{\alpha L}{2}+\frac{9L^2\alpha^2\gamma_1}{n\beta}\right]$, we have:

$$\begin{aligned}\frac{C_\mu}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]\leq&\frac{n}{\alpha T}U(\bar{\mathbf{x}}^1)-\frac{n}{\alpha T}U(\bar{\mathbf{x}}^{T+1})+\frac{3\gamma_1}{T\left(1-{\lambda_{\mathbf{W}}}^2\right)}\varepsilon^1+\frac{3\gamma_1\bar{\sigma}^2}{2T}\left(\frac{1}{b_0\beta n}+\frac{2\beta T}{n}\right.\\&\left.+\frac{2{\lambda_{\mathbf{W}}}^2\beta^2}{1-{\lambda_{\mathbf{W}}}^2}\left(\frac{4}{\left(1-{\lambda_{\mathbf{W}}}^2\right)b_0\beta}+\frac{8\beta T}{\left(1-{\lambda_{\mathbf{W}}}^2\right)}+3T\right)\right).\end{aligned}$$

Also, from the initialization of $\mathbf{z}_i^1$ and the update (5) of the D-MSSCA algorithm, we have:

$$\begin{aligned}\varepsilon^1&=\mathbb{E}\left\|(\mathbf{I}-\frac{1}{2}\mathbf{1}\mathbf{1}^{\mathsf{T}}\otimes\mathbf{I}_d)\mathbf{y}^1\right\|^2\leq\mathbb{E}\left\|\mathbf{y}^1\right\|^2=\mathbb{E}\left\|\nabla f(\mathbf{x}^1,\boldsymbol{\xi}^1)-\nabla u(\mathbf{x}^1)+\nabla u(\mathbf{x}^1)\right\|^2\\&\leq2\sum_{i=1}^{n}\mathbb{E}\left\|\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\nabla f_i(\mathbf{x}_i^1,\boldsymbol{\xi}_i^{1,r})-\nabla u_i(\mathbf{x}_i^1)\right)\right\|^2+2\mathbb{E}\left\|\nabla u(\mathbf{x}^1)\right\|^2,\\&\overset{\text{(i)}}{\leq}\frac{2}{b_0^2}\sum_{i=1}^{n}\sum_{r=1}^{b_0}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^1,\boldsymbol{\xi}_i^{1,r})-\nabla u_i(\mathbf{x}_i^1)\right\|^2+2\mathbb{E}\left\|\nabla u(\mathbf{x}^1)\right\|^2,\\&\overset{\text{(ii)}}{\leq}\frac{2\bar{\sigma}^2}{b_0^2}+2\left\|\nabla u(\mathbf{x}^1)\right\|^2,\end{aligned}\tag{19}$$

where in (i), we used the fact that $\boldsymbol{\xi}_i^{1,l},\boldsymbol{\xi}_j^{1,m}$ are independent for $l\neq m$ and in (ii), we applied (**A3**). On substituting the bound $\varepsilon^1\leq\frac{2\bar{\sigma}^2}{b_0^2}+2\left\|\nabla u(\mathbf{x}^1)\right\|^2$, and $U^\star<U(\bar{\mathbf{x}}^{T+1})$, we obtain:,

$$\begin{aligned}\frac{C_\mu}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]\leq&\frac{n}{\alpha T}U(\bar{\mathbf{x}}^1)-\frac{n}{\alpha T}U^\star+\frac{6\gamma_1}{T\left(1-{\lambda_{\mathbf{W}}}^2\right)}\left\|\nabla u(\mathbf{x}^1)\right\|^2+\frac{3\gamma_1\bar{\sigma}^2}{2T}\left(\frac{1}{b_0\beta n}+\frac{2\beta T}{n}+\frac{4}{b_0^2\left(1-{\lambda_{\mathbf{W}}}^2\right)}\right.\\&\left.+\frac{2{\lambda_{\mathbf{W}}}^2\beta^2}{1-{\lambda_{\mathbf{W}}}^2}\left(\frac{4}{\left(1-{\lambda_{\mathbf{W}}}^2\right)b_0\beta}+\frac{8\beta T}{\left(1-{\lambda_{\mathbf{W}}}^2\right)}+3T\right)\right).\end{aligned}$$

Further if we consider $\beta = \alpha^2$,$\gamma_1 = \frac{1}{\mu}$, and $0 < \alpha \leq \min\left\{\frac{1}{114}, \frac{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}{432\lambda_{\mathbf{W}}{}^2}, \left(\frac{\left(1-\lambda_{\mathbf{W}}{}^2\right)}{24\lambda_{\mathbf{W}}}\right)^{2/3}, \frac{\mu}{6L}, \frac{\mu^2\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}{48L^2\lambda_{\mathbf{W}}{}^2}\right\}$, and if $\mu \geq \frac{6\sqrt{3}L}{n}\left(1+\frac{8\lambda_{\mathbf{W}}{}^2}{(1-\lambda_{\mathbf{W}}{}^2)}\right)$, then after further simplification we get $C_\mu \geq \frac{\mu}{4} > 0$, utilizing which we get the desired result. ■

Finally, we are ready to state the main theorem regarding the existence of ϵ-KKT point. Specifically, we will bound $\frac{1}{n}\sum_{i=1}^n \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t) + \hat{\mathbf{w}}_i^t\right\|^2\right]$ (6).

Remark. It is remarked that combining the results of Lemma 6 and 10 proves that the consensus is achieved

Theorem 1. *Under the considered Assumptions **A1**-**A7**, if* $0 < \beta = \alpha^2 < 1, \mu \geq \frac{6\sqrt{3}L}{n}\left(1+\frac{8\lambda_{\mathbf{W}}{}^2}{(1-\lambda_{\mathbf{W}}{}^2)}\right)$*, and* $0 < \alpha \leq \min\left\{\frac{1}{116}, \frac{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}{432\lambda_{\mathbf{W}}{}^2}, \left(\frac{\left(1-\lambda_{\mathbf{W}}{}^2\right)}{24\lambda_{\mathbf{W}}}\right)^{2/3}, \frac{\mu}{6L}, \frac{\mu^2\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}{48L^2\lambda_{\mathbf{W}}{}^2}\right\}$*, then the mean squared stationary gap of the proposed D-MSSCA algorithm is upper bounded for all* $T \geq 2$ *as below:*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t) + \hat{\mathbf{w}}_i^t\right\|^2\right] \leq{}& \left(\frac{8L^2}{T}\left(2+\frac{9}{n}+\frac{72\lambda_{\mathbf{W}}{}^2}{n(1-\lambda_{\mathbf{W}}{}^2)^2}\right)\right)\frac{4}{\alpha\mu}\left(U(\bar{\mathbf{x}}^1)-U^\star\right) + \frac{48\left\|\nabla u(\mathbf{x}_1^1)\right\|^2}{nT(1-\lambda_{\mathbf{W}}{}^2)}P \\
&+ \frac{48\bar{\sigma}^2}{nTb_0^2(1-\lambda_{\mathbf{W}}{}^2)}P + \frac{12\bar{\sigma}^2}{Tn^2b_0\alpha^2}P + \frac{24\alpha^2\bar{\sigma}^2}{n}P\left(\frac{4\lambda_{\mathbf{W}}{}^2}{Tb_0(1-\lambda_{\mathbf{W}}{}^2)^2}+\frac{1}{n}\right) \\
&+ \frac{72\alpha^4\lambda_{\mathbf{W}}{}^2\bar{\sigma}^2}{n(1-\lambda_{\mathbf{W}}{}^2)}P + \frac{192\alpha^6\lambda_{\mathbf{W}}{}^2\bar{\sigma}^2}{n(1-\lambda_{\mathbf{W}}{}^2)^2}P,
\end{aligned} \tag{20}$$

where, $P = \frac{8L^2}{\mu^2} + \frac{36L^2}{n\mu^2} + \frac{288L^2\lambda_{\mathbf{W}}{}^2}{n\mu^2(1-\lambda_{\mathbf{W}}{}^2)^2} + 1$.

Proof: We will start the proof by using the optimality condition of (1) to bound the mean squared stationary gap (6).

From the update equation(1) and the definition of the surrogate function $\tilde{f}$, there exists $\hat{\mathbf{w}}_i^t \in \partial(h(\hat{\mathbf{x}}_i^t) + \mathbf{1}_{\mathcal{X}})$ for all $t \geq 1$ such that;

$$\nabla\hat{f}_i\left(\hat{\mathbf{x}}_i^t, \mathbf{x}_i^t, \xi_i^t\right) + (1-\beta)\left(\mathbf{v}_i^{t-1} - \nabla f_i(\mathbf{x}_i^{t-1}, \xi_i^t)\right) + \pi_i^t + \hat{\mathbf{w}}_i^t = 0.$$

On adding and subtracting $\nabla\hat{f}_i\left(\mathbf{x}_i^t, \mathbf{x}_i^t, \xi_i^t\right)$, applying the definition of π_i^t and update $\mathbf{z}_i^t$ (4), we obtain

$$\nabla\hat{f}_i\left(\hat{\mathbf{x}}_i^t, \mathbf{x}_i^t, \xi_i^t\right) - \nabla\hat{f}_i\left(\mathbf{x}_i^t, \mathbf{x}_i^t, \xi_i^t\right) + \mathbf{y}_i^t + \hat{\mathbf{w}}_i^t = 0.$$

By substituting $\hat{\mathbf{w}}_i^t = -\nabla\hat{f}_i\left(\hat{\mathbf{x}}_i^t, \mathbf{x}_i^t, \xi_i^t\right) + \nabla\hat{f}_i\left(\mathbf{x}_i^t, \mathbf{x}_i^t, \xi_i^t\right) - \mathbf{y}_i^t$, into the definition of mean squared stationary gap, we get

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t) + \hat{\mathbf{w}}_i^t\right\|^2\right] = \frac{1}{n}\sum_{i=1}^n \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t) - \nabla\hat{f}_i\left(\hat{\mathbf{x}}_i^t, \mathbf{x}_i^t, \xi_i^t\right) + \nabla\hat{f}_i\left(\mathbf{x}_i^t, \mathbf{x}_i^t, \xi_i^t\right) - \mathbf{y}_i^t\right\|^2\right].$$

Now, by adding and subtracting $\nabla u(\mathbf{x}_i^t) - \nabla u(\bar{\mathbf{x}}^t)$ and separating the terms using the properties of the norm, we obtain:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t) + \hat{\mathbf{w}}_i^t\right\|^2\right] \\
&\leq \frac{4}{n}\sum_{i=1}^n \frac{1}{T}\sum_{t=1}^T \Bigg(\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t) - \nabla u(\mathbf{x}_i^t)\right\|^2\right] + \mathbb{E}\left[\left\|\nabla u(\mathbf{x}_i^t) - \nabla u(\bar{\mathbf{x}}^t)\right\|^2\right] + \mathbb{E}\left[\left\|\nabla u(\bar{\mathbf{x}}^t) - \mathbf{y}_i^t\right\|^2\right]^2
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}\left[\left\|\nabla \hat{f}_i\left(\mathbf{x}_i^t,\mathbf{x}_i^t,\xi_i^t\right)-\nabla \hat{f}_i\left(\hat{\mathbf{x}}_i^t,\mathbf{x}_i^t,\xi_i^t\right)\right\|^2\right]\Bigg),\\
&\overset{\text{(i)}}{\leq} \frac{4}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\left(2L^2\mathbb{E}\left[\left\|\hat{\mathbf{x}}_i^t-\mathbf{x}_i^t\right\|^2\right]+L^2\mathbb{E}\left[\left\|\mathbf{x}_i^t-\bar{\mathbf{x}}^t\right\|^2\right]\right)\\
&\quad+\frac{4}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left\|(\mathbf{1}\otimes\mathbf{I}_d)\left(\nabla u(\bar{\mathbf{x}}^t)-\bar{\nabla} u(\mathbf{x}^t)+\bar{\nabla} u(\mathbf{x}^t)-\bar{\mathbf{y}}^t+\bar{\mathbf{y}}^t\right)-\mathbf{y}^t\right\|^2\right],\\
&\leq \frac{8L^2}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\hat{\mathbf{x}}^t-\mathbf{x}^t\right\|^2\right]+\frac{4L^2}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\mathbf{x}^t-(\mathbf{1}\otimes\mathbf{I}_d)\bar{\mathbf{x}}^t\right\|^2\right]+\frac{4}{nT}\sum_{t=1}^{T}\Bigg(3n\mathbb{E}\left[\left\|\bar{\nabla} u(\mathbf{x}^t)-\bar{\mathbf{y}}^t\right\|^2\right]\\
&\qquad+3\mathbb{E}\left[\left\|(\mathbf{1}\otimes\mathbf{I}_d)\left(\nabla u(\bar{\mathbf{x}}^t)-\bar{\nabla} u(\mathbf{x}^t)\right)\right\|^2\right]+3\mathbb{E}\left[\left\|\mathbf{y}^t-(\mathbf{1}\otimes\mathbf{I}_d)\bar{\mathbf{y}}^t\right\|^2\right]\Bigg),\\
&\overset{\text{(ii)}}{\leq} \frac{8L^2}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{4L^2}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]+\frac{12L^2}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)\right]+\frac{12}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\bar{\nabla} u(\mathbf{x}^t)-\bar{\mathbf{z}}^t\right\|^2\right]\\
&\quad+\frac{12}{nT}\sum_{t=1}^{T}\varepsilon^t.
\end{aligned}
$$

In (i), we have utilized assumption **A4**, and in (ii), applied (14b) and (14c). Now, by substituting the value of $\sum_{t=1}^{T}\mathbb{E}\left[\left\|\bar{\nabla} u(\mathbf{x}^t)-\bar{\mathbf{z}}^t\right\|^2\right]$, from Lemma 7, we get

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t)+\hat{\mathbf{w}}_i^t\right\|^2\right]\leq{}& \frac{8L^2}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{16L^2}{nT}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]+\frac{12}{nT}\sum_{t=1}^{T}\varepsilon^t\\
&+\frac{12}{T}\left(\frac{\bar{\sigma}^2}{n^2 b_0\beta}+\frac{2\beta\bar{\sigma}^2 T}{n^2}+\frac{6L^2\alpha^2}{n^2\beta}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{12L^2}{\beta n^2}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]\right).
\end{aligned}
$$

Further on combining the common terms and substituting the bound of ε^t from Lemma 8, we get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t)+\hat{\mathbf{w}}_i^t\right\|^2\right]\leq \frac{8L^2}{nT}\left(1+\frac{9\alpha^2}{n\beta}\right)\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{16L^2}{nT}\left(1+\frac{9}{\beta n}\right)\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]\\
&+\frac{12}{T}\left(\frac{\bar{\sigma}^2}{n^2 b_0\beta}+\frac{2\beta\bar{\sigma}^2 T}{n^2}\right)+\frac{12}{nT}\Bigg[\frac{24\alpha^2\lambda_{\mathbf{W}}{}^2L^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}(3+2\beta)\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]+\frac{2}{1-\lambda_{\mathbf{W}}{}^2}\varepsilon^1\\
&+\frac{24\lambda_{\mathbf{W}}{}^2L^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}(3+4\beta)\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]+\frac{2\lambda_{\mathbf{W}}{}^2\beta^2\bar{\sigma}^2}{1-\lambda_{\mathbf{W}}{}^2}\left(\frac{4}{\left(1-\lambda_{\mathbf{W}}{}^2\right)b_0\beta}+\frac{8\beta T}{\left(1-\lambda_{\mathbf{W}}{}^2\right)}+3T\right)\Bigg].
\end{aligned}
$$

On combining the common terms and substituting the bound of $\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]$ from Lemma 6, we can further simplify as follows:

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t)+\hat{\mathbf{w}}_i^t\right\|^2\right]\leq \frac{8L^2}{n}\left(1+\frac{9\alpha^2}{n\beta}+\frac{36\alpha^2\lambda_{\mathbf{W}}{}^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}(3+2\beta)\right)\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]\\
&+\frac{12}{T}\left(\frac{\bar{\sigma}^2}{n^2 b_0\beta}+\frac{2\beta\bar{\sigma}^2 T}{n^2}\right)+\frac{12}{nT}\left[\frac{2}{1-\lambda_{\mathbf{W}}{}^2}\varepsilon^1+\frac{2\lambda_{\mathbf{W}}{}^2\beta^2\bar{\sigma}^2}{1-\lambda_{\mathbf{W}}{}^2}\left(\frac{4}{\left(1-\lambda_{\mathbf{W}}{}^2\right)b_0\beta}+\frac{8\beta T}{\left(1-\lambda_{\mathbf{W}}{}^2\right)}+3T\right)\right]\\
&+\frac{16L^2}{nT}\left(1+\frac{9}{\beta n}+\frac{18\lambda_{\mathbf{W}}{}^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}(3+4\beta)\right)\frac{4\alpha^2\lambda_{\mathbf{W}}{}^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right].
\end{aligned}
$$

Further using the bound of $\frac{1}{T}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right]$ obtained in Lemma 10, and applying (19), we get

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t)+\hat{\mathbf{w}}_i^t\right\|^2\right]$$
$$\leq \frac{8L^2}{n}\left(1+\frac{9\alpha^2}{n\beta}+\frac{4\alpha^2\lambda_{\mathbf{W}}{}^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}\left(29+18\beta+\frac{18}{n\beta}+\frac{36\lambda_{\mathbf{W}}{}^2}{(1-\lambda_{\mathbf{W}}{}^2)^2}(3+4\beta)\right)\right)\left[\frac{4n}{\alpha T\mu}U(\bar{\mathbf{x}}^1)\right.$$
$$-\frac{4n}{\alpha T\mu}U^\star+\frac{24}{T\mu^2\left(1-\lambda_{\mathbf{W}}{}^2\right)}\left\|\nabla u(\mathbf{x}^1)\right\|^2+\frac{12\bar{\sigma}^2}{2T\mu^2}\left(\frac{1}{b_0\alpha^2 n}+\frac{2\alpha^2 T}{n}+\frac{4}{b_0^2\left(1-\lambda_{\mathbf{W}}{}^2\right)}\right.$$
$$\left.\left.+\frac{2\lambda_{\mathbf{W}}{}^2\alpha^4}{1-\lambda_{\mathbf{W}}{}^2}\left(\frac{4}{\left(1-\lambda_{\mathbf{W}}{}^2\right)b_0\alpha^2}+\frac{8\alpha^2 T}{\left(1-\lambda_{\mathbf{W}}{}^2\right)}+3T\right)\right)\right]+\frac{12}{T}\left(\frac{\bar{\sigma}^2}{n^2 b_0\beta}+\frac{2\beta\bar{\sigma}^2 T}{n^2}\right)$$
$$+\frac{12}{nT}\left[\frac{2\lambda_{\mathbf{W}}{}^2\beta^2\bar{\sigma}^2}{1-\lambda_{\mathbf{W}}{}^2}\left(\frac{4}{\left(1-\lambda_{\mathbf{W}}{}^2\right)b_0\beta}+\frac{8\beta T}{\left(1-\lambda_{\mathbf{W}}{}^2\right)}+3T\right)+\frac{2}{1-\lambda_{\mathbf{W}}{}^2}\left(\frac{2\bar{\sigma}^2}{b_0^2}+2\left\|\nabla u(\mathbf{x}^1)\right\|^2\right)\right].$$

Further on substituting $\beta=\alpha^2$, considering $\alpha\leq\min\left\{\frac{1}{116},\frac{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}{432\lambda_{\mathbf{W}}{}^2},\frac{\left(1-\lambda_{\mathbf{W}}{}^2\right)^{2/3}}{8\lambda_{\mathbf{W}}{}^{2/3}},\frac{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}{4\lambda_{\mathbf{W}}{}^2}\right\}$, and rearranging, we get the desired result. ■

Remark. It can be noted from the Theorem 1 that the mean squared stationary gap of the proposed D-MSSCA algorithm reaches to a steady state-error at a sublinear rate. Where the steady-state error is defined as

$$\limsup_{T\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t)+\hat{\mathbf{w}}_i^t\right\|^2\right]\leq\frac{24\alpha^2\bar{\sigma}^2}{n^2}P$$

From this expression, it is evident that the steady-state error can be reduced by selecting smaller values of α and β. Additionally, the error decreases as the number of nodes increases, which is expected since n nodes function as n oracles (SFO) for estimating the gradient.

Finally, the next corollary provides the convergence rate in terms of the SFO complexity of the proposed D-MSSCA algorithm for fixed values of α,β, and b_0.

Corollary 1. *Under the conditions such that Theorem 1 holds, if we further consider $\alpha=\mathcal{O}(T^{-1/3}),\beta=\mathcal{O}(T^{-2/3})$, and $b_0=\mathcal{O}(T^{1/3})$, The proposed D-MSSCA algorithm achieves an $\epsilon-$ KKT point in $\mathcal{O}(\epsilon^{-3/2})$ oracle calls.*

Proof: By substituting $\alpha=T^{-1/3},b_0=T^{1/3}$ in (20), we get

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla u(\hat{\mathbf{x}}_i^t)+\hat{\mathbf{w}}_i^t\right\|^2\right]\leq\left(\frac{8L^2}{T^{2/3}}\left(2+\frac{9}{n}+\frac{72\lambda_{\mathbf{W}}{}^2}{n(1-\lambda_{\mathbf{W}}{}^2)^2}\right)\right)\frac{4}{\mu}\left(U(\bar{\mathbf{x}}^1)-U^\star\right)$$
$$+\frac{48\left\|\nabla u(\mathbf{x}_1^1)\right\|^2}{nT(1-\lambda_{\mathbf{W}}{}^2)}P+\frac{48\bar{\sigma}^2}{nT^{5/3}(1-\lambda_{\mathbf{W}}{}^2)}P+\frac{12\bar{\sigma}^2}{n^2T^{2/3}}P$$
$$+\frac{24T^{-2/3}\bar{\sigma}^2}{n}P\left(\frac{4\lambda_{\mathbf{W}}{}^2}{T^{4/3}(1-\lambda_{\mathbf{W}}{}^2)^2}+\frac{1}{n}\right)+\frac{72T^{-4/3}\lambda_{\mathbf{W}}{}^2\bar{\sigma}^2}{n(1-\lambda_{\mathbf{W}}{}^2)}P$$
$$+\frac{192T^{-6/3}\lambda_{\mathbf{W}}{}^2\bar{\sigma}^2}{n(1-\lambda_{\mathbf{W}}{}^2)^2}P,$$

In order to reach $\epsilon-$KKT point, we require

$$\frac{1}{T^{2/3}}\left[\frac{32L^2}{\mu}\left(2+\frac{9}{n}+\frac{72\lambda_{\mathbf{W}}{}^2}{n(1-\lambda_{\mathbf{W}}{}^2)^2}\right)\left(U(\bar{\mathbf{x}}^1)-U^\star\right)+\frac{48P\left\|\nabla u(\mathbf{x}_1^1)\right\|^2}{nT^{1/3}(1-\lambda_{\mathbf{W}}{}^2)}+\frac{48P\bar{\sigma}^2}{nT(1-\lambda_{\mathbf{W}}{}^2)}+\frac{12P\bar{\sigma}^2}{n^2}\right.$$

$$+ \frac{24P\bar{\sigma}^2}{n}\left(\frac{4\lambda_{\mathbf{W}}{}^2}{T^{4/3}(1-\lambda_{\mathbf{W}}{}^2)^2}+\frac{1}{n}\right)+\frac{72P\lambda_{\mathbf{W}}{}^2\bar{\sigma}^2}{nT^{2/3}(1-\lambda_{\mathbf{W}}{}^2)}+\frac{192T^{-6/3}\lambda_{\mathbf{W}}{}^2\bar{\sigma}^2}{n(1-\lambda_{\mathbf{W}}{}^2)^2}P\Bigg] < \epsilon, \tag{21}$$

which gives $T = \mathcal{O}(\epsilon^{-3/2})$. ■

The SFO complexity of Corollary 1 matches that of DEEPSTORM [7] and ProxGT-SR-O/E [5]. It should be noted that, unlike DEEPSTORM and ProxGT-SR-O/E which require small and large batch sizes respectively, our algorithm is batchless and uses one sample at each iteration. Also, this rate matches the SFO complexity lower bound for centralized unconstrained stochastic non-convex optimization problems.

IV. Experimental data and results

In this section, we will demonstrate the applicability of D-MSSCA. Let us consider a simple distributed optimization problem, which is a stochastic version of the synthetic problem in [8], [33] over a network of $n = 3$ nodes:

$$U(x) = \min \sum_{i=1}^{3} \mathbb{E}\left[f_i(x, \xi_i)\right] \tag{22}$$

Each local objective function f_i is defined as

$$f_1(x,\xi_1) = \begin{cases} (x^3-16x)(x+2)+n_1x, & |x| \le 10 \\ 4248x-32400+n_1x, & x > 10 \\ -3112x-25040+n_1x, & x < -10 \end{cases} \tag{23}$$

$$f_2(x,\xi_2) = \begin{cases} (0.5x^3+x^2)(x-4)+n_2x, & |x| \le 10 \\ 1620x-12600+n_2x, & x > 10 \\ -2220x-16600+n_2x, & x < -10 \end{cases} \tag{24}$$

$$f_3(x,\xi_3) = \begin{cases} (x^3-16x)(x+2)+n_3x, & |x| \le 10 \\ 288x-2016+n_3x, & x > 10 \\ 228x-2624+n_3x, & x < -10 \end{cases} \tag{25}$$

where $\xi_i = n_i \sim \mathcal{N}(0,1)$. The objective function is non-convex, and its surrogate can be constructed using (2). First, D-MSSCA is used to demonstrate the effect of communication topology in Fig 1. We observe that the fully connected network performs better than the Tree network, which is how the behavior is expected. Next, setting $\alpha = 0.8$ and $\beta = 0.16$ and $\mu = 5000$, we plot the evolution of local variable x_i^t with different initial values in Fig 2 for a fully connected graph with $\lambda_{\mathbf{W}} = 0.5$. We observe the nodes converge to local minima. Finally, we plot the evolution of local variables of each node given a global constraint set $|x_i| \le 2.25$, when all the nodes are initialized at $x_i^t = 0$ in Fig 3. It can be observed that all the local variables get as close as possible to the true minima.

V. Conclusion and Future work

In this work, we consider decentralized consensus stochastic non-convex optimization to minimize the sum of non-convex (possibly smooth) and convex (possibly non-smooth) cost functions over a network of nodes. While

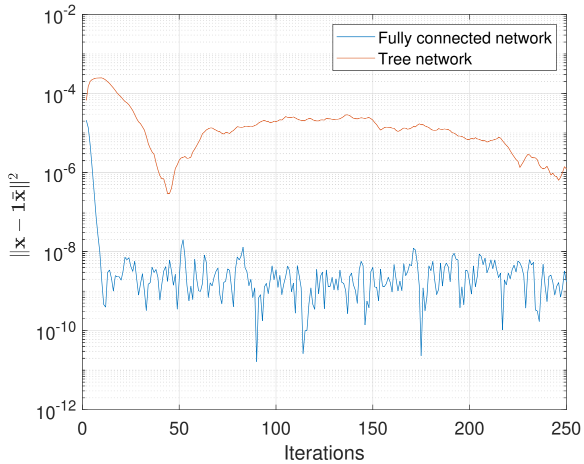

Fig. 1. Evolution of residual $\|\mathbf{x} - \mathbf{1}\bar{\mathbf{x}}\|^2$ over different networks.

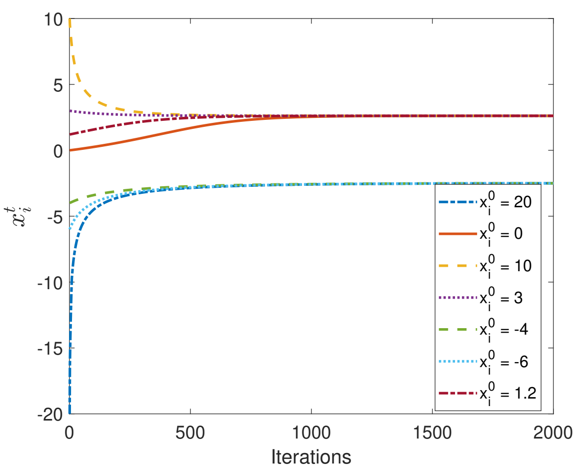

Fig. 2. Evolution of local variable for varying initialization.

this problem is well studied, comprehensive convergence analysis under the stochastic SCA rubric has remained an open problem. We have analyzed and proposed D-MSSCA that achieves the optimal rate of $\mathcal{O}(\epsilon^{-3/2})$. The rate matches the SFO rate of the state-of-the-art decentralized gradient-based algorithm while processing a single sample at each iteration. The algorithm uses strong convex surrogates and leverages recursive momentum-based updates at each node, achieving faster convergence. The applicability of D-MSSCA is demonstrated on a synthetic stochastic problem. One interesting future direction of this paper is under investigation, wherein we simplify the optimization problem (1) in Algorithm 1 by linearizing the global constraint g making it (1) easier to solve than proximal-based methods. By reducing the complexity of each iteration, this modification could improve convergence rates and expand the applicability of the D-MSSCA algorithm to broader classes of optimization problems, especially those where proximal methods face scalability and efficiency challenges.

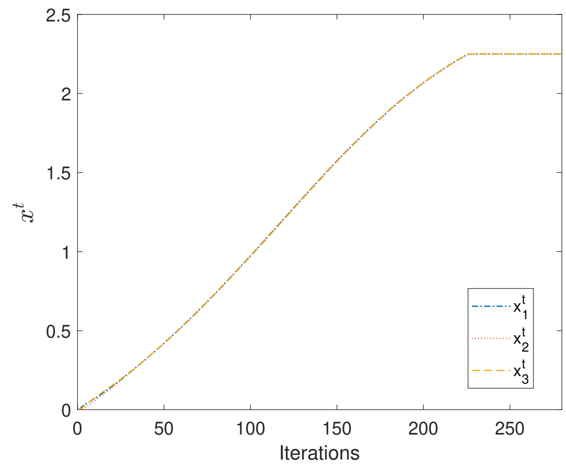

Fig. 3. Evolution of local variable when each node is initialized at $\mathbf{x}_i^t = 0$ given a global constraint $|x_i| \leq 2.25$.

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.

[2] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE transactions on automatic control*, vol. 58, no. 2, pp. 391–405, 2012.

[3] B. Swenson, R. Murray, H. V. Poor, and S. Kar, "Distributed stochastic gradient descent: Nonconvexity, nonsmoothness, and convergence to local minima," *Journal of Machine Learning Research*, vol. 23, no. 328, pp. 1–62, 2022.

[4] Z. Wang, J. Zhang, T.-H. Chang, J. Li, and Z.-Q. Luo, "Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4486–4501, 2021.

[5] R. Xin, S. Das, U. A. Khan, and S. Kar, "A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency," *arXiv preprint arXiv:2110.01594*, 2021.

[6] Y. Yan, J. Chen, P.-Y. Chen, X. Cui, S. Lu, and Y. Xu, "Compressed decentralized proximal stochastic gradient method for nonconvex composite problems with heterogeneous data," in *International Conference on Machine Learning*. PMLR, 2023, pp. 39 035–39 061.

[7] G. Mancino-Ball, S. Miao, Y. Xu, and J. Chen, "Proximal stochastic recursive momentum methods for nonconvex composite decentralized optimization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 9055–9063.

[8] L. Zheng, H. Li, J. Li, Z. Wang, Q. Lü, Y. Shi, H. Wang, T. Dong, L. Ji, and D. Xia, "A distributed nesterov-like gradient tracking algorithm for composite constrained optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 9, pp. 60–73, 2023.

[9] P. Di Lorenzo and S. Scardapane, "Distributed stochastic nonconvex optimization and learning based on successive convex approximation," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 1–5.

[10] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 641–656, 2013.

[11] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization—part I: Theory," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1929–1944, 2016.

[12] G. Scutari, F. Facchinei, L. Lampariello, S. Sardellitti, and P. Song, "Parallel and distributed methods for constrained nonconvex optimization-part II: Applications in communications and machine learning," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1945–1960, 2016.

[13] Y. Yang, G. Scutari, D. P. Palomar, and M. Pesavento, "A parallel decomposition method for nonconvex stochastic multi-agent optimization problems," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2949–2964, 2016.

[14] A. Liu, V. K. Lau, and B. Kananian, “Stochastic successive convex approximation for non-convex constrained stochastic optimization,” *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4189–4203, 2019.

[15] A. Liu, V. K. Lau, and M.-J. Zhao, “Online successive convex approximation for two-stage stochastic nonconvex optimization,” *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5941–5955, 2018.

[16] C. Ye and Y. Cui, “Stochastic successive convex approximation for general stochastic optimization problems,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 755–759, 2019.

[17] A. Liu, X. Chen, W. Yu, V. K. Lau, and M.-J. Zhao, “Two-timescale hybrid compression and forward for massive MIMO aided C-RAN,” *IEEE Trans. Signal Process.*, vol. 67, no. 9, pp. 2484–2498, 2019.

[18] A. Liu, R. Yang, T. Q. Quek, and M.-J. Zhao, “Two-stage stochastic optimization via primal-dual decomposition and deep unrolling,” *IEEE Trans. Signal Process.*, vol. 69, pp. 3000–3015, 2021.

[19] A. Mokhtari, A. Koppel, G. Scutari, and A. Ribeiro, “Large-scale nonconvex stochastic optimization by doubly stochastic successive convex approximation,” in *IEEE ICASSP*, 2017, pp. 4701–4705.

[20] A. Koppel, A. Mokhtari, and A. Ribeiro, “Parallel stochastic successive convex approximation method for large-scale dictionary learning,” in *IEEE ICASSP*, 2018, pp. 2771–2775.

[21] A. Mokhtari and A. Koppel, “High-dimensional nonconvex stochastic optimization by doubly stochastic successive convex approximation,” *IEEE Trans. Signal Process.*, vol. 68, pp. 6287–6302, 2020.

[22] B. M. Idrees, J. Akhtar, and K. Rajawat, “Practical precoding via asynchronous stochastic successive convex approximation,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 4177–4191, 2021.

[23] B. M. Idrees, L. Arora, and K. Rajawat, “Constrained stochastic recursive momentum successive convex approximation,” *arXiv preprint arXiv:2404.11790*, 2024.

[24] P. Di Lorenzo and G. Scutari, “Next: In-network nonconvex optimization,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[25] A. Cutkosky and F. Orabona, “Momentum-based variance reduction in non-convex sgd,” in *Advances in Neural Information Processing Systems*, 2019, pp. 15 236–15 245.

[26] R. Xin, U. Khan, and S. Kar, “A hybrid variance-reduced method for decentralized stochastic non-convex optimization,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 459–11 469.

[27] Q. Tran-Dinh, N. H. Pham, D. T. Phan, and L. M. Nguyen, “Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization,” *arXiv preprint arXiv:1905.05920*, 2019.

[28] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Tak’avc, “Sarah a novel method for machine learning problems using stochastic recursive gradient,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 2613–2621.

[29] S. D. Sharma and K. Rajawat, “Optimized gradient tracking for decentralized online learning,” *IEEE Transactions on Signal Processing*, vol. 72, pp. 1443–1459, 2024.

[30] R. Xin, U. A. Khan, and S. Kar, “An improved convergence analysis for decentralized online stochastic non-convex optimization,” *IEEE Transactions on Signal Processing*, vol. 69, pp. 1842–1858, 2021.

[31] ——, “Fast decentralized nonconvex finite-sum optimization with recursive variance reduction,” *SIAM Journal on Optimization*, vol. 32, no. 1, pp. 1–28, 2022.

[32] G. Qu and N. Li, “Harnessing smoothness to accelerate distributed optimization,” *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2017.

[33] T. Tatarenko and B. Touri, “Non-convex distributed optimization,” *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3744–3757, 2017.

Appendix A
Proof of Lemma 2

Proof: From the definition of θ^t(9) and $\mathbf{x}^t$-update (7), we have

$$\begin{aligned}
\left(\theta^t\right)^2 &= \left\|\mathbf{x}^t - \frac{1}{n}\left(\mathbf{1}_n \otimes \mathbf{I}_d\right)\bar{\mathbf{x}}^t\right\|^2, \\
&= \left\|\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\underline{\mathbf{W}}\left(\mathbf{x}^{t-1} + \alpha\left(\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right)\right)\right\|^2, \\
&\overset{\text{(a)}}{=} \left\|\left(\underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\mathbf{x}^{t-1} + \alpha\left(\underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\left(\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right)\right\|^2, \\
&\overset{\text{(b)}}{\leq} \left(1+\eta_1\right)\left\|\underline{\mathbf{W}}\mathbf{x}^{t-1} - \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\mathbf{x}^{t-1}\right\|^2 \\
&\quad + \left(1+\frac{1}{\eta_1}\right)\alpha^2\left(\lambda_{\max}\left(\underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\right)^2\left\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right\|^2, \\
&\overset{\text{(c)}}{\leq} \left(1+\eta_1\right)\lambda_{\mathbf{W}}{}^2\left\|\mathbf{x}^{t-1} - \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\mathbf{x}^{t-1}\right\|^2 + \left(1+\frac{1}{\eta_1}\right)\alpha^2\lambda_{\mathbf{W}}{}^2\left\|\hat{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right\|^2, \\
&= \left(1+\eta_1\right)\lambda_{\mathbf{W}}{}^2\left(\theta^{t-1}\right)^2 + \left(1+\frac{1}{\eta_1}\right)\alpha^2\lambda_{\mathbf{W}}{}^2\left\|\boldsymbol{\delta}^{t-1}\right\|^2.
\end{aligned}$$

In (a), we have applied the property of $\mathbf{W}$, i.e., $\mathbf{1}^\top\mathbf{W} = \mathbf{1}^\top$. In (b), Young's inequality and the property of norm are utilized, i.e., for any square matrix $\mathbf{A}$ and vector $\mathbf{x}$ of compatible size, $\|\mathbf{A}\mathbf{x}\|^2 \leq \lambda_{\max}(\mathbf{A})^2\|\mathbf{x}\|^2$. In (c), we have utilized (14a) and the definition of $\lambda_{\mathbf{W}}$.

■

Appendix B
Proof of Lemma 4

Proof: From the definition of ε^t and (8)

$$\begin{aligned}
\varepsilon^t &= \mathbb{E}\left[\left\|\underline{\mathbf{W}}\left(\mathbf{y}^{t-1} + \mathbf{z}^t - \mathbf{z}^{t-1}\right) - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\underline{\mathbf{W}}\left(\mathbf{y}^{t-1} + \mathbf{z}^t - \mathbf{z}^{t-1}\right)\right\|^2\right], \\
&\overset{\text{(i)}}{=} \mathbb{E}\left[\left\|\underline{\mathbf{W}}\left(\mathbf{y}^{t-1} + \mathbf{z}^t - \mathbf{z}^{t-1}\right) - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\left(\mathbf{y}^{t-1} + \mathbf{z}^t - \mathbf{z}^{t-1}\right)\right\|^2\right], \\
&= \mathbb{E}\Bigg[\left\|\underline{\mathbf{W}}\mathbf{y}^{t-1} - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\mathbf{y}^{t-1}\right\|^2 + \left\|\left(\underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\left(\mathbf{z}^t - \mathbf{z}^{t-1}\right)\right\|^2 \\
&\quad + \left\langle \underline{\mathbf{W}}\mathbf{y}^{t-1} - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\mathbf{y}^{t-1}, \left(\underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\left(\mathbf{z}^t - \mathbf{z}^{t-1}\right)\right\rangle\Bigg], \\
&\overset{\text{(14a)}}{\leq} \lambda_{\mathbf{W}}{}^2\mathbb{E}\left[\left\|\mathbf{y}^{t-1} - \frac{1}{n}\left(\mathbf{1}_n \otimes \mathbf{I}_d\right)\bar{\mathbf{y}}^{t-1}\right\|^2\right] + \lambda_{\mathbf{W}}{}^2\mathbb{E}\left[\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\left\langle \underline{\mathbf{W}}\mathbf{y}^{t-1} - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\mathbf{y}^{t-1}, \left(\underline{\mathbf{W}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_d\right)\left(\mathbf{z}^t - \mathbf{z}^{t-1}\right)\right\rangle\right].
\end{aligned} \tag{26}$$

Now we will simplify each term in the above equation separately, starting with $\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2$ we have

$$\left\|\mathbf{z}^t - \mathbf{z}^{t-1}\right\|^2 = \sum_{i=1}^{n}\left\|\mathbf{z}_i^t - \mathbf{z}_i^{t-1}\right\|^2,$$

$$\stackrel{(4)}{=}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}_i^t,\xi_i^t)+(1-\beta)\left(\mathbf{z}_i^{t-1}-\nabla f_i(\mathbf{x}_i^{t-1},\xi_i^t)\right)-\mathbf{z}_i^{t-1}\right\|^2,$$

$$=\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}_i^t,\xi_i^t)-\nabla f_i(\mathbf{x}_i^{t-1},\xi_i^t)-\beta\mathbf{z}_i^{t-1}+\beta\nabla f_i(\mathbf{x}_i^{t-1},\xi_i^t)\right\|^2. \tag{27}$$

On adding and subtraction by $\beta\nabla u_i(\mathbf{x}^{t-1})$ and simplifying further we get

$$\begin{aligned}&\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2\\&=\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}_i^t,\xi_i^t)-\nabla f_i(\mathbf{x}_i^{t-1},\xi_i^t)-\beta\left(\mathbf{z}_i^{t-1}-\nabla u_i(\mathbf{x}^{t-1})\right)+\beta\nabla f_i(\mathbf{x}_i^{t-1},\xi_i^t)-\beta\nabla u_i(\mathbf{x}^{t-1})\right\|^2,\\&\le\sum_{i=1}^{n}3\left\|\nabla f_i(\mathbf{x}_i^t,\xi_i^t)-\nabla f_i(\mathbf{x}_i^{t-1},\xi_i^t)\right\|^2+\sum_{i=1}^{n}3\beta^2\left\|\left(\mathbf{z}_i^{t-1}-\nabla u_i(\mathbf{x}^{t-1})\right)\right\|^2\\&\quad+\sum_{i=1}^{n}3\beta^2\left\|\nabla f_i(\mathbf{x}_i^{t-1},\xi_i^t)-\nabla u_i(\mathbf{x}^{t-1})\right\|^2.\end{aligned}$$

On taking expectation on both sides and further utilizing Assumptions **A3** and **A4** we get

$$\mathbb{E}\left[\left\|\mathbf{z}^t-\mathbf{z}^{t-1}\right\|^2\right]\le 3L^2\mathbb{E}\left[\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2\right]+3\beta^2\mathbb{E}\left[\left\|\left(\mathbf{z}^{t-1}-\nabla u(\mathbf{x}^{t-1})\right)\right\|^2\right]+3\beta^2\bar{\sigma}^2. \tag{28}$$

By using conditional expectation, we can further simplify the last term of (26) as

$$\begin{aligned}&2\mathbb{E}\left[\left\langle\underline{\mathbf{W}}\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\mathbf{y}^{t-1},\left(\underline{\mathbf{W}}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\left(\mathbf{z}^t-\mathbf{z}^{t-1}\right)\right\rangle\right],\\&\quad=2\mathbb{E}\left[\mathbb{E}\left[\left\langle\underline{\mathbf{W}}\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\mathbf{y}^{t-1},\left(\underline{\mathbf{W}}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\left(\mathbf{z}^t-\mathbf{z}^{t-1}\right)\right\rangle\middle|\mathcal{H}_k\right]\right],\\&\quad=2\mathbb{E}\left[\left\langle\underline{\mathbf{W}}\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\mathbf{y}^{t-1},\left(\underline{\mathbf{W}}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\mathbb{E}\left[\mathbf{z}^t-\mathbf{z}^{t-1}|\mathcal{H}_k\right]\right\rangle\right].\end{aligned}$$

By applying (27) and utilizing Assumption **A2**, we have $\mathbb{E}\left[\mathbf{z}^t-\mathbf{z}^{t-1}|\mathcal{H}_k\right]=\nabla u(\mathbf{x}^t)-\nabla u(\mathbf{x}^t)-\beta(\mathbf{z}^{t-1}-\nabla u(\mathbf{x}^{t-1}))$. Substituting this and applying the Cauchy-Schwarz inequality, we get

$$\begin{aligned}&2\mathbb{E}\left[\left\langle\underline{\mathbf{W}}\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\mathbf{y}^{t-1},\left(\underline{\mathbf{W}}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\left(\mathbf{z}^t-\mathbf{z}^{t-1}\right)\right\rangle\right]\\&\quad=2\mathbb{E}\Bigg[\left\|\underline{\mathbf{W}}\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\mathbf{y}^{t-1}\right\|\times\\&\qquad\left\|\left(\underline{\mathbf{W}}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\left(\nabla u(\mathbf{x}^t)-\nabla u(\mathbf{x}^t)-\beta(\mathbf{z}^{t-1}-\nabla u^{t-1})\right)\right\|\Bigg],\\&\quad\stackrel{(14a)}{\le}2\mathbb{E}\left[\lambda_{\mathbf{W}}{}^2\left\|\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\otimes\mathbf{I}_d\right)\bar{\mathbf{y}}^{t-1}\right\|\left\|\nabla u(\mathbf{x}^t)-\nabla u(\mathbf{x}^t)-\beta(\mathbf{z}^{t-1}-\nabla u^{t-1})\right\|\right].\end{aligned}$$

On further applying Young's inequality and a few mathematical simplifications, we get

$$\begin{aligned}&\mathbb{E}\left[\left\langle\underline{\mathbf{W}}\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\mathbf{y}^{t-1},\left(\underline{\mathbf{W}}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal\otimes\mathbf{I}_d\right)\left(\mathbf{z}^t-\mathbf{z}^{t-1}\right)\right\rangle\right]\\&\quad\le\lambda_{\mathbf{W}}{}^2\gamma_1\mathbb{E}\left[\left\|\mathbf{y}^{t-1}-\frac{1}{n}\left(\mathbf{1}_n\otimes\mathbf{I}_d\right)\bar{\mathbf{y}}^{t-1}\right\|^2\right]+\frac{2\lambda_{\mathbf{W}}{}^2L^2}{\gamma_1}\mathbb{E}\left[\left\|\mathbf{x}^t-\mathbf{x}^{t-1}\right\|^2\right]\\&\quad+\frac{2\lambda_{\mathbf{W}}{}^2\beta^2}{\gamma_1}\mathbb{E}\left[\left\|(\mathbf{z}^{t-1}-\nabla u^{t-1})\right\|^2\right].\end{aligned} \tag{29}$$

By substituting equations (28), (29) in (26), we can write

$$\varepsilon^t \leq \lambda_{\mathbf{W}}{}^2 \left(1+\gamma_1\right) \varepsilon^{t-1} + \lambda_{\mathbf{W}}{}^2 \beta^2 \left(\frac{2}{\gamma_1}+3\right) \upsilon^{t-1} + 3\lambda_{\mathbf{W}}{}^2 \beta^2 \bar{\sigma}^2$$
$$+ \lambda_{\mathbf{W}}{}^2 L^2 \left(3+\frac{2}{\gamma_1}\right) \mathbb{E}\left[\left\|\mathbf{x}^t - \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^t + \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^t - \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^{t-1} + \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right\|^2\right].$$

We can further simplify the last term as follows

$$\begin{aligned}
\left\|\mathbf{x}^t - \mathbf{x}^{t-1}\right\|^2 &= \left\|\mathbf{x}^t - \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^t + \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^t - \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^{t-1} + \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right\|^2, \\
&\leq 3\left\|\mathbf{x}^t - \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^t\right\|^2 + 3\left\|\mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^t - \mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^{t-1}\right\|^2 + 3\left\|\mathbf{1}\otimes\mathbf{I}_d\bar{\mathbf{x}}^{t-1} - \mathbf{x}^{t-1}\right\|^2, \\
&\overset{(7)}{\leq} 3\left(\theta^t\right)^2 + 3n\alpha^2\left\|\bar{\hat{\mathbf{x}}}^{t-1} - \bar{\mathbf{x}}^{t-1}\right\|^2 + 3\left(\theta^{t-1}\right)^2, \\
&\overset{\text{Lemma 2}}{\leq} 3\left[(1+\eta_1)\lambda_{\mathbf{W}}{}^2\left(\theta^{t-1}\right)^2 + \left(1+\frac{1}{\eta_1}\right)\alpha^2\lambda_{\mathbf{W}}{}^2\left\|\boldsymbol{\delta}^{t-1}\right\|^2\right] \\
&\qquad + 3\alpha^2\left\|\boldsymbol{\delta}^{t-1}\right\|^2 + 3\left(\theta^{t-1}\right)^2, \\
&= 3\alpha^2\left(\left(1+\frac{1}{\eta_1}\right)\lambda_{\mathbf{W}}{}^2 + 1\right)\left\|\boldsymbol{\delta}^{t-1}\right\|^2 + 3((1+\eta_1)\lambda^2+1)\left(\theta^{t-1}\right)^2.
\end{aligned}$$

Finally, we get

$$\varepsilon^t \leq \lambda_{\mathbf{W}}{}^2 \left(1+\gamma_1\right) \varepsilon^{t-1} + \lambda_{\mathbf{W}}{}^2 \beta^2 \left(\frac{2}{\gamma_1}+3\right) \upsilon^{t-1} + 3\lambda_{\mathbf{W}}{}^2 L^2 ((1+\eta_1)\lambda_{\mathbf{W}}{}^2+1)\left(\frac{2}{\gamma_1}+3\right)\mathbb{E}\left[(\theta^t)^2\right]$$
$$+ 3\lambda_{\mathbf{W}}{}^2\beta^2\bar{\sigma}^2 + 3\alpha^2\lambda_{\mathbf{W}}{}^2 L^2\left(\frac{2}{\gamma_1}+3\right)\left(\lambda_{\mathbf{W}}{}^2\left(1+\frac{1}{\eta_1}\right)+1\right)\mathbb{E}\left[\left\|\boldsymbol{\delta}^{t-1}\right\|^2\right],$$

where $\gamma_1, \eta_1 > 0$ are Young's parameters. By considering $\gamma_1 = \frac{1-\lambda_{\mathbf{W}}{}^2}{2\lambda_{\mathbf{W}}{}^2}, \eta_1 = 1$, we get the desired bound. ■

Appendix C

Proof of Lemma 6

Proof: On substituting $\eta_1 = \frac{1-\lambda_{\mathbf{W}}{}^2}{2\lambda_{\mathbf{W}}{}^2}$ in Lemma 2, we get

$$(\theta^t)^2 \leq \left(\frac{1+\lambda_{\mathbf{W}}{}^2}{2}\right)\left(\theta^{t-1}\right)^2 + \frac{2\alpha^2\lambda_{\mathbf{W}}{}^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)}\left\|\boldsymbol{\delta}^{t-1}\right\|^2.$$

Applying Lemma 5, we get

$$\sum_{t=1}^{T}(\theta^t)^2 \leq \frac{2}{1-\lambda_{\mathbf{W}}{}^2}(\theta^1)^2 + \frac{4\alpha^2\lambda_{\mathbf{W}}{}^2}{\left(1-\lambda_{\mathbf{W}}{}^2\right)^2}\sum_{t=1}^{T-1}\left\|\boldsymbol{\delta}^t\right\|^2.$$

From the initialization, we have $\theta^1 = 0$ and substituting this yields the desired result. ■

Appendix D

Proof of Lemma 7

Proof: On summing (15) for $1 \leq t \leq T$ and applying (18) with $\eta_2 = 1$, we obtain,

$$\sum_{t=1}^{T}\phi^t \leq \frac{\phi^1}{1-(1-\beta)^2} + \frac{6L^2(1-\beta)^2\alpha^2}{n^2(1-(1-\beta)^2)}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right] + \frac{12L^2(1-\beta)^2}{(1-(1-\beta)^2)n^2}\sum_{t=0}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]$$

$$+\frac{2\beta^2\bar{\sigma}^2T}{n^2(1-(1-\beta)^2)}.$$

Observing that $\frac{1}{1-(1-\beta)^2} \le \frac{1}{\beta}$ for $\beta \in (0,1)$ we have,

$$\sum_{t=1}^{T}\phi^t \le \frac{\phi^1}{\beta} + \frac{2\beta\bar{\sigma}^2T}{n^2} + \frac{6L^2\alpha^2}{n^2\beta}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right] + \frac{12L^2}{\beta n^2}\sum_{t=0}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right].$$

Also, based on the initialization of $\mathbf{z}_i^1$ and Assumption (**A3**), we have:

$$\phi^1 = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i^1 - \frac{1}{n}\sum_{i=1}^{n}\nabla u_i(\mathbf{x}_i^1)\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b_0}\sum_{r=1}^{b_0}\left(\nabla f_i(\mathbf{x}_i^1,\xi_i^{1,r}) - \nabla u_i(\mathbf{x}_i^1)\right)\right\|^2\right],$$
$$\overset{\text{(i)}}{\le} \frac{\bar{\sigma}^2}{n^2b_0}.$$

In (i), we have applied Assumption **A3** and the fact that stochastic local gradient oracles at each node are independent. Substituting ϕ^1 yields the desired result. The second result can also be obtained by starting from (16) and following similar steps. ■

APPENDIX E
PROOF OF LEMMA 8

Proof: On summing the bound obtained in Lemma 4 for $1 \le t \le T$ and applying (18), we have

$$\sum_{t=1}^{T}\varepsilon^t \le \frac{2}{1-\lambda_{\mathbf{W}}^2}\varepsilon^1 + \frac{8\beta^2\lambda_{\mathbf{W}}^2}{\left(1-\lambda_{\mathbf{W}}^2\right)^2}\sum_{t=2}^{T}v^{t-1} + \frac{6\lambda_{\mathbf{W}}^2\beta^2\bar{\sigma}^2T}{1-\lambda_{\mathbf{W}}^2} + \frac{72\lambda_{\mathbf{W}}^2L^2}{\left(1-\lambda_{\mathbf{W}}^2\right)^2}\sum_{t=2}^{T}\mathbb{E}\left[(\theta^{t-1})^2\right]$$
$$+\frac{72\alpha^2\lambda_{\mathbf{W}}^2L^2}{\left(1-\lambda_{\mathbf{W}}^2\right)^2}\sum_{t=2}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^{t-1}\right\|^2\right],$$
$$\overset{\text{Lemma (7)}}{\le} \frac{2}{1-\lambda_{\mathbf{W}}^2}\varepsilon^1 + \frac{72\alpha^2\lambda_{\mathbf{W}}^2L^2}{\left(1-\lambda_{\mathbf{W}}^2\right)^2}\sum_{t=2}^{T}\mathbb{E}\left[\left\|\boldsymbol{\delta}^{t-1}\right\|^2\right] + \frac{72\lambda_{\mathbf{W}}^2L^2}{\left(1-\lambda_{\mathbf{W}}^2\right)^2}\sum_{t=2}^{T}\mathbb{E}\left[(\theta^{t-1})^2\right]$$
$$+\frac{8\beta^2\lambda_{\mathbf{W}}^2}{\left(1-\lambda_{\mathbf{W}}^2\right)^2}\left(\frac{\bar{\sigma}^2}{b_0\beta} + 2\beta\bar{\sigma}^2T + \frac{6L^2\alpha^2}{\beta}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\boldsymbol{\delta}^t\right\|^2\right] + \frac{12L^2}{\beta}\sum_{t=1}^{T}\mathbb{E}\left[\left(\theta^t\right)^2\right]\right) + \frac{6\lambda_{\mathbf{W}}^2\beta^2\bar{\sigma}^2T}{1-\lambda_{\mathbf{W}}^2}.$$

On further combining the common terms, we get the desired result. ■

APPENDIX F
PROOF OF LEMMA 9

Proof: Since the surrogate $\tilde{f}$ is a strongly convex function, solving (1) is equivalent to solving a simple convex optimization problem. The optimality condition of convex optimization problem (1) implies:

$$\langle\nabla\tilde{f}\left(\hat{\mathbf{x}}_i^t,\mathbf{x}_i^t,\xi_i^t\right) + \pi_i^t + \hat{\mathbf{w}}_i^t, \mathbf{x}_i^t - \hat{\mathbf{x}}_i^t\rangle \ge 0,$$

where $\hat{\mathbf{w}}_i^t \in \partial(h + \mathbf{1}_{\mathcal{X}})\mid_{\mathbf{x}=\hat{\mathbf{x}}_i^t}$, and $\mathbf{1}_{\mathcal{X}}(\mathbf{x})$ is an indicator function. $\mathbf{1}_{\mathcal{X}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{X}$, otherwise $\mathbf{1}_{\mathcal{X}}(\mathbf{x}) = \infty$. As per the definition of $\tilde{f}\left(\mathbf{x}_i,\mathbf{x}_i^t,\xi_i^t\right)$ given in (2) and using (4), we have $\nabla\tilde{f}(\mathbf{x}_i^t,\mathbf{x}_i^t,\boldsymbol{\xi}_i^t) = \mathbf{z}_i^t$. Furthermore, by adding and subtracting $\nabla\tilde{f}(\mathbf{x}_i^t,\mathbf{x}_i^t,\boldsymbol{\xi}_i^t) = \mathbf{z}_i^t$ and substituting $\pi_i^t = \mathbf{y}_i^t - \mathbf{z}_i^t$, we get

$$\langle\nabla\tilde{f}\left(\hat{\mathbf{x}}_i^t,\mathbf{x}_i^t,\xi_i^t\right) + \mathbf{y}_i^t - \mathbf{z}_i^t + \mathbf{z}_i^t - \nabla\tilde{f}\left(\mathbf{x}_i^t,\mathbf{x}_i^t,\xi_i^t\right) + \hat{\mathbf{w}}_i^t, \mathbf{x}_i^t - \hat{\mathbf{x}}_i^t\rangle \ge 0.$$

From the definition of $\tilde{f}\left(\mathbf{x}_i, \mathbf{x}_i^t, \xi_i^t\right)$ (2), and utilizing (4), we have $\nabla \tilde{f}\left(\hat{\mathbf{x}}_i^t, \mathbf{x}_i^t, \xi_i^t\right) - \mathbf{z}_i^t = \mu_i(\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t)$, substituting this gives us:

$$\langle \mu_i(\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t) + \mathbf{y}_i^t + \hat{\mathbf{w}}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \rangle \leq 0,$$
$$\langle \mathbf{y}_i^t + \hat{\mathbf{w}}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \rangle \leq -\mu \left\| \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \right\|^2.$$

Further, on summing the above inequality over all i we get

$$\frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{y}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \rangle + \frac{1}{n}\sum_{i=1}^{n} \langle \hat{\mathbf{w}}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \rangle \leq \frac{-1}{n}\sum_{i=1}^{n} \mu \left\| \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \right\|^2. \tag{30}$$

From the update equation (3), convexity of $h + \mathbf{1}_{\mathcal{X}}$ and Assumption (**A7**) we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \left[h(\mathbf{x}_i^{t+1}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}_i^{t+1})\right] &= \frac{1}{n}\sum_{i=1}^{n} h\left(\sum_{j=1}^{n} W_{i,j}\left(\mathbf{x}_j^t + \alpha\left(\hat{\mathbf{x}}_j^t - \mathbf{x}_j^t\right)\right)\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\mathcal{X}}\left(\sum_{j=1}^{n} W_{i,j}\left(\mathbf{x}_j^t + \alpha\left(\hat{\mathbf{x}}_j^t - \mathbf{x}_j^t\right)\right)\right), \\
&\overset{(i)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{i,j} h\left(\mathbf{x}_j^t + \alpha\left(\hat{\mathbf{x}}_j^t - \mathbf{x}_j^t\right)\right) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{i,j}\mathbf{1}_{\mathcal{X}}\left(\mathbf{x}_j^t + \alpha\left(\hat{\mathbf{x}}_j^t - \mathbf{x}_j^t\right)\right), \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} W_{i,j}\right) h\left((1-\alpha)\mathbf{x}_j^t + \alpha\hat{\mathbf{x}}_j^t\right) \\
&\quad + \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} W_{i,j}\right) \mathbf{1}_{\mathcal{X}}\left((1-\alpha)\mathbf{x}_j^t + \alpha\hat{\mathbf{x}}_j^t\right), \\
&\overset{(ii)}{\leq} \frac{1}{n}\sum_{j=1}^{n}\left((1-\alpha)h\left(\mathbf{x}_j^t\right) + \alpha h\left(\hat{\mathbf{x}}_j^t\right)\right) + \frac{1}{n}\sum_{j=1}^{n}\left((1-\alpha)\mathbf{1}_{\mathcal{X}}\left(\mathbf{x}_j^t\right) + \alpha\mathbf{1}_{\mathcal{X}}\left(\hat{\mathbf{x}}_j^t\right)\right). \\
&= \frac{(1-\alpha)}{n}\sum_{j=1}^{n}\left(h\left(\mathbf{x}_j^t\right) + \mathbf{1}_{\mathcal{X}}\left(\mathbf{x}_j^t\right)\right) + \frac{\alpha}{n}\sum_{j=1}^{n}\left(h\left(\hat{\mathbf{x}}_j^t\right) + \mathbf{1}_{\mathcal{X}}\left(\hat{\mathbf{x}}_j^t\right)\right).
\end{aligned}$$

In (i) and (ii), we have utilized the convexity of $h + \mathbf{1}_{\mathcal{X}}$, property of $\mathbf{W}$ being doubly stochastic, and $W_{i,j} > 0$ for all $i, j \in \mathcal{V}$ ((**A7**)).

From the first order convexity condition of $h(\mathbf{x}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}_i^t)$, we have

$$h(\hat{\mathbf{x}}_i^t) + \mathbf{1}_{\mathcal{X}}(\hat{\mathbf{x}}_i^t) \leq h(\mathbf{x}_i^t) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}_i^t) + \langle \hat{\mathbf{w}}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \rangle,$$

Using this we can simplify further as below:

$$\frac{1}{n}\sum_{i=1}^{n} \left[h(\mathbf{x}_i^{t+1}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}_i^{t+1})\right] \leq \frac{(1-\alpha)}{n}\sum_{j=1}^{n}\left(h\left(\mathbf{x}_j^t\right) + \mathbf{1}_{\mathcal{X}}\left(\mathbf{x}_j^t\right)\right) + \frac{\alpha}{n}\sum_{j=1}^{n}\left(h(\mathbf{x}_j^t) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}_j^t) + \langle \hat{\mathbf{w}}_j^t, \hat{\mathbf{x}}_j^t - \mathbf{x}_j^t \rangle\right),$$

$$\frac{1}{n}\sum_{i=1}^{n} \left[h(\mathbf{x}_i^{t+1}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}_i^{t+1})\right] \leq \frac{1}{n}\sum_{j=1}^{n}\left(h\left(\mathbf{x}_j^t\right) + \mathbf{1}_{\mathcal{X}}\left(\mathbf{x}_j^t\right)\right) + \frac{\alpha}{n}\sum_{j=1}^{n}\langle \hat{\mathbf{w}}_j^t, \hat{\mathbf{x}}_j^t - \mathbf{x}_j^t \rangle. \tag{31}$$

On dividing (31) by α and adding in (30) we get

$$\frac{1}{\alpha n}\sum_{i=1}^{n} \left[h(\mathbf{x}_i^{t+1}) + \mathbf{1}_{\mathcal{X}}(\mathbf{x}_i^{t+1})\right] + \frac{1}{n}\sum_{i=1}^{n} \langle \mathbf{y}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \rangle + \frac{1}{n}\sum_{i=1}^{n} \langle \hat{\mathbf{w}}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t \rangle$$

$$\leq \frac{1}{\alpha n}\sum_{i=1}^{n}\left(h\left(\mathbf{x}_i^t\right)\right) + \frac{1}{n}\sum_{j=1}^{n}\langle \hat{\mathbf{w}}_j^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\rangle + \frac{-1}{n}\sum_{i=1}^{n}\mu\left\|\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\right\|^2.$$

Furthermore, as $\mathcal{X}$ is a convex set ($g(\mathbf{x})$ is convex), $\hat{\mathbf{x}}_i^t \in \mathcal{X}$ (1), Algorithm 1 is initialized with a feasible point, and update equation (7) is a convex combination of vectors within the set $\mathcal{X}$. Therefore, we conclude that $\mathbf{x}_i^t \in \mathcal{X}$ and $\mathbf{1}_{\mathcal{X}}\left(\mathbf{x}_i^t\right) = 0$ for all $i \in \mathcal{V}$ and $t > 0$. Using this, we obtain:

$$\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{y}_i^t, \hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\rangle \leq \frac{-1}{n}\sum_{i=1}^{n}\mu\left\|\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\right\|^2 - \frac{1}{\alpha n}\sum_{i=1}^{n}h(\mathbf{x}_i^{t+1}) + \frac{1}{\alpha n}\sum_{i=1}^{n}h\left(\mathbf{x}_i^t\right). \tag{32}$$

From the smoothness of $u(\mathbf{x})$ (Assumption **A4**), we have

$$u(\bar{\mathbf{x}}^{t+1}) \leq u(\bar{\mathbf{x}}^t) + \langle \nabla u(\bar{\mathbf{x}}^t), \left(\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right)\rangle + \frac{L}{2}\left\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\right\|^2$$

Furthermore, pre-multiplying the update equation (7) by $\frac{1}{n}\left(\mathbf{1}^\top \otimes \mathbf{I}_d\right)$ gives

$$\bar{\mathbf{x}}^{t+1} = \bar{\mathbf{x}}^t + \alpha\left(\bar{\hat{\mathbf{x}}}^t - \bar{\mathbf{x}}^t\right) \qquad \left(\text{As, } \tfrac{1}{n}\left(\mathbf{1}^\top \otimes \mathbf{I}_d\right)\underline{\mathbf{W}} = \tfrac{1}{n}\left(\mathbf{1}^\top \otimes \mathbf{I}_d\right)\right)$$

Using this we can further simplify the quadratic upper bound of $u(\mathbf{x})$, as below:

$$\begin{aligned}
u(\bar{\mathbf{x}}^{t+1}) &\leq u(\bar{\mathbf{x}}^t) + \alpha\left\langle \nabla u(\bar{\mathbf{x}}^t) - \mathbf{y}_i^t + \mathbf{y}_i^t, \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i^t - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^t\right)\right\rangle + \frac{\alpha^2 L}{2}\left\|\bar{\hat{\mathbf{x}}}^t - \bar{\mathbf{x}}^t\right\|^2, \\
&= u(\bar{\mathbf{x}}^t) + \frac{\alpha}{n}\sum_{i=1}^{n}\langle \nabla u(\bar{\mathbf{x}}^t) - \mathbf{y}_i^t, \left(\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\right)\rangle + \frac{\alpha}{n}\sum_{i=1}^{n}\langle \mathbf{y}_i^t, \left(\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\right)\rangle + \frac{\alpha^2 L}{2}\left\|\bar{\hat{\mathbf{x}}}^t - \bar{\mathbf{x}}^t\right\|^2, \\
&\overset{(32)}{\leq} u(\bar{\mathbf{x}}^t) + \frac{\alpha}{n}\langle (\mathbf{1}\otimes\mathbf{I}_d)\nabla u(\bar{\mathbf{x}}^t) - \mathbf{y}^t, \hat{\mathbf{x}}^t - \mathbf{x}^t\rangle + \frac{\alpha^2 L}{2}\left\|\bar{\hat{\mathbf{x}}}^t - \bar{\mathbf{x}}^t\right\|^2 \\
&\quad + \frac{\alpha}{n}\left(-\sum_{i=1}^{n}\mu\left\|\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\right\|^2 - \frac{1}{\alpha}\sum_{i=1}^{n}h(\mathbf{x}_i^{t+1}) + \frac{1}{\alpha}\sum_{i=1}^{n}h\left(\mathbf{x}_i^t\right)\right),
\end{aligned}$$

$$\begin{aligned}
&u(\bar{\mathbf{x}}^{t+1}) + \frac{1}{n}\sum_{i=1}^{n}h(\mathbf{x}_i^{t+1}) - u(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^{n}h\left(\mathbf{x}_i^t\right) \\
&\qquad \leq \frac{\alpha}{n}\langle (\mathbf{1}\otimes\mathbf{I}_d)\nabla u(\bar{\mathbf{x}}^t) - \mathbf{y}^t, \hat{\mathbf{x}}^t - \mathbf{x}^t\rangle + \frac{\alpha^2 L}{2}\left\|\bar{\hat{\mathbf{x}}}^t - \bar{\mathbf{x}}^t\right\|^2 - \frac{\alpha\mu}{n}\sum_{i=1}^{n}\left\|\hat{\mathbf{x}}_i^t - \mathbf{x}_i^t\right\|^2.
\end{aligned}$$

On further applying Cauchy Schwartz inequality and peter-paul's inequality for $\gamma_1 > 0$ we get,

$$\begin{aligned}
&u(\bar{\mathbf{x}}^{t+1}) + \frac{1}{n}\sum_{i=1}^{n}h(\mathbf{x}_i^{t+1}) - u(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^{n}h\left(\mathbf{x}_i^t\right) \\
&\leq \frac{\alpha\gamma_1}{2n}\left\|(\mathbf{1}\otimes\mathbf{I}_d)\nabla u(\bar{\mathbf{x}}^t) - \mathbf{y}^t\right\|^2 + \frac{\alpha}{2n\gamma_1}\left\|\hat{\mathbf{x}}^t - \mathbf{x}^t\right\|^2 + \frac{\alpha^2 L}{2}\left\|\bar{\hat{\mathbf{x}}}^t - \bar{\mathbf{x}}^t\right\|^2 - \frac{\alpha\mu}{n}\left\|\hat{\mathbf{x}}^t - \mathbf{x}^t\right\|^2 \\
&\overset{(14d)}{\leq} \frac{\alpha\gamma_1}{2n}\left\|(\mathbf{1}\otimes\mathbf{I}_d)\nabla u(\bar{\mathbf{x}}^t) - (\mathbf{1}\otimes\mathbf{I}_d)\bar{\nabla} u(\mathbf{x}^t) + (\mathbf{1}\otimes\mathbf{I}_d)\bar{\nabla} u(\mathbf{x}^t) - (\mathbf{1}\otimes\mathbf{I}_d)\bar{\mathbf{y}}^t + (\mathbf{1}\otimes\mathbf{I}_d)\bar{\mathbf{y}}^t - \mathbf{y}^t\right\| \\
&+ \frac{\alpha}{n}\left(-\mu + \frac{1}{2\gamma_1} + \frac{\alpha L}{2}\right)\left\|\hat{\mathbf{x}}^t - \mathbf{x}^t\right\|
\end{aligned}$$

By further applying the Cauchy-Schwarz inequality and the Peter-Paul inequality for $\gamma_1 > 0$, we obtain:

$$\begin{aligned}
&u(\bar{\mathbf{x}}^{t+1}) + \frac{1}{n}\sum_{i=1}^{n}h(\mathbf{x}_i^{t+1}) - u(\bar{\mathbf{x}}^t) - \frac{1}{n}\sum_{i=1}^{n}h\left(\mathbf{x}_i^t\right) \\
&\qquad\qquad \leq \frac{\alpha}{2n}\gamma_1\left(3L^2\left\|\mathbf{x}^t - \underline{\mathbf{1}}\bar{\mathbf{x}}^t\right\|^2 + 3n\left\|\bar{\nabla} u(\mathbf{x}^t) - \bar{\mathbf{z}}^t\right\|^2 + 3\left\|\mathbf{y}^t - \underline{\mathbf{1}}\bar{\mathbf{y}}^t\right\|^2\right) \\
&\qquad\qquad\quad + \frac{\alpha}{n}\left(-\mu + \frac{1}{2\gamma_1} + \frac{\alpha L}{2}\right)\left\|\hat{\mathbf{x}}^t - \mathbf{x}^t\right\|^2.
\end{aligned}$$

On taking the Expectation on both side and summing for all $1 \leq t \leq T$

$$\begin{aligned}
&\sum_{t=1}^{T} u(\bar{\mathbf{x}}^{t+1}) + \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} h(\mathbf{x}_i^{t+1}) - \sum_{t=1}^{T} u(\bar{\mathbf{x}}^{t}) - \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n} h\left(\mathbf{x}_i^{t}\right) \\
&\leq \frac{\alpha}{2n}\gamma_1 \left(3L^2 \sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{x}^t - \underline{\mathbf{1}}\bar{\mathbf{x}}^t\right\|^2\right] + 3n \sum_{t=1}^{T} \mathbb{E}\left[\left\|\bar{\nabla} u(\mathbf{x}^t) - \bar{\mathbf{z}}^t\right\|^2\right] + 3\sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{y}^t - \underline{\mathbf{1}}\bar{\mathbf{y}}^t\right\|^2\right]\right) \\
&\quad + \frac{\alpha}{n}\left(-\mu + \frac{1}{2\gamma_1} + \frac{\alpha L}{2}\right)\sum_{t=1}^{T} \mathbb{E}\left[\left\|\hat{\mathbf{x}}^t - \mathbf{x}^t\right\|^2\right].
\end{aligned}$$

which, on further simplifications, gives

$$\begin{aligned}
&u(\bar{\mathbf{x}}^{T+1}) + \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i^{T+1}) - u(\bar{\mathbf{x}}^{1}) - \frac{1}{n}\sum_{i=1}^{n} h\left(\mathbf{x}_i^{1}\right) \\
&\leq \frac{\alpha}{2n}\gamma_1 \left(3L^2 \sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{x}^t - \underline{\mathbf{1}}\bar{\mathbf{x}}^t\right\|^2\right] + 3n \sum_{t=1}^{T} \mathbb{E}\left[\left\|\bar{\nabla} u(\mathbf{x}^t) - \bar{\mathbf{z}}^t\right\|^2\right] + 3\sum_{t=1}^{T} \mathbb{E}\left[\left\|\mathbf{y}^t - \underline{\mathbf{1}}\bar{\mathbf{y}}^t\right\|^2\right]\right) \\
&\quad + \frac{\alpha}{n}\left(-\mu + \frac{1}{2\gamma_1} + \frac{\alpha L}{2}\right)\sum_{t=1}^{T} \mathbb{E}\left[\left\|\hat{\mathbf{x}}^t - \mathbf{x}^t\right\|^2\right].
\end{aligned}$$

Further from the zeroth order convexity condition of $h(\mathbf{x})$ we have $h(\bar{\mathbf{x}}^{T+1}) \leq \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i^{T+1})$ and from the initialization of D-MSSCA we have $\mathbf{x}_i^1 = \bar{\mathbf{x}}^1$ for all $i \in \mathcal{V}$, utilizing which we get the desired result.

■