Linear Convergence for Distributed Optimization Without Strong Convexity

Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H. Johansson

Abstract— **This paper considers the distributed optimization problem of minimizing a global cost function formed by a sum of local smooth cost functions by using local information exchange. Various distributed optimization algorithms have been proposed for solving such a problem. A standard condition for proving the linear convergence for existing distributed algorithms is the strong convexity of the cost functions. However, the strong convexity may not hold for many practical applications, such as least squares and logistic regression. In this paper, we propose a distributed primal-dual gradient descent algorithm and establish its linear convergence under the condition that the global cost function satisfies the Polyak–Łojasiewicz condition. This condition is weaker than strong convexity and the global minimizer is not necessarily unique. The theoretical result is illustrated by numerical simulations.**

I. Introduction

Distributed optimization has a long history, which can be traced back at least to [1]–[3]. It has gained renewed interests in recent years due to its wide applications in power systems, machine learning, and sensor networks, just to name a few [4], [5].

When the cost functions are convex, various distributed optimization algorithms have been developed in discrete and continuous time. Most existing algorithms are in discrete time and are based on consensus and the distributed gradient descent method [6]–[9]. Distributed gradient descent algorithms have at most sub-linear convergence rate for diminishing stepsizes. With a fixed stepsize, the distributed gradient descent algorithms converge faster, but only to a neighborhood of an optimal point [10], [11]. Recent accelerated algorithms with fixed stepsizes use some historical information in the updates [12]–[20].

Among these distributed optimization algorithms, a standard assumption for proving linear convergence is that (local or global) cost functions are strongly convex. For example, in [14]–[17], the authors assumed that each local cost function is strongly convex. Unfortunately, in many practical applications, such as least squares and logistic regression, the cost functions are not strongly convex [21]–[23]. Therefore, the recent literature focuses on investigating alternatives to strong convexity. There are some results in centralized optimization. For instance, in [21], the authors derived the linear convergence of several centralized first-order methods for solving the smooth convex constrained optimization problem under the quadratic functional growth condition and in [22], the authors showed the linear convergence of centralized proximal-gradient methods for solving the smooth optimization problem under the assumption that the cost function satisfies the Polyak–Łojasiewicz condition.

However, to the best of our knowledge, there are only limited results in distributed optimization. In [12], the authors proposed the distributed exact first-order algorithm (EXTRA) to solve the smooth convex optimization problem and proved the linear convergence under the assumptions that the global cost function is restricted strongly convex and the optimal set is a singleton. In [24], the authors established the exponential convergence of a continuous-time distributed primal-dual gradient descent algorithm for solving the smooth convex optimization problem under the assumption that the primal-dual gradient map is metrically subregular which is weaker than strict or strong convexity. In [25], the authors proposed the zeroth-order gradient tracking algorithm and established the linear convergence under the assumption that the cost function satisfies the Polyak–Łojasiewicz condition. In [26], the authors proposed a continuous-time distributed primal-dual gradient descent algorithm to solve the smooth nonconvex optimization problem and proved the exponential convergence under the assumptions that the global cost function satisfies the restricted secant inequality condition and the set of the gradients of each local cost function at optimal points is a singleton.

In this paper, we consider the distributed optimization problem. We propose a distributed primal-dual gradient descent algorithm and establish its linear convergence under the condition that the global cost function satisfies the Polyak–Łojasiewicz condition. This condition is weaker than the (restrict) strong convexity condition assumed in [12]–[18] since it does not require convexity and the global minimizer is not necessarily unique. This condition is also weaker than the restricted secant inequality condition assumed in [26]. Moreover, this condition is different from the metric subregularity criterion assumed in [24].

The rest of this paper is organized as follows. Section II introduces some preliminaries. Section III presents problem formulation and assumptions. The main results are stated in Section IV. Simulations are given in Section V. Finally, concluding remarks are offered in Section VI.

This work was supported by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, the Swedish Research Council, and the National Natural Science Foundation of China under grants 61991403, 61991404, and 61991400.

X. Yi and K. H. Johansson are with the Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44, Stockholm, Sweden. {xinleiy, kallej}@kth.se.

S. Zhang is with the Department of Electrical Engineering, University of North Texas, Denton, TX 76203 USA. ShengjunZhang@my.unt.edu.

T. Yang and T. Chai are with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, 110819, Shenyang, China. {yangtao, tychai}@mail.neu.edu.cn.

Notations: $[n]$ denotes the set $\{1,\dots,n\}$ for any positive integer n. $\mathrm{col}(z_1,\dots,z_k)$ is the concatenated column vector of vectors $z_i \in \mathbb{R}^{p_i}$, $i \in [k]$. $\mathbf{1}_n$ ($\mathbf{0}_n$) denotes the column one (zero) vector of dimension n. $\mathbf{I}_n$ is the n-dimensional identity matrix. Given a vector $[x_1,\dots,x_n]^\top \in \mathbb{R}^n$, $\mathrm{diag}([x_1,\dots,x_n])$ is a diagonal matrix with the i-th diagonal element being x_i. The notation $A \otimes B$ denotes the Kronecker product of matrices A and B. $\mathrm{null}(A)$ is the null space of matrix A. Given two symmetric matrices M, N, $M \geq N$ means that $M - N$ is positive semi-definite. $\rho(\cdot)$ stands for the spectral radius for matrices and $\rho_2(\cdot)$ indicates the minimum positive eigenvalue for matrices having positive eigenvalues. $\|\cdot\|$ represents the Euclidean norm for vectors or the induced 2-norm for matrices. For any square matrix A, $\|x\|_A^2 = x^\top A x$. Given a differentiable function f, ∇f denotes the gradient of f.

II. Preliminaries

In this section, we present some definitions from algebraic graph theory, smooth functions, and the Polyak–Łojasiewicz condition.

A. Algebraic Graph Theory

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ denote a weighted undirected graph with the set of vertices (nodes) $\mathcal{V} = [n]$, the set of links (edges) $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, and the weighted adjacency matrix $A = A^\top = (a_{ij})$ with nonnegative elements a_{ij}. A link of $\mathcal{G}$ is denoted by $(i,j) \in \mathcal{E}$ if $a_{ij} > 0$, i.e., if vertices i and j can communicate with each other. It is assumed that $a_{ii} = 0$ for all $i \in [n]$. Let $\mathcal{N}_i = \{j \in [n] : \ a_{ij} > 0\}$ and $\deg_i = \sum\limits_{j=1}^{n} a_{ij}$ denotes the neighbor set and weighted degree of vertex i, respectively. The degree matrix of graph $\mathcal{G}$ is $\mathrm{Deg} = \mathrm{diag}([\deg_1, \cdots, \deg_n])$. The Laplacian matrix is $L = (L_{ij}) = \mathrm{Deg} - A$. A path of length k between vertices i and j is a subgraph with distinct vertices $i_0 = i, \dots, i_k = j \in [n]$ and edges $(i_j, i_{j+1}) \in \mathcal{E}$, $j = 0, \dots, k-1$. An undirected graph is connected if there exists at least one path between any two distinct vertices. If the graph $\mathcal{G}$ is connected, then its Laplacian matrix L is positive semi-definite and $\mathrm{null}(L) = \{\mathbf{1}_n\}$, see [27].

B. Smooth Function

Definition 1. *The function $f(x):\ \mathbb{R}^p \mapsto \mathbb{R}$ is smooth with constant $L_f > 0$ if it is differentiable and*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \ \forall x, y \in \mathbb{R}^p. \tag{1}$$

From Lemma 1.2.3 in [28], an immediate consequence of (1) is the following inequality:

$$|f(y) - f(x) - (y - x)^\top \nabla f(x)| \leq \frac{L_f}{2}\|y - x\|^2, \ \forall x, y \in \mathbb{R}^p. \tag{2}$$

C. Polyak–Łojasiewicz Condition

Let $f(x):\ \mathbb{R}^p \mapsto \mathbb{R}$ be a differentiable function. Let $\mathbb{X}^* = \arg\min_{x \in \mathbb{R}^p} f(x)$ and $f^* = \min_{x \in \mathbb{R}^p} f(x)$. Moreover, we assume that $f^* > -\infty$.

Definition 2. *The function f satisfies the Polyak–Łojasiewicz condition with constant $\nu > 0$ if*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \nu (f(x) - f^*), \ \forall x \in \mathbb{R}^p. \tag{3}$$

It is straightforward to see that every (essentially or weakly) strongly convex function satisfies the Polyak–Łojasiewicz condition. The Polyak–Łojasiewicz condition implies that every stationary point is a global minimizer, i.e., $\mathbb{X}^* = \{x \in \mathbb{R}^p : \nabla f(x) = \mathbf{0}_p\}$. But unlike the (essentially or weakly) strong convexity, the Polyak–Łojasiewicz condition (3) alone does not even imply the convexity of f. Moreover, it does not imply that $\mathbb{X}^*$ is a singleton either.

In some practical applications, the cost functions may be not strongly convex but satisfy the Polyak–Łojasiewicz condition. For example, the cost function in least squares problems has the form

$$f(x) = \frac{1}{2}\|Ax - b\|^2,$$

where $A \in \mathbb{R}^{m \times p}$ and $b \in \mathbb{R}^m$. Note that if A has full column rank, then $f(x)$ is strongly convex. However, if A is rank deficient, then $f(x)$ is not strongly convex, but it is convex and satisfies the Polyak–Łojasiewicz condition. Examples of nonconvex functions which satisfy the Polyak–Łojasiewicz condition can be found in [22], [29].

Although it is difficult to precisely characterize the general class of functions for which the Polyak–Łojasiewicz condition is satisfied, in [22], one important special case was given as follows:

Lemma 1. *Let $f(x) = g(Ax)$, where $g : \mathbb{R}^p \to \mathbb{R}$ is a strongly convex function and $A \in \mathbb{R}^{p \times p}$ is a matrix, then f satisfies the Polyak–Łojasiewicz condition.*

Moreover, from Theorem 2 in [22] we know that the following property holds.

Lemma 2. *Suppose that the function f satisfies the Polyak–Łojasiewicz condition* (3) *and $\mathcal{P}_{\mathbb{X}^*}(x)$, $\forall x \in \mathbb{R}^p$ is well defined, where $\mathcal{P}_{\mathbb{X}^*}(x)$ is the projection of x onto the set $\mathbb{X}^*$, i.e., $\mathcal{P}_{\mathbb{X}^*}(x) = \arg\min_{y \in \mathbb{X}^*} \|x - y\|^2$ then*

$$f(x) - f^* \geq 2\nu \|\mathcal{P}_{\mathbb{X}^*}(x) - x\|^2, \ \forall x \in \mathbb{R}^p.$$

From Theorem 1.5.5 in [30], we know that $\mathcal{P}_{\mathbb{X}^*}(\cdot)$ is well defined if $\mathbb{X}^*$ is closed and convex.

III. Problem Formulation and assumptions

Consider a network of n agents, each of which has a local cost function $f_i : \mathbb{R}^p \to \mathbb{R}$. All agents collaborate together to solve the following optimization problem

$$\min_{x \in \mathbb{R}^p} f(x) = \sum_{i=1}^{n} f_i(x). \tag{4}$$

The communication among agents is described by a weighted undirected graph $\mathcal{G}$. Let $\mathbb{X}^*$ and f^* denote the optimal set and the minimum function value of the optimization problem (4), respectively. The following assumptions are made.

Assumption 1. *The undirected graph $\mathcal{G}$ is connected.*

Assumption 2. *The optimal set $\mathbb{X}^*$ is nonempty and $f^* > -\infty$.*

Assumption 3. *Each local cost function is smooth with constant $L_f > 0$.*

Assumption 4. *The global cost function $f(x)$ satisfies the Polyak–Łojasiewicz condition with constant $\nu > 0$.*

Remark 1. *It should be highlighted that the convexity of the cost functions and the boundedness of their gradients are not assumed. Assumptions 1–3 are standard. Assumption 4 is weaker than the assumption that the global or each local cost function is strongly convex, commonly assumed in the literature, since it does not implies convexity and global optimal solution is not necessarily unique.*

IV. Distributed primal-dual gradient descent algorithm

In this section, we propose a distributed primal-dual gradient descent algorithm and analyse its convergence rate.

For simplicity, denote $\boldsymbol{x} = \mathrm{col}(x_1, \dots, x_n)$, $\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i)$, and $\boldsymbol{L} = L \otimes \mathbf{I}_p$. The optimization problem (4) is equivalent to the following constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^{np}} \quad & \tilde{f}(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i) \\ \text{s.t.} \quad & x_i = x_j, \ \forall i, j \in [n]. \end{aligned} \tag{5}$$

Noting that the Laplacian matrix L is positive semi-definite and $\mathrm{null}(L) = \{\mathbf{1}_n\}$ when $\mathcal{G}$ is connected, we know that the optimization problem (5) is equivalent to the following constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^{np}} \quad & \tilde{f}(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{L}^{1/2}\boldsymbol{x} = \mathbf{0}_{np}. \end{aligned} \tag{6}$$

Here, we use $\boldsymbol{L}^{1/2}\boldsymbol{x} = \mathbf{0}_{np}$ rather than $\boldsymbol{L}\boldsymbol{x} = \mathbf{0}_{np}$ as the constraint since it is also equivalent to $\boldsymbol{x} = \mathbf{1}_n \otimes x$ and it has a good property which will be shown later in Remark 3.

Let $\boldsymbol{u} \in \mathbb{R}^{np}$ denote the dual variable, then the augmented Lagrangian function associated with (6) is

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} + \beta \boldsymbol{u}^\top \boldsymbol{L}^{1/2}\boldsymbol{x}, \tag{7}$$

where $\alpha > 0$ and $\beta > 0$ are constants.

Based on the primal-dual gradient method, a distributed first-order algorithm to solve (6) is

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta(\alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{L}^{1/2}\boldsymbol{u}_k + \nabla \tilde{f}(\boldsymbol{x}_k)), \tag{8a}$$

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \eta\beta \boldsymbol{L}^{1/2}\boldsymbol{x}_k, \ \forall \boldsymbol{x}_0, \ \boldsymbol{u}_0 \in \mathbb{R}^{np}, \tag{8b}$$

where $\eta > 0$ is a fixed stepsize. Denote $\boldsymbol{v}_k = \mathrm{col}(v_1, \dots, v_n) = \boldsymbol{L}^{1/2}\boldsymbol{u}_k$, then the algorithm (8) can be rewritten as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta(\alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{v}_k + \nabla \tilde{f}(\boldsymbol{x}_k)), \tag{9a}$$

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \eta\beta \boldsymbol{L}\boldsymbol{x}_k, \ \forall \boldsymbol{x}_0 \in \mathbb{R}^{np}, \ \boldsymbol{v}_0 = \mathbf{0}_{np}. \tag{9b}$$

Remark 2. *Compared with the EXTRA proposed in [12]*

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{W}\boldsymbol{x}_0 - \eta \nabla \tilde{f}(\boldsymbol{x}_0), \ \forall \boldsymbol{x}_0 \in \mathbb{R}^{np}, \\ \boldsymbol{x}_{k+1} &= (\mathbf{I}_{np} + \boldsymbol{W})\boldsymbol{x}_k - \tilde{\boldsymbol{W}}\boldsymbol{x}_{k-1} - \eta(\nabla \tilde{f}(\boldsymbol{x}_k) - \nabla \tilde{f}(\boldsymbol{x}_{k-1})), \end{aligned}$$

it is straightforward to verify that the algorithm (9) *is equivalent to the EXTRA with mixing matrices $\boldsymbol{W} = \mathbf{I}_{np} - \eta\alpha \boldsymbol{L}$ and $\tilde{\boldsymbol{W}} = \mathbf{I}_{np} - \eta\alpha \boldsymbol{L} + \eta^2\beta^2 \boldsymbol{L}$.*

Note that the distributed algorithm (9) can also be written agent-wise.

$$x_{i,k+1} = x_{i,k} - \eta(\alpha \sum_{j=1}^n L_{ij} x_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k})), \tag{10a}$$

$$\begin{aligned} v_{i,k+1} &= v_{i,k} + \eta\beta \sum_{j=1}^n L_{ij} x_{j,k}, \\ & \forall x_{i,0} \in \mathbb{R}^p, \ v_{i,0} = \mathbf{0}_p, \ \forall i \in [n]. \end{aligned} \tag{10b}$$

The following theorem establishes the convergence results for the distributed primal-dual gradient descent algorithm (10).

Theorem 1. *Each agent $i \in [n]$ runs the distributed primal-dual gradient descent algorithm* (10).

(i) If Assumptions 1–4 hold, $\beta + \kappa_1 \le \alpha \le \kappa_2\beta$, $\beta \ge \max\{\kappa_3, \ \kappa_4, \ \kappa_5\}$, and $0 < \eta < \min\{\frac{1}{\epsilon_1}, \ \frac{\epsilon_2}{\epsilon_3}\}$, then $\|x_{i,k} - \bar{x}_k\|^2$, $i \in [n]$ and $f(\bar{x}_k) - f^*$ linearly converge to 0 *with a rate no less than $1 - \epsilon$, where $\bar{x}_k = \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\boldsymbol{x}_k$, $\epsilon = \frac{\eta\epsilon_4}{\epsilon_5}$,*

$$\begin{aligned} \kappa_1 &= \frac{1}{\rho_2(L)}(4 + \frac{3}{2}L_f^2), \\ \kappa_2 &> 1, \\ \kappa_3 &= \frac{\kappa_1}{\kappa_2 - 1}, \\ \kappa_4 &= \frac{1}{4}(3 + (9 + 8\kappa_2 + \frac{8}{\rho_2(L)})^{\frac{1}{2}}), \\ \kappa_5 &= 2(\kappa_2 + \frac{1}{\rho_2(L)})L_f^2 + 2((\kappa_2 + \frac{1}{\rho_2(L)})^2 L_f^4 + L_f^2)^{\frac{1}{2}}, \\ \epsilon_1 &= \max\{\beta^2\rho(L) + (2\alpha^2 + \beta^2)\rho^2(L) + \frac{5}{2}L_f^2, \ 2\beta^2 + \frac{1}{2}\}, \\ \epsilon_2 &= \frac{1}{4} - \frac{1}{2\beta}(\frac{1}{\beta} + \frac{1}{\rho_2(L)} + \frac{\alpha}{\beta})L_f^2, \\ \epsilon_3 &= \frac{1}{\beta^2}(1 + \frac{1}{\rho_2(L)} + \frac{\alpha}{\beta})L_f^2 + \frac{L_f(1 + L_f)}{2}, \\ \epsilon_4 &= \min\{1 - \eta\epsilon_1, \ \frac{\nu}{2n}\}, \\ \epsilon_5 &= \frac{\alpha + \beta}{2\beta} + \frac{1}{2\rho_2(L)}. \end{aligned}$$

(ii) Moreover, if the projection operator $\mathcal{P}_{\mathbb{X}^*}(\cdot)$ is well defined, then $\|x_{i,k} - \mathcal{P}_{\mathbb{X}^*}(\bar{x}_k)\|^2$, $i \in [n]$ linearly converges to 0 *with a rate no less than $1 - \epsilon$.*

Proof: The proof is given in Appendix B. ■

Remark 3. *If we use $\boldsymbol{L}\boldsymbol{x} = \mathbf{0}_{np}$ as the constraint in* (6), *then we could construct an alternative distributed primal-*

dual gradient descent algorithm

$$x_{i,k+1} = x_{i,k} - \eta\Big(\sum_{j=1}^{n} L_{ij}(\alpha x_{j,k} + \beta v_{j,k}) + \nabla f_i(x_{i,k})\Big), \tag{11a}$$

$$v_{i,k+1} = v_{i,k} + \eta\beta \sum_{j=1}^{n} L_{ij} x_{j,k}, \forall x_{i,0},\ v_{i,0} \in \mathbb{R}^p. \tag{11b}$$

Similar results as shown in Theorem 1 could be obtained. We omit the details due to the space limitation. Different from the requirement that $v_{i,0} = \mathbf{0}_p$ in the algorithm (10)*, $v_{i,0}$ can be arbitrarily chosen in the algorithm* (11)*. In other words, the algorithm* (11) *is robust to the initial condition $v_{i,0}$. However, the algorithm* (11) *requires additional communication of $v_{j,k}$ in* (11a)*, compared to the algorithm* (10)*.*

Remark 4. *The linear convergence for the distributed first-order algorithms proposed in [12]–[18], [26] was also established. However, in [13]–[17], it was assumed that each local cost function is strongly convex. In [12], [18], it was assumed that the global cost function is restricted strongly convex and X^* is a singleton. In [26], it was assumed that the global cost function satisfies the restricted secant inequality condition and the set of the gradients of each local cost function at optimal points is a singleton. In contrast, the linear convergence result established in Theorem 1 only requires the assumption that the global cost function satisfies the Polyak–Łojasiewicz condition, but the convexity assumption on cost functions and the singleton assumption on the optimal set and the set of the gradients of each local cost function at optimal points are not required. Moreover, it should be highlighted that when executing algorithm* (10) *the Polyak–Łojasiewicz constant ν is not needed, while this constant needs to be known in advance for executing the algorithm proposed in [25]. However, the algorithm proposed in [25] is gradient-free. It is our ongoing work to extend the algorithm* (10) *to a gradient-free algorithm.*

V. Simulations

In this section, we evaluate the performance of the proposed distributed primal-dual gradient descent algorithm (10) in solving the phase retrieval problem considered in [25]. All settings for cost functions and the communication graph are adopted from [25] for the purpose of comparison. We implement the distributed first-order algorithm (10) with $\alpha = \beta = 10$ and $\eta = 0.03$. We also implement the distributed first-order algorithm proposed in [16], which is the first-order version of the zeroth-order algorithm proposed in [25], with $\eta = 0.03$. The evolutions of $\sum_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2$ and $\nabla f(\bar{x}_k)$ for these algorithms with the same initial conditions are plotted in Fig. 1 and Fig. 2, respectively. We see that the primal-dual gradient descent algorithm (10) exhibits better performance than the algorithm proposed in [16].

VI. Conclusions

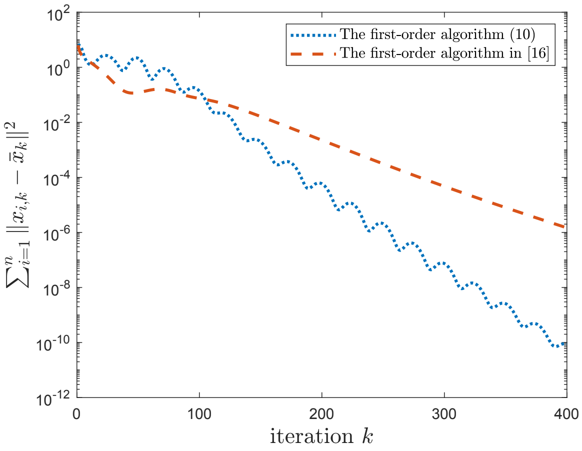

Fig. 1. Evolutions of $\sum_{i=1}^{n} ||x_{i,k} - \bar{x}_k||^2$.

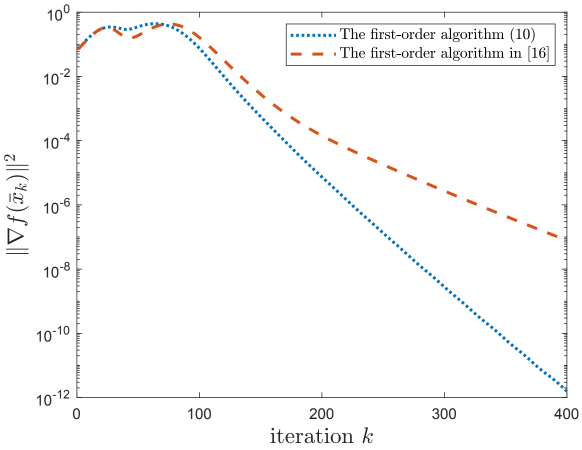

Fig. 2. Evolutions of $\|\nabla f(\bar{x}_k)\|^2$.

In this paper, we proposed a distributed primal-dual gradient descent algorithm and derived its linear convergence rate under the condition that the global cost function satisfies the Polyak–Łojasiewicz condition. This condition relaxes the standard strong convexity condition commonly assumed in the literature. Interesting directions for future work include proving the linear convergence rate for larger stepsizes, extending the first-order algorithm to the zeroth-order algorithm, considering asynchronous and dynamic network setting, and studying constraints.

Acknowledgments

The authors would like to thank Dr. Na Li and Dr. Yujie Tang for sharing their codes.

References

[1] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D. dissertation, MIT, Cambridge, MA, 1984.

[2] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, 1986.

[3] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[4] A. Nedić, “Convergence rate of distributed averaging dynamics and optimization in networks,” *Foundations and Trends in Systems and Control*, vol. 2, no. 1, pp. 1–100, 2015.

[5] 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.

[6] B. Johansson, T. Keviczky, M. Johansson, and K. H. Johansson, “Subgradient methods and consensus algorithms for solving convex optimization problems,” in *IEEE Conference on Decision and Control*, 2008, pp. 4185–4190.

[7] 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.

[8] M. Zhu and S. Martínez, “On distributed convex optimization under inequality and equality constraints,” *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 151–164, 2011.

[9] T. Yang, J. Lu, D. Wu, J. Wu, G. Shi, Z. Meng, and K. H. Johansson, “A distributed algorithm for economic dispatch over time-varying directed networks with delays,” *IEEE Transactions on Industrial Electronics*, vol. 64, no. 6, pp. 5095–5106, 2017.

[10] I. Matei and J. S. Baras, “Performance evaluation of the consensus-based distributed subgradient method under random communication topologies,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 754–771, 2011.

[11] K. Yuan, Q. Ling, and W. Yin, “On the convergence of decentralized gradient descent,” *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2015.

[12] 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.

[13] S. S. Kia, J. Cortés, and S. Martínez, “Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication,” *Automatica*, vol. 55, pp. 254–264, 2015.

[14] A. Nedić, A. Olshevsky, and W. Shi, “Achieving geometric convergence for distributed optimization over time-varying graphs,” *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[15] A. Nedić, A. Olshevsky, W. Shi, and C. A. Uribe, “Geometrically convergent distributed optimization with uncoordinated step-sizes,” in *American Control Conference*, 2017, pp. 3950–3955.

[16] G. Qu and N. Li, “Harnessing smoothness to accelerate distributed optimization,” *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1245–1260, 2018.

[17] ——, “Accelerated distributed Nesterov gradient descent,” *IEEE Transactions on Automatic Control*, vol. 65, no. 6, pp. 2566–2581, 2020.

[18] X. Yi, L. Yao, T. Yang, J. George, and K. H. Johansson, “Distributed optimization for second-order multi-agent systems with dynamic event-triggered communication,” in *IEEE Conference on Decision and Control*, 2018, pp. 3397–3402.

[19] W. Du, L. Yao, D. Wu, X. Li, G. Liu, and T. Yang, “Accelerated distributed energy management for microgrids,” in *IEEE Power and Energy Society General Meeting*, 2018.

[20] L. Yao, Y. Yuan, S. Sundaram, and T. Yang, “Distributed finite-time optimization,” in *IEEE International Conference on Control and Automation*, 2018, pp. 147–154.

[21] I. Necoara, Y. Nesterov, and F. Glineur, “Linear convergence of first order methods for non-strongly convex optimization,” *Mathematical Programming*, vol. 175, no. 1-2, pp. 69–107, 2019.

[22] H. Karimi, J. Nutini, and M. Schmidt, “Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 795–811.

[23] T. Yang, J. George, J. Qin, X. Yi, and J. Wu, “Distributed finite-time least squares solver for network linear equations,” *Automatica*, vol. 113, p. 108798, 2020.

[24] S. Liang, L. Y. Wang, and G. Yin, “Exponential convergence of distributed primal–dual convex optimization algorithm without strong convexity,” *Automatica*, vol. 105, pp. 298–306, 2019.

[25] Y. Tang and N. Li, “Distributed zero-order algorithms for nonconvex multi-agent optimization,” in *Annual Allerton Conference on Communication, Control, and Computing*, 2019, pp. 781–786.

[26] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, “Exponential convergence for distributed smooth optimization under the restricted secant inequality condition,” in *IFAC World Congress*, 2020.

[27] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press, 2010.

[28] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed. Springer International Publishing, 2018.

[29] H. Zhang and L. Cheng, “Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization,” *Optimization Letters*, vol. 9, no. 5, pp. 961–979, 2015.

[30] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag, New York, 2007.

[31] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, “Linear convergence of first- and zeroth-order algorithms for distributed nonconvex optimization under the Polyak-Łojasiewicz condition,” *arXiv preprint arXiv:1912.12110*, 2019.

Appendix

A. Useful Lemma

The following results are used in the proofs.

Lemma 3. *(Lemmas 1 and 2 in [18]) Let L be the Laplacian matrix of the connected graph $\mathcal{G}$ and $K_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Then L and K_n are positive semi-definite,* $\mathrm{null}(L) = \mathrm{null}(K_n) = \{\mathbf{1}_n\}$, $L \le \rho(L)\mathbf{I}_n$, $\rho(K_n) = 1$,

$$K_nL = LK_n = L, \tag{12}$$

$$0 \le \rho_2(L)K_n \le L \le \rho(L)K_n. \tag{13}$$

Moreover, there exists an orthogonal matrix $[r\ R] \in \mathbb{R}^{n\times n}$ with $r = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $R \in \mathbb{R}^{n\times(n-1)}$ such that

$$R\Lambda_1^{-1}R^\top L = LR\Lambda_1^{-1}R^\top = K_n, \tag{14}$$

$$\frac{1}{\rho(L)}K_n \le R\Lambda_1^{-1}R^\top \le \frac{1}{\rho_2(L)}K_n, \tag{15}$$

where $\Lambda_1 = \mathrm{diag}([\lambda_2, \dots, \lambda_n])$ with $0 < \lambda_2 \le \dots \le \lambda_n$ being the eigenvalues of the Laplacian matrix L.

B. Proof of Theorem 1

For space purposes, the detail reasoning of some steps in this proof is omitted, but can be found in [31].

(i) Denote $\boldsymbol{K} = K_n \otimes \mathbf{I}_p$, $\boldsymbol{H} = \frac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top \otimes \mathbf{I}_p)$, $\boldsymbol{Q} = R\Lambda_1^{-1}R^\top \otimes \mathbf{I}_p$, $\bar{\boldsymbol{x}}_k = \mathbf{1}_n \otimes \bar{x}_k$, $\bar{v}_k = \frac{1}{n}(\mathbf{1}_n^\top \otimes \mathbf{I}_p)\boldsymbol{v}_k$, $\boldsymbol{g}_k = \nabla\tilde{f}(\boldsymbol{x}_k)$, $\bar{\boldsymbol{g}}_k = \boldsymbol{H}\boldsymbol{g}_k$, $\boldsymbol{g}_k^0 = \nabla\tilde{f}(\bar{\boldsymbol{x}}_k)$, and $\bar{\boldsymbol{g}}_k^0 = \boldsymbol{H}\boldsymbol{g}_k^0 = \frac{1}{n}(\mathbf{1}_n \otimes \nabla f(\bar{x}_k))$.

From (9b), we know that

$$\bar{v}_{k+1} = \bar{v}_k. \tag{16}$$

Then, from (16), $\sum_{i=1}^n v_{i,0} = \mathbf{0}_p$, and (9a), we know that $\bar{v}_k = \mathbf{0}_p$ and

$$\bar{\boldsymbol{x}}_{k+1} = \bar{\boldsymbol{x}}_k - \eta\bar{\boldsymbol{g}}_k. \tag{17}$$

Noting that $\nabla\tilde{f}$ is Lipschitz-continuous with constant $L_f > 0$ as assumed in Assumption 3 and $\rho(\boldsymbol{H}) = 1$, we have that

$$\|\boldsymbol{g}_{k+1}^0 - \boldsymbol{g}_k^0\|^2 \le L_f^2\|\bar{\boldsymbol{x}}_{k+1} - \bar{\boldsymbol{x}}_k\|^2 = \eta^2L_f^2\|\bar{\boldsymbol{g}}_k\|^2, \tag{18}$$

$$\|\boldsymbol{g}_k^0 - \boldsymbol{g}_k\|^2 \le L_f^2\|\bar{\boldsymbol{x}}_k - \boldsymbol{x}_k\|^2 = L_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2, \tag{19}$$

$$\begin{aligned}&\|\bar{\boldsymbol{g}}_k^0 - \bar{\boldsymbol{g}}_k\|^2 = \|\boldsymbol{H}(\boldsymbol{g}_k^0 - \boldsymbol{g}_k)\|^2\\ &\le \|\boldsymbol{g}_k^0 - \boldsymbol{g}_k\|^2 \le L_f^2\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2.\end{aligned} \tag{20}$$

From Assumption 4 and (3), we have that

$$\|\bar{\boldsymbol{g}}_k^0\|^2 = \frac{1}{n}\|\nabla f(\bar{x}_k)\|^2 \geq \frac{2\nu}{n}(f(\bar{x}_k) - f^*). \tag{21}$$

Denote $V_{1,k} = \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}}$, $V_{2,k} = \frac{1}{2}\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}}$, $V_{3,k} = \boldsymbol{x}_k^\top\boldsymbol{K}(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0)$, $V_{4,k} = f(\bar{x}_k) - f^* = \tilde{f}(\bar{\boldsymbol{x}}_k) - f^*$, and $V_k = \sum_{i=1}^4 V_{i,k}$. Then, from (17)–(21), we have

$$\begin{aligned}
&V_{1,k+1}\\
&\leq \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} - \|\boldsymbol{x}_k\|^2_{\eta\alpha\boldsymbol{L}-\frac{\eta}{2}\boldsymbol{K}-\frac{3\eta^2\alpha^2}{2}\boldsymbol{L}^2-\frac{\eta}{2}(1+3\eta)L_f^2\boldsymbol{K}}\\
&\quad - \eta\beta\boldsymbol{x}_k^\top\boldsymbol{K}(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0) + \|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\|^2_{\frac{3\eta^2\beta^2}{2}\boldsymbol{K}}, \qquad (22)\\
&V_{2,k+1}\\
&\leq \frac{1}{2}\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}} + \eta\boldsymbol{x}_k^\top(\beta\boldsymbol{K} + \alpha\boldsymbol{L})(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0)\\
&\quad + \|\boldsymbol{x}_k\|^2_{\eta^2\beta(\beta\boldsymbol{L}+\alpha\boldsymbol{L}^2)} + \|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\|^2_{\frac{\eta}{2\beta}(\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K})}\\
&\quad + \eta(\frac{\eta}{\beta^2} + \frac{1}{2\beta})(\frac{1}{\rho_2(L)} + \frac{\alpha}{\beta})L_f^2\|\bar{\boldsymbol{g}}_k\|^2, \qquad (23)\\
&V_{3,k+1}\\
&\leq \boldsymbol{x}_k^\top\boldsymbol{K}(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0) - \eta\alpha\boldsymbol{x}_k^\top\boldsymbol{L}(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0)\\
&\quad + \|\boldsymbol{x}_k\|^2_{\eta(\beta\boldsymbol{L}+\frac{1}{2}\boldsymbol{K})+\eta^2(\frac{\alpha^2}{2}-\alpha\beta+\beta^2)\boldsymbol{L}^2+\frac{\eta}{2}(1+2\eta)L_f^2\boldsymbol{K}}\\
&\quad + \eta(\frac{1}{2\beta^2} + \frac{\eta}{\beta^2} + \frac{\eta}{2})L_f^2\|\bar{\boldsymbol{g}}_k\|^2\\
&\quad - \|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\|^2_{\eta(\beta-\frac{1}{2}-\frac{\eta}{2}-\frac{\eta\beta^2}{2})\boldsymbol{K}}, \qquad (24)\\
&V_{4,k+1}\\
&\leq f(\bar{x}_k) - f^* - \frac{\eta}{4}(1 - 2\eta L_f)\|\bar{\boldsymbol{g}}_k\|^2 + \|\boldsymbol{x}_k\|^2_{\frac{\eta}{2}L_f^2\boldsymbol{K}}\\
&\quad - \frac{\eta\nu}{2n}(f(\bar{x}_k) - f^*). \qquad (25)
\end{aligned}$$

From (22)–(25), we have

$$\begin{aligned}
&V_{k+1}\\
&\leq V_k - \|\boldsymbol{x}_k\|^2_{\eta\boldsymbol{M}_1-\eta^2\boldsymbol{M}_2} - \|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\|^2_{\eta\boldsymbol{M}_3-\eta^2\boldsymbol{M}_4}\\
&\quad - (\eta\epsilon_2 - \eta^2\epsilon_3)\|\bar{\boldsymbol{g}}_k\|^2 - \frac{\eta\nu}{2n}(f(\bar{x}_k) - f^*), \qquad (26)
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{M}_1 =&(\alpha - \beta)\boldsymbol{L} - \frac{1}{2}(2 + 3L_f^2)\boldsymbol{K},\\
\boldsymbol{M}_2 =&\beta^2\boldsymbol{L} + (2\alpha^2 + \beta^2)\boldsymbol{L}^2 + \frac{5}{2}L_f^2\boldsymbol{K},\\
\boldsymbol{M}_3 =&(\beta - \frac{1}{2} - \frac{\alpha}{2\beta^2})\boldsymbol{K} - \frac{1}{2\beta}\boldsymbol{Q},\\
\boldsymbol{M}_4 =&(2\beta^2 + \frac{1}{2})\boldsymbol{K}.
\end{aligned}$$

From $\beta + \kappa_1 \leq \alpha$ and $\kappa_1 = \frac{1}{\rho_2(L)}(4 + \frac{3}{2}L_f^2)$, we have

$$(\alpha - \beta)\rho_2(L) - \frac{1}{2}(2 + 3L_f^2) \geq 1. \tag{27}$$

From $\beta \geq \kappa_4$, we have

$$(\beta - \frac{1}{2} - \frac{\kappa_2}{2\beta}) - \frac{1}{2\beta\rho_2(L)} \geq 1. \tag{28}$$

From $\alpha \leq \kappa_2\beta$ and $\beta \geq \kappa_5$, we have

$$\begin{aligned}
\epsilon_2 =&\frac{1}{4} - \frac{1}{2\beta}(\frac{1}{\beta} + \frac{1}{\rho_2(L)} + \frac{\alpha}{\beta})L_f^2\\
\geq&\frac{1}{4} - \frac{1}{2\beta}(\frac{1}{\beta} + \frac{1}{\rho_2(L)} + \kappa_2)L_f^2 \geq \frac{1}{8}. \qquad (29)
\end{aligned}$$

From (29), and $0 < \eta < \frac{\epsilon_2}{\epsilon_3}$, we have

$$\eta\epsilon_2 - \eta^2\epsilon_3 > 0. \tag{30}$$

From (13), (15), $\alpha \leq \kappa_2\beta$, (27), and (28), we have

$$\begin{aligned}
\boldsymbol{M}_1 =&(\alpha - \beta)\boldsymbol{L} - \frac{1}{2}(2 + 3L_f^2)\boldsymbol{K}\\
\geq&(\alpha - \beta)\rho_2(L)\boldsymbol{K} - \frac{1}{2}(2 + 3L_f^2)\boldsymbol{K} \geq \boldsymbol{K}, \qquad (31)\\
\boldsymbol{M}_2 =&\beta^2\boldsymbol{L} + (2\alpha^2 + \beta^2)\boldsymbol{L}^2 + \frac{5}{2}L_f^2\boldsymbol{K} \leq \epsilon_1\boldsymbol{K}, \qquad (32)\\
\boldsymbol{M}_3 =&(\beta - \frac{1}{2} - \frac{\alpha}{2\beta^2})\boldsymbol{K} - \frac{1}{2\beta}\boldsymbol{Q}\\
\geq&(\beta - \frac{1}{2} - \frac{\kappa_2}{2\beta})\boldsymbol{K} - \frac{1}{2\beta\rho_2(L)}\boldsymbol{K} \geq \boldsymbol{K}, \qquad (33)\\
\boldsymbol{M}_4 =&(2\beta^2 + \frac{1}{2})\boldsymbol{K} \leq \epsilon_1\boldsymbol{K}. \qquad (34)
\end{aligned}$$

Denote $\hat{V}_k = \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\|^2_{\boldsymbol{K}} + f(\bar{x}_k) - f^*$. Then, from (26) and (30)–(34), we have

$$V_{k+1} \leq V_k - \eta\epsilon_4\hat{V}_k. \tag{35}$$

From the Cauchy-Schwarz inequality, we have

$$\epsilon_6\hat{V}_k \leq V_k \leq \epsilon_5\hat{V}_k, \tag{36}$$

where $\epsilon_6 = \min\{\frac{1}{2\rho(L)},\ \frac{\alpha-\beta}{2\alpha}\}$.

From (35), (36), and $0 < \eta < \frac{1}{\epsilon_1}$, we have

$$\begin{aligned}
V_{k+1} \leq&V_k - \frac{\eta\epsilon_4}{\epsilon_5}V_k = (1 - \epsilon)V_k\\
\leq&(1 - \epsilon)^{k+1}V_0. \qquad (37)
\end{aligned}$$

Hence, from (36) and (37), for all $i \in [n]$, we have

$$\begin{aligned}
&\|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^* \leq \|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + f(\bar{x}_k) - f^*\\
&\leq \hat{V}_k \leq \frac{1}{\epsilon_6}V_k \leq \frac{1}{\epsilon_6}(1 - \epsilon)^k V_0. \qquad (38)
\end{aligned}$$

In other words, $\|x_{i,k} - \bar{x}_k\|^2$, $i \in [n]$ and $f(\bar{x}_k) - f^*$ linearly converge to 0 with a rate no less than $1 - \epsilon$.

(ii) If the projection operator $\mathcal{P}_{\mathbb{X}^*}(\cdot)$ is well defined, then from Lemma 2 and (38), we know that $\|x_{i,k} - \mathcal{P}_{\mathbb{X}^*}(\bar{x}_k)\|^2$, $i \in [n]$ linearly converge to 0 with a rate no less than $1 - \epsilon$.