

# Gradient-Free Distributed Optimization with Exact Convergence

Yipeng Pang and Guoqiang Hu

**Abstract**—In this paper, a gradient-free distributed algorithm is introduced to solve a set constrained optimization problem under a directed communication network. Specifically, at each time-step, the agents locally compute a so-called pseudo-gradient to guide the updates of the decision variables, which can be applied in the fields where the gradient information is unknown, not available or non-existent. A surplus-based method is adopted to remove the doubly stochastic requirement on the weighting matrix, which enables the implementation of the algorithm in graphs having no associated doubly stochastic weighting matrix. For the convergence results, the proposed algorithm is able to obtain the exact convergence to the optimal value with any positive, non-summable and non-increasing step-sizes. Furthermore, when the step-size is also square-summable, the proposed algorithm is guaranteed to achieve the exact convergence to an optimal solution. In addition to the standard convergence analysis, the convergence rate of the proposed algorithm is also investigated. Finally, the effectiveness of the proposed algorithm is verified through numerical simulations.

**Index Terms**—Distributed optimization, gradient-free methods, multi-agent systems, directed graphs.

## I. INTRODUCTION

In recent years, with the prevalence of multi-agent systems, there has been a growing interest in solving the optimization problem in a distributed scheme. The advantage of doing so is that agents access local information and communicate with the neighbors only, making it suitable for the applications with large data size, huge computation and complex network structure, such as parameter estimation and detection [1], [2], source localization in sensor networks [3], [4], utility maximization [5], resource allocation [6], [7], and multi-robot coordination [8]–[11]. Distributed optimization of a sum of cost functions have been extensively studied over decades, such as the work in [12]–[21]. A common underlying assumption in all these methods is that the derivative term of the local cost functions and the constraints can be directly accessed. However, there are many applications in the fields of bio-chemistry, aircraft design, hydro-dynamics, earth sciences, *etc.*, where the relation between the variables and the objective functions are unknown, the gradient information is not available for usage, or the derivative is not possible to determine [22], these methods are no longer applicable. Hence, researchers start to draw attention to the gradient-free optimization.

Gradient-free optimization schemes can be traced back to the age of developing optimization theory, such as the work in [23]. Recent studies on this topic have been reported in

[24]–[31]. Shamir *et al.* in [24] investigated the performance of stochastic gradient descent method for non-smooth optimization problems. An averaging scheme was proposed to attain the minimax-optimal rates. On the other hand, Nesterov *et al.* in [25] provided an explicit way of computing the stochastic gradient information known as gradient-free oracle and investigated the convergence property for both convex and non-convex problems. This idea was extended to minimize a sum of non-smooth but Lipschitz continuous functions in [26]–[28], where the Gaussian smoothing technique was introduced to obtain the gradient-free oracle to replace the derivative in the standard subgradient methods. The same technique was applied to the algorithms in [29] and [30], [31], where the doubly stochastic requirement on the weighting matrix was removed by adopting a push-sum method [32] and a surplus-based method [33], respectively. It should be noted that these derivative-free methods are based on the Gaussian smoothing technique, where the introduced smoothing parameter imposes an additional penalty term along the iteration. Thus, only an inexact convergence to a neighborhood of the optimal value can be achieved. To achieve the exact convergence, Duchi *et al.* in [34] introduced a two point gradient estimation technique, and proved the exact convergence of the function value to the optimal value by choosing appropriate smoothing parameter sequences. This technique was extended to the distributed scenario in [35], [36] where an exact convergence of the function value to the optimal value was obtained.

In this paper, we aim to investigate gradient-free distributed optimization algorithms with exact convergence. Motivated by our work in [36], a distributed projected pseudo-gradient descent method is proposed to achieve an exact convergence with possibly a larger class of the step-sizes. The convergence properties of the proposed algorithm are carefully studied with different settings of the step-size. The main contributions of this work are summarized as follows.

- 1) Most gradient-free optimization algorithms, *e.g.*, [25]–[31] are based on Gaussian smoothing techniques, and hence can only achieve approximate convergence results. In terms of the exact convergence results, the work in [35] proved an exact convergence of the function value to the optimal value for a step-size of  $\alpha_k = \frac{1}{\sqrt{k}}$  ( $k$  is the iteration index), and our work in [36] proved the same convergence result for a non-summable and square-summable step-size. In this work, we introduce an optimal averaging scheme locally to trace a weighted average of the decision variable along the iteration. This averaging scheme is straightforward in terms of the implementation, and is able to obtain the exact convergence of the function value to the optimal value with any positive, non-increasing and non-summable

This work was supported by Singapore Ministry of Education Academic Research Fund Tier 1 RG180/17(2017-T1-002-158).

Y. Pang and G. Hu are with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore [ypang005@e.ntu.edu.sg](mailto:ypang005@e.ntu.edu.sg), [gqhu@ntu.edu.sg](mailto:gqhu@ntu.edu.sg).

step-sizes, hence increasing the range of the step-size selection.

- 2) The convergence of the agent's function value does not imply that its decision variable also converges. The square-summable step-size condition is a typical setting in subgradient descent algorithms, *e.g.*, [12]–[15], [32], [36]–[41] to establish the exact convergence of the agent's decision variable to an optimal solution. In this work, we show that this result also holds in distributed gradient-free algorithms. The proposed distributed projected pseudo-gradient descent method is guaranteed to achieve the exact convergence of the agent's decision variable to an optimal solution when the step-size also satisfies square-summable condition, which recovers the same convergence results in the literature.
- 3) The convergence rate has been widely studied in gradient-based distributed optimization literature, but received little attention in gradient-free distributed optimization literature. The only relevant works are [29], [31] and [27], where [29], [31] proved a rate of  $O(\frac{\ln t}{\sqrt{t}})$  for a diminishing step-size, and [27] showed a rate of  $O(\frac{1}{\sqrt{t}})$  for a constant step-size if the number of iterations  $t$  is known in advance. However, these rates were obtained for the algorithms with approximate convergence. In this work, the convergence rate of the proposed algorithm is studied, and we obtain the same convergence rate results as in [27], [29], [31] for the two settings of the step-size, but with exact convergence results.

The rest of the paper is organized as follows. The problem is defined in Section II. Section III introduces the proposed algorithm. The detailed convergence analysis is conducted in Section IV, where some auxiliary lemmas are introduced, followed by the main results of the paper. The numerical simulations are presented in Section V to illustrate the performance of the algorithm. Section VI concludes the paper.

## II. PROBLEM FORMULATION

For a directed graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ ,  $\mathcal{V} = \{1, 2, \dots, N\}$  is the set of agents, and  $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$  is the set of ordered pairs,  $(i, j)$ ,  $i, j \in \mathcal{V}$ , where agent  $i$  is able to send information to agent  $j$ . We denote the set of agent  $i$ 's in-neighbors by  $\mathcal{N}_i^{\text{in}} = \{j \in \mathcal{V} | (j, i) \in \mathcal{E}\}$  and out-neighbors by  $\mathcal{N}_i^{\text{out}} = \{j \in \mathcal{V} | (i, j) \in \mathcal{E}\}$ . Specifically, we allow both  $\mathcal{N}_i^{\text{in}}$  and  $\mathcal{N}_i^{\text{out}}$  to contain agent  $i$  itself, and  $\mathcal{N}_i^{\text{in}} \neq \mathcal{N}_i^{\text{out}}$  in general. The objective of the multi-agent system is to cooperatively solve the following set constrained optimization problem:

$$\min_{\mathbf{x}} f(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}, \quad (1)$$

where  $\mathcal{X} \subseteq \mathbb{R}^n$  is a convex and closed set, and  $f_i$  is a local cost function of agent  $i$  and  $\mathbf{x} = [x_1, \dots, x_n]^\top$  is a global decision vector. The explicit expression of the local cost function  $f_i$  is unknown, but the measurements can be made by agent  $i$  only. Denote the (non-empty) solution set to (1) by  $\mathcal{X}^*$ , *i.e.*,  $\mathcal{X}^* = \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$ .

Throughout this paper, we suppose the following assumptions hold:

*Assumption 1:* The directed graph is strongly connected.

*Assumption 2:* Each local cost function  $f_i$  is convex, but not necessarily differentiable. For  $\forall \mathbf{x} \in \mathcal{X}$ , the subgradient  $\partial f_i(\mathbf{x})$  exists and is bounded, *i.e.*, there exists a positive constant  $\hat{D}$  such that  $\|\partial f_i(\mathbf{x})\| \leq \hat{D}$ ,  $\forall \mathbf{x} \in \mathcal{X}$ .

## III. ALGORITHM

In this section, we will develop the distributed projected pseudo-gradient descent method for the optimization problem defined in (1) as follows.

At time-step  $k$ , each agent  $j$  broadcasts its state information  $\mathbf{x}_k^j$  with a weighted auxiliary variable  $[A_c]_{ij} \mathbf{y}_k^j$  to all of the nodes  $i$  in its out-neighborhood. Then, for each agent  $i$ , on receiving the information  $\mathbf{x}_k^j$ , and  $[A_c]_{ij} \mathbf{y}_k^j$  from all of the nodes  $j$  in its in-neighborhood, it updates its variables  $\mathbf{x}_{k+1}^i$  and  $\mathbf{y}_{k+1}^i$ <sup>1</sup>. Finally, each agent  $i$  adopts an optimal averaging scheme to trace the average of  $\mathbf{x}_k^i$ ,  $\ell = 0, 1, \dots, k+1$  weighted by the step-size sequence, defined by  $\hat{\mathbf{x}}_{k+1}^i$ . The updating law is given as follows.

$$\mathbf{x}_{k+1}^i = \mathcal{P}_{\mathcal{X}} \left[ \sum_{j=1}^N [A_r]_{ij} \mathbf{x}_k^j + \epsilon \mathbf{y}_k^i - \alpha_k \mathbf{g}^i(\mathbf{x}_k^i) \right], \quad (2a)$$

$$\mathbf{y}_{k+1}^i = \mathbf{x}_k^i - \sum_{j=1}^N [A_r]_{ij} \mathbf{x}_k^j + \sum_{j=1}^N [A_c]_{ij} \mathbf{y}_k^j - \epsilon \mathbf{y}_k^i, \quad (2b)$$

$$\hat{\mathbf{x}}_{k+1}^i = \hat{\mathbf{x}}_k^i + \frac{\alpha_{k+1}}{\sum_{\ell=0}^{k+1} \alpha_\ell} (\mathbf{x}_{k+1}^i - \hat{\mathbf{x}}_k^i), \quad (2c)$$

where  $A_r, A_c$  are the row stochastic and column stochastic weighting matrices, respectively, *i.e.*,  $A_r \mathbf{1}_n = \mathbf{1}_n$ , and  $\mathbf{1}_n^\top A_c = \mathbf{1}_n^\top$ .  $\alpha_k > 0$  is a non-increasing step-size.  $\epsilon$  is a small positive number. The auxiliary variable  $\mathbf{y}_k^i$  is used to offset the shift caused by the unbalanced (non-doubly stochastic) weighting matrices ( $A_r, A_c$ ), known as ‘‘surplus’’. The parameter  $\epsilon$  is to specify the amount of surplus during the update (see [33] for the details).  $\mathbf{g}^i(\mathbf{x}_k^i)$  is a pseudo-gradient motivated from [34], given as

$$\mathbf{g}^i(\mathbf{x}_k^i) = \frac{1}{\beta_{2,k}} [f_i(\mathbf{x}_k^i + \beta_{1,k} \xi_{1,k}^i + \beta_{2,k} \xi_{2,k}^i) - f_i(\mathbf{x}_k^i + \beta_{1,k} \xi_{1,k}^i)] \xi_{2,k}^i, \quad (3)$$

$\beta_{1,k}, \beta_{2,k}$  are two positive non-increasing sequences with their ratio defined as

$$\tilde{\beta}_k = \beta_{2,k} / \beta_{1,k}. \quad (4)$$

$\xi_{1,k}^i$  and  $\xi_{2,k}^i \in \mathbb{R}^n$  are two random variables satisfying the following assumption:

*Assumption 3:* (Assumption F in [34]) The random variables  $\xi_{1,k}^i$  and  $\xi_{2,k}^i \in \mathbb{R}^n$  are generated by any one of the following: (a) both  $\xi_{1,k}^i$  and  $\xi_{2,k}^i$  are standard normal in  $\mathbb{R}^n$  with identity covariance; (b) both  $\xi_{1,k}^i$  and  $\xi_{2,k}^i$  are uniform on the  $\ell_2$ -ball of radius  $\sqrt{n+2}$ ; (c) the distribution of  $\xi_{1,k}^i$  is uniform on

<sup>1</sup>The update process does not require each agent to know the state information from its out-neighbors. but we assume agent  $i$  knows the number of its in-neighbors and out-neighbors to design the weights in  $A_r$  and  $A_c$ .

the  $\ell_2$ -ball of radius  $\sqrt{n+2}$  and the distribution of  $\xi_{2,k}^i$  is uniform on the  $\ell_2$ -ball of radius  $\sqrt{n}$ .

Similar to the gradient-free oracle in [25], at each time  $k$ , the pseudo-gradient operator (3) estimates the gradient in a random direction  $\xi_{2,k}^i$  with a parameter  $\beta_{2,k}$ , but the function difference is taken at a perturbed point  $\mathbf{x}_k^i + \beta_{1,k}\xi_{1,k}^i$  instead of  $\mathbf{x}_k^i$ , where the amount of perturbation is determined by the parameter  $\beta_{1,k}$  and the random variable  $\xi_{1,k}^i$ . As compared to the gradient-free oracle where the function difference is evaluated at  $\mathbf{x}_k^i$  which may not be differentiable for non-smooth problems, the extra perturbation step in pseudo-gradient operator allows the function difference to be evaluated at a point which is less likely to be non-smooth. In fact, we can define a smoothed function of  $f_i(\mathbf{x})$  based on the convolution of this perturbation, given by [34],

$$\begin{aligned} f_{i,\beta_{1,k}}(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x} + \beta_{1,k}\xi_{1,k}^i)] \\ &= \int_{\mathbb{R}^n} f_i(\mathbf{x} + \beta_{1,k}\xi_{1,k}^i) d\mu(\xi_{1,k}^i), \end{aligned}$$

with the random variable  $\xi_{1,k}^i \in \mathbb{R}^n$  having density  $\mu$  with respect to Lebesgue measure.  $\beta_{1,k}$  is a positive non-increasing sequence.

In fact, algorithms (2a) and (2b) can be written into an equivalent form

$$\mathbf{z}_{k+1}^i = \sum_{j=1}^{2N} [A]_{ij} \mathbf{z}_k^j + g_k^i, \quad (5)$$

where  $g_k^i$ ,  $i \in \{1, \dots, 2N\}$  is an augmented pseudo-gradient defined by  $g_k^i = \mathbf{x}_{k+1}^i - \sum_{j=1}^{2N} [A_r]_{ij} \mathbf{x}_k^j - \epsilon \mathbf{y}_k^i$  for  $i \in \{1, \dots, N\}$ ,  $g_k^i = \mathbf{0}_n$  for  $i \in \{N+1, \dots, 2N\}$ ; matrix  $A \in \mathbb{R}^{2N \times 2N}$  is an augmented weighting matrix defined by  $A = \begin{bmatrix} A_r & \epsilon I \\ I - A_r & A_c - \epsilon I \end{bmatrix}$ ; and decision variable  $\mathbf{z}_k^i$ ,  $i \in \{1, \dots, 2N\}$  is defined by  $\mathbf{z}_k^i = \mathbf{x}_k^i$  for  $i \in \{1, \dots, N\}$ ,  $\mathbf{z}_k^i = \mathbf{y}_k^{i-N}$  for  $i \in \{N+1, \dots, 2N\}$ .

#### IV. CONVERGENCE ANALYSIS

In this section, the detailed convergence analysis of our proposed algorithm is provided. We first introduce some auxiliary lemmas in Subsection IV-A, followed by the main results in Subsection IV-B.

##### A. Auxiliary Lemmas

In this part, we introduce some auxiliary results, which will be helpful in the analysis of the main theorems. We denote the  $\sigma$ -field generated by the entire history of the random variables from step 0 to  $k-1$  by  $\mathcal{F}_k$ , i.e.,  $\mathcal{F}_k = \{(\mathbf{x}_0^i, i \in \mathcal{V}); (\xi_{1,s}^i, \xi_{2,s}^i, i \in \mathcal{V}); 0 \leq s \leq k-1\}$  with  $\mathcal{F}_0 = \{\mathbf{x}_0^i, i \in \mathcal{V}\}$ .

The following lemma summarizes some properties of function  $f_{i,\beta_{1,k}}(\mathbf{x})$  and the pseudo-gradient  $\mathbf{g}^i(\mathbf{x}_k^i)$ .

*Lemma 1:* (see [34]) Suppose Assumptions 2 and 3 hold. Then, for each  $i \in \mathcal{V}$ , the following properties of the function  $f_{i,\beta_{1,k}}(\mathbf{x})$  are satisfied:

- 1)  $f_{i,\beta_{1,k}}(\mathbf{x})$  is convex and differentiable, and it satisfies

$$f_i(\mathbf{x}) \leq f_{i,\beta_{1,k}}(\mathbf{x}) \leq f_i(\mathbf{x}) + \beta_{1,k} \hat{D} \sqrt{n+2},$$

- 2) the pseudo-gradient  $\mathbf{g}^i(\mathbf{x}_k^i)$  satisfies

$$\mathbb{E}[\mathbf{g}^i(\mathbf{x}_k^i) | \mathcal{F}_k] = \nabla f_{i,\beta_{1,k}}(\mathbf{x}_k^i) + \tilde{\beta}_k \hat{D} \mathbf{v},$$

- 3) there is a universal constant  $Q$  such that

$$\mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\|^2 | \mathcal{F}_k] \leq \sqrt{\mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\|^2 | \mathcal{F}_k]} \leq Q \mathcal{T}_k,$$

where  $\beta_{1,k}$  and  $\tilde{\beta}_k$  are defined in (4),  $\mathbf{v} \in \mathbb{R}^n$  is a vector satisfying  $\|\mathbf{v}\| \leq n\sqrt{3n}/2$ , and  $\mathcal{T}_k = \hat{D} \sqrt{n[n\sqrt{\tilde{\beta}_k} + 1 + \ln n]}$ . If  $\tilde{\beta}_k$  is bounded, then  $\mathcal{T}_k$  is bounded. In this case, we denote the upper bound of  $Q\mathcal{T}_k$  by  $\mathcal{K}_1$ .

Following the results in [33], [37], we have the following lemma on the convergence of the augmented weighting matrix  $A$  in (5).

*Lemma 2:* Suppose Assumption 1 holds. Let  $\epsilon$  be the constant in the augmented weighting matrix  $A$  in (5) such that  $\epsilon \in (0, \bar{\epsilon})$  with  $\bar{\epsilon} = (\frac{1-|\lambda_3|}{20+8N})^N$ , where  $\lambda_3$  is the third largest eigenvalue of matrix  $A$  with  $\epsilon = 0$ . Then  $\forall i, j \in \{1, \dots, 2N\}$ , the entries  $[A^k]_{ij}$  converge to their limits as  $k \rightarrow \infty$  at a geometric rate, i.e.,

$$\left\| A^k - \begin{bmatrix} \frac{\mathbf{1}_N \mathbf{1}_N^T}{N} & \frac{\mathbf{1}_N \mathbf{1}_N^T}{N} \\ \mathbf{0}_{N \times N} & \mathbf{0}_{N \times N} \end{bmatrix} \right\|_{\infty} \leq \Gamma \gamma^k, \quad k \geq 1,$$

where  $\Gamma > 0$  and

$$\gamma = \max\{|\lambda_3| + (20+8N)\epsilon^{\frac{1}{N}}, |\lambda_2(\epsilon)|\} \in (0, 1)$$

are some constants, and  $\lambda_2(\epsilon)$  is the eigenvalue of the weighting matrix  $A$  corresponding to the second largest eigenvalue  $\lambda_2$  of matrix  $A$  with  $\epsilon = 0$ .

**Proof.** The first part of the result follows directly from the proof of Lemma 1 in [37], where constant  $\gamma$  is determined by the magnitude of the second largest eigenvalue of matrix  $A$ . Next we aim to characterize the second largest eigenvalue of matrix  $A$ . To do so, we denote  $A$  by  $A(\epsilon)$  to represent the dependency of  $A$  on parameter  $\epsilon$ . Then, matrix  $A(\epsilon)$  can be viewed as matrix  $A(0)$  with some perturbations on  $\epsilon$ , where matrix  $A(0)$  is matrix  $A(\epsilon)$  by setting  $\epsilon = 0$ . Denote the eigenvalues of matrix  $A(0)$  by  $\lambda_1, \lambda_2, \dots, \lambda_{2N}$  with  $|\lambda_1| \geq \dots \geq |\lambda_{2N}|$ . From the proof of Theorem 4 in [33], it holds that  $1 = \lambda_1 = \lambda_2 > |\lambda_3| \geq \dots \geq |\lambda_{2N}|$ . After perturbation, we denote by  $\lambda_i(\epsilon)$  the eigenvalues of matrix  $A(\epsilon)$  corresponding to  $\lambda_i$ ,  $i = \{1, \dots, N\}$ . It should be noted that the eigenvalues of the perturbed matrix  $A(\epsilon)$  do not necessarily satisfy  $|\lambda_1(\epsilon)| \geq \dots \geq |\lambda_{2N}(\epsilon)|$  given that  $|\lambda_1| \geq \dots \geq |\lambda_{2N}|$ . From Lemmas 10 and 11 in [33], when  $\epsilon \in (0, \bar{\epsilon})$ , we have the following inequality characterizing the distance between the corresponding eigenvalues  $\lambda_i(\epsilon)$  and  $\lambda_i$ ,  $i = \{1, \dots, N\}$

$$|\lambda_i(\epsilon) - \lambda_i| < 4(4+2N+\epsilon)\epsilon^{\frac{1}{N}} < (20+8N)\epsilon^{\frac{1}{N}},$$

which gives  $|\lambda_i(\epsilon)| < |\lambda_i| + (20+8N)\epsilon^{\frac{1}{N}}$ . Hence, for  $i = \{3, \dots, N\}$ , the above inequality yields  $|\lambda_i(\epsilon)| < |\lambda_3| + (20+8N)\epsilon^{\frac{1}{N}} < 1$  since  $|\lambda_3| \geq \dots \geq |\lambda_{2N}|$  and  $\epsilon \in (0, \bar{\epsilon})$ . Moreover, from the proof of Theorem 4 and Lemma 12 in [33], when  $\epsilon \in (0, \bar{\epsilon})$ , we have  $\lambda_1(\epsilon) = 1$  and  $|\lambda_2(\epsilon)| < 1$ . Hence  $\gamma$  can be selected as  $\max\{|\lambda_3| + (20+8N)\epsilon^{\frac{1}{N}}, |\lambda_2(\epsilon)|\}$ , which completes the proof.  $\blacksquare$

*Remark 1:* The work in [42] has proposed solutions on the design of the weighting matrix to guarantee the fastest averaging speed when the weighting matrix is symmetric and doubly-stochastic. For the weighting matrix  $A$  in this work, Lemma 2 shows that the averaging speed depends on constant  $\gamma$ . From the proof of Lemma 2, we can infer the effects of parameter  $\epsilon$ , the communication topology, and the number of agents  $N$  on constant  $\gamma$ . For the effect of parameter  $\epsilon$ , noting that  $|\lambda_2(\epsilon)| = 1$  when  $\epsilon = 0$  and  $|\lambda_3| + (20+8N)\epsilon^{\frac{1}{N}} = 1$  when  $\epsilon = \bar{\epsilon}$ , hence  $\gamma$  is dominant by  $|\lambda_2(\epsilon)|$  when  $\epsilon$  is small, and then dominant by  $|\lambda_3| + (20+8N)\epsilon^{\frac{1}{N}}$  when  $\epsilon$  is large. That implies there is an optimal value of  $\epsilon$  such that  $\gamma$  is minimized (when  $|\lambda_2(\epsilon)| = |\lambda_3| + (20+8N)\epsilon^{\frac{1}{N}}$ ). For the effect of the communication topology, suppose  $\epsilon$  is set at the optimal value, then a graph with a smaller  $|\lambda_3|$  leads to a smaller  $\gamma$ . For the effect of the number of agents  $N$ , since  $|\lambda_3| + (20+8N)\epsilon^{\frac{1}{N}}$  is smaller for a smaller  $N$ , hence  $\gamma$  is smaller for a smaller number of agents.

Define  $\bar{\mathbf{z}}_k = \frac{1}{N} \sum_{i=1}^{2N} \mathbf{z}_k^i = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_k^i + \frac{1}{N} \sum_{i=1}^N \mathbf{y}_k^i$ , which is an average of  $\mathbf{x}_k^i + \mathbf{y}_k^i$  over all agents at time-step  $k$ ; and

$$\hat{\mathbf{z}}_k = \frac{\sum_{\ell=0}^k \alpha_\ell \bar{\mathbf{z}}_\ell}{\sum_{\ell=0}^k \alpha_\ell}, \quad (6)$$

which is an average of  $\bar{\mathbf{z}}$  weighted by the step-size sequence  $\alpha_\ell$  over time duration  $k$ . Then, we can quantify the bounds of the terms  $\mathbf{x}_k^i - \bar{\mathbf{z}}_k$  and  $\mathbf{y}_k^i - \mathbf{0}_n$  as shown in the following lemma. For easy representation, we denote the aggregated norm of the augmented pseudo-gradient  $\sum_{j=1}^N \|g_k^j\|$  by  $\mathbf{G}_k$  in the rest of the paper.

*Lemma 3:* Suppose Assumptions 1, 2 and 3 hold. Let  $\epsilon$  be the constant such that  $\epsilon \in (0, \bar{\epsilon})$ , where  $\bar{\epsilon}$  is defined in Lemma 2. Let  $\{\mathbf{x}_k^i\}_{k \geq 0}$  and  $\{\mathbf{y}_k^i\}_{k \geq 0}$  be the sequences generated by (2a) and (2b), respectively. Then, it holds that for  $k \geq 1$

$$\begin{aligned} 1) \quad & \|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\| \leq 2N\varsigma\Gamma\gamma^k + \Gamma \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1} + \mathbf{G}_{k-1}; \\ 2) \quad & \|\mathbf{y}_k^i\| \leq 2N\varsigma\Gamma\gamma^k + \Gamma \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1}, \end{aligned}$$

where  $\varsigma = \max\{\|\mathbf{x}_0^i\|, \|\mathbf{y}_0^i\|, i \in \mathcal{V}\}$ ,  $\Gamma$  and  $\gamma$  are the constants defined in Lemma 2.

**Proof.** For  $k \geq 1$ , we have

$$\mathbf{z}_k^i = \sum_{j=1}^{2N} [A^k]_{ij} \mathbf{z}_0^j + \sum_{r=1}^{k-1} \sum_{j=1}^{2N} [A^{k-r}]_{ij} g_{r-1}^j + g_{k-1}^i. \quad (7)$$

by applying (5) recursively. Then we can obtain that

$$\bar{\mathbf{z}}_k = \frac{1}{N} \sum_{j=1}^{2N} \mathbf{z}_0^j + \frac{1}{N} \sum_{r=1}^{k-1} \sum_{j=1}^{2N} g_{r-1}^j + \frac{1}{N} \sum_{j=1}^{2N} g_{k-1}^j, \quad (8)$$

where we used column stochastic property of  $A$ .

For part (1), subtracting (8) from (7) and taking the norm, we have that for  $1 \leq i \leq N$  and  $k \geq 1$ ,

$$\|\mathbf{z}_k^i - \bar{\mathbf{z}}_k\| \leq \sum_{j=1}^{2N} \left| [A^k]_{ij} - \frac{1}{N} \right| \varsigma + \sum_{r=1}^{k-1} \sum_{j=1}^N [A^{k-r}]_{ij}$$

$$- \frac{1}{N} \left\| g_{r-1}^j \right\| + \frac{N-1}{N} \|g_{k-1}^i\| + \frac{1}{N} \sum_{j \neq i} \|g_{k-1}^j\|. \quad (9)$$

Noting that  $\frac{N-1}{N} \|g_{k-1}^i\| + \frac{1}{N} \sum_{j \neq i} \|g_{k-1}^j\| \leq \mathbf{G}_{k-1}$ , and applying the property of  $[A^k]_{ij}$  from Lemma 2 to (9), we complete the proof of part (1).

For part (2), taking the norm in (7) for  $N+1 \leq i \leq 2N$  and  $k > 1$ , and applying the property of  $[A^k]_{ij}$  from Lemma 2, we complete the proof of part (2). ■

It can be seen from Lemma 3 that the bound for the consensus terms is a function of the aggregated norm of the augmented pseudo-gradient term  $\mathbf{G}_k$ . Hence, in the following lemma, we provide some properties on this term  $\mathbf{G}_k$ .

*Lemma 4:* Suppose Assumptions 1, 2 and 3 hold. Let  $\epsilon$  be the constant such that  $0 < \epsilon < \min(\bar{\epsilon}, \frac{1-\gamma}{2\sqrt{3N\Gamma}\gamma})$ , where  $\bar{\epsilon}$ ,  $\Gamma$  and  $\gamma$  are the constants defined in Lemma 2. Let  $\tilde{\beta}_k$  defined in (4) be bounded. Then, for any  $K \geq 1$ , the aggregated norm of the augmented pseudo-gradient term  $\mathbf{G}_k$  satisfies that

$$\begin{aligned} 1) \quad & \sum_{k=1}^K \alpha_k \mathbb{E}[\mathbf{G}_k] \leq \Phi_1 \sum_{k=1}^K \alpha_k^2 + \Psi_1, \\ 2) \quad & \sum_{k=1}^K \mathbb{E}[\mathbf{G}_k^2] \leq \Phi_2 \sum_{k=1}^K \alpha_k^2 + \Psi_2, \\ 3) \quad & \sum_{k=1}^K \sum_{i=1}^N \alpha_k \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_k] \leq \Phi_3 \sum_{k=1}^K \alpha_k^2 + \Psi_3, \end{aligned}$$

where  $\Phi_1, \Psi_1, \Phi_2, \Psi_2, \Phi_3$  and  $\Psi_3$  are positive bounded constants, and  $\alpha_k > 0$  is a non-increasing step-size.

**Proof.** See Appendix A. ■

In addition, we will frequently utilize the Stolz-Cesaro Theorem [43] to facilitate the analysis, which is quoted below for completeness.

*Lemma 5:* (Stolz-Cesaro Theorem) If  $\{b_k\}_{k \geq 1}$  is a sequence of positive real numbers, such that  $\sum_{k=1}^{\infty} b_k = \infty$ , then for any sequence  $\{a_k\}_{k \geq 1}$  one has the inequality:

$$\begin{aligned} \liminf_{k \rightarrow \infty} \frac{a_k}{b_k} & \leq \liminf_{k \rightarrow \infty} \frac{a_1 + a_2 + \dots + a_k}{b_1 + b_2 + \dots + b_k} \\ & \leq \limsup_{k \rightarrow \infty} \frac{a_1 + a_2 + \dots + a_k}{b_1 + b_2 + \dots + b_k} \leq \limsup_{k \rightarrow \infty} \frac{a_k}{b_k}. \end{aligned}$$

In particular, if the sequence  $\{a_k/b_k\}_{k \geq 1}$  has a limit, then

$$\lim_{k \rightarrow \infty} \frac{a_1 + a_2 + \dots + a_k}{b_1 + b_2 + \dots + b_k} = \lim_{k \rightarrow \infty} \frac{a_k}{b_k}.$$

With the above lemmas, we are able to establish a one-step iteration and a consensus result under only non-summable step-size condition.

*Proposition 1:* Suppose Assumptions 1, 2 and 3 hold. Let  $\{\mathbf{x}_k^i\}_{k \geq 0}$ ,  $\{\mathbf{y}_k^i\}_{k \geq 0}$  and  $\{\tilde{\mathbf{x}}_k^i\}_{k \geq 0}$  be the sequences generated by (2) with a non-increasing step-size sequence  $\{\alpha_k\}_{k \geq 0}$  satisfying

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \lim_{k \rightarrow \infty} \alpha_k = \alpha_\infty.$$

Let  $\epsilon$  be the constant such that  $0 < \epsilon < \min(\bar{\epsilon}, \frac{1-\gamma}{2\sqrt{3N\Gamma}\gamma})$ , where  $\bar{\epsilon}$ ,  $\Gamma$  and  $\gamma$  are the constants defined in Lemma 2. Let  $\tilde{\beta}_k$  defined in (4) be bounded. Then

(1)  $\widehat{\mathbf{x}}_k^i$  holds that

$$\mathbb{E}[\|\widehat{\mathbf{x}}_k^i - \widehat{\mathbf{z}}_k\|] \leq \frac{\sum_{\ell=0}^k \alpha_\ell^2}{\sum_{\ell=0}^k \alpha_\ell} \left[ N\varsigma\Gamma + \Phi_1 \left( 1 + \frac{\Gamma\gamma}{1-\gamma} \right) \right] + \frac{1}{\sum_{\ell=0}^k \alpha_\ell} \left[ B_0 + \frac{N\varsigma\Gamma\gamma^2}{1-\gamma^2} + \Psi_1 \left( 1 + \frac{\Gamma\gamma}{1-\gamma} \right) \right],$$

where  $B_0 = \max_i \alpha_i \|\mathbf{x}_0^i - \bar{\mathbf{z}}_0\|$ ,  $\Phi_1 > 0$ ,  $\Psi_1 > 0$  are constants defined in Lemma 4, and  $\widehat{\mathbf{z}}_k$  is defined in (6).

(2) for any  $\mathbf{z}^* \in \mathcal{X}^*$ , the following relation holds

$$\mathbb{E}[\|\bar{\mathbf{z}}_{k+1} - \mathbf{z}^*\|^2 | \mathcal{F}_k] \leq \|\bar{\mathbf{z}}_k - \mathbf{z}^*\|^2 - \frac{2\alpha_k}{N} (f(\bar{\mathbf{z}}_k) - f^*) + Z_k,$$

where

$$\begin{aligned} Z_k &= 2\alpha_k\beta_{1,k}\hat{D}\sqrt{n+2} + 4N\varsigma(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2)\Gamma\gamma^k\alpha_k \\ &+ 2\mathcal{K}_2 \left( \frac{2N(2N+\epsilon)\varsigma\Gamma\gamma}{1-\gamma} + \max_{i \in \mathcal{V}} \|\mathbf{x}_0^i - \mathbf{z}^*\| + B_1 \right) \alpha_k \tilde{\beta}_k \\ &+ 2\mathcal{K}_2 \left( \frac{(2N+\epsilon)\Gamma\gamma}{1-\gamma} + 2N \right) \alpha_k \tilde{\beta}_k \sum_{r=1}^{k-1} \mathbf{G}_{r-1} \\ &+ \frac{2\mathcal{K}_2}{N} \alpha_k \tilde{\beta}_k \sum_{r=0}^{k-1} \sum_{i=1}^N \|\mathbf{g}^i(\mathbf{x}_r^i)\| \\ &+ 2(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2)\Gamma\alpha_k \sum_{r=1}^k \gamma^{k-r+1} \mathbf{G}_{r-1} \\ &+ 2(2\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2)\alpha_k \mathbf{G}_{k-1} + 4\varsigma\Gamma\gamma^{k+1} \mathbb{E}[\mathbf{G}_k | \mathcal{F}_k] \\ &+ \frac{2\Gamma}{N} \sum_{r=1}^k \gamma^{k-r+1} \mathbb{E}[\mathbf{G}_k | \mathcal{F}_k] \mathbf{G}_{r-1} \\ &+ \frac{4\alpha_k}{N} \sum_{i=1}^N \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_k | \mathcal{F}_k] + \frac{5}{N} \mathbb{E}[\mathbf{G}_k^2 | \mathcal{F}_k], \end{aligned}$$

$B_1 = \max_{i \in \mathcal{V}} \epsilon \|\mathbf{y}_0^i\| + 2 \sum_{i=1}^N \|\mathbf{x}_0^i - \bar{\mathbf{z}}_0\|$ , and  $\bar{\beta}$  is the upper bound of  $\beta_k$ .

**Proof.** For part (1), by the definitions of  $\widehat{\mathbf{x}}_k^i$  and  $\widehat{\mathbf{z}}_k$ , we know that  $\|\widehat{\mathbf{x}}_k^i - \widehat{\mathbf{z}}_k\| \leq \frac{\sum_{\ell=0}^k \alpha_\ell \|\mathbf{x}_\ell^i - \bar{\mathbf{z}}_\ell\|}{\sum_{\ell=0}^k \alpha_\ell}$ . Taking the total expectation and applying Lemma 3-1), we obtain that for  $k \geq 1$

$$\begin{aligned} \mathbb{E}[\|\widehat{\mathbf{x}}_k^i - \widehat{\mathbf{z}}_k\|] &\leq \frac{B_0 + \sum_{\ell=1}^k \alpha_\ell \mathbb{E}[\|\mathbf{x}_\ell^i - \bar{\mathbf{z}}_\ell\|]}{\sum_{\ell=0}^k \alpha_\ell} \\ &\leq \frac{1}{\sum_{\ell=0}^k \alpha_\ell} \left( B_0 + 2N\varsigma\Gamma \sum_{\ell=1}^k \gamma^\ell \alpha_\ell \right. \\ &\quad \left. + \Gamma \sum_{\ell=1}^k \alpha_\ell \sum_{r=1}^{\ell-1} \gamma^{\ell-r} \mathbb{E}[\mathbf{G}_{r-1}] + \sum_{\ell=1}^k \alpha_\ell \mathbb{E}[\mathbf{G}_{\ell-1}] \right), \end{aligned}$$

where  $B_0 = \max_i \alpha_i \|\mathbf{x}_0^i - \bar{\mathbf{z}}_0\|$  is bounded. Since  $\alpha_k$  is non-increasing, it follows from Lemma 4 that

$$\begin{aligned} \sum_{\ell=1}^k \alpha_\ell \gamma^\ell &\leq \frac{1}{2} \sum_{\ell=1}^k \alpha_\ell^2 + \frac{\gamma^2}{2(1-\gamma^2)}, \\ \Gamma \sum_{\ell=1}^k \alpha_\ell \sum_{r=1}^{\ell-1} \gamma^{\ell-r} \mathbb{E}[\mathbf{G}_{r-1}] &\leq \Gamma \sum_{\ell=1}^k \sum_{r=1}^{\ell-1} \gamma^{\ell-r} \alpha_{r-1} \mathbb{E}[\mathbf{G}_{r-1}] \end{aligned}$$

$$\begin{aligned} &\leq \frac{\Gamma\gamma}{1-\gamma} \sum_{\ell=1}^k \alpha_\ell \mathbb{E}[\mathbf{G}_\ell] \leq \frac{\Phi_1\Gamma\gamma}{1-\gamma} \sum_{\ell=1}^k \alpha_\ell^2 + \frac{\Psi_1\Gamma\gamma}{1-\gamma}, \\ \sum_{\ell=1}^k \alpha_\ell \mathbb{E}[\mathbf{G}_{\ell-1}] &\leq \sum_{\ell=0}^k \alpha_\ell \mathbb{E}[\mathbf{G}_\ell] \leq \Phi_1 \sum_{\ell=0}^k \alpha_\ell^2 + \Psi_1. \end{aligned} \quad (10)$$

Substituting (10) to the preceding relation completes the proof of part (1).

For part (2), considering (5), and the fact that  $A$  is column-stochastic, we have  $\bar{\mathbf{z}}_{k+1} = \bar{\mathbf{z}}_k + \frac{1}{N} \sum_{i=1}^N g_k^i$ . Then, for any  $\mathbf{z}^* \in \mathcal{X}^*$ , it follows that

$$\begin{aligned} \|\bar{\mathbf{z}}_{k+1} - \mathbf{z}^*\|^2 &\leq \|\bar{\mathbf{z}}_k - \mathbf{z}^*\|^2 + \frac{\mathbf{G}_k^2}{N^2} + \frac{2}{N} \sum_{i=1}^N \langle g_k^i, \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle \\ &= \|\bar{\mathbf{z}}_k - \mathbf{z}^*\|^2 + \frac{\mathbf{G}_k^2}{N^2} \end{aligned} \quad (11a)$$

$$+ \frac{2}{N} \sum_{i=1}^N \langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle \quad (11b)$$

$$- \frac{2\alpha_k}{N} \sum_{i=1}^N \langle \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle. \quad (11c)$$

For the second term in (11a), we have that  $\mathbb{E}[\frac{\mathbf{G}_k^2}{N^2} | \mathcal{F}_k] \leq \frac{1}{N} \mathbb{E}[\mathbf{G}_k^2 | \mathcal{F}_k]$ .

For term (11b), it can be expanded as

$$\begin{aligned} &\sum_{i=1}^N \langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle \\ &= \sum_{i=1}^N \langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \bar{\mathbf{z}}_{k+1} \rangle \end{aligned} \quad (12a)$$

$$+ \sum_{i=1}^N \langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_{k+1} - \mathbf{x}_{k+1}^i \rangle \quad (12b)$$

$$+ \sum_{i=1}^N \langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \mathbf{x}_{k+1}^i - \mathbf{z}^* \rangle. \quad (12c)$$

For (12a), we have

$$\begin{aligned} &\sum_{i=1}^N \mathbb{E}[\langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \bar{\mathbf{z}}_{k+1} \rangle | \mathcal{F}_k] \\ &\leq \frac{1}{N} \mathbb{E}[\mathbf{G}_k^2 | \mathcal{F}_k] + \frac{\alpha_k}{N} \sum_{i=1}^N \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_k | \mathcal{F}_k] \\ &\leq \mathbb{E}[\mathbf{G}_k^2 | \mathcal{F}_k] + \alpha_k \sum_{i=1}^N \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_k | \mathcal{F}_k]. \end{aligned} \quad (13)$$

For (12b), we have  $\sum_{i=1}^N \langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_{k+1} - \mathbf{x}_{k+1}^i \rangle \leq (\mathbf{G}_k + \alpha_k \sum_{i=1}^N \|\mathbf{g}^i(\mathbf{x}_k^i)\|)(2N\varsigma\Gamma\gamma^{k+1} + \Gamma \sum_{r=1}^k \gamma^{k-r+1} \mathbf{G}_{r-1} + \mathbf{G}_k)$ , where Lemma 3-1) was substituted. Hence, we obtain

$$\begin{aligned} &\sum_{i=1}^N \mathbb{E}[\langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_{k+1} - \mathbf{x}_{k+1}^i \rangle | \mathcal{F}_k] \\ &\leq 2N^2\varsigma\mathcal{K}_1\Gamma\gamma^{k+1}\alpha_k + 2N\varsigma\Gamma\gamma^{k+1} \mathbb{E}[\mathbf{G}_k | \mathcal{F}_k] \\ &\quad + N\mathcal{K}_1\Gamma\alpha_k \sum_{r=1}^k \gamma^{k-r+1} \mathbf{G}_{r-1} \end{aligned}$$

$$\begin{aligned}
& + \Gamma \sum_{r=1}^k \gamma^{k-r+1} \mathbb{E}[\mathbf{G}_k | \mathcal{F}_k] \mathbf{G}_{r-1} \\
& + \alpha_k \sum_{i=1}^N \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| | \mathbf{G}_k | \mathcal{F}_k] + \mathbb{E}[\mathbf{G}_k^2 | \mathcal{F}_k]. \quad (14)
\end{aligned}$$

For (12c), it follows from [13, Lemma 1-(a)] that

$$\langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \mathbf{x}_{k+1}^i - \mathbf{z}^* \rangle \leq 0. \quad (15)$$

Thus, taking the conditional expectation on  $\mathcal{F}_k$  in (12) and substituting (13), (14) and (15), we obtain

$$\begin{aligned}
& \sum_{i=1}^N \mathbb{E}[\langle g_k^i + \alpha_k \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle | \mathcal{F}_k] \\
& \leq 2N^2 \zeta \mathcal{K}_1 \Gamma \gamma^{k+1} \alpha_k + 2N \zeta \Gamma \gamma^{k+1} \mathbb{E}[\mathbf{G}_k | \mathcal{F}_k] \\
& \quad + N \mathcal{K}_1 \Gamma \alpha_k \sum_{r=1}^k \gamma^{k-r+1} \mathbf{G}_{r-1} \\
& \quad + \Gamma \sum_{r=1}^k \gamma^{k-r+1} \mathbb{E}[\mathbf{G}_k | \mathcal{F}_k] \mathbf{G}_{r-1} \\
& \quad + 2\alpha_k \sum_{i=1}^N \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| | \mathbf{G}_k | \mathcal{F}_k] + 2\mathbb{E}[\mathbf{G}_k^2 | \mathcal{F}_k]. \quad (16)
\end{aligned}$$

For (11c), from Lemma 1-(2),  $\sum_{i=1}^N \mathbb{E}[\langle \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle | \mathcal{F}_k] = \sum_{i=1}^N \langle \nabla f_{i,\beta_{1,k}}(\mathbf{x}_k^i) + \tilde{\beta}_k \hat{D} \mathbf{v}, \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle$ . Denote  $\hat{D} \|\mathbf{v}\|$  by  $\mathcal{K}_2$ , we have

$$\begin{aligned}
& \langle \nabla f_{i,\beta_{1,k}}(\mathbf{x}_k^i) + \tilde{\beta}_k \hat{D} \mathbf{v}, \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle \\
& = \langle \nabla f_{i,\beta_{1,k}}(\mathbf{x}_k^i) + \tilde{\beta}_k \hat{D} \mathbf{v}, \bar{\mathbf{z}}_k - \mathbf{x}_k^i \rangle \\
& \quad + \langle \nabla f_{i,\beta_{1,k}}(\mathbf{x}_k^i) + \tilde{\beta}_k \hat{D} \mathbf{v}, \mathbf{x}_k^i - \mathbf{z}^* \rangle \\
& \geq -\|\nabla f_{i,\beta_{1,k}}(\mathbf{x}_k^i)\| \|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\| - \tilde{\beta}_k \mathcal{K}_2 \|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\| \\
& \quad + f_{i,\beta_{1,k}}(\mathbf{x}_k^i) - f_{i,\beta_{1,k}}(\mathbf{z}^*) - \tilde{\beta}_k \mathcal{K}_2 \|\mathbf{x}_k^i - \mathbf{z}^*\| \\
& \geq f_{i,\beta_{1,k}}(\bar{\mathbf{z}}_k) - f_{i,\beta_{1,k}}(\mathbf{z}^*) - \tilde{\beta}_k \mathcal{K}_2 \|\mathbf{x}_k^i - \mathbf{z}^*\| \\
& \quad - (2\mathcal{K}_1 + \tilde{\beta}_k \mathcal{K}_2) \|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\| \\
& \geq f_i(\bar{\mathbf{z}}_k) - f_i(\mathbf{z}^*) - \beta_{1,k} \hat{D} \sqrt{n+2} - \tilde{\beta}_k \mathcal{K}_2 \|\mathbf{x}_k^i \\
& \quad - \mathbf{z}^*\| - (2\mathcal{K}_1 + \tilde{\beta}_k \mathcal{K}_2) \|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\|. \quad (17)
\end{aligned}$$

Considering the term  $\|\mathbf{x}_k^i - \mathbf{z}^*\|$ , it follows that  $\|\mathbf{x}_k^i - \mathbf{z}^*\| \leq \sum_{j=1}^N [A_r]_{ij} \|\mathbf{x}_{k-1}^j - \mathbf{z}^*\| + \epsilon \|\mathbf{y}_{k-1}^i\| + \alpha_{k-1} \|\mathbf{g}^i(\mathbf{x}_{k-1}^i)\| \leq \sum_{j=1}^N [A_r]_{ij} \|\mathbf{x}_{k-1}^j - \mathbf{z}^*\| + \epsilon \|\mathbf{y}_{k-1}^i\| + \alpha_{k-1} \|\mathbf{g}^i(\mathbf{x}_{k-1}^i)\| + 2 \sum_{i=1}^N \|\mathbf{x}_{k-1}^i - \bar{\mathbf{z}}_{k-1}\|$ .

Applying the above relation recursively yields  $\|\mathbf{x}_k^i - \mathbf{z}^*\| \leq \|\mathbf{x}_0^i - \mathbf{z}^*\| + \epsilon \sum_{\tau=0}^{k-1} \|\mathbf{y}_\tau^i\| + \sum_{\tau=0}^{k-1} \alpha_\tau \|\mathbf{g}^i(\mathbf{x}_\tau^i)\| + 2 \sum_{\tau=0}^{k-1} \sum_{i=1}^N \|\mathbf{x}_\tau^i - \bar{\mathbf{z}}_\tau\|$ .

Thus, substituting the above result to (17) gives  $\langle \nabla f_{i,\beta_{1,k}}(\mathbf{x}_k^i) + \tilde{\beta}_k \hat{D} \mathbf{v}, \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle \geq f_i(\bar{\mathbf{z}}_k) - f_i(\mathbf{z}^*) - \tilde{\beta}_k \mathcal{K}_2 (\|\mathbf{x}_0^i - \mathbf{z}^*\| + \epsilon \sum_{\tau=0}^{k-1} \|\mathbf{y}_\tau^i\| + \sum_{\tau=0}^{k-1} \alpha_\tau \|\mathbf{g}^i(\mathbf{x}_\tau^i)\|) + 2 \sum_{\tau=0}^{k-1} \sum_{i=1}^N \|\mathbf{x}_\tau^i - \bar{\mathbf{z}}_\tau\| - (2\mathcal{K}_1 + \tilde{\beta}_k \mathcal{K}_2) \|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\| - \beta_{1,k} \hat{D} \sqrt{n+2}$ . Applying Lemma 3 and noting that  $\tilde{\beta}_k$  is bounded (where its upper bound is denoted by  $\beta$ ), we obtain that

$$\frac{2\alpha_k}{N} \sum_{i=1}^N \mathbb{E}[\langle \mathbf{g}^i(\mathbf{x}_k^i), \bar{\mathbf{z}}_k - \mathbf{z}^* \rangle | \mathcal{F}_k] \geq \frac{2\alpha_k}{N} (f(\bar{\mathbf{z}}_k) - f^*)$$

$$\begin{aligned}
& - 2\alpha_k \beta_{1,k} \hat{D} \sqrt{n+2} - 2\mathcal{K}_2 \left( \frac{2N(2N+\epsilon)\zeta\Gamma\gamma}{1-\gamma} \right. \\
& \quad \left. + \max_{i \in \mathcal{V}} \|\mathbf{x}_0^i - \mathbf{z}^*\| + B_1 \right) \alpha_k \tilde{\beta}_k \\
& - 2\mathcal{K}_2 \left( \frac{(2N+\epsilon)\Gamma\gamma}{1-\gamma} + 2N \right) \alpha_k \tilde{\beta}_k \sum_{r=1}^{k-1} \mathbf{G}_{r-1} \\
& - \frac{2\mathcal{K}_2}{N} \alpha_k \tilde{\beta}_k \sum_{r=0}^{k-1} \alpha_r \sum_{i=1}^N \|\mathbf{g}^i(\mathbf{x}_r^i)\| \\
& - 4N\zeta(2\mathcal{K}_1 + \tilde{\beta}_k \mathcal{K}_2) \Gamma \gamma^k \alpha_k - 2(2\mathcal{K}_1 + \tilde{\beta}_k \mathcal{K}_2) \alpha_k \mathbf{G}_{k-1} \\
& - 2(2\mathcal{K}_1 + \tilde{\beta}_k \mathcal{K}_2) \Gamma \alpha_k \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1}, \quad (18)
\end{aligned}$$

where  $B_1 = \max_{i \in \mathcal{V}} \epsilon \|\mathbf{y}_0^i\| + 2 \sum_{i=1}^N \|\mathbf{x}_0^i - \bar{\mathbf{z}}_0\|$ .

Taking the conditional expectation on  $\mathcal{F}_k$  in (11), and substituting (16) and (18) gives the result of part (2). ■

## B. Main Results

In this subsection, we present the main convergence results of our proposed algorithm, including convergence under non-summable and square-summable step-size condition (Theorem 1), convergence under non-summable step-size condition only (Theorem 2), and the convergence rate analysis (Corollary 1).

Our first result demonstrates the standard convergence results under non-summable and square-summable step-size condition.

**Theorem 1:** Suppose Assumptions 1, 2 and 3 hold. Let  $\{\hat{\mathbf{x}}_k^i\}_{k \geq 0}$  be the sequence generated by (2) with a non-increasing step-size sequence  $\{\alpha_k\}_{k \geq 0}$  satisfying

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty.$$

Let  $\epsilon$  be the constant such that  $0 < \epsilon < \min(\bar{\epsilon}, \frac{1-\gamma}{2\sqrt{3}N\Gamma\gamma})$ , where  $\Gamma$  and  $\gamma$  are some constants, and  $\bar{\epsilon} = (\frac{1-\lambda_3}{20+8N})^N$  with  $\lambda_3$  being the third largest eigenvalue of the weighting matrix  $A$  in (5) by setting  $\epsilon = 0$ . Let  $\beta_{1,k}$  and  $\tilde{\beta}_k$  defined in (4) satisfy  $\sum_{k=0}^{\infty} \beta_{1,k} \alpha_k < \infty$  and  $\sum_{k=0}^{\infty} \tilde{\beta}_k < \infty$ . Then, for  $\forall i \in \mathcal{V}$ , we have  $\{\hat{\mathbf{x}}_k^i\}_{k \geq 0}$  converges a.s. to an optimizer  $\mathbf{x}^* \in \mathcal{X}^{*2}$ .

**Proof.** we proceed to the proof by showing (A) the convergence of  $\hat{\mathbf{x}}_k^i$  to  $\mathbf{x}_k^i$ , (B) the convergence of  $\mathbf{x}_k^i$  to  $\bar{\mathbf{z}}_k$ , and (C) the convergence of  $\bar{\mathbf{z}}_k$  to an optimizer  $\mathbf{x}^* \in \mathcal{X}^*$  under appropriate conditions.

(A) Convergence of  $\hat{\mathbf{x}}_k^i$  to  $\mathbf{x}_k^i$ :

Suppose  $\{\mathbf{x}_k^i\}_{k \geq 0}$  converges a.s. to some point  $\tilde{\mathbf{x}}$ , i.e.,  $\mathbb{P}(\lim_{k \rightarrow \infty} \mathbf{x}_k^i = \tilde{\mathbf{x}}) = 1$ . By definition of  $\hat{\mathbf{x}}_k^i = \frac{\sum_{\ell=0}^k \alpha_\ell \mathbf{x}_\ell^i}{\sum_{\ell=0}^k \alpha_\ell}$  in (2c), it follows from Lemma 5 with  $a_k = \alpha_k \mathbf{x}_k^i$  and  $b_k = \alpha_k$  that  $\mathbb{P}(\lim_{k \rightarrow \infty} \hat{\mathbf{x}}_k^i = \lim_{k \rightarrow \infty} \mathbf{x}_k^i = \tilde{\mathbf{x}}) = 1$ . Hence, we obtain  $\{\hat{\mathbf{x}}_k^i\}_{k \geq 0}$  converges a.s. to the same point  $\tilde{\mathbf{x}}$ .

(B) Convergence of  $\mathbf{x}_k^i$  to  $\bar{\mathbf{z}}_k$ :

<sup>2</sup>In this paper, ‘a.s.’ is meant for ‘almost surely’. For a sequence of random vectors  $\{a_k\}_{k \geq 0}$ , we say that  $a_k$  converges to  $a$  almost surely, if  $\mathbb{P}(\lim_{k \rightarrow \infty} a_k = a) = 1$ , i.e., the probability of  $\lim_{k \rightarrow \infty} a_k = a$  is 1.

Squaring both sides of Lemma 3-(1), taking the total expectation, and summing over from  $k = 1$  to infinity, we obtain

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbb{E}[\|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\|^2] &\leq \frac{12N^2\zeta^2\Gamma^2\gamma^2}{1-\gamma^2} + \left( \frac{3\Psi_2\Gamma^2\gamma^2}{(1-\gamma)^2} + 3\Psi_2 \right) \\ &\quad + \left( \frac{3\Phi_2\Gamma^2\gamma^2}{(1-\gamma)^2} + 3\Phi_2 \right) \sum_{k=0}^{\infty} \alpha_k^2, \end{aligned}$$

where  $(\sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1})^2 \leq \frac{\gamma}{1-\gamma} \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1}^2$ ,  $\sum_{k=1}^{\infty} \sum_{r=1}^{k-1} \gamma^{k-r} \mathbb{E}[\mathbf{G}_{r-1}^2] \leq \frac{\gamma}{1-\gamma} \sum_{k=0}^{\infty} \mathbb{E}[\mathbf{G}_k^2]$  and Lemma 4 have been applied. As the step-size is square-summable, we obtain  $\sum_{k=1}^{\infty} \mathbb{E}[\|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\|^2] < \infty$ . By the monotone convergence theorem, it follows that  $\mathbb{E}[\sum_{k=1}^{\infty} \|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\|^2] = \sum_{k=1}^{\infty} \mathbb{E}[\|\mathbf{x}_k^i - \bar{\mathbf{z}}_k\|^2] < \infty$ , which implies  $\{\mathbf{x}_k^i - \bar{\mathbf{z}}_k\}_{k \geq 0}$  converges a.s. to 0.

(C) Convergence of  $\bar{\mathbf{z}}_k$  to an optimizer  $\mathbf{x}^* \in \mathcal{X}^*$ :

Finally, we will show that  $\{\bar{\mathbf{z}}_k\}_{k \geq 0}$  indeed has a limit, and converges to an optimizer  $\mathbf{x}^* \in \mathcal{X}^*$ . The proof of this part is based on the Robbins-Siegmund's Lemma [44] as quoted below for completeness.

*Lemma 6: (Robbins-Siegmund's Lemma)* Let  $u_k, v_k, w_k, \eta_k$  be non-negative random variables satisfying that

$$\begin{aligned} \mathbb{E}[u_{k+1}|\mathcal{F}_k] &\leq (1 + \eta_k)u_k - v_k + w_k \quad \text{a.s.}, \\ \sum_{k=0}^{\infty} \eta_k &< \infty \quad \text{a.s.}, \quad \sum_{k=0}^{\infty} w_k < \infty \quad \text{a.s.}, \end{aligned}$$

where  $\mathbb{E}[u_{k+1}|\mathcal{F}_k]$  is the conditional expectation for the given  $u_0, \dots, u_k, v_0, \dots, v_k, w_0, \dots, w_k, \eta_0, \dots, \eta_k$ . Then

- 1)  $\{u_k\}_{k \geq 0}$  converges a.s.;
- 2)  $\sum_{k=0}^{\infty} v_k < \infty$  a.s.

From Proposition 1-(2), we have that for any  $\mathbf{z}^* \in \mathcal{X}^*$ ,  $\mathbb{E}[\|\bar{\mathbf{z}}_{k+1} - \mathbf{z}^*\|^2|\mathcal{F}_k] \leq \|\bar{\mathbf{z}}_k - \mathbf{z}^*\|^2 - \frac{2\alpha_k}{N}(f(\bar{\mathbf{z}}_k) - f^*) + Z_k$ . To invoke Lemma 6, it suffices to show that  $\sum_{k=0}^{\infty} Z_k < \infty$ , a.s.

Now, taking the total expectation for  $Z_k$  and summing over from  $k = 1$  to infinity, we have

$$\begin{aligned} \sum_{k=1}^{\infty} \mathbb{E}[Z_k] &\leq 2\hat{D}\sqrt{n} + 2 \sum_{k=1}^{\infty} \alpha_k \beta_{1,k} + 2\alpha_0 \mathcal{K}_2 \left( B_1 \right. \\ &\quad \left. + \frac{2N(2N + \epsilon)\zeta\Gamma\gamma}{1-\gamma} + \max_{i \in \mathcal{V}} \|\mathbf{x}_0^i - \mathbf{z}^*\| \right) \sum_{k=0}^{\infty} \tilde{\beta}_k \\ &\quad + 2\mathcal{K}_2 \left[ \mathcal{K}_1 + \Phi_1 \left( \frac{(2N + \epsilon)\Gamma\gamma}{1-\gamma} + 2N \right) \right] \sum_{k=0}^{\infty} \tilde{\beta}_k \sum_{k=0}^{\infty} \alpha_k^2 \\ &\quad + 2\Phi_1(2\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) \sum_{k=0}^{\infty} \alpha_k^2 + \frac{5\Phi_2 + 4\Phi_3}{N} \sum_{k=1}^{\infty} \alpha_k^2 \\ &\quad + 2\zeta(\Phi_2 + 3N\mathcal{K}_1 + N\bar{\beta}\mathcal{K}_2)\Gamma \sum_{k=1}^{\infty} \alpha_k^2 \\ &\quad + 2 \left( \Phi_1(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + \frac{\Phi_2}{N} \right) \frac{\Gamma\gamma}{1-\gamma} \sum_{k=1}^{\infty} \alpha_k^2 \\ &\quad + 2\mathcal{K}_2\Psi_1 \left( \frac{(2N + \epsilon)\Gamma\gamma}{1-\gamma} + 2N \right) \sum_{k=0}^{\infty} \tilde{\beta}_k + 2\Psi_1(2\mathcal{K}_1 \end{aligned}$$

$$\begin{aligned} &\quad + \bar{\beta}\mathcal{K}_2) + 2\zeta\Gamma \left( \frac{(1 + 3N\mathcal{K}_1 + N\bar{\beta}\mathcal{K}_2)\gamma^2}{1-\gamma^2} + \Psi_2 \right) \\ &\quad + 2 \left( \Psi_1(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + \frac{\Psi_2}{N} \right) \frac{\Gamma\gamma}{1-\gamma} + \frac{5\Psi_2 + 4\Psi_3}{N}, \end{aligned}$$

where we applied  $\mathbb{E}[\mathbb{E}[\mathbf{G}_k|\mathcal{F}_k]\mathbf{G}_{r-1}] \leq \sqrt{\mathbb{E}[\mathbf{G}_k^2]\mathbb{E}[\mathbf{G}_{r-1}^2]} \leq \frac{1}{2}(\mathbb{E}[\mathbf{G}_k^2] + \mathbb{E}[\mathbf{G}_{r-1}^2])$  based on Cauchy-Schwarz inequality, the results in (10), and  $\sum_{k=0}^{\infty} \alpha_k \tilde{\beta}_k \sum_{r=0}^{k-1} \mathbb{E}[\mathbf{G}_r] \leq \sum_{k=0}^{\infty} \tilde{\beta}_k \sum_{r=0}^{k-1} \alpha_r \mathbb{E}[\mathbf{G}_r] \leq \sum_{k=0}^{\infty} \tilde{\beta}_k \sum_{k=0}^{\infty} \alpha_k \mathbb{E}[\mathbf{G}_k] \leq \Phi_1 \sum_{k=0}^{\infty} \tilde{\beta}_k \sum_{k=0}^{\infty} \alpha_k^2 + \Psi_1 \sum_{k=0}^{\infty} \tilde{\beta}_k$ . Since  $\sum_{k=0}^{\infty} \beta_{1,k} \alpha_k < \infty$ ,  $\sum_{k=0}^{\infty} \tilde{\beta}_k < \infty$  and  $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ , by the monotone convergence theorem, we have  $\mathbb{E}[\sum_{k=1}^{\infty} Z_k] = \sum_{k=1}^{\infty} \mathbb{E}[Z_k] < \infty$ , which proves that  $\sum_{k=1}^{\infty} Z_k < \infty$  a.s.

Invoking Lemma 6, we obtain that

$$\forall \mathbf{z}^* \in \mathcal{X}^*, \{\|\bar{\mathbf{z}}_k - \mathbf{z}^*\|^2\}_{k \geq 0} \text{ converges a.s.} \quad (19a)$$

$$\sum_{k=0}^{\infty} \alpha_k (f(\bar{\mathbf{z}}_k) - f^*) < \infty \text{ a.s.} \quad (19b)$$

Since  $f(\bar{\mathbf{z}}_k) - f^* \geq 0$ , and the step-size is non-summable, it follows from (19b) that  $\liminf_{k \rightarrow \infty} f(\bar{\mathbf{z}}_k) = f^*$  a.s. Let  $\{\bar{\mathbf{z}}_{k_1}\}_{k_1 \geq 0}$  be a subsequence of  $\{\bar{\mathbf{z}}_k\}_{k \geq 0}$  such that

$$\lim_{k_1 \rightarrow \infty} f(\bar{\mathbf{z}}_{k_1}) = \liminf_{k \rightarrow \infty} f(\bar{\mathbf{z}}_k) = f^* \text{ a.s.} \quad (20)$$

From (19a), the sequence  $\{\bar{\mathbf{z}}_k\}_{k \geq 0}$  is bounded a.s. Without loss of generality, we may assume that  $\{\bar{\mathbf{z}}_{k_1}\}_{k_1 \geq 0}$  converges a.s. to some  $\mathbf{x}^*$  (if not, we may choose one such subsequence). Due to the continuity of  $f$ , we have  $f(\bar{\mathbf{z}}_{k_1})$  converges to  $f(\mathbf{x}^*)$  a.s., which by (20) implies that  $f(\mathbf{x}^*) = f^*$ , i.e.,  $\mathbf{x}^* \in \mathcal{X}^*$ . Then we let  $\mathbf{z}^* = \mathbf{x}^*$  in (19a) and consider the sequence  $\{\|\bar{\mathbf{z}}_k - \mathbf{x}^*\|^2\}_{k \geq 0}$ . It converges a.s., and its subsequence  $\{\|\bar{\mathbf{z}}_{k_1} - \mathbf{x}^*\|^2\}_{k_1 \geq 0}$  converges a.s. to 0. Thus, we have  $\{\bar{\mathbf{z}}_k\}_{k \geq 0}$  converges a.s. to  $\mathbf{x}^*$ .

Therefore, combining the arguments of (A), (B) and (C), we complete the proof of Theorem 1.  $\blacksquare$

Our second result removes the square-summable step-size condition, and shows the convergence of  $\mathbb{E}[f(\hat{\mathbf{x}}_k^i)]$  to the optimal value.

*Theorem 2:* Suppose Assumptions 1, 2 and 3 hold. Let  $\{\hat{\mathbf{x}}_k^i\}_{k \geq 0}$  be the sequence generated by (2) with a non-increasing step-size sequence  $\{\alpha_k\}_{k \geq 0}$  satisfying

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \lim_{k \rightarrow \infty} \alpha_k = \alpha_{\infty}.$$

Let  $\epsilon$  be the constant such that  $0 < \epsilon < \min(\bar{\epsilon}, \frac{1-\gamma}{2\sqrt{3}N\Gamma\gamma})$ , where  $\Gamma$  and  $\gamma$  are some constants, and  $\bar{\epsilon} = (\frac{1-|\lambda_3|}{20+8N})^N$  with  $\lambda_3$  being the third largest eigenvalue of the weighting matrix  $A$  in (5) by setting  $\epsilon = 0$ . Let  $\beta_{1,k}$  and  $\tilde{\beta}_k$  defined in (4) satisfy  $\lim_{k \rightarrow \infty} \beta_{1,k} = 0$  and  $\sum_{k=0}^{\infty} \tilde{\beta}_k < \infty$ . Then, for any  $\mathbf{z}^* \in \mathcal{X}^*$ , we have

$$\begin{aligned} \limsup_{k \rightarrow \infty} \mathbb{E}[f(\hat{\mathbf{x}}_k^i)] - f^* &\leq \alpha_{\infty} \sum_{k=0}^{\infty} \tilde{\beta}_k N \mathcal{K}_2 \left[ \mathcal{K}_1 \right. \\ &\quad \left. + \Phi_1 \left( \frac{(2N + \epsilon)\Gamma\gamma}{1-\gamma} + 2N \right) \right] + \alpha_{\infty} \left[ 2.5\Phi_2 + 2\Phi_3 \right. \\ &\quad \left. + \frac{(N\Phi_1(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + \Phi_2 + \hat{D}\Phi_1)\Gamma\gamma}{1-\gamma} + \hat{D}\Phi_1 \right] \end{aligned}$$

$$+ N\varsigma(\Phi_2 + N(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + \hat{D})\Gamma],$$

where  $f^*$  is the optimal value of the problem, *i.e.*,  $f^* = \min_{\mathbf{z}^* \in \mathcal{X}^*} f(\mathbf{z}^*)$ ,  $\mathcal{K}_1$ ,  $\mathcal{K}_2$ ,  $\Phi_1$ ,  $\Phi_2$ ,  $\Phi_3$ ,  $\Gamma$  and  $\gamma$  are positive constants,  $\bar{\beta}$  is the upper bound of  $\tilde{\beta}_k$ , and  $\varsigma = \max\{\|\mathbf{x}_0^i\|, \|\mathbf{y}_0^i\|, i \in \mathcal{V}\}$ .

**Proof.** Taking the total expectation for the result in Proposition 1-(2) and re-arranging the terms, we have  $\alpha_k(\mathbb{E}[f(\bar{\mathbf{z}}_k)] - f^*) \leq \frac{N}{2}(\mathbb{E}[\|\bar{\mathbf{z}}_k - \mathbf{z}^*\|^2] - \mathbb{E}[\|\bar{\mathbf{z}}_{k+1} - \mathbf{z}^*\|^2]) + \frac{N}{2}\mathbb{E}[Z_k]$ . Summing over from  $k = 0$  to  $t - 1$ , we have

$$\begin{aligned} \sum_{k=0}^{t-1} \alpha_k(\mathbb{E}[f(\bar{\mathbf{z}}_k)] - f^*) &\leq \frac{N}{2} \sum_{k=0}^{t-1} \mathbb{E}[Z_k] + \frac{N}{2} \|\bar{\mathbf{z}}_0 - \mathbf{z}^*\|^2 \\ &\leq N\hat{D}\sqrt{n} + 2 \sum_{k=0}^{t-1} \alpha_k \beta_{1,k} + N\mathcal{K}_2 \left( \frac{2N(2N + \epsilon)\varsigma\Gamma\gamma}{1 - \gamma} \right. \\ &\quad \left. + \max_{i \in \mathcal{V}} \|\mathbf{x}_0^i - \mathbf{z}^*\| + B_1 \right) \sum_{k=0}^{t-1} \alpha_k \tilde{\beta}_k + N\mathcal{K}_2 \left[ \mathcal{K}_1 \right. \\ &\quad \left. + \Phi_1 \left( \frac{(2N + \epsilon)\Gamma\gamma}{1 - \gamma} + 2N \right) \right] \sum_{k=0}^{t-1} \tilde{\beta}_k \sum_{k=0}^{t-1} \alpha_k^2 \\ &\quad + \left[ \frac{(N\Phi_1(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + \Phi_2)\Gamma\gamma}{1 - \gamma} + 2.5\Phi_2 + 2\Phi_3 \right. \\ &\quad \left. + N\varsigma(\Phi_2 + N(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2))\Gamma \right] \sum_{k=0}^{t-1} \alpha_k^2 \\ &\quad + N\mathcal{K}_2\Psi_1 \left( \frac{(2N + \epsilon)\Gamma\gamma}{1 - \gamma} + 2N \right) \sum_{k=0}^{\infty} \tilde{\beta}_k \\ &\quad + N\varsigma\Gamma \left( \frac{(1 + 3N\mathcal{K}_1 + N\bar{\beta}\mathcal{K}_2)\gamma^2}{1 - \gamma^2} + \Psi_2 \right) \\ &\quad + \frac{(N\Psi_1(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + \Psi_2)\Gamma\gamma}{1 - \gamma} + \frac{N}{2} \|\bar{\mathbf{z}}_0 - \mathbf{z}^*\|^2 \\ &\quad + N\Psi_1(2\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + 2.5\Psi_2 + 2\Psi_3. \end{aligned} \quad (21)$$

Dividing both sides of (21) by  $\sum_{k=0}^{t-1} \alpha_k$  and taking the limit superior as  $t \rightarrow \infty$ , it follows from Jensen's inequality that  $\mathbb{E}[f(\hat{\mathbf{z}}_t)] \leq \frac{\sum_{k=0}^{t-1} \alpha_k \mathbb{E}[f(\bar{\mathbf{z}}_k)]}{\sum_{k=0}^{t-1} \alpha_k}$ , and Lemma 5 that  $\frac{\sum_{k=0}^{\infty} \alpha_k \beta_{1,k}}{\sum_{k=0}^{\infty} \alpha_k} = \lim_{k \rightarrow \infty} \beta_{1,k} = 0$ ,  $\frac{\sum_{k=0}^{\infty} \alpha_k^2}{\sum_{k=0}^{\infty} \alpha_k} = \alpha_{\infty}$ , we obtain  $\limsup_{k \rightarrow \infty} \mathbb{E}[f(\hat{\mathbf{z}}_k)] - f^* \leq \alpha_{\infty} \sum_{k=0}^{\infty} \tilde{\beta}_k N\mathcal{K}_2 [\mathcal{K}_1 + \Phi_1 \left( \frac{(2N + \epsilon)\Gamma\gamma}{1 - \gamma} + 2N \right) + \alpha_{\infty} \left[ \frac{(N\Phi_1(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2) + \Phi_2)\Gamma\gamma}{1 - \gamma} + 2.5\Phi_2 + 2\Phi_3 + N\varsigma(\Phi_2 + N(3\mathcal{K}_1 + \bar{\beta}\mathcal{K}_2))\Gamma \right]]$ .

It follows from Assumption 2 and Proposition 1-(1) that  $\limsup_{k \rightarrow \infty} (\mathbb{E}[f(\hat{\mathbf{x}}_k^i)] - \mathbb{E}[f(\hat{\mathbf{z}}_k)]) \leq \hat{D} \limsup_{k \rightarrow \infty} \mathbb{E}[\|\hat{\mathbf{x}}_k^i - \hat{\mathbf{z}}_k\|] \leq \hat{D}(N\varsigma\Gamma + \Phi_1 + \frac{\Phi_1\Gamma\gamma}{1 - \gamma})\alpha_{\infty}$ . The desired result follows by combining the preceding two relations. ■

*Remark 2:* Theorem 2 shows that the cost value of the multi-agent system will finally converge to a neighborhood of its optimal value with an error bounded by some terms, which are dependent on the step-size  $\alpha_k$  and parameters  $\beta_{1,k}, \beta_{2,k}$ . Appropriate choice of the step-size and parameters will lead to the exact convergence to the optimal value. In particular, if the step-size  $\alpha_k$  is set to  $1/(k+1)^a$ , where  $a \in (0, 1)$ ; the parameters  $\beta_{1,k}, \beta_{2,k}$  are set to  $1/(k+1)^{p_1}$  and  $1/(k+1)^{p_2}$ , respectively, where  $p_1 > 0$  and  $p_2 - p_1 > 1$ ; then  $\alpha_{\infty} = 0$  and  $\sum_{k=0}^{\infty} \tilde{\beta}_k < \infty$ , which means all the error terms will

converge to 0. On the other hand, Theorem 2 only proves the convergence of  $\mathbb{E}[f(\hat{\mathbf{x}}_k^i)]$ , but cannot state anything about the convergence of the sequence  $\hat{\mathbf{x}}_k^i$ , for  $i \in \mathcal{V}$ . We remark that achieving the exact convergence to the optimal value (*i.e.*,  $f(\hat{\mathbf{x}}_k^i) \rightarrow f^*$ ) is theoretically weaker than the exact convergence to an optimal solution (*i.e.*,  $\hat{\mathbf{x}}_k^i \rightarrow x^*$ ). The exact convergence of the sequence  $\hat{\mathbf{x}}_k^i$  to the optimal solution can be guaranteed based on the square-summable step-size condition, by using the Robbins-Siegmund's Lemma [44], see the proof of Theorem 1.

In the following corollary, we characterize the convergence rate of the proposed algorithm for both a diminishing step-size of  $\alpha_k = \frac{\alpha}{\sqrt{k+2}}$  and a constant step-size of  $\alpha_k = \frac{\alpha}{\sqrt{t+2}}$  if the number of iterations  $t$  is known in advance.

*Corollary 1:* Suppose Assumptions 1, 2 and 3 hold. Let  $\{\hat{\mathbf{x}}_k^i\}_{k \geq 0}$  be the sequence generated by (2) with a step-size sequence  $\alpha_k$ . Let  $\epsilon$  be the constant such that  $0 < \epsilon < \min(\bar{\epsilon}, \frac{1-\gamma}{2\sqrt{3}N\Gamma\gamma})$ , where  $\Gamma$  and  $\gamma$  are some constants, and  $\bar{\epsilon} = (\frac{1-\lambda_3}{20+8N})^N$  with  $\lambda_3$  being the third largest eigenvalue of the weighting matrix  $A$  in (5) by setting  $\epsilon = 0$ . Let the parameters  $\beta_{1,k}, \beta_{2,k}$  be set to  $\frac{1}{(k+2)^{p_1}}$  and  $\frac{1}{(k+2)^{p_2}}$ , respectively, where  $p_1 > 1$  and  $p_2 - p_1 > 1$ . Then

1) if the step-size  $\alpha_k = \frac{\alpha}{\sqrt{k+2}}$ ,  $k = 0, \dots, t-1$ , we have

$$\mathbb{E}[f(\hat{\mathbf{x}}_t^i)] - f^* \leq O(\ln t / \sqrt{t}),$$

2) if the step-size  $\alpha_k = \frac{\alpha}{\sqrt{t+2}}$ ,  $k = 0, \dots, t-1$ , we have

$$\mathbb{E}[f(\hat{\mathbf{x}}_t^i)] - f^* \leq O(1/\sqrt{t}).$$

**Proof.** Following the proof of Theorem 2, we can obtain that

$$\begin{aligned} \mathbb{E}[f(\hat{\mathbf{x}}_t^i)] - f^* &\leq \frac{1}{\sum_{k=0}^{t-1} \alpha_k} \left[ C_0 + C_1 \sum_{k=0}^{t-1} \alpha_k^2 + C_2 \sum_{k=0}^{t-1} \alpha_k \beta_{1,k} \right. \\ &\quad \left. + C_3 \sum_{k=0}^{t-1} \alpha_k \tilde{\beta}_k + C_4 \left( \sum_{k=0}^{t-1} \alpha_k^2 \right) \left( \sum_{k=0}^{t-1} \tilde{\beta}_k \right) \right], \end{aligned}$$

where  $C_0, C_1, C_2, C_3$  and  $C_4$  some constants.

For (1),  $\alpha_k = \frac{\alpha}{\sqrt{k+2}}$ ,  $k = 0, \dots, t-1$ , we have

$$\begin{aligned} \mathbb{E}[f(\hat{\mathbf{x}}_t^i)] - f^* &\leq \frac{C_0}{2\alpha[\sqrt{t+2} - \sqrt{2}]} + \frac{\alpha C_1 \ln(t+1)}{2(\sqrt{t+2} - \sqrt{2})} \\ &\quad + \frac{C_2(1 - \frac{1}{(t+1)^{p_1-0.5}})}{[\sqrt{t+2} - \sqrt{2}](2p_1 - 1)} + \frac{C_3(1 - \frac{1}{(t+1)^{p-0.5}})}{[\sqrt{t+2} - \sqrt{2}](2p - 1)} \\ &\quad + \frac{\alpha C_4 \ln(t+1)(1 - \frac{1}{(t+1)^{p-1}})}{2(\sqrt{t+2} - \sqrt{2})(p-1)} \\ &= O(1/\sqrt{t}) + O(\ln t / \sqrt{t}) = O(\ln t / \sqrt{t}). \end{aligned}$$

Likewise for (2),  $\alpha_k = \frac{\alpha}{\sqrt{t+2}}$ ,  $k = 0, \dots, t-1$ , we have

$$\begin{aligned} \mathbb{E}[f(\hat{\mathbf{x}}_t^i)] - f^* &\leq \frac{C_0\sqrt{t+2}}{t\alpha} + \frac{\alpha C_1}{\sqrt{t+2}} \\ &\quad + \frac{C_2(1 - \frac{1}{(t+1)^{p_1-1}})}{t(p_1 - 1)} + \frac{C_3(1 - \frac{1}{(t+1)^{p-1}})}{t(p - 1)} \\ &\quad + \frac{\alpha C_4(1 - \frac{1}{(t+1)^{p-1}})}{\sqrt{t+2}(p-1)} \end{aligned}$$

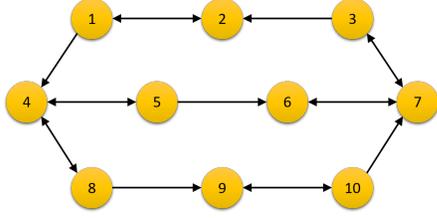


Fig. 1. Communication topology.

$$= O(1/\sqrt{t}) + O(1/t) = O(1/\sqrt{t}).$$

which gives the desired convergence rate results. ■

## V. NUMERICAL SIMULATION

In this section, we investigate the performance of the proposed algorithm through a numerical example. In particular, we consider a non-smooth test problem in a multi-agent system with  $N$  agents originated from [25]:

$$\min f(\mathbf{x}) = \sum_{i=1}^N \left( l_i |x_1 - 1| + \sum_{d=1}^{n-1} |1 + x_{d+1} - 2x_d|^2 \right), \mathbf{x} \in \mathcal{X},$$

where  $\mathbf{x} = [x_1, \dots, x_n]^\top \in \mathcal{X} \subseteq \mathbb{R}^n$ ,  $l_i, i = 1, 2, \dots, N$  is a positive constant.

In the simulation, the performance of the proposed algorithm is investigated from the following perspectives: the step-size and parameters selections, and comparison with both state-of-the-art gradient-free algorithm and gradient-based algorithm. Throughout the simulation, we let  $[A_r]_{ij} = 1/|\mathcal{N}_i^{\text{in}}|$  and  $[A_c]_{ij} = 1/|\mathcal{N}_j^{\text{out}}|$ , where  $|\mathcal{N}|$  denotes the number of elements in  $\mathcal{N}$ .  $l_i$  is randomly set in  $[0.5, 1.5]$ .

### A. Influence of Step-Size $\alpha_k$ and Parameters $\beta_{1,k}, \beta_{2,k}$

In this part, we set the dimension of the problem  $n = 1$ , the number of agents  $N = 10$  under the directed graph  $\mathcal{G}$  shown in Fig. 1. Then, we investigated the performance of the algorithm for the cases of different step-size  $\alpha_k$  and two positive parameter sequences  $\beta_{1,k}, \beta_{2,k}$ , respectively.

To test the influence of the step-size on the convergence, we set the step-size  $\alpha_k = 0.1/(1+k)^a$ , where  $a = 0, 0.2, 0.5, 0.7$  and 1. It should be noted that the step-size  $\alpha_k$  is not square-summable for  $a = 0, 0.2, 0.5$ . Two positive sequences were set to  $\beta_{1,k} = 1/(1+k)^{1.5}$  and  $\beta_{2,k} = 1/(1+k)^{2.5}$ . The convergence result was shown in Fig. 2. As can be seen, both the optimality gap decreases for diminishing step-sizes, which is consistent with our findings in Theorem 2. Moreover, it can be observed that faster convergence result is attained with slower diminishing step-size (*i.e.*, smaller  $a$ ), but larger errors (oscillations in the plot) are incurred.

To test the influence of the two positive parameter sequences on the convergence, we set  $\beta_{1,k} = 1/(1+k)^{1.5}$ ,  $\beta_k = \beta_{2,k}/\beta_{1,k} = 1/(1+k)^b$ , where  $b = 1, 3, 5, 7$  and 9. The step-size  $\alpha_k$  was set to  $0.1/\sqrt{k+1}$ . The convergence result under these five cases was plotted in Fig. 3. As can be seen, typical  $b$  values (ranging from 1 to 3) do not have much influence on the convergence rate. However, it can also be observed that when  $b$  is increasing, the convergence performance is downgraded.

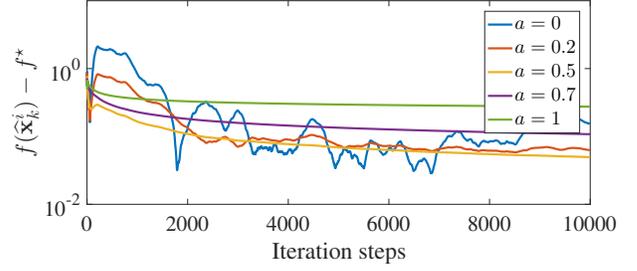


Fig. 2. Influence of step-size  $\alpha_k$  on the convergence property.

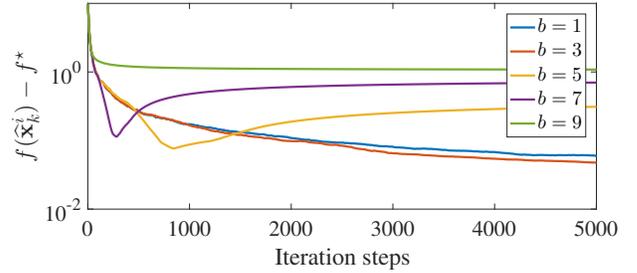


Fig. 3. Influence of  $\beta_k$  on the convergence property.

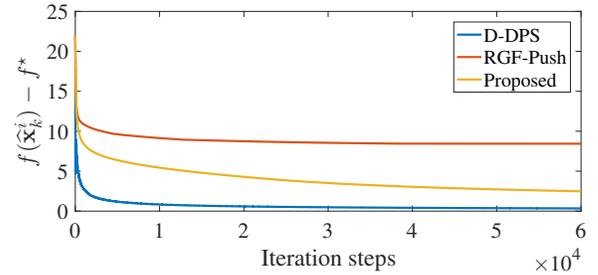


Fig. 4. Comparison between D-DPS, RGF-Push and the proposed method.

### B. Comparison with the State-Of-The-Art Algorithms

In this part, we compared our proposed method with the state-of-the-art algorithms, including the randomized gradient-free push-sum protocol (RGF-Push) proposed in [29] using diminishing smoothing parameter and a subgradient-based method (D-DPS) proposed in [37]. All these three methods can work for directed graphs. We set the dimension of the problem  $n = 2$ , the number of agents  $N = 10$  under the directed graph  $\mathcal{G}$  shown in Fig. 1. The step-size was set to  $\alpha_k = 0.1/(k+1)^{0.5}$ . The convergence results of all three methods were shown in Fig. 4. As can be seen, our proposed method shows a similar performance to the RGF-Push protocol, where both methods exhibit a theoretical convergence rate of  $\ln k/\sqrt{k}$ . The gradient-based algorithm (D-DPS) outperforms the two gradient-free methods as expected due to the use of the true gradient information.

## VI. CONCLUSIONS

This paper has considered a set constrained distributed optimization problem with possibly non-smooth cost functions. A distributed projected pseudo-gradient descent algorithm with an optimal averaging scheme has been proposed to solve the problem. The proposed algorithm has been shown to achieve

the exact convergence to the optimal value with any positive, non-summable and non-increasing step-size sequence. When the step-size is also square-summable, the exact convergence to an optimal solution has been guaranteed. Theoretical analysis on the convergence rate of the proposed algorithm has also been provided. To illustrate its performance, the proposed algorithm has been tested in a non-smooth problem. The convergence properties have been investigated, and the effectiveness has been verified by comparing with the state-of-the-art algorithms.

## APPENDIX

### A. Proof of Lemma 4

For part (1), by definition of  $g_k^i$  in (5)  $\|g_k^i\| \leq \|\epsilon \mathbf{y}_k^i\| + \|\mathbf{x}_{k+1}^i - \sum_{j=1}^N [A_r]_{ij} \mathbf{x}_k^j\| \leq \epsilon \|\mathbf{y}_k^i\| + \|\epsilon \mathbf{y}_k^i - \alpha_k \mathbf{g}^i(\mathbf{x}_k^i)\| \leq 2\epsilon \|\mathbf{y}_k^i\| + \alpha_k \|\mathbf{g}^i(\mathbf{x}_k^i)\|$ , where the second inequality follows from the projection's nonexpansive property. Summing over  $i = 1, \dots, N$ , and applying Lemma 3-(2),

$$\mathbf{G}_k \leq 4N^2 \zeta \epsilon \Gamma \gamma^k + \alpha_k \sum_{i=1}^N \|\mathbf{g}^i(\mathbf{x}_k^i)\| + 2N\epsilon \Gamma \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1}. \quad (22)$$

Multiplying both sides by  $\alpha_k$ , summing over from  $k = 1$  to  $K$ , and noting that  $\sum_{k=1}^K \alpha_k \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1} \leq \sum_{k=1}^K \sum_{r=1}^{k-1} \gamma^{k-r} \alpha_r \mathbf{G}_{r-1} \leq \frac{\gamma}{1-\gamma} \sum_{k=1}^K \alpha_k \mathbf{G}_k$ , we obtain

$$\begin{aligned} \sum_{k=1}^K \alpha_k \mathbf{G}_k &\leq 2N^2 \zeta \epsilon \Gamma \sum_{k=1}^K \gamma^{2k} + 2N^2 \zeta \epsilon \Gamma \sum_{k=1}^K \alpha_k^2 \\ &+ \sum_{k=1}^K \alpha_k^2 \sum_{i=1}^N \|\mathbf{g}^i(\mathbf{x}_k^i)\| + \frac{2N\epsilon \Gamma \gamma}{1-\gamma} \sum_{k=1}^K \alpha_k \mathbf{G}_k, \end{aligned}$$

Taking the total expectation and invoking Lemma 1-(3), we have  $\sum_{k=1}^K \alpha_k \mathbb{E}[\mathbf{G}_k] \leq 2N^2 \zeta \epsilon \Gamma \sum_{k=1}^K \gamma^{2k} + N\mathcal{K}_1 \sum_{k=1}^K \alpha_k^2 + \frac{2N\epsilon \Gamma \gamma}{1-\gamma} \sum_{k=1}^K \alpha_k \mathbb{E}[\mathbf{G}_k]$ . Re-arranging the term and noticing that  $\epsilon < \frac{1-\gamma}{2\sqrt{3}N\Gamma\gamma} < \frac{1-\gamma}{2N\Gamma\gamma}$ , we obtain the desired result by denoting  $\Phi_1 = \frac{N\mathcal{K}_1}{1-\gamma(2N\epsilon\Gamma+1)}$ , and  $\Psi_1 = \frac{2N^2 \zeta \epsilon \Gamma \gamma^2}{1-\gamma(2N\epsilon\Gamma+1)}$ .

For part (2), squaring both sides of (22), summing over from  $k = 1$  to  $K$ , and taking the total expectation, we have

$$\begin{aligned} \sum_{k=1}^K \mathbb{E}[\mathbf{G}_k^2] &\leq 48N^4 \zeta^2 \epsilon^2 \Gamma^2 \sum_{k=1}^K \gamma^{2k} + 3N^2 \mathcal{K}_1^2 \sum_{k=1}^K \alpha_k^2 \\ &+ 12N^2 \epsilon^2 \Gamma^2 \sum_{k=1}^K \mathbb{E} \left[ \left( \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1} \right)^2 \right]. \end{aligned}$$

Applying Cauchy-Schwarz inequality on the last term that

$$\left( \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1} \right)^2 \leq \frac{\gamma}{1-\gamma} \sum_{r=1}^{k-1} \gamma^{k-r} \mathbf{G}_{r-1}^2,$$

we obtain

$$\sum_{k=1}^K \mathbb{E}[\mathbf{G}_k^2] \leq 48N^4 \zeta^2 \epsilon^2 \Gamma^2 \sum_{k=1}^K \gamma^{2k} + 3N^2 \mathcal{K}_1^2 \sum_{k=1}^K \alpha_k^2$$

$$\begin{aligned} &+ \frac{12N^2 \epsilon^2 \Gamma^2 \gamma}{1-\gamma} \sum_{k=1}^K \sum_{r=1}^{k-1} \gamma^{k-r} \mathbb{E}[\mathbf{G}_{r-1}^2] \\ &\leq 48N^4 \zeta^2 \epsilon^2 \Gamma^2 \sum_{k=1}^K \gamma^{2k} + 3N^2 \mathcal{K}_1^2 \sum_{k=1}^K \alpha_k^2 \\ &+ \frac{12N^2 \epsilon^2 \Gamma^2 \gamma^2}{(1-\gamma)^2} \sum_{k=1}^K \mathbb{E}[\mathbf{G}_k^2]. \end{aligned}$$

Re-arranging the term and noticing that  $\epsilon < \frac{1-\gamma}{2\sqrt{3}N\Gamma\gamma}$ , we obtain the desired result by denoting  $\Phi_2 = \frac{3N^2 \mathcal{K}_1^2}{(1-\gamma)^2 - 12N^2 \epsilon^2 \Gamma^2 \gamma^2}$ , and  $\Psi_2 = \frac{48N^4 \zeta^2 \epsilon^2 \Gamma^2 \gamma^2}{(1-\gamma)^2 - 12N^2 \epsilon^2 \Gamma^2 \gamma^2}$ .

For part (3), multiplying both sides of (22) by  $\sum_{i=1}^N \alpha_k \|\mathbf{g}^i(\mathbf{x}_k^i)\|$ , summing over from  $k = 1$  to  $K$ , and taking the total expectation, we have

$$\begin{aligned} \sum_{k=1}^K \sum_{i=1}^N \alpha_k \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_k] &\leq 2N^3 \zeta \epsilon \mathcal{K}_1 \Gamma \sum_{k=1}^K \gamma^{2k} \\ &+ 2N^3 \zeta \epsilon \mathcal{K}_1 \Gamma \sum_{k=1}^K \alpha_k^2 + N^2 \mathcal{K}_1^2 \sum_{k=1}^K \alpha_k^2 \\ &+ 2N\epsilon \Gamma \sum_{k=1}^K \sum_{r=1}^{k-1} \gamma^{k-r} \alpha_k \sum_{i=1}^N \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_{r-1}]. \quad (23) \end{aligned}$$

Based on Cauchy-Schwarz inequality that

$$\begin{aligned} \sum_{i=1}^N \mathbf{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_{r-1}] &\leq \sum_{i=1}^N \sqrt{\mathbf{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\|^2] \mathbf{E}[\mathbf{G}_{r-1}^2]} \\ &\leq N\mathcal{K}_1 \sqrt{\mathbf{E}[\mathbf{G}_{r-1}^2]}, \end{aligned}$$

the last term of (23) holds that

$$\begin{aligned} 2N\epsilon \Gamma \sum_{k=1}^K \sum_{r=1}^{k-1} \gamma^{k-r} \alpha_k \sum_{i=1}^N \mathbb{E}[\|\mathbf{g}^i(\mathbf{x}_k^i)\| \mathbf{G}_{r-1}] \\ &\leq 2N^2 \epsilon \mathcal{K}_1 \Gamma \sum_{k=1}^K \sum_{r=1}^{k-1} \gamma^{k-r} \alpha_k \sqrt{\mathbf{E}[\mathbf{G}_{r-1}^2]} \\ &\leq N^2 \epsilon \mathcal{K}_1 \Gamma \sum_{k=1}^K \sum_{r=1}^{k-1} \gamma^{k-r} (\alpha_k^2 + \mathbf{E}[\mathbf{G}_{r-1}^2]) \\ &\leq \frac{N^2 \epsilon \mathcal{K}_1 \Gamma \gamma}{1-\gamma} \sum_{k=1}^K \alpha_k^2 + \frac{N^2 \epsilon \mathcal{K}_1 \Gamma \gamma}{1-\gamma} \sum_{k=1}^K \mathbf{E}[\mathbf{G}_k^2] \\ &\leq \frac{N^2 \epsilon \mathcal{K}_1 \Gamma \gamma}{1-\gamma} \sum_{k=1}^K \alpha_k^2 + \frac{N^2 \epsilon \mathcal{K}_1 \Gamma \gamma}{1-\gamma} \left( \Phi_2 \sum_{k=1}^K \alpha_k^2 + \Psi_2 \right). \end{aligned}$$

Combining the above relation with (23), we obtain the desired result by denoting  $\Phi_3 = 2N^3 \zeta \epsilon \mathcal{K}_1 \Gamma + N^2 \mathcal{K}_1^2 + \frac{N^2 \epsilon \mathcal{K}_1 \Gamma \gamma (1+\Phi_2)}{1-\gamma}$ , and  $\Psi_3 = \frac{2N^3 \zeta \epsilon \mathcal{K}_1 \Gamma \gamma^2}{(1-\gamma)^2 - 12N^2 \epsilon^2 \Gamma^2 \gamma^2} + \frac{N^2 \epsilon \mathcal{K}_1 \Gamma \gamma \Psi_2}{1-\gamma}$ .

## REFERENCES

- [1] R. Nowak, "Distributed EM algorithms for density estimation and clustering in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2245–2253, 2003.
- [2] S. S. Ram, V. V. Veeravalli, and A. Nedic, "Distributed and Recursive Parameter Estimation in Parametrized Linear State-Space Models," *IEEE Transactions on Automatic Control*, vol. 55, no. 2, pp. 488–492, 2010.

- [3] V. Lesser, C. L. Ortiz, and M. Tambe, *Distributed Sensor Networks : a Multiagent Perspective*. Springer US, 2003.
- [4] M. Rabbat and R. Nowak, "Decentralized source localization and tracking [wireless sensor networks]," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, pp. iii–921–4.
- [5] D. P. Palomar and M. Chiang, "Alternative Distributed Algorithms for Network Utility Maximization: Framework and Applications," *IEEE Transactions on Automatic Control*, vol. 52, no. 12, pp. 2254–2269, 2007.
- [6] M. Chiang, P. Hande, T. Lan, and C. W. Tan, "Power Control in Wireless Cellular Networks," *Foundations and Trends® in Networking*, vol. 2, no. 4, pp. 381–533, 2007.
- [7] C. Shen, T.-H. Chang, K.-Y. Wang, Z. Qiu, and C.-Y. Chi, "Distributed Robust Multicell Coordinated Beamforming With Imperfect CSI: An ADMM Approach," *IEEE Transactions on Signal Processing*, vol. 60, no. 6, pp. 2988–3003, 2012.
- [8] X. Dong and G. Hu, "Time-Varying Formation Tracking for Linear Multiagent Systems With Multiple Leaders," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3658–3664, 2017.
- [9] Z. Feng, C. Sun, and G. Hu, "Robust Connectivity Preserving Rendezvous of Multirobot Systems Under Unknown Dynamics and Disturbances," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 4, pp. 725–735, 2017.
- [10] C. Sun, G. Hu, L. Xie, and M. Egerstedt, "Robust finite-time connectivity preserving coordination of second-order multi-agent systems," *Automatica*, vol. 89, pp. 21–27, 2018.
- [11] Y. Wu, W. Xia, M. Cao, and X.-M. Sun, "Reach control problem for affine multi-agent systems on simplices," *Automatica*, vol. 107, pp. 264–271, 2019.
- [12] T.-H. Chang, A. Nedic, and A. Scaglione, "Distributed Constrained Optimization by Consensus-Based Primal-Dual Perturbation Method," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1524–1538, 2014.
- [13] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained Consensus and Optimization in Multi-Agent Networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.
- [14] I. Masubuchi, T. Wada, T. Asai, T. H. L. Nguyen, Y. Ohta, and Y. Fujisaki, "Distributed Multi-Agent Optimization Based on an Exact Penalty Method with Equality and Inequality Constraints," *SICE Journal of Control, Measurement, and System Integration*, vol. 9, no. 4, pp. 179–186, 2016.
- [15] M. Zhu and S. Martinez, "On Distributed Convex Optimization Under Inequality and Equality Constraints," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2012.
- [16] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.
- [17] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of Asynchronous Distributed Gradient Methods over Stochastic Networks," *IEEE Transactions on Automatic Control*, 2017.
- [18] G. Qu and N. Li, "Accelerated distributed Nesterov Gradient Descent for convex and smooth functions," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017, pp. 2260–2267.
- [19] D. Yuan, Y. Hong, D. W. Ho, and G. Jiang, "Optimal distributed stochastic mirror descent for strongly convex optimization," *Automatica*, vol. 90, pp. 196–203, 2018.
- [20] P. Lin, W. Ren, C. Yang, and W. Gui, "Distributed optimization with nonconvex velocity constraints, nonuniform position constraints and nonuniform stepsizes," *IEEE Transactions on Automatic Control*, vol. 64, no. 6, pp. 2575–2582, 2018.
- [21] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.
- [22] O. Kramer, D. E. Ciaurri, and S. Kozziel, "Derivative-Free Optimization." Springer, Berlin, Heidelberg, 2011, pp. 61–83.
- [23] J. Matyas, "Random Optimization," *Automation and Remote control*, vol. 26, no. 2, pp. 246–253, 1965.
- [24] O. Shamir and T. Zhang, "Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes," in *Proceedings of the 30th International Conference on Machine Learning*, 2013, pp. 71–79.
- [25] Y. Nesterov and V. Spokoiny, "Random Gradient-Free Minimization of Convex Functions," *Foundations of Computational Mathematics*, vol. 17, no. 2, pp. 527–566, 2017.
- [26] D. Yuan and D. W. C. Ho, "Randomized Gradient-Free Method for Multiagent Optimization Over Time-Varying Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1342–1347, 2015.
- [27] J. Li, C. Wu, Z. Wu, and Q. Long, "Gradient-free method for nonsmooth distributed optimization," *Journal of Global Optimization*, vol. 61, no. 2, pp. 325–340, 2015.
- [28] X.-M. Chen and C. Gao, "Strong consistency of random gradient-free algorithms for distributed optimization," *Optimal Control Applications and Methods*, vol. 38, no. 2, pp. 247–265, 2017.
- [29] D. Yuan, S. Xu, and J. Lu, "Gradient-free method for distributed multi-agent optimization via push-sum algorithms," *International Journal of Robust and Nonlinear Control*, vol. 25, no. 10, pp. 1569–1580, 2015.
- [30] Y. Pang and G. Hu, "A distributed optimization method with unknown cost function in a multi-agent system via randomized gradient-free method," in *2017 11th Asian Control Conference (ASCC)*, 2017, pp. 144–149.
- [31] Y. Pang and G. Hu, "Randomized Gradient-Free Distributed Optimization Methods for a Multi-Agent System with Unknown Cost Function," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 333–340, 2020.
- [32] A. Nedic and A. Olshevsky, "Distributed Optimization Over Time-Varying Directed Graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2015.
- [33] K. Cai and H. Ishii, "Average consensus on general strongly connected digraphs," *Automatica*, vol. 48, no. 11, pp. 2750–2761, 2012.
- [34] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.
- [35] D. Yuan, D. W. C. Ho, and S. Xu, "Zeroth-Order Method for Distributed Optimization With Approximate Projections," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 2, pp. 284–294, 2016.
- [36] Y. Pang and G. Hu, "Exact Convergence of Gradient-Free Distributed Optimization Method in a Multi-Agent System," in *2018 IEEE 58th Conference on Decision and Control (CDC)*, 2018, pp. 5728–5733.
- [37] C. Xi and U. A. Khan, "Distributed Subgradient Projection Algorithm over Directed Graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3986–3992, 2016.
- [38] I. Lobel and A. Ozdaglar, "Distributed Subgradient Methods for Convex Optimization Over Random Networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1291–1306, 2011.
- [39] V. S. Mai and E. H. Abed, "Distributed optimization over weighted directed graphs using row stochastic matrix," in *2016 American Control Conference (ACC)*, 2016, pp. 7165–7170.
- [40] C. Xi, Q. Wu, and U. A. Khan, "On the distributed optimization over directed networks," *Neurocomputing*, vol. 267, pp. 508–515, 2017.
- [41] L. Xiao, "Distributed Subgradient Algorithm for Multi-Agent Convex Optimization with Global Inequality and Equality Constraints," *Applied and Computational Mathematics*, vol. 5, no. 5, p. 213, 2016.
- [42] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems & Control Letters*, vol. 53, pp. 65–78, 2004.
- [43] G. Nagy, "The Stolz-Cesaro Theorem," *Preprint*, pp. 1–4.
- [44] B. T. Polyak, "Introduction to Optimization," *Optimization Software, Inc, New York*, 1987.