2024 Tenth Indian Control Conference (ICC)
9 - 11 December, 2024
IISER Bhopal, India

On Linear Convergence of PI Consensus Algorithm under the Restricted Secant Inequality

Kushal Chakrabarti Mayank Baranwal

***Abstract*— This paper considers solving distributed optimization problems in peer-to-peer multi-agent networks. The network is synchronous and connected. By using the proportional-integral (PI) control strategy, various algorithms with fixed stepsize have been developed. Two notable among them are the PI algorithm and the PI consensus algorithm. Although the PI algorithm has provable linear or exponential convergence without the standard requirement of (strong) convexity, a similar guarantee for the PI consensus algorithm is unavailable. In this paper, using Lyapunov theory, we guarantee exponential convergence of the PI consensus algorithm for global cost functions that satisfy the restricted secant inequality, with rate-matching discretization, without requiring convexity. To accelerate the PI consensus algorithm, we incorporate local preconditioning in the form of constant positive definite matrices and numerically validate its efficiency compared to the prominent distributed convex optimization algorithms. Unlike classical pre-conditioning, where only the gradients are multiplied by a pre-conditioner, the proposed pre-conditioning modifies both the gradients and the consensus terms, thereby controlling the effect of the communication graph on the algorithm.**

***Index Terms*— Agents-based systems, Distributed optimization algorithms, Lyapunov methods**

I. Introduction

We consider solving the multi-agent distributed convex optimization problem over a peer-to-peer network of m agents. Each agent $i \in \{1, 2, \ldots, m\}$ in the network can communicate with a certain set of other agents called its *neighbors*, denoted by $\mathcal{N}_i$. The inter-agent communication topology is represented by an undirected *graph* $\mathbb{G} = (\{1, \ldots, m\}, \mathcal{E})$, with an *edge* $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$ if agent i and agent j are *neighbors*, for any $i, j \in \{1, \ldots, m\}$, $i \neq j$. Each agent has a *local and private cost function* $f_i : \mathbb{R}^d \to \mathbb{R}$. The agents aim to compute a common vector $\mathbf{x}_* \in \mathbb{R}^d$ that minimizes the aggregate cost function held by all the agents:

$$\mathbf{x}_* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{m} f_i(\mathbf{x}). \tag{1}$$

We let $f : \mathbb{R}^d \to \mathbb{R}$ and $F : \mathbb{R}^{md} \to \mathbb{R}$, respectively, denote the *aggregate cost* and the *cumulative cost*, i.e., $f(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x})$ and $F(\mathbf{x}) = \sum_{i=1}^{m} f_i(\mathbf{x}_i)$ for $\mathbf{x} \in \mathbb{R}^{md}$.

A class of distributed gradient-based continuous-time algorithms that solve (1) is based on proportional-integral (PI) control law [1]. Here, an integral error term [2], [3] or a consensus term of integral errors [4] is fed back to the state dynamics to guarantee consensus and convergence [5]. We focus on PI consensus algorithm, because we noticed that the PI consensus algorithm has convergence speed comparable or favorable to competing algorithms, e.g., EXTRA, DIGing, PI, that also have provable exponential rate. However, PI consensus can be slower than the more recent double-loop Accelerated-EXTRA [6] algorithm (see Figure 2). We find that the PI consensus algorithm can be accelerated by a suitable preconditioning strategy. Details of the preconditioning matrix is presented in Section IV.

Existing works [2]–[5] used Lyapunov theory to prove asymptotic stability of PI-based algorithms for convex local costs. However, achieving accelerated convergence to a solution of (1) requires a stronger form than just asymptotic stability. Exponential stability of the continuous-time PI algorithm has been proved in [2] for directed graphs when the local costs are *strongly convex* and in [3] when the cumulative cost F satisfy the *restricted secant inequality* (RSI) condition. RSI is one of the strong-convexity relaxation classes for guaranteeing linear convergence of centralized algorithms [7]. Quasi-strongly convex or composition of strongly convex function with linear map plus a linear term satisfy RSI. Despite utilizing PI control, the PI algorithm [2], [3] and the PI consensus [4], [5] algorithm are different in their estimate updates and the initialization of the integral error states [1]. To highlight this, we briefly present their update strategies below. Each agent i in PI consensus updates its state $\mathbf{x}_i$ and the integral consensus $\mathbf{v}_i$ as [4]

$$\dot{\mathbf{x}}_i = \sum_{j\in\mathcal{N}_i}(\mathbf{x}_j - \mathbf{x}_i) - \beta\sum_{j\in\mathcal{N}_i}(\mathbf{v}_j - \mathbf{v}_i) - \alpha\nabla f_i(\mathbf{x}_i), \tag{2}$$

$$\dot{\mathbf{v}}_i = \beta \sum_{j\in\mathcal{N}_i}(\mathbf{x}_j - \mathbf{x}_i), \tag{3}$$

where $\alpha, \beta > 0$ are algorithm parameters. Whereas the PI algorithm is described by [2]

$$\dot{\mathbf{x}}_i = \sum_{j\in\mathcal{N}_i}(\mathbf{x}_j - \mathbf{x}_i) - \beta\mathbf{v}_i - \alpha\nabla f_i(\mathbf{x}_i), \tag{4}$$

$$\dot{\mathbf{v}}_i = -\beta \sum_{j\in\mathcal{N}_i}(\mathbf{x}_j - \mathbf{x}_i). \tag{5}$$

Specifically, the PI consensus algorithm (2)-(3) can be interpreted as a primal-dual algorithm for solving

$$\min_{\{\mathbf{x}_i\in\mathbb{R}^d, \forall i\}} \sum_{i=1}^{m} f_i(\mathbf{x}_i),\ s.t.\ (L \otimes I)\mathbf{x} = \mathbf{0_{md}},$$

K. Chakrabarti and M. Baranwal are with the Division of Data & Decision Sciences, Tata Consultancy Services Research, Mumbai, 400607 India. e-mails: {chakrabarti.k, baranwal.mayank}@tcs.com. M. Baranwal is also with the Faculty of Systems & Control Engineering, Indian Institute of Technology, Bombay.

and the PI algorithm (4)-(5) is a primal-dual algorithm for

$$\min_{\{\mathbf{x}_i\in\mathbb{R}^d,\forall i\}} \sum_{i=1}^{m} f_i(\mathbf{x}_i),\ s.t.\ (L\otimes I)^{1/2}\mathbf{x}=\mathbf{0_{md}},$$

where I denotes the $(d\times d)$-dimensional identity matrix, L denotes the Laplacian matrix of the graph $\mathbb{G}$, $\otimes$ denotes the Kronecker product, and $\mathbf{x}=[\mathbf{x}_1^\top\dots\mathbf{x}_m^\top]^\top$. So, each agent in PI consensus (2)-(3) shares both $\mathbf{x}_i$ and $\mathbf{v}_i$ with its neighbors, which allows the agents to incorporate the integral consensus term $\sum_{j\in\mathcal{N}_i}(\mathbf{v}_j-\mathbf{v}_i)$ in its $\mathbf{x}_i$-update. On the other hand, each agent in PI (4)-(5) shares only $\mathbf{x}_i$ with its neighbors, and does not use $\mathbf{v}_j$ to update its $\mathbf{x}_i$. Moreover, the convergence of the PI algorithm (4)-(5) requires each $\mathbf{v}_i(0)$ to be initialized as the zero vector, whereas the PI consensus (2)-(3) allows arbitrary initialization of each $\mathbf{v}_i(0)$. So, PI consensus (2)-(3) is robust to the initialization of $\mathbf{v_i(0)}$ [3], with both algorithms having communication complexity $\mathcal{O}(\boldsymbol{d})$. Exponential stability of the continuous-time PI consensus (2)-(3) has been proved in [8] for metrically subregular primal-dual gradient maps and convex local costs and in [9] for strong convex f.

Hence, while exponential stability of the PI algorithm (4)-(5) has been proved in [3] when F satisfies the RSI condition, exponential stability of the PI consensus algorithm (2)-(3) for the same class of cost functions has not been proved in the literature. We prove exponential stability of the continuous-time PI consensus algorithm for undirected graphs when F satisfies the RSI condition with Lipschitz continuous local gradients without requiring convexity of the cost function. Note that an even stronger notion of convergence is fixed-time convergence [10] in continuous-time.

Despite key insights from studying optimization algorithms as continuous-time dynamics, there is substantial literature on discrete-time distributed algorithms. In the seminal distributed gradient-descent (DGD) algorithm, each agent combines its local gradient and consensus terms to update the local estimate of $\mathbf{x}_*$ [11]. Notable discrete-time algorithms that are built upon DGD include ADMM [12], EXTRA [13], DIGing [14], PI [2], APM-C [15], Mudag [16], Accelerated EXTRA [6], DAccGD [17], and ACC-SONATA [18]. For convex local costs, asymptotic convergence of DGD, ADMM, EXTRA, DIGing, PI, and Accelerated EXTRA is proved, assuming Lipschitz continuous or bounded local gradients. Linear convergence of these algorithms requires stronger assumptions, such as restricted strong convexity of the aggregate cost for EXTRA [13], *restricted secant inequality* for the PI algorithm [3], strong convexity of each local cost [12], [14], [19]. The implementation itself of APM-C, Mudag, DAccGD, and ACC-SONATA requires the value of the strong-convexity coefficient of the cumulative or the aggregate cost function. For the first time, we prove linear convergence of the discrete-time PI consensus algorithm for undirected graphs when F satisfies the RSI condition with Lipschitz continuous local gradients without requiring convexity of individual local costs. Thus, compared to the analysis of PI algorithm (4)-(5) in [3], we have a similar contribution for the PI consensus algorithm (2)-(3).

The key contributions of our work are summarized below.

- Although the PI algorithm (4)-(5) has provable linear or exponential convergence when the cost function satisfies the *restricted secant inequality*, without the standard requirement of (strong) convexity, the distributed optimization literature lacks a similar guarantee for the PI consensus algorithm (2)-(3) when solving (1) with the same class of cost functions. We aim to address this problem. For the first time, we rigorously prove exponential convergence of the continuous-time and discrete-time PI consensus algorithm when F satisfies the *restricted secant inequality* condition with Lipschitz continuous local gradients without requiring convexity. Our analyses also apply to the case of local pre-conditioning with constant positive-definite matrices. The details are in Section III.
- In Section IV, we propose a choice of pre-conditioning and numerically show its efficacy. Pre-conditioning potentially increases the effective graph connectivity by modifying the edge weights, leading to faster consensus.

Due to their difference in local estimate update, as described in (2)-(5), analysis of the PI algorithm (4)-(5) in [3] does not trivially apply to PI consensus (2)-(3). Thus, the novelty of our paper lies in proving exponential or linear convergence of PI consensus (2)-(3), under relaxed assumption than the existing analyses of (2)-(3). Furthermore, we propose a local pre-conditioning that can potentially reduce the convergence time of the existing PI consensus method.

II. Assumptions and Preliminaries

Notation: Let N be any natural number. We use ∇g to denote the gradient of a function $g:\mathbb{R}^N\to\mathbb{R}$. We let $\|\mathbf{v}\|$ denote the Euclidean norm of $\mathbf{v}\in\mathbb{R}^N$ and $\|M\|$ denote the induced 2-norm of a matrix M. We let $\mathbf{0_N}$ denote the N-dimensional zero vector. We denote a block-diagonal matrix of appropriate dimensions by $Diag(.)$. We use the abbreviation SPD for symmetric positive definite. We let $\overline{\lambda}_L$ and $\underline{\lambda}_L$ denote the largest and the smallest non-zero eigenvalue of the Laplacian L. We let $\mathbf{X}_*=[(\mathbf{x}_*)^\top,\dots,(\mathbf{x}_*)^\top]^\top$. Finally, we let $\tilde{\mathcal{L}}=L\otimes I$.

Definition 1. *A differentiable function $g:\mathbb{R}^N\to\mathbb{R}$ satisfies the restricted secant inequality (RSI) with respect to $\mathbf{x}_*$ with constant $\mu>0$ if $(\nabla g(\mathbf{x})-\nabla g(\mathbf{x}_*))^\top(\mathbf{x}-\mathbf{x}_*)\geq\mu\|\mathbf{x}-\mathbf{x}_*\|^2\ \forall\mathbf{x}\in\mathbb{R}^N$, where $\mathbf{x}_*$ is the unique global minimizer of g [3], [13].*

Definition 2. *A sequence $\{\mathbf{x}(k)\}\subset\mathbb{R}^N$ converges linearly to $\mathbf{x}_*\in\mathbb{R}^N$ with rate $\rho\in(0,1)$ if there exist constants $C,K>0$ such that $\|\mathbf{x}(k)-\mathbf{x}_*\|\leq C\rho^{k-K},\forall k\geq K$.*

Assumption 1. *$|\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})|<\infty$ and the solution set of problem (1) is non-empty.*

Assumption 2. *Each f_i is continuously differentiable.*

Assumption 3. *$\mathbb{G}$ is undirected and connected.*

Assumption 4. *F satisfies the restricted secant inequality with respect to $\mathbf{x}_*$ with a constant $\mu>0$, and its gradient ∇F is L_f-Lipcshitz continuous (F is smooth).*

Assumption 4 is weaker than the existing works on the analysis of PI consensus algorithm (2)-(3), such as [5], [20], in the sense that it does not require convexity of the cost function. It further implies that every stationary point is a global minimizer, i.e., the solution set is $\{\mathbf{x}_* \in \mathbb{R}^d | \sum_{i=1}^m \nabla f_i(\mathbf{x}_*) = 0_d\}$.

The following result is standard in Lyapunov stability theory of autonomous systems [21].

Lemma 1. *Consider the system* $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x})$ *where* $\mathbf{x} \in \mathbb{R}^N, \mathbf{g} : \mathbb{R}^N \to \mathbb{R}^N, \mathbf{g}(\mathbf{0_N}) = \mathbf{0_N}$, *and* $\mathbf{g}$ *is Lipschitz continuous over* $\mathbb{R}^N$. *Let* $V : \mathbb{R}^N \to \mathbb{R}$ *be a continuously differentiable positive definite function such that* $k_1 \|\mathbf{x}\|^a \leq V(\mathbf{x}) \leq k_2 \|\mathbf{x}\|^a$ *and* $\dot{V}(\mathbf{x}) \leq -k_3 \|\mathbf{x}\|^a$ *for any* $\mathbf{x} \in \mathbb{R}^N$, *where* k_1, k_2, k_3, a *are positive constants. Then, the origin is globally exponentially stable, i.e.,* $\|\mathbf{x}(t)\| \leq (\frac{k_2}{k_1})^{1/a} e^{-\frac{k_3}{k_2 a}t} \|\mathbf{x}(0)\|, \forall \mathbf{x}(0) \in \mathbb{R}^N$.

III. Convergence of PI Consensus Algorithm

For simplicity in presenting our results, we assume that the solution of (1), defined in Remark 1, is unique, denoted by $\mathbf{x}_*$. Later, we discuss the applicability of our results in the case when the solution is not unique.

The algorithm presented next is built upon the distributed PI consensus algorithm [4], [5] for solving (1). However, the difference is that each agent $i \in \{1, 2, \ldots, m\}$ in this algorithm multiplies the local gradient and the consensus terms with a fixed pre-conditioner matrix K_i. Each K_i is SPD and chosen by the agents independently before the algorithm begins. Later, in the next section, we propose a choice of pre-conditioning and numerically show its efficacy in improving the convergence time.

At each $t \geq 0$, each agent $i \in \{1, 2, \ldots, m\}$ in the (pre-conditioned) PI consensus algorithm updates its state $\mathbf{x}_i$ and the integral consensus $\mathbf{v}_i$ according to

$$\dot{\mathbf{x}}_i = K_i \left(\sum_{j \in \mathcal{N}_i} (\mathbf{x}_j - \mathbf{x}_i) - \beta \sum_{j \in \mathcal{N}_i} (\mathbf{v}_j - \mathbf{v}_i) - \alpha \nabla f_i(\mathbf{x}_i) \right), \quad (6)$$

$$\dot{\mathbf{v}}_i = \beta K_i \sum_{j \in \mathcal{N}_i} (\mathbf{x}_j - \mathbf{x}_i). \quad (7)$$

Here, α and β are positive scalar parameters of the algorithm, whose values are presented later in this section. For convenience, we combine the agents' dynamics in a matrix form as follows. We let $\mathbf{x} = [\mathbf{x}_1^\top \ldots \mathbf{x}_m^\top]^\top$ and $\mathbf{v} = [\mathbf{v}_1^\top \ldots \mathbf{v}_m^\top]^\top$ respectively denote the agents' combined state vector and the integral consensus at each time $t \geq 0$. We define the combined pre-conditioner $K = Diag\left(\{K^i\}_{i=1}^m\right)$. Note that K is SPD, as each K_i is SPD. Then, (6)-(7) can be rewritten:

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{v}} \end{bmatrix} = - \begin{bmatrix} K\tilde{\mathcal{L}}\mathbf{x} - \beta K\tilde{\mathcal{L}}\mathbf{v} + \alpha K \nabla F(\mathbf{x}) \\ \beta K \tilde{\mathcal{L}} \mathbf{x} \end{bmatrix}. \quad (8)$$

Recall the definition of $\mathbf{X}_*$ from Section II. Since K is SPD, it can be verified that, under Assumptions 1-4, if $(\mathbf{X}_*, \mathbf{v}_*)$ is an equilibrium point of (8), then $\mathbf{x}_*$ solves (1).

Upon utilizing LaSalle's invariance principle [21], under Assumptions 1-4, it can be proved that each agent's estimate in Algorithm (6)-(7) with $\mathbf{x}(0), \mathbf{v}(0) \in \mathbb{R}^{md}$ asymptotically converges to the same solution $\mathbf{x}_*$ of (1). It does not require the convexity of individual local costs. The proof for asymptotic stability uses the Lyapunov function $V : \mathbb{R}^{2md} \to \mathbb{R}$ such that $V(\mathbf{z}, \mathbf{y}) = \frac{1}{2}\left(\mathbf{z}^\top K^{-1}\mathbf{z} + \mathbf{y}^\top K^{-1}\mathbf{y}\right)$ for $\mathbf{z}, \mathbf{y} \in \mathbb{R}^{md}$ and follows the proof of Theorem 1 in [5]. So, LaSalle's principle guarantees asymptotic stability of (8), but with an unknown rate. The following analysis implies that Lyapunov theory (Lemma 1) can guarantee stronger stability for the class of F that satisfies Assumption 4.

Theorem 1. *Consider algorithm* (8) *with initial condition* $\mathbf{x}(0), \mathbf{v}(0) \in \mathbb{R}^{md}$. *Suppose that Assumptions 1-4 hold true. Then there exists* $\overline{\alpha}, \overline{\beta}, \underline{\beta} \in (0, \infty)$ *such that, for* $\alpha < \overline{\alpha}$ *and* $\beta \in (\overline{\beta}, \underline{\beta})$, *the local estimate* $\mathbf{x}_i$ *exponentially converges to the same* $\mathbf{x}_*$ *for each agent* $i \in \{1, 2, \ldots, m\}$.

Proof. We define the estimation errors at time $t \geq 0$ as $\mathbf{z} = \mathbf{x} - \mathbf{X}_*$, $\mathbf{y} = \mathbf{v} - \mathbf{v}_*$. Then, from (8),

$$\begin{bmatrix} \dot{\mathbf{z}} \\ \dot{\mathbf{y}} \end{bmatrix} = - \begin{bmatrix} K(\tilde{\mathcal{L}}\mathbf{z} - \beta\tilde{\mathcal{L}}\mathbf{y} + \alpha \nabla F(\mathbf{x}) - \alpha \nabla F(\mathbf{X}_*)) \\ \beta K \tilde{\mathcal{L}} \mathbf{z} \end{bmatrix}. \quad (9)$$

Consider the case $\tilde{\mathcal{L}}\mathbf{y} \neq \mathbf{0_{md}}$. We define Lyapunov candidate $V : \mathbb{R}^{2md} \to \mathbb{R}$ such that

$$V(\mathbf{z}, \mathbf{y}) = \frac{c_1}{2}\mathbf{z}^\top K^{-1}\mathbf{z} + \frac{c_2}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - c_3 \mathbf{z}^\top K^{-1}\mathbf{y},$$

where $c_1, c_2, c_3 > 0$. Along the trajectories of (9), $\dot{V} = c_1\mathbf{z}^\top K^{-1}\dot{\mathbf{z}} + c_2\mathbf{y}^\top K^{-1}\dot{\mathbf{y}} - c_3\mathbf{z}^\top K^{-1}\dot{\mathbf{y}} - c_3\mathbf{y}^\top K^{-1}\dot{\mathbf{z}}$. Upon substituting from (9), $\dot{V} = -(c_1 - \beta c_3)\mathbf{z}^\top\tilde{\mathcal{L}}\mathbf{z} - \alpha c_1 \mathbf{z}^\top(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*)) + (\beta c_1 - \beta c_2 + c_3)\mathbf{y}^\top\tilde{\mathcal{L}}\mathbf{z} - c_3(\beta\mathbf{y}^\top\tilde{\mathcal{L}}\mathbf{y} - \alpha\mathbf{y}^\top(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*)))$. Suppose, $(c_1 - \beta c_3) > 0$. We use the following result from Lemma 2 of [3]. Under Assumptions 1-4, there exist $\overline{\alpha} > 0, \mu_1 = \min\{\frac{\mu}{2m}\alpha c_1, (c_1 - \beta c_3)\underline{\lambda}_L - \frac{2mL_f^2 + \mu\alpha c_1 L_f}{\mu\alpha c_1}\}$ s.t. if $\alpha < \overline{\alpha}$,

$$\alpha c_1(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*))^\top(\mathbf{x} - \mathbf{X}_*) + (c_1 - \beta c_3)\mathbf{x}^\top\tilde{\mathcal{L}}\mathbf{x} \geq \mu_1 \|\mathbf{x} - \mathbf{X}_*\|^2, \forall \mathbf{x} \in \mathbb{R}^{md}. \quad (10)$$

Following the proof of Theorem 1 in [3], from above, $\dot{V} \leq -\mu_1 \|\mathbf{z}\|^2 + (\beta c_1 - \beta c_2 + c_3)\mathbf{y}^\top\tilde{\mathcal{L}}\mathbf{z} - c_3(\beta\mathbf{y}^\top\tilde{\mathcal{L}}\mathbf{y} - \alpha\mathbf{y}^\top(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*)))$. If $(\beta c_1 - \beta c_2 + c_3) = 0$,

$$\begin{aligned} \dot{V} &\leq -\mu_1 \|\mathbf{z}\|^2 - \beta c_3 \underline{\lambda}_L \|\mathbf{y}\|^2 + \alpha c_3 \mathbf{y}^\top(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*)) \\ &\leq -\mu_1 \|\mathbf{z}\|^2 - \beta c_3 \underline{\lambda}_L \|\mathbf{y}\|^2 + \alpha c_3 \|\mathbf{y}\| \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*)\| \\ &\leq -\mu_1 \|\mathbf{z}\|^2 - \beta c_3 \underline{\lambda}_L \|\mathbf{y}\|^2 + \alpha c_3 p \|\mathbf{y}\|^2 \\ &\quad + \frac{\alpha c_3}{p} \|\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*)\|^2 \\ &\leq -(\mu_1 - \frac{\alpha c_3 L_f^2}{p}) \|\mathbf{z}\|^2 - c_3(\beta\underline{\lambda}_L - \alpha p) \|\mathbf{y}\|^2, \end{aligned}$$

where $p > 0$ and the third inequality follows from ∇F being L_f-Lipschitz continuous under Assumption 4. Suppose that $(\mu_1 - \frac{\alpha c_3 L_f^2}{p}) > 0$ and $(\beta\underline{\lambda}_L - \alpha p) > 0$. From the definition of V, $c_1 c_2 > c_3^2$ implies that $V(\mathbf{z}, \mathbf{y}) > 0$ for any $\begin{bmatrix} \mathbf{z}^\top & \mathbf{y}^\top \end{bmatrix}^\top \neq \mathbf{0_{2md}}$ and $V(\mathbf{0_{2md}}) = 0$. Hence, if the four positive scalar parameters c_1, c_2, c_3, β satisfy

$$c_1 c_2 > c_3^2, c_1 > \beta c_3, \alpha^2 L_f^2 c_3 < \beta\mu_1\underline{\lambda}_L, \beta c_1 - \beta c_2 + c_3 = 0, \quad (11)$$

then one can choose $p \in (\frac{\alpha L_f^2 c_3}{\mu_1}, \frac{\beta\underline{\lambda}_L}{\alpha})$, so that

$$\dot{V} \leq -\min\{\mu_1 - \frac{\alpha c_3 L_f^2}{p}, c_3(\beta\underline{\lambda}_L - \alpha p)\}(\|\mathbf{z}\|^2 + \|\mathbf{y}\|^2), \quad (12)$$

$$V \geq \underline{\epsilon}(\|\mathbf{z}\|^2 + \|\mathbf{y}\|^2), \quad (13)$$

for some $\underline{\epsilon} > 0$. By definition, $V(\mathbf{z},\mathbf{y}) \leq \left\|K^{-1}\right\| (\frac{c_1}{2}\|\mathbf{z}\|^2 + \frac{c_2}{2}\|\mathbf{y}\|^2 + \frac{c_3}{2}(\|\mathbf{z}\|^2 + \|\mathbf{y}\|^2))$, implying that

$$V(\mathbf{z},\mathbf{y}) \leq \bar{\epsilon}(\|\mathbf{z}\|^2 + \|\mathbf{y}\|^2), \quad (14)$$

for some $\bar{\epsilon} > 0$. So, the conditions of Lemma 1 hold.

Next, we consider the other case $\tilde{\mathcal{L}}\mathbf{y} = \mathbf{0_{md}}$. The error dynamics (9) is reduced to $\dot{\mathbf{z}} = -K\tilde{\mathcal{L}}\mathbf{z} - \alpha K(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*))$. Consider the Lyapunov candidate $V_2 : \mathbb{R}^{md} \to \mathbb{R}$ such that $V_2(\mathbf{z}) = \frac{1}{2}\mathbf{z}^\top K^{-1}\mathbf{z}$ for $\mathbf{z} \in \mathbb{R}^{md}$. Along the above trajectories, $\dot{V}_2(\mathbf{z}) = \mathbf{z}^\top K^{-1}\dot{\mathbf{z}} = -\mathbf{z}^\top\tilde{\mathcal{L}}\mathbf{z} - \alpha\mathbf{z}^\top(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*))$. Under Assumptions 1-4, we have similar result as (10): there exist $\overline{\overline{\alpha}}, \mu_2 > 0$ such that if $\alpha < \overline{\overline{\alpha}}$ then,

$$\alpha(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*))^\top(\mathbf{x} - \mathbf{X}_*) + \mathbf{x}^\top\tilde{\mathcal{L}}\mathbf{x} \geq \mu_2\|\mathbf{x} - \mathbf{X}_*\|^2.$$

Then, $\dot{V}_2(\mathbf{z}) \leq -\mu_2\|\mathbf{z}\|^2$. Moreover, $V_2(\mathbf{z})$ is quadratic in $\mathbf{z} \in \mathbb{R}^{md}$. So, the conditions of Lemma 1 hold.

Since we have a switched dynamics of (9) depending on whether $\mathbf{y}(t)$ is in the null-space of $\tilde{\mathcal{L}}$, we make the following argument to conclude exponential stability of (9). Consider any time $t = t_1$ when the Lyapunov function V_2 is active. Since $V_2(\mathbf{z})$ is quadratic for any $\mathbf{z}$, $V_2(t_1) \geq k_1\|\mathbf{z}(t_1)\|^2$ and $V_2 \leq k_2\|\mathbf{z}\|^2$ for some $k_1, k_2 > 0$ and for all t. Moreover, $\mathbf{z}$ exponentially decreases when either V or V_2 is active. So, there exists $t_2 \in [t_1, \infty)$ such that $V_2(t_2) \leq k_2\|\mathbf{z}(t_2)\|^2 \leq k_1\|\mathbf{z}(t_1)\|^2 \leq V_2(t_1)$. Hence, there exists a subsequence of $\mathbb{R}_{>0}$ over which V_2 is exponentially decreasing. Owing to Lemma 1, the proof is complete. ■

Next, we verify the feasibility of (11). We have

$$\beta c_1 - \beta c_2 + c_3 = 0, c_1 c_2 > c_3^2 \iff \beta c_1^2 + c_1 c_3 > \beta c_3^2,$$
$$c_1 > \beta c_3 \iff \beta c_1^2 + c_1 c_3 > (\beta^2 + 1)c_1 c_3.$$

$(\beta^2+1)c_1 > c_1$ for any $c_1 > 0$. We choose c_1 and c_3 such that $c_1 > \beta c_3$. Then, $(\beta^2+1)c_1 > c_1 > \beta c_3$, which implies that $(\beta^2+1)c_1 c_3 > \beta c_3^2$. The above and $c_1 > \beta c_3$ together imply that $\beta c_1^2 + c_1 c_3 > \beta c_3^2$. So, with $\beta > 0$, the conditions $c_1 c_2 > c_3^2, c_1 > \beta c_3, \beta c_1 - \beta c_2 + c_3 = 0$ are feasible. We have two cases for

$$\mu_1 = \min\{\frac{\mu}{2m}\alpha c_1, (c_1 - \beta c_3)\underline{\lambda}_L - \frac{2mL_f^2 + \mu\alpha c_1 L_f}{\mu\alpha c_1}\}.$$

Case-I: $\mu_1 = \frac{\mu}{2m}\alpha c_1$. In this case,

$$\alpha^2 L_f^2 c_3 < \beta\mu_1\underline{\lambda}_L \iff \beta c_3 \leq \frac{\beta^2 c_1\underline{\lambda}_L\mu}{2mL_f^2\alpha}.$$

There exists $\overline{\alpha}$ such that, for $\alpha < \overline{\alpha}$ we have $\frac{2mL_f^2\alpha}{\beta^2\underline{\lambda}_L\mu} < 1$, i.e., $\frac{\beta^2 c_1\underline{\lambda}_L\mu}{2mL_f^2\alpha} > c_1$. Thus, $c_1 > \beta c_3, \alpha^2 L_f^2 c_3 < \beta\mu_1\underline{\lambda}_L$ are feasible for $\alpha < \overline{\alpha}$.

Case-II: $\mu_1 = (c_1 - \beta c_3)\underline{\lambda}_L - \frac{2mL_f^2 + \mu\alpha c_1 L_f}{\mu\alpha c_1}$. We denote $q = \frac{2mL_f^2 + \mu\alpha c_1 L_f}{\mu\alpha c_1}$. Then, $\mu_1 = (c_1 - \beta c_3)\underline{\lambda}_L - q$ and

$$\alpha^2 L_f^2 c_3 < \beta\mu_1\underline{\lambda}_L \iff c_3 < \frac{\beta\underline{\lambda}_L((c_1 - \beta c_3)\underline{\lambda}_L - q)}{\alpha^2 L_f^2}$$
$$\iff \beta c_3 < \frac{\beta^2\underline{\lambda}_L(c_1\underline{\lambda}_L - q)}{\alpha^2 L_f^2 + \beta^2\underline{\lambda}_L^2}.$$

We denote $p = \frac{1+\beta^2\underline{\lambda}_L^2}{\beta^2\underline{\lambda}_L^2}$. Note that $p > 1$. Now,

$$p = \frac{1+\beta^2\underline{\lambda}_L^2}{\beta^2\underline{\lambda}_L^2} \implies c_1(1 - p\frac{\beta^2\underline{\lambda}_L^2}{1+\beta^2\underline{\lambda}_L^2}) = 0$$
$$\implies c_1(1 - p\frac{\beta^2\underline{\lambda}_L^2}{1+\beta^2\underline{\lambda}_L^2}) > -p\frac{\beta^2\underline{\lambda}_L q}{1+\beta^2\underline{\lambda}_L^2}$$
$$\iff c_1 > p\frac{\beta^2\underline{\lambda}_L(c_1\underline{\lambda}_L - q)}{1+\beta^2\underline{\lambda}_L^2}.$$

Then, $c_1 > p\frac{\beta^2\underline{\lambda}_L(c_1\underline{\lambda}_L - q)}{\alpha^2 L_f^2 + \beta^2\underline{\lambda}_L^2} > \frac{\beta^2\underline{\lambda}_L(c_1\underline{\lambda}_L - q)}{\alpha^2 L_f^2 + \beta^2\underline{\lambda}_L^2}$ for $\alpha < \frac{1}{L_f}$. Thus, $c_1 > \beta c_3, \alpha^2 L^2 c_3 < \beta\mu_1\underline{\lambda}_L$ are feasible for $\alpha < \frac{1}{L_f}$. From the above arguments, we conclude that (11) is feasible under the conditions of Theorem 1.

Next, we use a first-order explicit Euler discretization of (6)-(7) with fixed stepsize $h > 0$ to obtain the following algorithm from (6)-(7). For each $k \geq 0$ and each agent i,

$$\mathbf{x}_i(k+1) = \mathbf{x}_i(k) + hK_i\sum_{j\in\mathcal{N}_i}(\mathbf{x}_j(k) - \mathbf{x}_i(k))$$
$$- h\beta K_i\sum_{j\in\mathcal{N}_i}(\mathbf{v}_j(k) - \mathbf{v}_i(k)) - h\alpha K_i\nabla f_i(\mathbf{x}_i(k)), \quad (15)$$

$$\mathbf{v}_i(k+1) = \mathbf{v}_i(k) + h\beta K_i\sum_{j\in\mathcal{N}_i}(\mathbf{x}_j(k) - \mathbf{x}_i(k)). \quad (16)$$

The above algorithm can be compactly written as

$$\begin{bmatrix}\mathbf{x}(k+1)\\ \mathbf{v}(k+1)\end{bmatrix} = \begin{bmatrix}\mathbf{x}(k)\\ \mathbf{v}(k)\end{bmatrix}$$
$$- h\begin{bmatrix}K\tilde{\mathcal{L}}\mathbf{x}(k) - \beta K\tilde{\mathcal{L}}\mathbf{v}(k) + \alpha K\nabla F(\mathbf{x}(k))\\ \beta K\tilde{\mathcal{L}}\mathbf{x}(k)\end{bmatrix}. \quad (17)$$

The following result makes use of Lyapunov stability theory to prove linear convergence of the locally pre-conditioned algorithm (17) under the same Assumptions of Theorem 1, leading to rate-matching or consistent discretization of (8).

Theorem 2. *Consider algorithm (17) with initial condition $\mathbf{x}(0), \mathbf{v}(0) \in \mathbb{R}^{md}$. Suppose that Assumptions 1-4 hold true. Then there exists $\overline{\alpha}, \overline{\beta}, \underline{\beta}, \overline{h} \in (0,\infty)$ such that, for $\alpha < \overline{\alpha}$, $\beta \in (\overline{\beta}, \underline{\beta})$, and $h < \overline{h}$, the local estimate $\mathbf{x}_i(k)$ linearly converges to the same $\mathbf{x}_*$ for each agent $i \in \{1, 2, \ldots, m\}$.*

Proof. We define the error at iteration $k \geq 0$ as $\mathbf{Z}(k) = \begin{bmatrix}\mathbf{z}(k)\\ \mathbf{y}(k)\end{bmatrix} = \begin{bmatrix}\mathbf{x}(k) - \mathbf{X}_*\\ \mathbf{v}(k) - \mathbf{v}_*\end{bmatrix}$. For $\mathbf{Z} = [\mathbf{z}^\top, \mathbf{y}^\top]^\top \in \mathbb{R}^{2md}$, let

$$\tilde{F}(\mathbf{Z}) = \begin{bmatrix}K\tilde{\mathcal{L}}\mathbf{z} - \beta K\tilde{\mathcal{L}}\mathbf{y} + \alpha K(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*))\\ \beta K\tilde{\mathcal{L}}\mathbf{z}\end{bmatrix}.$$

Then, from (17), the error dynamics is

$$\mathbf{Z}(k+1) = \mathbf{Z}(k) - h\tilde{F}(\mathbf{Z}(k)). \tag{18}$$

Consider the case $\tilde{\mathcal{L}}\mathbf{y}(k) \neq 0_{md}$. As in the proof of Theorem 1, we define the Lyapunov function candidate $V : \mathbb{R}^{2md} \to \mathbb{R}$ such that $V(\mathbf{Z}) = \frac{c_1}{2}\mathbf{z}^\top K^{-1}\mathbf{z} + \frac{c_2}{2}\mathbf{y}^\top K^{-1}\mathbf{y} - c_3\mathbf{z}^\top K^{-1}\mathbf{y}$ for $\mathbf{Z} = [\mathbf{z}^\top \mathbf{y}^\top]^\top \in \mathbb{R}^{2md}$, where $c_1, c_2, c_3 > 0$. Clearly, $\nabla V(\mathbf{Z})$ is Lipschitz continuous with some Lipschitz constant $\eta > 0$. Then, from Lemma 5 in [3], we get $V(\mathbf{Z}(k+1)) - V(\mathbf{Z}(k)) \leq (\mathbf{Z}(k+1) - \mathbf{Z}(k))^\top \nabla V(\mathbf{Z}(k)) + \frac{\eta}{2}\left\|\mathbf{Z}(k+1) - \mathbf{Z}(k)\right\|^2$. Along the trajectories of (18),

$$\begin{aligned} &V(\mathbf{Z}(k+1)) - V(\mathbf{Z}(k)) \\ &\leq -h\tilde{F}(\mathbf{Z}(k))^\top \nabla V(\mathbf{Z}(k)) + \frac{h^2\eta}{2}\left\|\tilde{F}(\mathbf{Z}(k))\right\|^2. \end{aligned} \tag{19}$$

Following the proof of Theorem 1, similar to (12)-(14):

$$-\tilde{F}(\mathbf{Z}(k))^\top \nabla V(\mathbf{Z}(k)) \leq -\epsilon_1 V(\mathbf{Z}(k)), \tag{20}$$

$$V(\mathbf{Z}(k)) \geq \underline{\epsilon}\left\|\mathbf{Z}(k)\right\|^2. \tag{21}$$

From the definition, $\|\tilde{F}(\mathbf{Z})\|^2 \leq \|K\|^2(\beta^2\overline{\lambda}_L^2\|\mathbf{z}\|^2 + \left\|(\tilde{\mathcal{L}}\mathbf{z} - \beta\tilde{\mathcal{L}}\mathbf{y} + \alpha(\nabla F(\mathbf{x}) - \nabla F(\mathbf{X}_*)))\right\|^2$. Upon expanding the second term on the RHS above and utilizing AM-GM inequality, the definition of induced 2-norm of matrices, and Lipschitz continuity of ∇F under Assumption 4, we obtain that there exists $\epsilon_2 > 0$ such that $\left\|\tilde{F}(\mathbf{Z})\right\|^2 \leq \epsilon_2\|\mathbf{Z}\|^2$. Upon substituting above from (21), there exists $\epsilon_3 = \frac{\epsilon_2}{\underline{\epsilon}} > 0$ such that $\left\|\tilde{F}(\mathbf{Z})\right\|^2 \leq \epsilon_3 V(\mathbf{Z})$. Upon substituting from above and (20) in (19), $V(\mathbf{Z}(k+1)) \leq (1 - h\frac{2\epsilon_1 - h\eta\epsilon_3}{2})V(\mathbf{Z}(k))$. If $h \in (0, \frac{2\epsilon_1}{\eta\epsilon_3})$, then $\rho := (1 - h\frac{2\epsilon_1 - h\eta\epsilon_3}{2}) < 1$.

For $\tilde{\mathcal{L}}\mathbf{y}(k) = 0_{md}$, we consider the Lyapunov candidate $V_2 : \mathbb{R}^{md} \to \mathbb{R}$ such that $V_2(\mathbf{z}) = \frac{1}{2}\mathbf{z}^\top K^{-1}\mathbf{z}$ for $\mathbf{z} \in \mathbb{R}^{md}$. Proceeding similarly as V above, we obtain $V(\mathbf{Z}(k+1)) \leq \rho_2 V_2(\mathbf{Z}(k))$ for some $\rho_2 < 1$. The proof follows from the argument in the last paragraph of Theorem 1's proof. ■

Remark 1. *The aforementioned analysis has assumed a unique solution $\mathbf{x}_*$ of (1). When the solution set is not singleton, we denote the solution set of (1) by $\mathbf{x}_{sol}$, and define $\mathbf{X}_{sol} = [(\mathbf{x}_{sol})^\top, \ldots, (\mathbf{x}_{sol})^\top]^\top$. In this case, F is said to satisfy the RSI condition if $(\nabla F(\mathbf{x}) - \nabla F(\mathcal{P}_{\mathbf{X}_{sol}}(\mathbf{x}))^\top(\mathbf{x} - \mathcal{P}_{\mathbf{X}_{sol}}(\mathbf{x})) \geq \mu\left\|\mathbf{x} - \mathcal{P}_{\mathbf{X}_{sol}}(\mathbf{x})\right\|^2 \forall \mathbf{x} \in \mathbb{R}^{md}$, where $\mathcal{P}_{\mathbf{X}_{sol}}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathbf{X}_{sol}}\left\|\mathbf{x} - \mathbf{y}\right\|^2$ is the projection of $\mathbf{x}$ onto $\mathbf{X}_{sol}$ [3]. We define $\mathbf{x}^0 = [(\mathcal{P}_{\mathbf{X}_{sol}}(\mathbf{x}))^\top, \ldots, (\mathcal{P}_{\mathbf{X}_{sol}}(\mathbf{x}))^\top]^\top$. We assume that the gradients of the cumulative cost F at any point on the solution set, i.e., $\{\nabla F(\mathbf{x}) | \mathbf{x} \in \mathbf{X}_{sol}\}$ is a singleton. Under this assumption, the proofs of Theorem 1-2 are valid upon replacing $\mathbf{X}_*$ with $\mathbf{x}^0$. This is possible by invoking the properties of the projection operator from Theorem 1.5.5 in [22].*

IV. Numerical Results

A suitable choice of K_i could impact the convergence rate of PI consensus. For example, when f_i's are convex, consider $K_i = \left(\nabla^2 f_i(\mathbf{x}_i(0)) + \gamma I\right)^{-1}$ where $\gamma > 0$. In the case of single-agent, $K_i' = \left(\nabla^2 f_i(\mathbf{x}_i(t))\right)^{-1}$ results in Newton's method which is fast. However, it requires computing the expensive inverse Hessian at each t. Instead, we limit computing the inverse Hessian only once, before initiating the algorithm. Since the matrix inverse for K_i is computed once by each agent, it does not increase the per-iteration computational complexity of PI consensus. The parameter γ extends this pre-conditioner to non-strongly convex f_i's. Hence, we consider this choice of K^i.

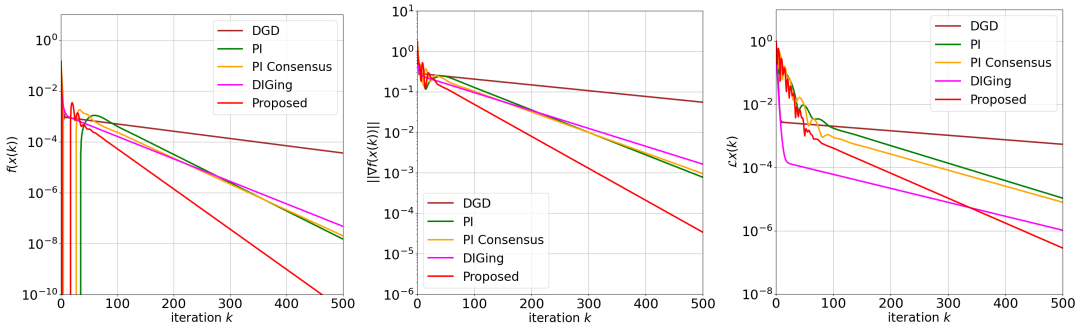
To show the efficiency of (15)-(16) with the proposed choice of pre-conditioner, we conduct two numerical simulations. First, we consider the problem from [3, Section VI] with $m = 5$ agents. Each f_i is non-convex and satisfies the RSI condition (see [3]). Next, we consider the binary classification problem between the digits one and five using the MNIST training dataset. We consider the logistic regression model and conduct experiments to minimize the cross-entropy error on the training data. The data points are distributed equally among $m = 5$ agents. The cumulative cost function does not satisfy Assumption 4 for this problem, and there are multiple solutions of (1). In both problems, the communication topology is a ring. For a fair comparison, the parameters in each algorithm are tuned such that the respective algorithm converges in fewer iterations. $\mathbf{x}_i(0)$ in all the algorithms are the same, and each entry is chosen from the Normal distribution with zero mean and 0.1 standard deviation. $\mathbf{v}_i(0)$ in the proposed algorithm and PI consensus algorithm are chosen similarly, and $\mathbf{v}_i(0)$ for the PI algorithm is according to [3]. Since F is not strongly convex, the APM-C, Mudag, DAccGD, and ACC-SONATA algorithms are not applicable in both the problems, as their implementation requires the positive value of the strong-convexity coefficient.

From Figure 1 and Figure 2, the PI consensus algorithm with the proposed K_i converges much faster than the other algorithms. In the first example, the condition number of the Hessian of F is of order 10. In contrast, in the second example, the ratio between the largest and the smallest non-zero singular value of the Hessian of F is of order 10^{11}. Thus, even for ill-conditioned problems, the proposed algorithm can converge significantly faster than the other algorithms. Further, from Figure 1, the PI consensus algorithm converges linearly under restricted secant inequality even though none of the local cost functions is convex, as proved in Theorem 2.

In (17), the effect of *connectivity* on the algorithm is represented by $\underline{\lambda}_{\min}(h\beta K\tilde{\mathcal{L}})$. In the first problem, this value is $0.98\underline{\lambda}_L$. Whereas, for PI consensus without preconditioning [4], this value is $0.1\underline{\lambda}_L$. Thus, introducing K increases the "effective connectivity" of $\mathbb{G}$, which helps in faster consensus of the proposed algorithm (Fig. 1(c)).

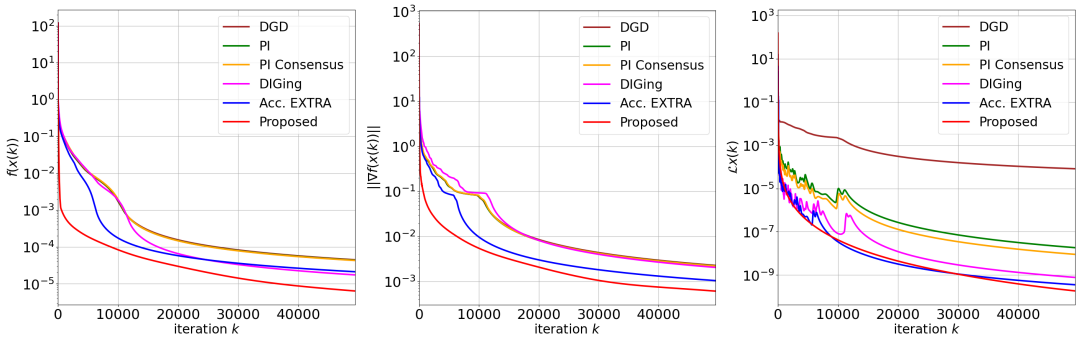
V. Conclusion

For the first time, we presented convergence guarantee at an exponential rate of the PI consensus algorithm for cumulative cost functions satisfying smoothness and the RSI condition, without requiring convexity. Using Lyapunov stability theory, we proved the convergence of this algorithm in the continuous-time domain and obtained its rate-matching discretization. Further, we introduced a local pre-



(a) aggregate cost (b) norm of gradient of aggregate cost (c) consensus term

Fig. 1: *Comparison of different distributed optimization algorithms for solving the problem from [3, Section VI].*



(a) aggregate cost (b) norm of gradient of aggregate cost (c) consensus term

Fig. 2: *Comparison of different distributed optimization algorithms for solving binary logistic regression on MNIST dataset.*

conditioning technique, locally computed by each agent, to accelerate the PI consensus algorithm. We demonstrated the efficacy of the proposed pre-conditioning compared to the existing distributed optimization algorithms, especially for ill-conditioned problems.

REFERENCES

[1] Tao Yang, Xinlei Yi, et al. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[2] Solmaz S Kia, Jorge Cortés, and Sonia Martínez. Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica*, 55:254–264, 2015.

[3] Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H Johansson. Exponential convergence for distributed optimization under the restricted secant inequality condition. *IFAC-PapersOnLine*, 53(2):2672–2677, 2020.

[4] Jing Wang and Nicola Elia. Control approach to distributed optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 557–561. IEEE, 2010.

[5] Takeshi Hatanaka, Nikhil Chopra, Takayuki Ishizaki, and Na Li. Passivity-based distributed optimization with communication delays using PI consensus algorithm. *IEEE Transactions on Automatic Control*, 63(12):4421–4428, 2018.

[6] Huan Li and Zhouchen Lin. Revisiting extra for smooth distributed optimization. *SIAM Journal on Optimization*, 30(3):1795–1821, 2020.

[7] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

[8] Shu Liang, George Yin, et al. Exponential convergence of distributed primal–dual convex optimization algorithm without strong convexity. *Automatica*, 105:298–306, 2019.

[9] Boya Sun. *Control Theoretic Approach to Study Accelerated First-Order Optimization Algorithms*. PhD thesis, University of California, Irvine, 2020.

[10] Kunal Garg, Mayank Baranwal, and Dimitra Panagou. A fixed-time convergent distributed algorithm for strongly convex functions in a time-varying network. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 4405–4410. IEEE, 2020.

[11] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[12] Wei Shi, Qing Ling, et al. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[13] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[14] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[15] Huan Li, Cong Fang, Wotao Yin, and Zhouchen Lin. A sharp convergence rate analysis for distributed accelerated gradient methods. *arXiv preprint arXiv:1810.01053*, 2018.

[16] Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *arXiv preprint arXiv:2005.00797*, 2020.

[17] Alexander Rogozin, Vladislav Lukoshkin, et al. Towards accelerated rates for distributed optimization over time-varying networks. In *Optimization and Applications: 12th International Conference, OPTIMA 2021, Petrovac, Montenegro, September 27–October 1, 2021, Proceedings 12*, pages 258–272. Springer, 2021.

[18] Ye Tian, Gesualdo Scutari, Tianyu Cao, and Alexander Gasnikov. Acceleration in distributed optimization under similarity. In *International Conference on Artificial Intelligence and Statistics*, pages 5721–5756. PMLR, 2022.

[19] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

[20] Bahman Gharesifard and Jorge Cortés. Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, 59(3):781–786, 2013.

[21] Hassan K Khalil. *Nonlinear systems; 3rd ed.* Prentice-Hall, Upper Saddle River, NJ, 2002.

[22] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.