2024 IEEE 63rd Conference on Decision and Control (CDC)
December 16-19, 2024. MiCo, Milan, Italy

Zeroth-order Algorithm Design with Orthogonal Direction for Distributed Weakly Convex Optimization*

Renyi Wang, Yuan Fan, and Songsong Cheng

***Abstract*— This paper investigates a zeroth-order algorithm to solve a distributed weakly convex optimization problem over a multi-agent network, where each agent in the network has access to a local weakly convex objective function. We utilize a pseudo-gradient estimation scheme with orthogonal random directions to estimate the gradient information, which is more general than the existing coordinate descent, discretized gradient descent, and spherical smoothing methods. Moreover, we design a projected pseudo-gradient algorithm with a diminishing step size to achieve the optimal solution. Furthermore, we show the proposed algorithm converges to the optimal solution with an $\mathcal{O}(\ln k/\sqrt{k})$ convergence rate from the perspective of the Moreau envelope. Finally, we provide a numerical example to illustrate the effectiveness of the proposed algorithm.**

I. Introduction

Many practical challenges from engineering and academic fields, can be modeled as optimization problems, such as the optimal resource allocation in smart grid [1], the distributed policy evaluation in reinforcement learning [2], and the framework matching in multi-robot systems [3]. With the increase of the scale and complexity of the optimization problem, centralized optimization algorithms are difficult to address these problems because of the limited computation and communication resources. Distributed optimization is an effective tool to overcome these challenges [4]. There are many excellent works for solving distributed optimization problems, such as primal-dual [5], gradient tracking [6], and dual averaging [7], just to name a few.

In many circumstances, the optimization problem is weakly convex, such as robust phase retrieval, low-rank matrix completion, and sparse dictionary learning [8]. For unconstrained weakly convex optimization, [9] characterized a distributed primal-dual algorithm converges to the stationary point with an $\mathcal{O}(1/\sqrt{k})$ rate with respect to the norm of the gradient of the objective function. Considering the weakly convex optimization problem with feasible set constraints, [10] proposed a momentum-based Frank-Wolfe algorithm but with an $\mathcal{O}(1/\ln k)$ convergence rate. [11] investigated the projected subgradient algorithm for weakly convex optimization with the aid of the Moreau envelope and showed an $\mathcal{O}(\ln k/\sqrt{k})$ convergence rate.

*This work was supported in part by the National Natural Science Foundation of China under Grant 62103003, 61973002, and in part by the Anhui Provincial Natural Science Foundation under Grant 2008085J32, and in part by the Anhui Provincial Science and Technology Innovation Key Project 202423i08050033. (*Corresponding author: Songsong Cheng.*)

Renyi Wang, Songsong Cheng, and Yuan Fan are with the School of Electrical Engineering and Automation, Anhui University, Hefei 230601, Anhui, China. E-mail: rywang@stu.ahu.edu.cn; sscheng@ahu.edu.cn; yuanf@ahu.edu.cn.

In some cases, the exact gradient information, which is necessary for optimization algorithm design, is difficult or even infeasible to access. As an alternative, one can design the pseudo-gradient estimation by the difference of the objective function value. For the unconstrained optimization problem, [12] designed a random gradient estimation scheme based on the right-sided difference for the primal-dual algorithm with an $\mathcal{O}(1/\sqrt{k})$ convergence rate. [13] extended random gradient estimation scheme in [12] for constrained convex optimization problems with an $\mathcal{O}(1/\sqrt{k})$ convergence rate. Moreover, [14] modified the random gradient estimation with two-sided differences for constrained convex optimization problems. [15] characterized the distributed zeroth-order algorithms as having a comparative convergence performance with the centralized counterparts.

Inspired by existing works, we consider a weakly convex optimization problem to encounter a feasible set constraint and the absence of gradient information. We present a general pseudo-gradient estimation scheme utilizing orthogonal random directions. Subsequently, we propose a projected zeroth-order algorithm for the considered problem. The main contributions of this paper are summarized as follows.

1) Most distributed zeroth-order algorithms for optimization problems, e.g., [13]–[16], need the objective function being convex. To relax this constraint, [12], [17] designed zeroth-order algorithms for nonconvex optimization problems with Polyak–Łojasiewicz (P-Ł) condition. In this work, we propose a distributed zeroth-order algorithm for a weakly convex optimization problem, which needs only the objective function being Lipschitz smooth and is more general than that of [12]–[17].
2) Compared with traditional distributed algorithms for convex optimization problems in [5]–[7] and weakly convex optimization problems in [11], [18], [19], we design a pseudo-gradient algorithm for the weakly convex optimization problem. Moreover, our estimation scheme is proposed by orthogonal random directions, which includes the spherical smoothing of [12], [13], [15], [16], discrete coordinate descent of [17], and discrete gradient descent of [20] as special cases of our method.

The remainder of this paper is organized as follows: Section II presents some preliminary settings on notation and graph theory. In Section III, we formulate a distributed weakly convex optimization problem with a feasible set constraint. We design a distributed zeroth-order algorithm to solve the considered problem and analyze its convergence in Section IV. Section V provides a numerical example to

TABLE I

RELATED WORKS ON DISTRIBUTED WEAKLY OPTIMIZATION AND ZEROTH-ORDER OPTIMIZATION

Related works	Objective functions	Zeroth-order	Orthogonal Direction	Constrained	Convergence
[9]	weakly convex	×	×	×	$\mathcal{O}(1/k)$
[10], [11]	weakly convex	×	×	√	$\mathcal{O}(1/\ln k)$ in [10]; $\mathcal{O}(\ln k/\sqrt{k})$ in [11]
[12]	weakly convex	√	×	×	$\mathcal{O}(1/\sqrt{k})$
[13]–[15]	convex	√	×	√	$\mathcal{O}(1/\sqrt{k})$
This work	weakly convex	√	√	√	$\mathcal{O}(\ln k/\sqrt{k})$

illustrate the effectiveness of our algorithm and Section VI ends this paper with some conclusions. The proofs of some technical lemmas are placed in the Appendix Section.

II. PRELIMINARIES

A. Notations

In this paper, $\mathbb{N}_+$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$ are the sets of positive integer, n dimensional real column vector, and $n\times m$ dimensional real matrix. We utilize $\langle\nabla h(\mathbf{x}),\mathbf{y}\rangle$ to denote the Euclidean inner product of vectors $\nabla h(\mathbf{x})$ and $\mathbf{y}$, where $\nabla h(\mathbf{x})$ denotes the gradient of the function $h(\mathbf{x})$. Besides, we denote $\mathbf{x}=\mathrm{col}\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\}$ is the concatenated column vector of $\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N$ and use $\|\mathbf{x}\|$ to denote the Euclidean norm of $\mathbf{x}$. $[\cdot]_j$ means the j-th entry of a given vector and $\mathbf{p}^{(j)}\in\mathbb{R}^m$ is the j-th column of matrix $P\in\mathbb{R}^{m\times l}$. $\mathbf{1}_N\in\mathbb{R}^N$ denotes the N-dimensional column vector with all elements 1.

B. Graph theory

An undirected time-varying graph of a network is denoted by $\mathcal{G}(k)=(\mathcal{V},\mathcal{E}(k))$, where $\mathcal{V}=\{1,\cdots,N\}$ and $\mathcal{E}(k)\subseteq\mathcal{V}\times\mathcal{V}$ are sets of nodes and edges at time k, respectively. We mark $j\in\mathcal{N}_i(k)$ if $\{j,i\}\in\mathcal{E}(k)$, namely, node j is the neighbor of node i and exchanges information with node i at time k. $W(k)=[w_{ij}(k)]\in\mathbb{R}^{N\times N}$ is the weighted matrix such that $w_{ij}(k)=w_{ji}(k)>0$ for $\{j,i\}\in\mathcal{E}(k)$ and $w_{ij}(k)=0$ otherwise. Moreover, $\Phi(k,l)=W(k)W(k-1)\cdots W(l)$ for any $k>l$, $\Phi(k,k)=W(k)$, and $\Phi(k,l)=I$ for $k<l$. A graph is strongly connected if there exists at least one directed path between any two distinct nodes. We take the following assumption on the graph $\mathcal{G}(k)$.

Assumption 1: The graph $\mathcal{G}(k)$ is undirected and U-uniformly jointly strongly connected. Namely, one can construct doubly stochastic matrix sequences $\{W(k)\}$ and there exists a positive constant U such that $\mathcal{G}(k)\bigcup\mathcal{G}(k-1)\bigcup\cdots\bigcup\mathcal{G}(k-U+1)$ is strongly connected for any $k\in\mathbb{N}_+$.

By Assumption 1, we have the following technique lemma.

Lemma 1: [21, Proposition 1] If Assumption 1 holds, then $\|\Phi(k,l)-\frac{1}{N}\mathbf{1}\mathbf{1}^\top\|\le c\lambda^{k-l}$ with $c>0$ and $\lambda\in(0,1)$.

III. PROBLEM FORMULATION

Consider the following distributed optimization problem

$$\begin{aligned}\min_{\mathbf{x}}\quad & f(\mathbf{x}):=\tfrac{1}{N}\sum_{i=1}^{N}f_i(\mathbf{x}_i)\\ \text{s.t.}\quad & \mathbf{x}_i=\mathbf{x}_j\in\mathcal{X},\quad i,j\in\mathcal{V},\end{aligned}\tag{1}$$

where $\mathbf{x}_i\in\mathbb{R}^m,\forall i\in\mathcal{V}$ and $\mathcal{X}\subset\mathbb{R}^m$ is a closed convex set for all agents. We take the following mild assumptions for the optimization problem in (1).

Assumption 2: Each local objective function $f_i(\mathbf{x}_o)$ is L-Lipschtiz continuous and λ-Lipschtiz smooth. Namely, there exist constants L and λ such that $\|f_i(\mathbf{x}_o)-f_i(\mathbf{y}_o)\|\le L\|\mathbf{x}_o-\mathbf{y}_o\|$ and $\|\nabla f_i(\mathbf{x}_o)-\nabla f_i(\mathbf{y}_o)\|\le\lambda\|\mathbf{x}_o-\mathbf{y}_o\|$ hold.

Assumption 3: Each local objective function $f_i(\mathbf{x}_o)$ is τ-weakly convex, namely there exists a constant τ such that $h_i(\mathbf{x}_o)=f_i(\mathbf{x}_o)+\frac{\tau}{2}\|\mathbf{x}_o\|^2$ is convex.

Remark 1: This paper considers a distributed smooth optimization problem. Under Assumption 3, the considered optimization problem is nonconvex, which raises more challenges for algorithm design than the convex optimization in [2], [14], [15], [22], [23].

For the weakly convex objective function, we have the following lemma.

Lemma 2: [11, Lemma II.1.*]* If Assumption 3 holds, then

$$f\Big(\sum_{i=1}^{N}\xi_i\mathbf{x}_i\Big)\le\sum_{i=1}^{N}\xi_i f(\mathbf{x}_i)+\tfrac{\tau}{2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\xi_i\xi_j\|\mathbf{x}_i-\mathbf{x}_j\|^2,\tag{2}$$

where $\xi_i\in[0,1],\ \forall i\in\mathcal{V}$ and $\sum_{i=1}^{N}\xi_i=1$.

The intrinsic complexity of characterizing weakly convex problems predominantly stems from the presence of various stationary points, making precise measurement challenging. Therefore, we redefine the objective function of problem (1) as $\vartheta(\mathbf{x})=f(\mathbf{x})+\mathbb{I}_{\mathcal{X}}(\mathbf{x})$, where $\mathbb{I}_{\mathcal{X}}(\mathbf{x})=0$ if $\mathbf{x}\in\mathcal{X}$, and $\mathbb{I}_{\mathcal{X}}(\mathbf{x})=\infty$ otherwise. Based on $\vartheta(\mathbf{x})$, we analyze the proposed algorithm hereinafter by using the following Moreau envelope

$$\vartheta_t(\mathbf{x})=\min_{\mathbf{y}\in\mathbb{R}^m}\vartheta(\mathbf{y})+\tfrac{1}{2t}\|\mathbf{y}-\mathbf{x}\|^2,\quad 0<t<\tfrac{1}{\tau}.$$

Moreover, we define the proximal mapping $\mathrm{prox}_{tf}(\mathbf{x})=\hat{\mathbf{x}}$, where

$$\hat{\mathbf{x}}=\arg\min_{\mathbf{y}\in\mathbb{R}^m}\vartheta(\mathbf{y})+\tfrac{1}{2t}\|\mathbf{y}-\mathbf{x}\|^2.$$

Remark 2: It can be noted that $f_i(\cdot)$ is τ-weakly convex, but after defining a surrogate stationary measure by the Moreau envelope, $\vartheta(\mathbf{y})+\frac{1}{2t}\|\mathbf{y}-\mathbf{x}\|^2$ is strictly convex, which facilitates our analysis of the problem.

We provide the following technical lemma for the Moreau envelope and weakly convex objective functions.

Lemma 3: [24, Theorem 6.60] If $\vartheta(\mathbf{x})$ is a proper closed and convex function and $t>0$, then $\|\nabla\vartheta_t(\mathbf{x})\|=\frac{1}{t}\|\hat{\mathbf{x}}-\mathbf{x}\|$.

IV. Main Results

A. Zeroth-order Algorithm Design

For the optimization problem in (1), we design the following distributed zeroth-order algorithm

$$\begin{cases} \mathbf{v}_{i,k} = \sum\limits_{j=1}^{N} w_{ij}(k)\mathbf{x}_{j,k}, \\ \big[\nabla_k f_i(\mathbf{v}_{i,k})\big]_j = \frac{1}{h_k}\big[f_i(\mathbf{v}_{i,k} + h_k \mathbf{p}_{i,k}^{(j)}) - f_i(\mathbf{v}_{i,k})\big], \\ \mathbf{x}_{i,k+1} = \mathrm{Proj}_{\mathcal{X}}\big\{\mathbf{v}_{i,k} - \alpha_k P_{i,k}\nabla_k f_i(\mathbf{v}_{i,k})\big\}. \end{cases} \tag{3}$$

In (3), $\mathbf{v}_{i,k}$ is the weighted average of decision variables of agent i and its neighbors, $P_{i,k}\nabla_k f_i(\mathbf{v}_{i,k}) \in \mathbb{R}^m$ is the estimate of the gradient information $\nabla f_i(\mathbf{v}_{i,k})$. $\big[\nabla_k f_i(\mathbf{v}_{i,k})\big]_j$ is the j-th element of $\nabla_k f_i(\mathbf{v}_{i,k})$ and generated by the difference between $f_i(\mathbf{v}_{i,k}+h_k\mathbf{p}_{i,k}^{(j)})$ and $f_i(\mathbf{v}_{i,k})$, where $\mathbf{p}_{i,k}^{(j)}$ is the j-th column of matrix $P_{i,k}$. We make the following assumption for the matrix $P_{i,k}$.

Assumption 4: Each randomized matrix $P_{i,k} \in \mathbb{R}^{m\times l}$, $\forall i \in \mathcal{V}$, $k \in \mathbb{N}_+$ satisfies the following two conditions.

1) $P_{i,k}^\top P_{i,k} \overset{\text{a.s.}}{=} \frac{m}{l} I_l$;
2) $\mathbb{E}\{P_{i,k}P_{i,k}^\top\} = I_m$.

Remark 3: In our algorithm, we design the gradient estimation scheme with the aid of the differences based on the columns of the orthogonal matrix $P_{i,k}$. The matrix $P_{i,k}$ can be generated by coordinate descent, random orthogonal matrices, and spherical smoothing (see [25] for details).

Remark 4: Compared with existing works on weakly convex optimization with exact gradient information in [11], [19], we proposed a gradient estimation scheme for solving this class optimization problem, which raises more challenges in dealing with the gradient estimation error and the design of difference factor h_k.

Moreover, we have the following technical lemmas based on the proposed algorithm.

Lemma 4: [11, Lemma II.4.*]* If step size α_k satisfies $\sum_{k=0}^{\infty}\alpha_k = \infty$, $\lim_{k\to\infty}\alpha_k = 0$, and $\lim_{k\to\infty}\frac{\alpha_{k+1}}{\alpha_k} = 1$, then $\sum_{k=0}^{T-1}\varsigma^k\alpha_{T-k-1} = \mathcal{O}(\frac{\alpha_{T-1}}{1-\varsigma})$ with $\varsigma \in (0,1)$.

Lemma 5: If Assumptions 2 and 4 hold, then

1) $\mathbb{E}\big\{\|\nabla_k f_i(\mathbf{v}_{i,k}) - P_{i,k}^\top\nabla f_i(\mathbf{v}_{i,k})\|\big\} \le \frac{\lambda m h_k}{2\sqrt{l}}$;
2) $\mathbb{E}\big\{\|\nabla_k f_i(\mathbf{v}_{i,k})\|^2\big\} \le 2(\frac{\lambda^2 h_k^2 m^2}{4l} + L^2)$.

Remark 5: Lemma 5 indicates that a larger value of l yields a more accurate gradient estimation. Therefore, our scheme has a better estimation performance than the existing methods in [12], [15]–[17], [26], [27], which is the special case of our scheme with $l = 1$.

Define $\mathbf{e}_{i,k} = \mathrm{Proj}_{\mathcal{X}}\{\mathbf{v}_{i,k} - \alpha_k P_{i,k}\nabla_k f_i(\mathbf{v}_{i,k})\} - \mathbf{v}_{i,k}$ and $\Delta_k = \mathbf{x}_k - J\mathbf{x}_k$ with $J = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$. We have the following primary conclusion.

Lemma 6: Let $\mathbf{x}_{i,k}$ be generated by the algorithm in (3). if Assumptions 1-4 hold and $\lim_{k\to\infty} h_k = 0$, then

1) $\mathbb{E}\big\{\|\mathbf{e}_k\|^2\big\} \le N B_k^2\alpha_k^2$;
2) $\mathbb{E}\big\{\|\Delta_k\|\big\} = \mathcal{O}(\frac{\sqrt{N}B_k}{1-\lambda}\alpha_k)$,

where $\mathbf{e}_k = \mathrm{col}\{\mathbf{e}_{1,k}, \cdots, \mathbf{e}_{N,k}\}$ and $B_k^2 = \frac{2\lambda^2 h_k^2 m^3}{4l^2} + \frac{2mL^2}{l}$.

B. Converegnce Analysis

The following theorem provides the consensus performance of the proposed algorithm from the perspective of the expectation of penalty function $\vartheta(\cdot)$.

Theorem 1: Under Assumptions 1-4, let $\mathbf{x}_{i,k}$ be generated by the algorithm in (3). If $\sum_{k=0}^{\infty}\alpha_k^2 < \infty$, then

$$\lim_{k\to\infty}\big|\mathbb{E}\{\vartheta(\mathbf{x}_{i,k})\} - \mathbb{E}\{\vartheta(\bar{\mathbf{x}}_k)\}\big| = 0,$$

where $\bar{\mathbf{x}}_k = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i,k}$ is the average of local variables.

Proof. Define $\hat{\mathbf{v}}_{i,k} = \arg\min_{\mathbf{x}_o\in\mathcal{X}}\big\{f(\mathbf{x}_o) + \frac{1}{2t}\|\mathbf{x}_o - \mathbf{v}_{i,k}\|^2\big\}$ and according to the dynamics of $\mathbf{x}_{i,k}$ in (3),

$$\begin{aligned} &\|\mathbf{x}_{i,k+1} - \hat{\mathbf{v}}_{i,k}\|^2 \\ &\le \|\mathbf{v}_{i,k} - \alpha_k P_{i,k}\nabla_k f_i(\mathbf{v}_{i,k}) - \hat{\mathbf{v}}_{i,k}\|^2 \\ &= \|\mathbf{v}_{i,k} - \hat{\mathbf{v}}_{i,k}\|^2 + \alpha_k^2\|P_{i,k}\nabla_k f_i(\mathbf{v}_{i,k})\|^2 \\ &\quad +2\alpha_k\langle\nabla f_i(\mathbf{v}_{i,k}), \hat{\mathbf{v}}_{i,k} - \mathbf{v}_{i,k}\rangle \\ &\quad +2\alpha_k\langle P_{i,k}P_{i,k}^\top\nabla f_i(\mathbf{v}_{i,k}) - \nabla f_i(\mathbf{v}_{i,k}), \hat{\mathbf{v}}_{i,k} - \mathbf{v}_{i,k}\rangle \\ &\quad +2\alpha_k\langle P_{i,k}\nabla_k f_i(\mathbf{v}_{i,k}) - P_{i,k}P_{i,k}^\top\nabla f_i(\mathbf{v}_{i,k}), \hat{\mathbf{v}}_{i,k} - \mathbf{v}_{i,k}\rangle, \end{aligned} \tag{4}$$

where the inequality follows from the nonexpanding of the projected operator, the equality holds by the fact $\|\mathbf{a}+\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + 2\langle\mathbf{a},\mathbf{b}\rangle$. Taking expectation on (4),

$$\begin{aligned} &\mathbb{E}\big\{\|\mathbf{x}_{i,k+1} - \hat{\mathbf{v}}_{i,k}\|^2\big\} \\ &\le \mathbb{E}\big\{\|\mathbf{v}_{i,k} - \hat{\mathbf{v}}_{i,k}\|^2\big\} + 2\alpha_k\mathbb{E}\big\{(f_i(\hat{\mathbf{v}}_{i,k}) - f_i(\mathbf{v}_{i,k}) \\ &\quad +\tfrac{\tau}{2}\|\mathbf{v}_{i,k+1} - \hat{\mathbf{v}}_{i,k}\|^2)\big\} + \tfrac{\lambda m\sqrt{m}(\alpha_k h_k)}{l} \\ &\quad \times\mathbb{E}\big\{\|\mathbf{v}_{i,k} - \hat{\mathbf{v}}_{i,k}\|\big\} + B_k^2\alpha_k^2 \\ &\le \mathbb{E}\big\{\|\mathbf{v}_{i,k} - \hat{\mathbf{v}}_{i,k}\|^2\big\} + 2\alpha_k\mathbb{E}\big\{(f_i(\hat{\mathbf{v}}_{i,k}) - f_i(\mathbf{v}_{i,k}) \\ &\quad +\tfrac{\tau}{2}\|\mathbf{v}_{i,k+1} - \hat{\mathbf{v}}_{i,k}\|^2)\big\} + \tfrac{h_k^2}{2}\mathbb{E}\big\{\|\mathbf{v}_{i,k} - \hat{\mathbf{v}}_{i,k}\|^2\big\} \\ &\quad +B_k^2\alpha_k^2 + \tfrac{\lambda^2 m^3\alpha_k^2}{2l^2}, \end{aligned} \tag{5}$$

where the first inequality holds by Assumption 4, Lemma 5 and the weak convexity of $f_i(\cdot)$, the second inequality follows from Young's inequality. Note that

$$\begin{aligned} &f_i(\hat{\mathbf{v}}_{i,k}) - f_i(\mathbf{v}_{i,k}) \\ &\le L\|\hat{\mathbf{v}}_{i,k} - \mathbf{s}_k\| + f_i(\bar{\mathbf{x}}_k) - f_i(\mathbf{v}_{i,k}) + f_i(\mathbf{s}_k) - f_i(\bar{\mathbf{x}}_k) \\ &\le L(\tfrac{1}{1-t\tau} + 1)\|\mathbf{v}_{i,k} - \bar{\mathbf{x}}_k\| + f_i(\mathbf{s}_k) - f_i(\bar{\mathbf{x}}_k) \\ &\le \tfrac{L(2-t\tau)}{1-t\tau}\sum_{j=1}^{N} w_{ij}(k)\|\mathbf{x}_{j,k} - \bar{\mathbf{x}}_k\| + f_i(\mathbf{s}_k) - f_i(\bar{\mathbf{x}}_k), \end{aligned} \tag{6}$$

where $\mathbf{s}_k = \arg\min_{\mathbf{x}_o\in\mathcal{X}}\big\{f(\mathbf{x}_o) + \frac{1}{2t}\|\mathbf{x}_o - \bar{\mathbf{x}}_k\|^2\big\}$, the first inequality follows from the Lipschitz continuity of $f_i(\cdot)$, the second one holds because of the continuity of proximal operator (See *[11, Lemma* II.8.*]*), and the last one is deduced by the convexity of the norm function. Moreover,

$$\begin{aligned} &\tfrac{\tau}{2}\|\mathbf{v}_{i,k} - \hat{\mathbf{v}}_{i,k}\|^2 \\ &\le \tau\|\bar{\mathbf{x}}_k - \mathbf{s}_k\|^2 + \tau\|\mathbf{v}_{i,k} - \bar{\mathbf{x}}_k + \mathbf{s}_k - \hat{\mathbf{v}}_{i,k}\|^2 \\ &\le \tau\|\bar{\mathbf{x}}_k - \mathbf{s}_k\|^2 + 2\tau(1 + \tfrac{1}{(1-t\tau)^2})\|\mathbf{v}_{i,k} - \bar{\mathbf{x}}_k\|^2 \\ &\le \tau\|\bar{\mathbf{x}}_k - \mathbf{s}_k\|^2 + 2\tau(1 + \tfrac{1}{(1-t\tau)^2})\sum_{j=1}^{N} w_{ij}(k) \\ &\quad \times\|\mathbf{x}_{j,k} - \bar{\mathbf{x}}_k\|^2. \end{aligned} \tag{7}$$

According to (7), we further rewrite (5) as follows

$$\begin{aligned}
&\mathbb{E}\big\{\|\mathbf{x}_{i,k+1}-\hat{\mathbf{v}}_{i,k}\|^2\big\}\\
&\le \mathbb{E}\big\{\|\mathbf{v}_{i,k}-\hat{\mathbf{v}}_{i,k}\|^2\big\}+2\alpha_k\big(\mathbb{E}\big\{f_i(\hat{\mathbf{v}}_{i,k})-f_i(\mathbf{v}_{i,k})\big\}\\
&\quad+\tfrac{\tau}{2}\mathbb{E}\big\{\|\mathbf{v}_{i,k+1}-\hat{\mathbf{v}}_{i,k}\|^2\big\}\big)+h_k^2\mathbb{E}\big\{\|\bar{\mathbf{x}}_k-\mathbf{s}_k\|^2\big\}\\
&\quad+2h_k^2\big(1+\tfrac{1}{(1-t\tau)^2}\big)\mathbb{E}\big\{\sum_{j=1}^{N}w_{ij}(k)\|\mathbf{x}_{j,k}-\bar{\mathbf{x}}_k\|^2\big\}\\
&\quad+G_k^2\alpha_k^2,
\end{aligned}\tag{8}$$

where $G_k^2=\frac{2\lambda^2(1+h_k^2)m^3+8m}{4l}$. Combining (6) and (7) yields

$$\begin{aligned}
&\sum_{i=1}^{N}\big[f_i(\hat{\mathbf{v}}_{i,k})-f_i(\mathbf{v}_{i,k})+\tfrac{\tau}{2}\|\hat{\mathbf{v}}_{i,k}-\mathbf{v}_{i,k}\|^2\big]\\
&\le \tfrac{L(2-t\tau)}{1-t\tau}\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|+N\big[f(\mathbf{s}_k)-f(\bar{\mathbf{x}}_k)\\
&\quad+\tau\|\mathbf{s}_k-\bar{\mathbf{x}}_k\|^2\big]+2\tau\big[1+\tfrac{1}{(1-t\tau)^2}\big]\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|^2.
\end{aligned}\tag{9}$$

For the term $f(\mathbf{s}_k)-f(\bar{\mathbf{x}}_k)+\tau\|\mathbf{s}_k-\bar{\mathbf{x}}_k\|^2$, we obtain

$$\begin{aligned}
&f(\mathbf{s}_k)-f(\bar{\mathbf{x}}_k)+\tau\|\mathbf{s}_k-\bar{\mathbf{x}}_k\|^2\\
&=f(\mathbf{s}_k)-f(\bar{\mathbf{x}}_k)+\big(\tfrac{1}{2t}-\tfrac{1}{2t}+\tau\big)\|\bar{\mathbf{x}}_k-\mathbf{s}_k\|^2\\
&\le\big(\tau-\tfrac{1}{2t}\big)\|\bar{\mathbf{x}}_k-\mathbf{s}_k\|^2,
\end{aligned}\tag{10}$$

where the inequality follows from the definition of $\mathbf{s}_k$. Moreover, combining (8)-(10) and $h_k^2\le\alpha_k$ yields

$$\begin{aligned}
&\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k+1}-\hat{\mathbf{v}}_{i,k}\|^2\big\}\\
&\le\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{v}_{i,k}-\hat{\mathbf{v}}_{i,k}\|^2\big\}+2\alpha_k\Big[N(-\tfrac{1}{2t}+\tau+1)\\
&\quad\times\mathbb{E}\|\bar{\mathbf{x}}_k-\mathbf{s}_k\|^2+\tfrac{L(2-t\tau)}{1-t\tau}\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|\big\}\\
&\quad+2(\tau+1)(1+\tfrac{1}{(1-t\tau)^2})\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|^2\big\}\Big]\\
&\quad+NG_k^2\alpha_k^2.
\end{aligned}\tag{11}$$

From the definition of $\vartheta_t(\mathbf{x}_{i,k+1})$, we get

$$\vartheta_t(\mathbf{x}_{i,k+1})\le f(\mathbf{z})+\tfrac{1}{2t}\|\mathbf{x}_{i,k+1}-\mathbf{z}\|^2,\quad\forall\mathbf{z}\in\mathcal{X}.\tag{12}$$

Substituting $\mathbf{z}=\hat{\mathbf{v}}_{i,k}$ into (12) and taking its expectation,

$$\mathbb{E}\{\vartheta_t(\mathbf{x}_{i,k+1})\}\le\mathbb{E}\big\{f(\hat{\mathbf{v}}_{i,k})\big\}+\tfrac{1}{2t}\mathbb{E}\{\|\mathbf{x}_{i,k+1}-\hat{\mathbf{v}}_{i,k}\|^2\}.\tag{13}$$

Taking the summation of (13) from $i=1$ to N yields

$$\begin{aligned}
&\mathbb{E}\big\{\sum_{i=1}^{N}\vartheta_t(\mathbf{x}_{i,k+1})\big\}\\
&\le\mathbb{E}\big\{\sum_{i=1}^{N}f(\hat{\mathbf{v}}_{i,k})\big\}+\tfrac{1}{2t}\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k+1}-\hat{\mathbf{v}}_{i,k}\|^2\big\}\\
&\le\mathbb{E}\big\{\sum_{i=1}^{N}\vartheta_t(\mathbf{v}_{i,k})\big\}+\tfrac{\alpha_k}{t}\Big[N(\tau+1-\tfrac{1}{2t})\mathbb{E}\big\{\|\bar{\mathbf{x}}_k-\mathbf{s}_k\|^2\big\}\\
&\quad+\tfrac{L(2-t\tau)}{1-t\tau}\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|\big\}+\tfrac{NG_k^2\alpha_k^2}{2t}\\
&\quad+2(\tau+1)(1+\tfrac{1}{(1-t\tau)^2})\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|^2\big\}\Big].
\end{aligned}\tag{14}$$

By $\mathbf{v}_{i,k}=\sum_{j=1}^{N}w_{ij}(k)\mathbf{x}_{j,k}$, we achieve

$$\begin{aligned}
&\vartheta_t(\mathbf{v}_{i,k})\\
&=f\Big(\sum_{j=1}^{N}w_{ij}(k)\hat{\mathbf{v}}_{i,k}\Big)+\tfrac{1}{2t}\Big\|\sum_{j=1}^{N}w_{ij}(k)(\hat{\mathbf{v}}_{j,k}-\mathbf{x}_{j,k})\Big\|^2\\
&\le f\Big(\sum_{j=1}^{N}w_{ij}(k)\hat{\mathbf{x}}_{i,k}\Big)+\tfrac{1}{2t}\Big\|\sum_{j=1}^{N}w_{ij}(k)(\hat{\mathbf{x}}_{j,k}-\mathbf{x}_{j,k})\Big\|^2\\
&\le\sum_{j=1}^{N}w_{ij}(k)f(\hat{\mathbf{x}}_{j,k})+\tfrac{\tau}{2}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N}w_{ij}(k)w_{il}(k)\\
&\quad\times\|\hat{\mathbf{x}}_{j,k}-\hat{\mathbf{x}}_{l,k}\|^2+\sum_{j=1}^{N}w_{ij}(k)\tfrac{1}{2t}\|\hat{\mathbf{x}}_{j,k}-\mathbf{x}_{j,k}\|^2\\
&\le\sum_{j=1}^{N}w_{ij}(k)\vartheta_t(\mathbf{x}_{j,k})+\tfrac{\tau}{2(1-t\tau)^2}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N}w_{ij}(k)\\
&\quad\times w_{il}(k)\|\mathbf{x}_{j,k}-\mathbf{x}_{l,k}\|^2,
\end{aligned}\tag{15}$$

where the first inequality holds due to the definition of $\hat{\mathbf{v}}_{i,k}$ and $\sum_{j=1}^{N}w_{ij}(k)\hat{\mathbf{x}}_{j,k}\in\mathcal{X}$, the second one follows from the convexity of $f(\cdot)$ and Lemma 2, and the last one is deduced by the smoothness of proximal mapping (see *[11, Lemma* II.8.*]*). By (14) and defining $\bar{\vartheta}_{t,k+1}=\frac{1}{N}\sum_{i=1}^{N}\vartheta_t(\mathbf{x}_{i,k+1})$,

$$\mathbb{E}\{\bar{\vartheta}_{t,k+1}\}\le\mathbb{E}\big\{\bar{\vartheta}_{t,k}\big\}+H_k+\tfrac{G_k^2\alpha_k^2}{2t},\tag{16}$$

where

$$\begin{aligned}
H_k=&\tfrac{\tau}{2N(1-t\tau)^2}\mathbb{E}\big\{\sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N}w_{ij}(k)w_{il}(k)\|\mathbf{x}_{j,k}\\
&-\mathbf{x}_{l,k}\|^2\big\}+\tfrac{\alpha_k}{t}\Big[\tfrac{L(2-t\tau)}{N(1-t\tau)}\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|\big\}\\
&+\tfrac{2(\tau+1)}{N}(1+\tfrac{1}{(1-t\tau)^2})\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|^2\big\}\Big].
\end{aligned}$$

Note that,

$$\begin{cases}
\mathbb{E}\big\{\sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{l=j+1}^{N}w_{ij}(k)w_{il}(k)\|\mathbf{x}_{j,k}-\mathbf{x}_{l,k}\|^2\big\}=\mathcal{O}(\frac{NB_k^2\alpha_k^2}{(1-\lambda)^2}),\\
\alpha_k\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|\big\}=\mathcal{O}(\frac{NB_k\alpha_k^2}{1-\lambda}),\\
\alpha_k\mathbb{E}\big\{\sum_{i=1}^{N}\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|^2\big\}=\mathcal{O}(\frac{NB_k^2\alpha_k^3}{(1-\lambda)^2}).
\end{cases}\tag{17}$$

By (17), $H_k=\mathcal{O}(\frac{B_k^2\alpha_k^2}{(1-\lambda)^2})$. Noting that $\vartheta_t(x)$ is lower bounded on $\mathcal{X}$ and $B_k^2\le B_0^2$, we have

$$\begin{aligned}
&\mathbb{E}\{\bar{\vartheta}_{t,k+1}\}-\inf\mathbb{E}\big\{\vartheta_t(\mathbf{x})\big\}\\
&\le\mathbb{E}\big\{\bar{\vartheta}_{t,k}\big\}-\inf\mathbb{E}\big\{\vartheta_t(\mathbf{x})\big\}+\mathcal{O}(\alpha_k^2).
\end{aligned}\tag{18}$$

Since $\sum_{k=0}^{\infty}\alpha_k^2<\infty$, using Lemma 2 [1] in *[28, Chapter 2.2]*, $\{\bar{\vartheta}_{t,k}\}$ converges to some value $\bar{\vartheta}_t$.

Recalling the continuity of $\vartheta_t(\mathbf{x})$ and $\lim_{k\to\infty}\mathbb{E}\big\{\|\mathbf{x}_{i,k}-\bar{\mathbf{x}}_k\|\big\}=0$, it follows that

$$\begin{aligned}
&\lim_{k\to\infty}|\mathbb{E}\big\{\vartheta_t(\mathbf{x}_{i,k})\big\}-\mathbb{E}\big\{\vartheta_t(\bar{\mathbf{x}}_k)\big\}|^2=0,\\
&\text{and}\\
&\lim_{k\to\infty}|\mathbb{E}\big\{\bar{\vartheta}_{t,k}\big\}-\mathbb{E}\big\{\vartheta_t(\bar{\mathbf{x}}_k)\big\}|^2\\
&\le\tfrac{1}{N}\lim_{k\to\infty}\sum_{i=1}^{N}|\mathbb{E}\big\{\vartheta_t(\mathbf{x}_{i,k})\big\}-\mathbb{E}\big\{\vartheta_t(\bar{\mathbf{x}}_k)\big\}|^2=0.
\end{aligned}\tag{19}$$

This completes the proof. □

[1] If sequences $\{c_k\}$, $\{\iota_k\}$, and $\{\gamma_k\}$ satisfy $c_{k+1}\ge0$, $c_{k+1}\le(1+\iota_k)c_k+\gamma_k$, $\sum_{k=0}^{\infty}\iota_k<\infty$, and $\gamma_k<\infty$, then $\lim_{k\to\infty}c_k=c\ge0$.

Theorem 1 states that the function values of the Moreau envelope converge almost surely to the counterparts at the mean $\bar{\mathbf{x}}_k$. To achieve the explicit convergence performance of the proposed algorithm, it is necessary to investigate the convergence of $\|\nabla\vartheta_t(\bar{\mathbf{x}}_k)\|^2$, which is characterized in the following theorem.

Theorem 2: Under Assumptions 1-4, if $\sum_{k=0}^{\infty}\alpha_k=\infty$, $\sum_{k=0}^{\infty}\alpha_k^2<\infty$, and $h_k^2\le\alpha_k$, then

$$\begin{aligned}&\inf_{k\in\mathbb{N}_+}\mathbb{E}\{\|\nabla\vartheta_t(\bar{\mathbf{x}}_k)\|^2\}\\&\le\tfrac{2}{(1-2t\tau-2t)\sum_{k=0}^{\infty}\alpha_k}\Big[\mathbb{E}\{\bar{\vartheta}_{t,0}\}-\mathbb{E}\{\bar{\vartheta}_{t,\infty}\}\\&\quad+\sum_{k=0}^{\infty}H_k+\sum_{k=0}^{\infty}\tfrac{G_k^2\alpha_k^2}{2t}\Big].\end{aligned}$$

Proof. We rewrite the inequality (16) as follows

$$\begin{aligned}&\tfrac{\alpha_k}{t}(\tfrac{1}{2t}-\tau-1)\mathbb{E}\{\|\bar{\mathbf{x}}_k-\mathbf{s}_k\|^2\}\\&\le\mathbb{E}\{\bar{\vartheta}_{t,k}\}-\mathbb{E}\{\bar{\vartheta}_{t,k+1}\}+H_k+\tfrac{G_k^2\alpha_k^2}{2t}.\end{aligned}\tag{20}$$

Taking the summation of (20) from $k=0$ to ∞,

$$\begin{aligned}&\sum_{k=0}^{\infty}\tfrac{\alpha_k}{t}(\tfrac{1}{2t}-\tau-1)\mathbb{E}\{\|\bar{\mathbf{x}}_k-\mathbf{s}_k\|^2\}\\&\le\mathbb{E}\{\bar{\vartheta}_{t,0}\}-\mathbb{E}\{\bar{\vartheta}_{t,\infty}\}+\sum_{k=0}^{\infty}H_k+\sum_{k=0}^{\infty}\tfrac{G_k^2\alpha_k^2}{2t}.\end{aligned}\tag{21}$$

By Lemma 3, the definition of $\mathbf{s}_k$, and rearranging related terms of (21) yields

$$\begin{aligned}&\inf_{k\in\mathbb{N}_+}\mathbb{E}\{\|\nabla\vartheta_t(\bar{\mathbf{x}}_k)\|^2\}\\&\le\tfrac{2}{(1-2t\tau-2t)\sum_{k=0}^{\infty}\alpha_k}\Big[\mathbb{E}\{\bar{\vartheta}_{t,0}\}-\mathbb{E}\{\bar{\vartheta}_{t,\infty}\}\\&\quad+\sum_{k=0}^{\infty}H_k+\sum_{k=0}^{\infty}\tfrac{G_k^2\alpha_k^2}{2t}\Big].\end{aligned}\tag{22}$$

According to $\sum_{k=0}^{\infty}\alpha_k=\infty$, $\sum_{k=0}^{\infty}H_k<\infty$, and $\sum_{k=0}^{\infty}G_k^2\alpha_k^2<\infty$, we achieve the conclusion. □

By Theorem 2, when $\alpha_k=\mathcal{O}(\frac{1}{\sqrt{k}})$ and $h_k^2=\alpha_k$, we directly obtain the following corollary.

Corollary 1: Under Assumptions 1-4, if $\alpha_k=\mathcal{O}(\frac{1}{\sqrt{k}})$ and $h_k^2=\alpha_k$, then $\inf_{1\le k\le T}\mathbb{E}\{\|\nabla\vartheta_t(\bar{\mathbf{x}}_k)\|^2\}=\mathcal{O}(\frac{\ln k}{\sqrt{k}})$ for a sufficiently large T.

Remark 6: In contrast to the convergence rate of $\mathcal{O}(\ln k/\sqrt{k})$ in [11], our method attains the same counterpart even in the absence of exact gradient information. By relaxing the convexity assumption in [13], [14] as weak convexity, we obtain a similar convergence rate to that in [13], [14]. Moreover, the convergence rate of our method is faster than the $\mathcal{O}(1/\ln k)$ rate in [10], where the objective function is also weakly convex.

V. Numerical Simulation

We consider the following weakly convex problem to illustrate the effectiveness of the proposed algorithm

$$\min_{\mathbf{x}\in\mathbb{R}^m}\quad f(\mathbf{x})=\tfrac{1}{N}\sum_{i=1}^{N}\Big(\tfrac{1}{d}\sum_{j=1}^{d}|\langle\mathbf{u}_{i,j},\mathbf{x}\rangle^2-\langle\mathbf{u}_{i,j},\tilde{\mathbf{x}}\rangle^2|\Big),\tag{23}$$

which is yielded from the robust phase retrieval problem [11]. In this example, we set $N=5$, $d=10$ and randomly choose the entry of $\tilde{\mathbf{x}}$ from $[-6,-4]$. Moreover, the elements of the measurements $\mathbf{u}_{i,j}$ are drawn from the standard normal distribution $\mathcal{N}(0,1)$. Specifically, we design $\alpha_k=0.01l/\sqrt{k}$, $h_k=\sqrt{\alpha_k}$ and the network topology as an undirected circle.

The simulation results are shown in Fig. 1 with $l=1$, 3, and 5 cases, where $\mathrm{redist}(\bar{\mathbf{x}}_k,\mathcal{X}^*)=\mathrm{dist}(\bar{\mathbf{x}}_k,\mathcal{X}^*)/\mathrm{dist}(\bar{\mathbf{x}}_1,\mathcal{X}^*)$ and $\mathrm{dist}(\bar{\mathbf{x}}_k,\mathcal{X}^*)$ is the distance between $\bar{\mathbf{x}}_k$ and the optimal solution set $\mathcal{X}^*$. Notably, the methods in [15], [26] both are the special case of our method with $l=1$. From Fig. 1, for a larger l, our algorithm achieves a better convergence. Especially, the $l=3$ and 5 cases provide significant outperformance compared to that of [15], [26].

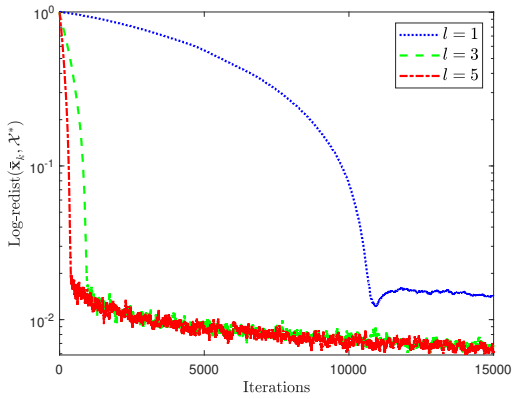

Fig. 1. The results of numerical simulation

VI. Conclusions

In this paper, we propose a distributed projected pseudo-gradient algorithm for solving a weakly convex optimization problem with a feasible set constraint. We provide a thorough analysis of our algorithm's convergence, focusing on the perspective of the Moreau envelope, under the premise of employing an unsummable step size. Notably, employing the design choices $\alpha_k=\alpha_0 l/\sqrt{k}$ and $h_k^2=\alpha_k$, we demonstrate an $\mathcal{O}(\ln k/\sqrt{k})$ convergence rate for our algorithm. Finally, we illustrated the effectiveness of the proposed algorithm by a numerical simulation. Future work will include extending the proposed approach to nonsmooth optimization problems with quantized and delayed communication.

Appendix

A. Proof of Lemma 5

Denoting $\mathbf{p}_{i,k}^{(j)}$ as the j-th column of $P_{i,k}$,

$$\begin{aligned}&\|\nabla_k f_i(\mathbf{v}_{i,k})-P_{i,k}^{\top}\nabla f_i(\mathbf{v}_{i,k})\|^2\\&=\sum_{j=1}^{l}([\nabla_k f_i(\mathbf{v}_{i,k})]_j-\langle\nabla f_i(\mathbf{v}_{i,k}),\mathbf{p}_{i,k}^{(j)}\rangle)^2.\end{aligned}\tag{24}$$

By the descent lemma (see [24, Lemma 5.7]), we have

$$\begin{aligned}&|f_i(\mathbf{v}_{i,k}+h_k\mathbf{p}_{i,k}^{(j)})-f_i(\mathbf{v}_{i,k})-h_k\langle\nabla f_i(\mathbf{v}_{i,k}),\mathbf{p}_{i,k}^{(j)}\rangle|\\&\le\tfrac{\lambda h_k^2}{2}\|\mathbf{p}_{i,k}^{(j)}\|^2.\end{aligned}\tag{25}$$

Then divide (25) by h_k,

$$\big|[\nabla_k f_i(\mathbf{v}_{i,k})]_j-\langle\nabla f_i(\mathbf{v}_{i,k}),\mathbf{p}_{i,k}^{(j)}\rangle\big|\le\tfrac{\lambda h_k}{2}\|\mathbf{p}_{i,k}^{(j)}\|^2,\text{a.s.}\tag{26}$$

Combining (24) and (26), and Assumption 4 yields

$$\|\nabla_k f_i(\mathbf{v}_{i,k}) - P_{i,k}^\top \nabla f_i(\mathbf{v}_{i,k})\| \quad \leq \frac{\lambda h_k m}{2\sqrt{l}}, \quad \text{a.s..} \tag{27}$$

Based on (27) and the L-Lipschitz continuity of $f_i(\cdot)$,

$$\begin{aligned} &\mathbb{E}\big\{\|\nabla_k f_i(\mathbf{v}_{i,k})\|^2\big\} \\ &\leq 2\mathbb{E}\big\{\|\nabla_k f_i(\mathbf{v}_{i,k}) - P_{i,k}^\top \nabla f_i(\mathbf{v}_{i,k})\|^2\big\} \\ &\quad +2\mathbb{E}\big\{\|P_{i,k}^\top \nabla f_i(\mathbf{v}_{i,k})\|^2\big\} \\ &\leq 2(\tfrac{\lambda^2 h_k^2 m^2}{4l} + L^2). \end{aligned} \tag{28}$$

This completes the proof. □

B. Proof of Lemma 6

Based on the definition of $\mathbf{e}_k$, we have

$$\begin{aligned} \|\mathbf{e}_k\|^2 &\leq \sum_{i=1}^{N} \|\mathbf{v}_{i,k} - \alpha_k P_{i,k} \nabla_k f_i(\mathbf{v}_{i,k}) - \mathbf{v}_{i,k}\|^2 \\ &= \sum_{i=1}^{N} \tfrac{\alpha_k^2 m}{l} \|\nabla_k f_i(\mathbf{v}_{i,k})\|^2. \end{aligned} \tag{29}$$

By Lemma 5 and taking expectation on (29),

$$\mathbb{E}\big\{\|\mathbf{e}_k\|^2\big\} \leq \tfrac{2m}{l} N \alpha_k^2 (\tfrac{\lambda^2 h_k^2 m^2}{4l} + L^2) = N B_k^2 \alpha_k^2. \tag{30}$$

According to the dynamic of $\mathbf{x}_k$ in (3) yields

$$\begin{aligned} \Delta_{k+1} &= (I-J)A(k)\mathbf{x}_k + (I-J)\mathbf{e}_k \\ &= A(k)\Delta_k + (I-J)\mathbf{e}_k \\ &= \Phi(k,0)\Delta_0 + \sum_{l=0}^{k-1} \Phi(k,l+1)(I-J)\mathbf{e}_l + (I-J)\mathbf{e}_k, \end{aligned} \tag{31}$$

where the second equality follows from $JA(k) = J = A(k)J$ and the last one holds by expanding the second equality. Since $\mathbf{1}^\top \Delta_l = \mathbf{1}^\top (I-J)\mathbf{x}_l = 0$, $\forall l \in \mathbb{N}_+$,

$$\begin{aligned} \Delta_{k+1} = \big[\Phi(k,s) - J\big]\Delta_s + \sum_{l=s}^{k-1} \big[\Phi(k,l+1) - J\big] \\ \times (I-J)\mathbf{e}_l + (I-J)\mathbf{e}_k. \end{aligned} \tag{32}$$

By Lemmas 1 and 4 and taking expectation on (32),

$$\begin{aligned} \mathbb{E}\big\{\|\Delta_{k+1}\|\big\} &\leq c\lambda^k \|\Delta_0\| + c\sqrt{N} B_k \sum_{l=0}^{k-1} \lambda^{k-l-1} \alpha_l \\ &\quad + \sqrt{N} B_k \alpha_k \\ &= \mathcal{O}(\tfrac{\sqrt{N} B_k}{1-\lambda} \alpha_k). \end{aligned} \tag{33}$$

This completes the proof. □

REFERENCES

[1] P. Yi, Y. Hong, and F. Liu, "Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and application to economic dispatch of power systems," *Automatica*, vol. 74, pp. 259–269, 2016.

[2] X. Zhao, P. Yi, and L. Li, "Distributed policy evaluation via inexact ADMM in multi-agent reinforcement learning," *Control Theory Technol*, vol. 18, pp. 362–378, 2020.

[3] K. Cao, X. Li, and L. Xie, "Distributed framework matching," *IEEE Transactions on Robotics*, vol. 39, pp. 823–838, 2023.

[4] 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.

[5] J. Lei, H. Chen, and H. Fang, "Primal–dual algorithm for distributed constrained optimization," *Systems & Control Letters*, vol. 96, pp. 110–117, 2016.

[6] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push–pull gradient methods for distributed optimization in networks," *IEEE Transactions on Automatic Control*, pp. 1–16, 2021.

[7] S. Liang, L. Wang, and G. Yin, "Dual averaging push for distributed convex optimization over time-varying directed graph," *IEEE Transactions on Automatic Control*, vol. 65, no. 4, pp. 1785–1791, 2020.

[8] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 207–239, 2019.

[9] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Communication compression for distributed nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 68, no. 9, pp. 5477–5492, 2023.

[10] J. Hou, X. Zeng, G. Wang, J. Sun, and J. Chen, "Distributed momentum-based Frank-Wolfe algorithm for stochastic optimization," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 3, pp. 685–699, 2022.

[11] S. Chen, A. Garcia, and S. Shahrampour, "On distributed nonconvex optimization: Projected subgradient method for weakly convex problems in networks," *IEEE Transactions on Automatic Control*, vol. 67, no. 2, pp. 662–675, 2021.

[12] X. Yi, S. Zhang, T. Yang, and K. H. Johansson, "Zeroth-order algorithms for stochastic distributed nonconvex optimization," *Automatica*, vol. 142, p. 110353, 2022.

[13] Y. Pang and G. Hu, "Gradient-free distributed optimization with exact convergence," *Automatica*, vol. 144, p. 110474, 2022.

[14] Y. Wang, W. Zhao, Y. Hong, and M. Zamani, "Distributed subgradient-free stochastic optimization algorithm for nonsmooth convex functions over time-varying networks," *SIAM Journal on Control and Optimization*, vol. 57, no. 4, pp. 2821–2842, 2019.

[15] D. Yuan, L. Wang, A. Proutiere, and G. Shi, "Distributed zeroth-order optimization: Convergence rates that match centralized counterpart," *Automatica*, vol. 159, p. 111328, 2024.

[16] Y. Pang and G. Hu, "Randomized gradient-free distributed optimization methods for a multiagent system with unknown cost function," *IEEE Transactions on Automatic Control*, vol. 65, no. 1, pp. 333–340, 2020.

[17] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "Linear convergence of first-and zeroth-order primal-dual algorithms for distributed nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 67, no. 8, pp. 4194–4201, 2021.

[18] W. Gao and Q. Deng, "Delayed algorithms for distributed stochastic weakly convex optimization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] M. Wei, W. Yu, H. Liu, and Q. Xu, "Distributed weakly convex optimization under random time-delay interference," *IEEE Transactions on Network Science and Engineering*, vol. 11, no. 1, pp. 212–224, 2024.

[20] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

[21] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[22] D. Yuan, D. W. 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.

[23] S. Cheng, X. Yu, Y. Fan, and G. Xiao, "Zeroth-order gradient tracking for distributed constrained optimization," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 5197–5202, 2023.

[24] A. Beck, *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

[25] D. Kozak, C. Molinari, L. Rosasco, L. Tenorio, and S. Villa, "Zeroth-order optimization with orthogonal random directions," *Mathematical Programming*, vol. 199, no. 1, pp. 1179–1219, 2023.

[26] D. Yuan and D. W. 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] Y. Pang and G. Hu, "A gradient-free distributed optimization method for convex sum of nonconvex cost functions," *International Journal of Robust and Nonlinear Control*, vol. 32, no. 14, pp. 8086–8101, 2022.

[28] B. T. Polyak, *Introduction to optimization*. Optimization Software, 1987.