DISTRIBUTED GRADIENT TRACKING ACHIEVING THE OPTIMAL WORST-CASE CONVERGENCE RATE OF CENTRALIZED GRADIENT DESCENT: NECESSARY AND SUFFICIENT CONDITIONS AND BEYOND

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper characterizes the optimal worst-case convergence rate of distributed gradient tracking, revealing its explicit dependence on both the objective functions and the communication network topology. We consider a modified version of the DIGing algorithm that incorporates the adapt-then-combine (ATC) strategy. In the algorithm, a parameter is set to tune the self-loop gain. For μ-strongly convex and L-smooth objective functions, we show that $2\varrho(\sqrt{\varrho}+1)/(3\varrho+1)$ is a threshold in the network topology that governs the worst-case convergence performance of the proposed DIGing-ATC algorithm, where $\varrho = L/\mu$. We prove that the algorithm can achieve the optimal worst-case convergence rate $(\varrho-1)/(\varrho+1)$ of centralized gradient descent (CGD) if and only if $\beta/\alpha \leq 2\varrho(\sqrt{\varrho}+1)/(3\varrho+1)$, where α and β are the smallest nonzero eigenvalue and the largest eigenvalue of the graph Laplacian matrix, respectively. We provide explicit formulas for the corresponding optimal parameters. For connected networks with $\beta/\alpha > 2\varrho(\sqrt{\varrho}+1)/(3\varrho+1)$, the optimal rate of CGD can never be reached. In this case, we further show that the achievable optimal worst-case convergence rate can be characterized as a root of a nonlinear equation depending only on ϱ and β/α, which can be computed efficiently using a bisection method. Our derivations and proofs employ nontrivial techniques, including the small gain theorem, Routh stability criteria, and analytic solutions of inequality-constrained optimization. Simulation results are presented to validate the theoretical convergence rates and demonstrate the practical effectiveness of the proposed algorithm.

1 INTRODUCTION

Distributed optimization plays an important role in large-scale collaborative training and learning, modern signal processing and multi-robotic systems, driven by the explosive growth of data from mobile devices, sensors, and edge platforms Xin et al. (2020); Shorinwa et al. (2024). In distributed optimization algorithms, a network of agents cooperatively solve a global optimization problem, with each agent performing local computations and exchanging information exclusively with its neighbors. A fundamental trade-off in distributed optimization lies between convergence speed, communication efficiency, and robustness to network uncertainties, all of which depend critically on algorithmic parameters—most notably the step size, the momentum factors, and in gradient-tracking or dual-based methods, auxiliary gain parameters.

Early distributed optimization algorithms emerged as natural extensions of centralized iterative methods. The Distributed Gradient Descent (DGD) algorithm Nedić & Ozdaglar (2009) adapts the standard gradient descent update by replacing global gradient averaging with local consensus over a communication graph. To address the sublinear convergence of DGD, algorithms including EXTRA Shi et al. (2015), DIGing Nedić et al. (2017a); Qu & Li (2017), and AugDGM Xu et al. (2015); Nedić et al. (2017b) were developed to guarantee linear convergence under fixed step sizes. These seminal works have served as the fundamental basis for distributed optimization, giving rise to a series of subsequent studies that develop variants of the gradient based schemes Li et al. (2019); Jakovetić (2018).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

For μ-strongly convex and L-smooth objective functions, it is well known that the optimal worst-case convergence rate of centralized gradient decent (CGD) is $(\varrho-1)/(\varrho+1)$, where $\varrho = L/\mu$. In this work, we address the following questions:

Can we design a distributed gradient tracking algorithm that achieve the optimal worst-case convergence rate of CGD? If yes, what are the necessary and sufficient conditions and the corresponding parameters?

In this paper, we answer the questions comprehensively by considering a modified version of the DIGing algorithm that incorporates the adapt-then-combine (ATC) strategy. In the algorithm, a parameter ε is set to tune the self-loop gain. Throughout this paper, the convergence rate of an algorithm refers exclusively to the worst-case convergence rate unless otherwise specified.

Our main contributions are summarized as follows:

- We characterize the optimal worst-case convergence rate of distributed gradient tracking, revealing its explicit dependence on both the objective functions and the communication network topology.
- We we show that $\mathcal{T}(\varrho) = \frac{2\varrho(\sqrt{\varrho}+1)}{3\varrho+1}$ is a threshold curve of the network topology that governs the worst-case convergence performance of the proposed DIGing-ATC algorithm. We prove that the DIGing-ATC algorithm can achieve the optimal worst-case convergence rate $\frac{\varrho-1}{\varrho+1}$ if and only if $\kappa := \frac{\beta}{\alpha} \leq \mathcal{T}(\varrho)$, where α and β are the smallest nonzero eigenvalue and the largest eigenvalue of the graph Laplacian matrix, respectively. We provide explicit formulas for the corresponding optimal parameters.
- For connected networks with $\kappa > \mathcal{T}(\varrho)$, the optimal rate of CGD can never be reached. In this case, we present the achievable optimal worst-case convergence rate, which is characterized as a root in an interval within $(0, 1)$ of a nonlinear equation depending only on ϱ and κ. The root can be computed efficiently using a bisection method. Besides the implicit characterization of the optimal rate, we derive explicit formulas for some achievable sub-optimal convergence rates and the corresponding parameters.
- We conduct numerical experiments to validate our theoretical results and demonstrate their practical effectiveness.

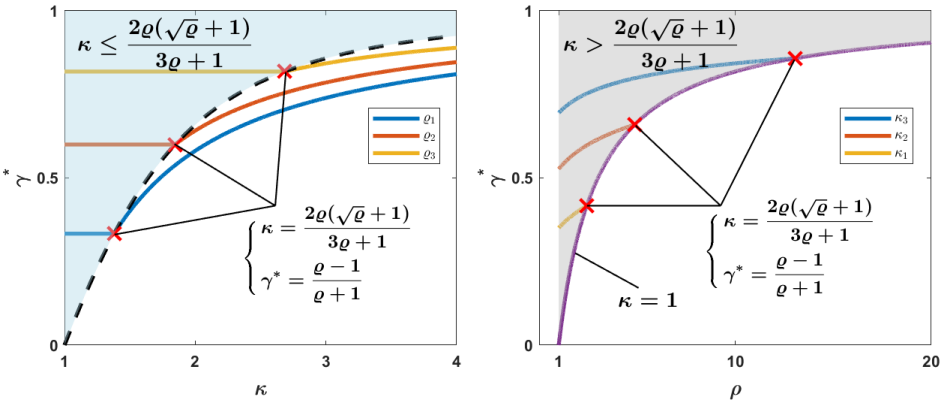


Figure 1: Illustration of the threshold curve of the optimal worst-case convergence rate γ^*

The key contributions can be clearly illustrated by Figure 1. In general, we know that the smaller κ is, the better the network connectivity is. Given a set of objective functions $\mathcal{F}_{\mu,L}$, the optimal worst-case convergence rate decreases as κ decreases if $\kappa > \mathcal{T}(\varrho)$. As the network connectivity becomes better and κ decreases further to the threshold $\kappa = \mathcal{T}(\varrho)$, the optimal convergence rate reaches the optimal convergence rate of CGD. Then the convergence rate cannot be decreased if we further improve the network connectivity, i.e., for any networks with $\kappa \leq \mathcal{T}(\varrho)$, the rate remains the same as that of CGD. This is of great importance in practical implementation of distributed optimization algorithms. For instance, if we know that the objective functions have a large ϱ, we can employ

a network with few communication connections to save the communication overhead while still keeping the optimal rate of CGD. To the best of our knowledge, this is the first characterization that explicitly reveals the network-topology threshold for a distributed algorithm to attain the optimal convergence rate of its centralized counterpart.

1.1 Related Work

Asymptotic Convergence Analysis of Distributed Optimization Algorithms. The comparison of convergence rates between distributed and centralized optimization has been a central concern in the research field. Qu & Li (2017); Li & Yan (2021); França & Bento (2020) derived asymptotic convergence rates of distributed optimization algorithms. Qu & Li (2017) observed in the experiments that the distributed algorithm requires more iterations although it has the same asymptotic convergence rate as the centralized GD. To achieve the same iteration count as centralized ADMM, FCD-ADMM in He et al. (2025) adds one global communication round per iteration. Scaman et al. (2019) demonstrates that an accelerated distributed algorithm achieves the optimal asymptotic rate with iteration complexity on par with centralized methods; and Yuan et al. (2024) shows that a zeroth-order distributed approach can match the asymptotic rate of centralized algorithms. Most recently, Kong et al. (2025) investigates the transient convergence performance for bilevel distributed optimization and provides an upper bound of the transient convergence rate. There have been many works devoted to improving the convergence rates of distributed algorithms. Qu & Li (2019); Li & Lin (2024) both adopt Nesterov's acceleration structure in distributed gradient tracking methods and provide asymptotic convergence rates. However, it is difficult to tune the parameters for the convergence. The theoretically guaranteed parameters are often conservative, necessitating empirical tuning or trial-and-error to achieve satisfactory experimental performance, as noted by the authors. Song et al. (2024) presents an optimal gradient tracking method that incorporates snapshot blocks and Chebyshev acceleration, and it attains optimality in both computation and communication complexity.

Worst-case Convergence Analysis of Optimization Algorithms via Robust Control Methods. It is well-known that iterative optimization algorithms can be viewed as feedback interconnections of discrete-time Lur'e systems Lessard (2022). The analysis and design of centralized optimization algorithms can be solved by different robust control methods, including integral quadratic constraint (IQC) theory Lessard et al. (2016); Fazlyab et al. (2018), dissipativity theory Lessard (2022), and small-gain theorem Zhou & Doyle (1998); Khalil & Grizzle (2002). For distributed optimization, robust control methods have also been adopted in various analysis and synthesis problems Hatanaka et al. (2018); Bin et al. (2019); Van Scoy & Lessard (2023); Liu et al. (2022). A unified framework is presented for the analysis of distributed optimization algorithms based on semidefinite programming (SDP) Sundararajan et al. (2020). Van Scoy & Lessard (2022) presented a universal decomposition structure with a centralized optimizer and a second-order consensus estimator, which encompasses many distributed optimization algorithms as special cases, including DIGing, EXTRA, and AugDGM (DIGing-ATC). The above frameworks require solving a small set of matrix inequalities by numerical methods, which cannot show the explicit relation between the convergence rate, the graph structure and the property of objective functions.

2 Preliminary and Problem Statements

Consider the distributed optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \triangleq \sum_{i=1}^{N} f_i(x), \tag{1}$$

where $x \in \mathbb{R}^d$ is the global decision variable, and each function $f_i : \mathbb{R}^d \to \mathbb{R}$ is a smooth and strongly convex objective function privately held by agent i. Let $x_i = \left[x_i^1, x_i^2, \cdots, x_i^d\right]^T$ be a local copy of the variable x held by agent i. The derivative of $f(\cdot)$ is denoted as $\nabla f(\cdot) = \sum_{i=1}^{N} \nabla f_i(\cdot)$. We introduce the aggregate function $\boldsymbol{f}(\boldsymbol{x}) \triangleq \sum_{i=1}^{N} f_i(x_i)$, where its argument and gradient are defined as $\boldsymbol{x} \triangleq \left[x_1^T, x_2^T, \cdots, x_N^T\right]^T$, $\nabla \boldsymbol{f}(\boldsymbol{x}) \triangleq \left[\nabla f_1^T(x_1), \nabla f_2^T(x_2), \cdots, \nabla f_N^T(x_N)\right]^T$.

The communication network is modeled as an undirected graph $\mathcal{G} = (\mathcal{E}, \mathcal{V}, W)$ with nodes $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$, edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ and the adjacency matrix $W \in \mathbb{R}^{N\times N}$. The set of neighbors of vertex v_i is denoted by $\mathcal{N}_i = \{v_j \in \mathcal{V} : (v_i, v_j) \in \mathcal{E}\}$. For any $v_i, v_j \in \mathcal{V}$, $w_{ij} > 0 \Leftrightarrow v_j \in \mathcal{N}_i$. The degree of vertex v_i is represented by $D_i = \sum_{j=1}^{N} w_{ij}$. The Laplacian matrix is defined as $\mathcal{L} \triangleq D - W$, where $D = \operatorname{diag}\{D_1, D_2, \cdots, D_N\}$.

For an undirected graph $\mathcal{G}$, the Laplacian matrix $\mathcal{L}$ is symmetric and positive semidefinite, and has the singular value decomposition of $\mathcal{L} = U\Lambda U^T$, where $\Lambda = \operatorname{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$, with $0 = \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_N$. Furthermore, the graph $\mathcal{G}$ is connected if and only if zero is a simple eigenvalue and the associated eigenvector (the first column of U) is $\frac{1}{\sqrt{N}}\mathbf{1}$.

We make the following two assumptions throughout the paper.

Assumption 1 *The objective function f_i is μ-strongly convex and L-smooth, i.e.,*

$$f_i(y) \geqslant f_i(x) + \nabla f_i(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2,$$
$$f_i(y) \leqslant f_i(x) + \nabla f_i(x)^T (y - x) + \frac{L}{2}\|y - x\|^2$$

hold for any $x, y \in \mathbb{R}^d$.

Assumption 2 *The communication graph $\mathcal{G}$ is undirected and connected.*

Denote $\mathcal{F}_{\mu,L}$ as the set of all μ-strongly convex and L-Lipschitz smooth functions. Denote $\{\mathcal{G}\}_{\alpha,\beta}$ as the set of all undirected connected graphs with $\lambda_2, \cdots, \lambda_N \in [\alpha, \beta]$. Define $\varrho = L/\mu$, $\kappa = \beta/\alpha$. By introducing a parameter ε to adjust the self-loop gain, we consider the following DIGing-ATC algorithm, which is a slightly generalized version of the algorithms in Xu et al. (2015); Nedić et al. (2017b).

$$\begin{aligned} x_i(k+1) &= (1 - \varepsilon D_i)[x_i(k) - \eta y_i(k)] + \varepsilon \sum_{j\in\mathcal{N}_i} w_{ij}[x_j(k) - \eta y_j(k)], \\ y_i(k+1) &= (1 - \varepsilon D_i)[y_i(k) + \nabla f_j(x_j(k+1)) - \nabla f_j(x_j(k))] \\ &\quad + \varepsilon \sum_{j\in\mathcal{N}_i} w_{ij}[y_j(k) + \nabla f_j(x_j(k+1)) - \nabla f_j(x_j(k))]. \end{aligned} \tag{2}$$

For each agent i, the initialization of (2) uses an arbitrary $x_i(0) \in \mathbb{R}^d$ and sets $y_i(0) = \nabla f_i(x_i(0))$ for $i = 1, \cdots, N$. The original algorithm in Xu et al. (2015); Nedić et al. (2017b) can be viewed as a special case of (2), where $\varepsilon = 1$ and the adjacency matrix $W = [w_{ij}]$ is doubly stochastic.

The aggregated formulation of DIGing-ATC can be written as

$$\begin{cases} \boldsymbol{x}(k+1) = ((I_N - \varepsilon\mathcal{L}) \otimes I_d)(\boldsymbol{x}(k) - \eta\boldsymbol{y}(k)) \\ \boldsymbol{y}(k+1) = ((I_N - \varepsilon\mathcal{L}) \otimes I_d)(\boldsymbol{y}(k) + \nabla\boldsymbol{f}(\boldsymbol{x}(k+1)) - \nabla\boldsymbol{f}(\boldsymbol{x}(k))) \end{cases}. \tag{3}$$

We are now ready to give the definition of the worst-case convergence rate and the problem statement.

Definition 1 *The algorithm (2) is said to converge to a stationary point $\boldsymbol{x}^*$ at the worst-case convergence rate γ if*

$$\lim_{k\to\infty} r^{-k}(\boldsymbol{x}(k) - \boldsymbol{x}^*) = \mathbf{0}, \ \forall r \in (\gamma, 1], \tag{4}$$

holds for any $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$, and any $\mathcal{G} \in \{\mathcal{G}\}_{\alpha,\beta}$.

Definition 2 *The optimal worst-case convergence rate of DIGing-ATC (2) is defined as follows*

$$\gamma^* \triangleq \inf_{\eta,\varepsilon} \sup_{\boldsymbol{f}\in\mathcal{F}_{\mu,L}, \mathcal{G}\in\{\mathcal{G}\}_{\alpha,\beta}} \gamma(\eta, \varepsilon), \tag{P1}$$

where γ is given by Definition 1.

3 MAIN RESULTS

In this section, we present the necessary and sufficient condition that the DIGing-ATC algorithm (2) achieves the optimal convergence rate of centralized gradient descent.

Theorem 1 *Consider the DIGing-ATC algorithm (2) for $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$ over the network $\mathcal{G} \in \{\mathcal{G}\}_{\alpha,\beta}$. The optimal worst-case convergence rate of (P1) is $\gamma^* = \frac{\varrho-1}{\varrho+1}$ if and only if κ and ϱ satisfy the following inequality*

$$\kappa \leqslant \frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1} \triangleq \mathcal{T}\left(\varrho\right). \tag{5}$$

The corresponding optimal parameters are set as $\eta^* = \frac{2}{L+\mu}$ and $\varepsilon^* \in \left[\frac{2}{\alpha(\sqrt{\varrho}+1)}, \frac{4\varrho}{\beta(3\varrho+1)}\right]$.

Remark 1 *The necessary and sufficient condition (5) depends only on ϱ and κ. In general , we know that the smaller ϱ is, the faster the CGD converges, and the smaller κ is, the better the network connectivity. Since $\frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1} = \frac{2\left(\sqrt{\varrho}+1\right)}{3+1/\varrho}$, the right-hand side of the inequality (5) monotonically increases to infinity as ϱ tends to infinity. Hence, for any objective function, there exist communication networks (not necessarily fully connected when $\varrho > 1$) such that the DIGing-ATC algorithm achieves the optimal convergence rate of CGD, provided the network connectivity is sufficiently strong. Conversely, for any communication network, there exist objective functions with sufficiently large ϱ such that DIGing-ATC achieves the CGD convergence rate.*

When $\kappa > \mathcal{T}(\varrho)$, the optimal rate of centralized CGD cannot be reached. In the following, we present the optimal achievable convergence rate. We need some notations. Define

$$g_1\left(\gamma\right) = \frac{\beta}{\alpha}\frac{\sqrt{\varrho}+1}{\sqrt{\varrho}+\frac{1}{1-\gamma}},\ g_2\left(\gamma\right) = 1 + \frac{\gamma}{1+\sqrt{\varrho\left(1-\gamma^2\right)}}. \tag{6}$$

Theorem 2 *Consider the DIGing-ATC algorithm (2) for $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$ over the network $\mathcal{G} \in \{\mathcal{G}\}_{\alpha,\beta}$. Suppose $\kappa \geqslant \frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1}$. Let $r_1 \in (0,1)$ be the root of $g_1(\gamma) = g_2(\gamma)$, where $g_1(\gamma)$ and $g_2(\gamma)$ are defined in (6). Then the worst-case optimal convergence rate of (P1) is $\gamma^* = r_1$, and the optimal parameters are set as $\eta^* = \frac{1-\gamma^*}{\mu}$ and $\varepsilon^* = \frac{1}{\beta}g_1(\gamma^*) = \frac{1}{\beta}g_2(\gamma^*)$.*

Remark 2 *There is no explicit formula to compute the root in (0,1) of the nonlinear equation $g_1(r) = g_2(r)$ in Theorem 2. However, it is easy to see that $g_1(r)$ is monotonically decreasing and $g_2(r)$ is monotonically increasing for $r \in [0,1]$. Since $g_1(0) = \frac{\beta}{\alpha} > 1 = g_2(0)$ and $g_1(1) = 0 < 2 = g_2(1)$, a bisection method given by Algorithm 1 can be easily employed to find the root r_1.*

Algorithm 1 Bisection Method

Initialize: $\gamma_0 = \frac{\varrho-1}{\varrho+1}, \gamma_1 = 1$
1: **while** $\gamma_1 - \gamma_0 > tolerance$ **do**
2: $\quad\gamma = \frac{1}{2}\left(\gamma_0+\gamma_1\right)$
3: $\quad$**if** $g_1\left(\gamma\right) < g_2\left(\gamma\right)$ **then** $\gamma_1 = \gamma$
4: $\quad$**else** $\gamma_0 = \gamma$
5: $\quad$**end if**
6: **end whilereturn** $\gamma, \varepsilon = \frac{1}{\beta}g_1\left(\gamma\right) = \frac{1}{\beta}g_2\left(\gamma\right), \eta = \frac{1-\gamma}{\mu}$.

In the following, we provide some suboptimal worst-case convergence rate for DIGing-ATC when $\kappa > \varrho$. Define

$$\gamma_1\left(b\right) = \frac{\sqrt{\varrho}+1}{\sqrt{\varrho}+\frac{1}{b}},\ \gamma_2\left(b\right) = \frac{b+b\sqrt{\varrho+b^2\varrho^2-b^2\varrho}}{1+b^2\varrho}. \tag{7}$$

Let $b_1 = 1-\frac{4\varrho}{\kappa(3\varrho+1)}$, $b_2 = \frac{2\kappa}{\sqrt{\varrho}+1}-1$, $b_3 = \frac{\kappa-1}{\kappa+1}$. We can verify that $b_1, b_2, b_3 \in (0,1)$ for $\kappa, \varrho > 1$ and $\kappa > \frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1}$.

Theorem 3 *Consider the DIGing-ATC algorithm (2) for $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$ over the network $\mathcal{G} \in \{\mathcal{G}\}_{\alpha,\beta}$. Suppose that $\kappa > \frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1}$. Then, we have the following statements:*

i) The worst-case convergence rate is $\gamma_1^* = \gamma_1(b_1)$ with $\eta_1^* = \frac{1-\gamma_1^*}{\mu}$, $\varepsilon_1^* = \frac{4\varrho}{\beta(3\varrho+1)}$;

ii) The worst-case convergence rate is $\gamma_2^* = \gamma_2(b_2)$ with $\eta_2^* = \frac{1-\gamma_2^*}{\mu}$, $\varepsilon_2^* = \frac{2}{\alpha\left(\sqrt{\varrho}+1\right)}$;

iii) The worst-case convergence rate is $\gamma_3^* = \gamma_2(b_3)$ with $\eta_3^* = \frac{1-\gamma_3^*}{\mu}$, $\varepsilon_3^* = \frac{\kappa-1}{\kappa+1}$.

Remark 3 *In the implementation of DIGing-ATC, the communication matrix $I - \varepsilon^* L$ is a doubly stochastic matrix. However, this matrix is unknown since ε has to be obtained following the above Theorems. On the other hand, for a given doubly stochastic matrix, we can use $(1-\varepsilon^*)I+\varepsilon^* W$ as an alternative communication matrix, which generally provides a better convergence rate even for other accelerated distributed algorithms Qu & Li (2019)Li & Lin (2024), as shown in the experiments.*

In the end of this section, we present the convergence result for the original DIGing-ATC proposed in Xu et al. (2015); Nedić et al. (2017b)

Theorem 4 *For the original DIGing-ATC, suppose that the adjacency matrix W is a doubly stochastic matrix with $0 < \alpha < \beta < 2$. Then the centralized rate $\gamma^* = \frac{\varrho-1}{\varrho+1}$ is achieved if and only if $\beta \leqslant \frac{4\varrho}{3\varrho+1}$ and $\alpha \geqslant \frac{\sqrt{\varrho}+1}{2}$, and the step size is set as $\eta^* = \frac{2}{L+\mu}$. Otherwise, the optimal worst-case convergence rate of $\min_{\eta} \gamma$ is given by $\gamma^* = \max\{\gamma_1(1-\alpha), \gamma_2(\beta-1)\}$, and the opitmal step size is set as $\eta^* = \frac{1-\gamma^*}{\mu}$, where $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ are defined as (7).*

4 Proof of Main Results

By using unitary transformation, we convert the general MIMO system (3) to a Lur'e system with diagonal linear components. Then we consider a subset of quadratic objective functions, under which the feedback system is linear and a lower bound of the convergence rate can be derived by using Routh stability criteria. Finally, we show that this lower bound is attainable for the whole set of objective functions $\mathcal{F}_{\mu,L}$ by small gain theorem. All theorems in Section 3 can be proved by the following three steps.

Step 1. Decomposition by unitary transformation. Let $\mathcal{L} = U\Lambda U^T$ be the singular value decomposition of the Laplacian matrix, where U is the unitary matrix and $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_N\}$. Define

$$\tilde{\boldsymbol{x}}(k) = U^T \otimes I_d \boldsymbol{x}(k), \ \tilde{\boldsymbol{y}}(k) = U^T \otimes I_d \boldsymbol{y}(k), \ \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}) = U^T \otimes I_d \nabla \boldsymbol{f}(\boldsymbol{x}). \tag{8}$$

Then system (3) can be written as

$$\begin{cases} \tilde{\boldsymbol{x}}(k+1) = (I_N - \varepsilon\Lambda) \otimes I_d (\tilde{\boldsymbol{x}}(k) - \eta\tilde{\boldsymbol{y}}(k)) \\ \tilde{\boldsymbol{y}}(k+1) = (I_N - \varepsilon\Lambda) \otimes I_d (\tilde{\boldsymbol{y}}(k) + \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k+1)) - \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k))) \end{cases}. \tag{9}$$

Denote $\tilde{\boldsymbol{u}}(k) = \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k+1)) - \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k))$, and $\tilde{\boldsymbol{v}}(k) = \tilde{\boldsymbol{x}}(k+1) - \tilde{\boldsymbol{x}}(k)$. Define the nonlinear operator $\tilde{\Delta}_{\boldsymbol{f}}(\cdot)$ as $\tilde{\boldsymbol{u}}(k) = \tilde{\Delta}_{\boldsymbol{f}}(\tilde{\boldsymbol{v}}(k))$. Then the dynamic system (9) can be represented by Fig. 2(a), where

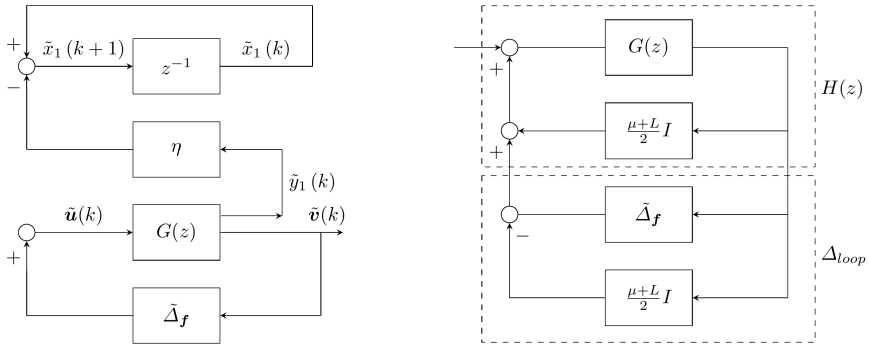
$$G(z) = \operatorname{diag}\{g_1(z), \cdots, g_N(z)\} \otimes I_d, \tag{10}$$

and $g_1(z) = -\frac{\eta}{z-1}$ and $g_i(z) = -\frac{\eta(z-1)(1-\varepsilon\lambda_i)^2}{(z-1+\varepsilon\lambda_i)^2}$ for $i = 2, \cdots, N$.

If the closed-loop Lur'e system is γ-stable, i.e. the feedback system with $G(\gamma z)$ and $\tilde{\Delta}_{\boldsymbol{f}}$ is stable, then the state variables converge to zero at the rate γ Khalil & Grizzle (2002). Therefore the problem (P1) becomes to find the minimal γ so that the closed-loop Lur'e system is γ-stable for all $\lambda_i \in [\alpha, \beta]$, and $\tilde{\Delta}_{\boldsymbol{f}}$ induced by $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$.

Step 2. Deriving the lower bound. Consider a subset of quadratic objective functions $Q_{\mu,L} \subseteq \mathcal{F}_{\mu,L}$, defined as $Q_{\mu,L} \triangleq \left\{\boldsymbol{f} : \tilde{\Delta}_{\boldsymbol{f}} = qI_{Nd}, q \in [\mu, L]\right\}$. For $\boldsymbol{f} \in Q_{\mu,L}$, the Lur'e system turns out to be linear, of which the closed-loop transfer matrix is given by

$$\Phi(z) = (I - qG(z))^{-1} G(z) = \operatorname{diag}\{\Phi_1(z), \cdots, \Phi_N(z)\} \otimes I_d.$$



(a) (b)

Figure 2: (a) The discrete dynamic system of DIGing-ATC. (b) Loop transformation of the high-frequency subsystem.

Hence $\underset{\eta,\varepsilon}{inf} \underset{Q_{\mu,L},\{\mathcal{G}\}_{\alpha,\beta}}{sup} \gamma(\eta,\varepsilon)$ is equivalent to the following minimization problem

$$\begin{aligned} &\min_{\eta>0,\varepsilon>0} \gamma \\ s.t.\ &\Phi_1(\gamma z) \text{ is stable for any } q\in[\mu,L], \\ &\Phi_i(\gamma z) \text{ is stable for any } q\in[\mu,L] \text{ and } \lambda_i\in[\alpha,\beta], i=2,\cdots,N. \end{aligned} \tag{P2}$$

By using Routh stability criteria, we can show that (P2) is equivalent to $\min_{\eta>0,\varepsilon>0} \gamma$ subject to the following inequalities

$$\begin{cases} L\eta \leqslant 1+\gamma, & \text{(11a)} \\ \mu\eta \geqslant 1-\gamma, & \text{(11b)} \\ (\gamma-1+\varepsilon\lambda)^2 - L\eta(1-\varepsilon\lambda)^2(1-\gamma) \geqslant 0,\ \lambda\in[\alpha,\beta], & \text{(11c)} \\ (\gamma+1-\varepsilon\lambda)^2 - L\eta(1-\varepsilon\lambda)^2(1+\gamma) \geqslant 0,\ \lambda\in[\alpha,\beta], & \text{(11d)} \\ \gamma^2 - (1-\varepsilon\lambda)^2 + \mu\eta(1-\varepsilon\lambda)^2 \geqslant 0,\ \lambda\in[\alpha,\beta]. & \text{(11e)} \end{cases}$$

The following lemma enables the derivation of the worst-case convergence rate γ^* in all theorems.

Lemma 1 *Minimizing γ under constraints (11a)-(11e) is equivalent to minimizing γ under constraints (11a)-(11b) and the following inequalities*

$$\begin{cases} \gamma-1+\varepsilon\alpha \geqslant \sqrt{L\eta(1-\gamma)}(1-\varepsilon\alpha), & \text{(12a)} \\ 1+\gamma-\varepsilon\beta \geqslant \sqrt{L\eta(1+\gamma)}(\varepsilon\beta-1). & \text{(12b)} \end{cases}$$

Step 3. Verifying the attainability of the lower bound via small gain theorem. Recall that the state variables converge to zero at the rate γ if the closed-loop Lur'e system is γ-stable Khalil & Grizzle (2002). In this step, we show that the system $G(z)$ with parameters η^*, ε^* is γ^*-stable for all $\tilde{\Delta}_{\boldsymbol{f}}$.

By loop transformation shown in Figure 2(b), we obtain an equivalent feedback system with $H(z)$ and Δ_{loop}, where

$$H(z) = \left(I - \frac{\mu+L}{2}G(z)\right)^{-1} G(z) = \operatorname{diag}\{H_1(z),\cdots,H_N(z)\}\otimes I_d. \tag{13}$$

It is shown in Lessard (2022) that $\tilde{\Delta}_{\boldsymbol{f}}(\cdot)$ is $[\mu,L]$ sector bounded for $\boldsymbol{f}\in\mathcal{F}_{\mu,L}$, which implies $\mu\|\tilde{\boldsymbol{v}}\| \leqslant \left\|\tilde{\Delta}_{\boldsymbol{f}}(\tilde{\boldsymbol{v}})\right\| \leqslant L\|\tilde{\boldsymbol{v}}\|$. Then it is obvious that $\|\Delta_{loop}\|_\infty = \left\|\tilde{\Delta}_{\boldsymbol{f}} - \frac{L+\mu}{2}\right\| \leqslant \frac{L-\mu}{2}$. $\|H(\gamma^* z)\|_\infty$ is given by the following Lemma.

Lemma 2 *Consider $H(z)$ defined in (13) with γ^*, η^*, and ε^* given by Theorem 1-3, we have*

$$\|H(\gamma^* z)\|_\infty = \max_{\substack{i\in\{1,\cdots,N\},\theta\in[0,2\pi),\\ \lambda_2,\cdots,\lambda_N\in[\alpha,\beta]}} \left|H_i\left(\gamma^* e^{j\theta}\right)\right| = \frac{2}{L-\mu}.$$

It follows from small gain theorem that the whole system is γ^*-stable since

$$\|H(\gamma^* z)\|_\infty \|\Delta_{loop}\|_\infty \leqslant \frac{2}{L-\mu}\cdot\frac{L-\mu}{2} = 1.$$

Key step in the proof of Theorem 1 Substituting $\gamma_c = \frac{\varrho-1}{\varrho+1}$, $\eta^* = \frac{2}{L+\mu}$ into (12a)-(12b), we have

$$-\frac{\sqrt{\varrho}-1}{\sqrt{\varrho}+1} \leqslant \varepsilon\alpha - 1 \leqslant \varepsilon\beta - 1 \leqslant \frac{\varrho-1}{3\varrho+1},$$

which implies $\frac{1}{\alpha}\frac{2}{\sqrt{\varrho}+1} \leqslant \varepsilon \leqslant \frac{1}{\beta}\frac{4\varrho}{3\varrho+1}$, i.e. $\frac{\beta}{\alpha} \leqslant \frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1}$.

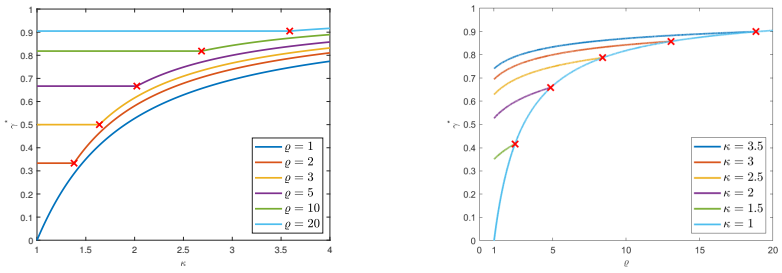
Key step in the proof of Theorem 2 (12a)-(12b) yield that

$$\frac{1}{\alpha}\left(1-\frac{\gamma}{\sqrt{\varrho}(1-\gamma)+1}\right) \leqslant \varepsilon \leqslant \frac{1}{\beta}\left(1+\frac{\gamma}{\sqrt{\varrho(1-\gamma^2)}+1}\right).$$

By equating both sides, we obtain the nonlinear equation $g_1(\gamma) = g_2(\gamma)$ in Theorem 2.

5 Numerical Experiments

In the first experiment, we plot the worst-case optimal convergence rate γ^* w.r.t κ and $\varrho \geqslant 1$. In Figure 3, the red cross marks the turning point $\kappa = \mathcal{T}(\varrho)$. The results clearly show that γ^* decreases monotonically as either κ or ϱ decreases, indicating a smaller and hence better convergence rate. In Figure 3(a), the horizontal segment of each line indicates that the optimal convergence rate $\frac{\varrho-1}{\varrho+1}$ is independent of κ. In Figure 3(b), the light blue line represents $\gamma^* = \frac{\varrho-1}{\varrho+1}$, which is achievable for all objective functions on the fully connected graph $\kappa = 1$. Clearly, to the right of the red cross marker, all lines coincide with the $\kappa = 1$ line, indicating that the optimal CGD rate $\frac{\varrho-1}{\varrho+1}$ is achieved when condition $\kappa \leq \mathcal{T}(\varrho)$ holds.



(a) (b)

Figure 3: Curves of γ^* w.r.t. κ and ϱ.

In the second experiemnt, we consider the following logistic regression objective function

$$f(x) = \sum_{i=1}^{N} f_i(x) = \sum_{i=1}^{N}\sum_{j=1}^{m} \ln\left(1+e^{-\langle l_j c_j, x\rangle}\right) + \frac{\delta}{2}\|x\|_2^2, \tag{14}$$

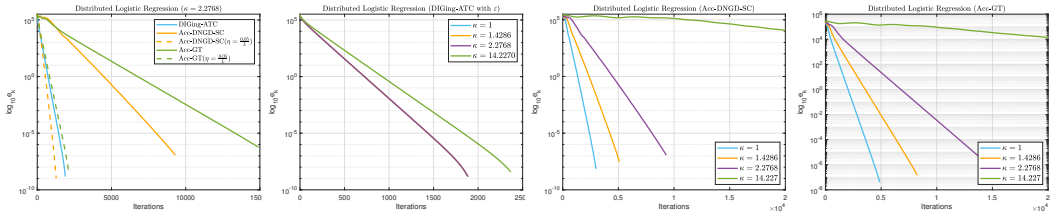
where m is the number of samples in each agent, $c_j \in \mathbb{R}^d$ is the feature vector with a corresponding label $l_j \in \{-1, 1\}$ and δ is the regularization coefficient. In this experiment, we use the $a9a$ dataset[1]

[1]The LIBSVM Chang & Lin (2011) dataset collection is available at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

from the LIBSVM dataset collection. Set $N = 50$, $m = 150$, $d = 123$, $\delta = 0.005$. The estimations of the objective function are $\mu = 0.005$, $L = 1.2518$.

We compare the ε-weight-adjusted DIGing-ATC with two distributed acceleration algorithms: Acc-DNGD-SC Qu & Li (2019) and Acc-GT Li & Lin (2024). We evaluate all distributed algorithms on a randomly generated graph with 50 nodes and $\kappa = 2.2768$. We use a Metropolis-Hastings weight matrix as the communication matrix, which is doubly stochastic. The step size η and the parameter ε of the DIGing-ATC algorithm are as stated in Theorem 1-2. For the two acceleration algorithms, we consider two choices of step size. The first is the theoretically guaranteed maximum step size $\eta = (1-\sigma)^3/119L$, derived from Theorem 10 in Li & Lin (2024). The second is $\eta = 0.05/L$, obtained through parameter tuning and recommended in Li & Lin (2024). Note that the tuned step size $\eta = 0.05/L$ may cause the algorithm to diverge, especially for poorly connected graphs and objective functions with a large condition number. Figure 4(a) show that the weight-adjusted DIGing-ATC algorithm with optimal parameter design performs comparably to the accelerated algorithms using the tuned step size. Moreover, it achieves a significantly faster convergence rate than the accelerated algorithm with the theoretically guaranteed step size.

Further, we demonstrate the turning point of the DIGing-ATC algorithm. We conduct experiments on four graphs of $N = 50$ nodes with different connectivity properties, including a fully-connected graph $\kappa = 1$, and three randomly generated graphs with $\kappa = 1.4286, 2.2768, 14.2270$, respectively. Results are shown in Figure 4(b), 4(c), 4(d) for DIGing-ATC, Acc-DNGD-SC and Acc-GT, respectively. We adopt the theoretically guaranteed maximum step size $\eta = (1-\sigma)^3/119L$ for acceleration methods. In Figure 4(b) for DIGing-ATC, it is clear that lines coincide when $\kappa \leqslant \mathcal{T}(\varrho) = 11.2003$, which confirms that a threshold exists. However, in Figure 4(c), 4(d), the number of iterations gradually increases as the network connectivity parameter κ increases without a threshold.



(a) (b) (c) (d)

Figure 4: Algorithm comparison and threshold verification.

6 Conclusion and Future Work

We present the necessary and sufficient condition $\kappa \leq 2\varrho(\sqrt{\varrho}+1)/(3\varrho+1))$, under which the DIGing-ATC can achieve the optimal worst-case convergence rate of CGD. Furthermore, for networks violating this condition, we derive the achievable optimal rate and the corresponding parameters. We show that the rate is a unique root in $(0,1)$ of a nonlinear equation depending solely on ϱ and κ, which can be easily computed via a bisection method. Our derivation and proof integrates a variety of non-trivial techniques, such as the small gain theorem, Routh stability criteria, and analytic solutions of inequality-constrained optimization. Numerical experiments validate the theoretical results and illustrate the practical efficacy.

Future direction includes extending the methodology to other distributed optimization algorithms, such as EXTRA and accelerated variants of distributed gradient descent. While the techniques are largely applicable, we would like to make two cautions. First, the lower bound derived from the subset of quadratic functions may not be sufficient to guarantee convergence when extended to general objective functions. Second, the inequalities derived from Routh stability criteria may become increasingly intricate, potentially precluding the derivation of explicit formulas and necessitating numerical characterizations.

AI USE STATEMENT

In this work, we used generative AI tools to help polish the writing and enhance readability. We have not used generative AI tools for other tasks with required disclosure.

REFERENCES

Michelangelo Bin, Ivano Notarnicola, Lorenzo Marconi, and Giuseppe Notarstefano. A system theoretical perspective to gradient-tracking algorithms for distributed quadratic optimization. *IEEE Conference on Decision and Control (CDC)*, pp. 2994–2999, 2019.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

Mahyar Fazlyab, Alejandro Ribeiro, Manfred Morari, and Victor M Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018.

Guilherme França and José Bento. Distributed optimization, averaging via admm, and network topology. *Proceedings of the IEEE*, 108(11):1939–1952, 2020.

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.

Jeannie He, Ming Xiao, and Mikael Skoglund. Fast-converging decentralized alternating direction method of multipliers for consensus optimization. *EURASIP Journal on Advances in Signal Processing*, 2025(1):27, 2025.

Jinrui Huang, Xueqin Wang, Dong Liu, Jingguo Lan, and Runxiong Wu. Communication-efficient decentralized optimization via double-communication symmetric admm. *International Conference on Learning Representations*, pp. 12508–12527, 2026.

Dušan Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2018.

Hassan K Khalil and Jessy W Grizzle. *Nonlinear Systems*. Prentice Hall Upper Saddle River, NJ, 2002.

Boao Kong, Shuchen Zhu, Songtao Lu, Xinmeng Huang, and Kun Yuan. Decentralized bilevel optimization: A perspective from transient iteration complexity. *Journal of Machine Learning Research*, 26(240):1–64, 2025.

Laurent Lessard. The analysis of optimization algorithms: A dissipativity approach. *IEEE Control Systems Magazine*, 42(3):58–72, 2022.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Huan Li and Zhouchen Lin. Accelerated gradient tracking over time-varying graphs for decentralized optimization. *Journal of Machine Learning Research*, 25(274):1–52, 2024.

Yao Li and Ming Yan. On the linear convergence of two decentralized algorithms. *Journal of Optimization Theory and Applications*, 189(1):271–290, 2021.

Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67 (17):4494–4506, 2019.

Bing Liu, Li Chai, and Jingwen Yi. Convergence analysis of distributed gradient descent algorithms with one and two momentum terms. *IEEE Transactions on Cybernetics*, 54(3):1511–1522, 2022.

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

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

Angelia Nedić, Alex Olshevsky, Wei Shi, and César A Uribe. Geometrically convergent distributed optimization with uncoordinated step-sizes. *American Control Conference (ACC)*, pp. 3950–3955, 2017b.

Yurii Nesterov et al. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

Norman S Nise. *Control Systems Engineering*. John Wiley & Sons, 2019.

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

Guannan Qu and Na Li. Accelerated distributed nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2019.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

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.

Ola Shorinwa, Trevor Halsted, Javier Yu, and Mac Schwager. Distributed optimization methods for multi-robot systems: Part 2—a survey. *IEEE Robotics & Automation Magazine*, 31(3):154–169, 2024.

Zhuoqing Song, Lei Shi, Shi Pu, and Ming Yan. Optimal gradient tracking for decentralized optimization. *Mathematical Programming*, 207(1-2):1–53, 2024.

Akhil Sundararajan, Bryan Van Scoy, and Laurent Lessard. Analysis and design of first-order distributed optimization algorithms over time-varying graphs. *IEEE Transactions on Control of Network Systems*, 7(4):1597–1608, 2020.

Bryan Van Scoy and Laurent Lessard. A universal decomposition for distributed optimization algorithms. *IEEE Control Systems Letters*, 6:3044–3049, 2022.

Bryan Van Scoy and Laurent Lessard. A tutorial on the structure of distributed optimization algorithms. *IEEE Conference on Decision and Control (CDC)*, pp. 3009–3014, 2023.

Lin Xiao and Stephen Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

Ran Xin, Shi Pu, Angelia Nedić, and Usman A Khan. A general framework for decentralized optimization with first-order methods. *Proceedings of the IEEE*, 108(11):1869–1889, 2020.

Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. *IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060, 2015.

Deming Yuan, Lei Wang, Alexandre Proutiere, and Guodong Shi. Distributed zeroth-order optimization: Convergence rates that match centralized counterpart. *Automatica*, 159:111328, 2024.

Kemin Zhou and John Comstock Doyle. *Essentials of Robust Control*, volume 104. Prentice Hall Upper Saddle River, NJ, 1998.

A Proof of Main Results

The proof is presented in three parts. First, we provide the partial decomposition of the dynamical system. Second, we derive the optimal worst-case convergence rate γ_Q^* over the function subset $Q_{\mu,L}$. Third, we use the small gain theorem to show that the rate γ_Q^* can be achieved for the whole set $\mathcal{F}_{\mu,L}$.

A.1 The Dynamical System Perspective of the Iteration Algorithm

Let $\mathcal{L} = U\Lambda U^T$ be the singular value decomposition of the Laplacian matrix, where U is the unitary matrix and $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_N\}$. Use the definition in (8) as

$$\tilde{\boldsymbol{x}}(k) = U^T \otimes I_d \boldsymbol{x}(k),\ \tilde{\boldsymbol{y}}(k) = U^T \otimes I_d \boldsymbol{y}(k),\ \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}) = U^T \otimes I_d \nabla \boldsymbol{f}(\boldsymbol{x}).$$

Then system (3) can be written as (9)

$$\begin{cases} \tilde{\boldsymbol{x}}(k+1) = (I_N - \varepsilon\Lambda) \otimes I_d (\tilde{\boldsymbol{x}}(k) - \eta\tilde{\boldsymbol{y}}(k)) \\ \tilde{\boldsymbol{y}}(k+1) = (I_N - \varepsilon\Lambda) \otimes I_d (\tilde{\boldsymbol{y}}(k) + \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k+1)) - \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k))) \end{cases}.$$

Denote $\tilde{\boldsymbol{u}}(k) = \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k+1)) - \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k))$, and $\tilde{\boldsymbol{v}}(k) = \tilde{\boldsymbol{x}}(k+1) - \tilde{\boldsymbol{x}}(k)$. Define the nonlinear operator $\tilde{\Delta}_{\boldsymbol{f}}(\cdot)$ as $\tilde{\boldsymbol{u}}(k) = \tilde{\Delta}_{\boldsymbol{f}}(\tilde{\boldsymbol{v}}(k))$. We know that $\tilde{\Delta}_{\boldsymbol{f}}(\cdot)$ is $[\mu, L]$ sector bounded for $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$ Lessard (2022), i.e., $\left(\tilde{\Delta}_{\boldsymbol{f}}(z) - \mu z\right)^T \left(\tilde{\Delta}_{\boldsymbol{f}}(z) - Lz\right) \leqslant 0$ for all $z \in \mathbb{R}^{Nd}$, which implies $\mu \|z\| \leqslant \left\|\tilde{\Delta}_{\boldsymbol{f}}(z)\right\| \leqslant L\|z\|$.

Lemma 3 *The dynamic system (9) has the following partially decoupled structure*

$$\tilde{x}_1(k+1) = \tilde{x}_1(k) - \eta\tilde{y}_1(k), \tag{15a}$$

$$\begin{bmatrix} \tilde{\boldsymbol{x}}_{2:N}(k+1) \\ \tilde{\boldsymbol{y}}(k+1) \end{bmatrix} = A \begin{bmatrix} \tilde{\boldsymbol{x}}_{2:N}(k) \\ \tilde{\boldsymbol{y}}(k) \end{bmatrix} + B\tilde{\boldsymbol{u}}(k), \tag{15b}$$

$$\tilde{\boldsymbol{v}}(k) = C \begin{bmatrix} \tilde{\boldsymbol{x}}_{2:N}(k) \\ \tilde{\boldsymbol{y}}(k) \end{bmatrix}, \tag{15c}$$

$$\tilde{\boldsymbol{u}}(k) = \tilde{\Delta}_{\boldsymbol{f}}(\tilde{\boldsymbol{v}}(k)), \tag{15d}$$

where $A = \begin{bmatrix} I_{N-1} - \varepsilon\Lambda_{2:N} & \left[0_{(N-1)\times 1} \quad -\eta(I_{N-1} - \varepsilon\Lambda_{2:N})\right] \\ 0_{N\times(N-1)} & I_N - \varepsilon\Lambda \end{bmatrix} \otimes I_d$, $B = \begin{bmatrix} 0_{(N-1)\times N} \\ I_N - \varepsilon\Lambda \end{bmatrix} \otimes I_d$, $C = \left[\begin{array}{c|c} 0_{1\times(N-1)} & -\eta(I_N - \varepsilon\Lambda) \\ -\varepsilon\Lambda_{2:N} & \end{array}\right] \otimes I_d$, $\Lambda_{2:N} = \operatorname{diag}\{\lambda_2, \cdots, \lambda_N\}$. *Furthermore, the transfer matrix from* $\tilde{\boldsymbol{u}}(k)$ *to* $\tilde{\boldsymbol{v}}(k)$ *is given by*

$$G(z) \quad = C(zI - A)^{-1} B \otimes I_d \quad = \operatorname{diag}\{g_1(z), \cdots, g_N(z)\} \otimes I_d,$$

where $g_1(z) = -\frac{\eta}{z-1}$ *and* $g_i(z) = -\frac{\eta(z-1)(1-\varepsilon\lambda_i)^2}{(z-1+\varepsilon\lambda_i)^2}$ *for* $i = 2, \cdots, N$.

The Lure system with partially decoupled structure can be shown in Figure 2(a). The subsystem of $\lambda_1 = 0$ is called zero-frequency in graph filter processing, while the subsystem with λ_i, $i = 2, \cdots, N$ is called high-frequency in graph filter processing. Figure 2(a) shows clearly that the convergence of the zero-frequency subsystem is totally governed by the high-frequency subsystem. The following lemma further shows that if the high-frequency subsystem (15b)-(15d) converges to zero at rate γ, then the zero-frequency subsystem (15a) converges to $\sqrt{N}x^*$ at the rate γ, where x^* is the optimal solution of problem.

Lemma 4 *If the high-frequency subsystem in (15b)-(15d) converges to zero at rate* γ*, i.e.,* $\lim_{k\to\infty} \gamma^{-k}\tilde{\boldsymbol{x}}_{2:N}(k) = \mathbf{0}$, $\lim_{k\to\infty} \gamma^{-k}\tilde{\boldsymbol{y}}(k) = \mathbf{0}$*, then the zero-frequency subsystem converges to* $\sqrt{N}x^*$ *at the rate* γ*, where* x^* *is the optimal solution of (1), i.e.,* $\lim_{k\to\infty} \gamma^{-k}\left(\tilde{x}_1(k) - \sqrt{N}x^*\right) = 0$.

Proof. *U has the form of $U = [\frac{1}{\sqrt{N}}\mathbf{1}, U_2, ..., U_N]$ since the graph is connected. It follows from $\lim_{k\to\infty} \gamma^{-k}\tilde{y}_1(k) = 0$ that $\lim_{k\to\infty} \gamma^{-k}(\tilde{x}_1(k+1) - \tilde{x}_1(k)) = 0$. Hence, $\tilde{x}_1(k)$ converges as $k \to \infty$, and denote $\lim_{k\to\infty} \tilde{x}_1(k) \triangleq \bar{x}$. Then, we have*

$$\begin{aligned} &\lim_{k\to\infty} \gamma^{-k}\left(\boldsymbol{x}(k) - \frac{1}{\sqrt{N}}\mathbf{1}_N \otimes \bar{x}\right) = \lim_{k\to\infty} \gamma^{-k}\left(U \otimes I_d \tilde{\boldsymbol{x}}(k) - \frac{1}{\sqrt{N}}\mathbf{1}_N \otimes \bar{x}\right) \\ &= \lim_{k\to\infty} \gamma^{-k}\left(\left[\tfrac{1}{\sqrt{N}}\mathbf{1}_N \quad U_2 \quad \cdots \quad U_N\right] \otimes I_d \begin{bmatrix} \tilde{x}_1(k) \\ \tilde{\boldsymbol{x}}_{2:N}(k) \end{bmatrix} - \frac{1}{\sqrt{N}}\mathbf{1}_N \otimes \bar{x}\right) \\ &= \lim_{k\to\infty} \gamma^{-k}\left(\frac{1}{\sqrt{N}}\mathbf{1}_N \otimes I_d \tilde{x}_1(k) - \frac{1}{\sqrt{N}}\mathbf{1}_N \otimes \bar{x} + [U_2 \quad \cdots \quad U_N] \otimes I_d \tilde{\boldsymbol{x}}_{2:N}(k)\right) = \mathbf{0}. \end{aligned}$$

Therefore $\lim_{k\to\infty} \boldsymbol{x}(k) = \frac{1}{\sqrt{N}}\mathbf{1}_N \otimes \bar{x}$, *which implies that* $\lim_{k\to\infty} x_i(k) = \frac{1}{\sqrt{N}}\bar{x}$ *for* $i = 1, \cdots, N$.

In the following, we prove that $\frac{1}{\sqrt{N}}\bar{x}$ *is the optimal solution of problem (1) satisfying* $\sum_{i=1}^{N} \nabla f_i\left(\frac{1}{\sqrt{N}}\bar{x}\right) = 0$. *To this end, we check the dynamics of* $\tilde{y}_1(k)$, *which is extracted from (15b) as* $\tilde{y}_1(k+1) = \tilde{y}_1(k) + \tilde{u}_1(k)$.

Notice that

$$\tilde{\boldsymbol{u}}(k) = U^T \otimes I_d \boldsymbol{u}(k) = \left[\frac{1}{\sqrt{N}}\mathbf{1}_N \quad \cdots \quad U_N\right]^T \otimes I_d\left(\nabla \boldsymbol{f}(\boldsymbol{x}(k+1)) - \nabla \boldsymbol{f}(\boldsymbol{x}(k))\right).$$

Then, $\tilde{u}_1(k) = \frac{1}{\sqrt{N}}\mathbf{1}_N^T \otimes I_d\left(\nabla \boldsymbol{f}(\boldsymbol{x}(k+1)) - \nabla \boldsymbol{f}(\boldsymbol{x}(k))\right)$, *and*

$$\begin{aligned}\sum_{l=0}^{k} \tilde{u}_1(l) &= \frac{1}{\sqrt{N}}\mathbf{1}_N^T \otimes I_d \sum_{l=0}^{k}\left(\nabla \boldsymbol{f}(\boldsymbol{x}(l+1)) - \nabla \boldsymbol{f}(\boldsymbol{x}(l))\right) \\ &= \frac{1}{\sqrt{N}}\mathbf{1}_N^T \otimes I_d\left(\nabla \boldsymbol{f}(\boldsymbol{x}(k+1)) - \nabla \boldsymbol{f}(\boldsymbol{x}(0))\right).\end{aligned} \tag{16}$$

Hence we have

$$\begin{aligned}&\tilde{y}_1(k+1) = \tilde{y}_1(k) + \tilde{u}_1(k) = \tilde{y}_1(0) + \sum_{l=0}^{k} \tilde{u}_1(l) \\ &= \frac{1}{\sqrt{N}}\mathbf{1}_N^T \otimes I_d\left(\boldsymbol{y}(0) - \nabla \boldsymbol{f}(\boldsymbol{x}(0)) + \nabla \boldsymbol{f}(\boldsymbol{x}(k+1))\right) = \frac{1}{\sqrt{N}}\mathbf{1}_N^T \otimes I_d \nabla \boldsymbol{f}(\boldsymbol{x}(k+1)).\end{aligned}$$

The last equality comes from the initial condition that $y_i(0) = \nabla f_i(x_i(0))$. *Combining the fact that* $\lim_{k\to\infty} \tilde{y}_1(k) = 0$ *and* $\lim_{k\to\infty} \boldsymbol{x}(k) = \frac{1}{\sqrt{N}}\mathbf{1}_N \otimes \bar{x}$, *we obtain* $\sum_{i=1}^{N} \nabla f_i\left(\frac{1}{\sqrt{N}}\bar{x}\right) = 0$. *Therefore, each agent's state* $x_i(k)$ *converges to* $x^* = \frac{1}{\sqrt{N}}\bar{x}$, *and* x^* *is the optimal solution of problem (1).*

It follows from Lemma 4 that the convergence of the zero-frequency subsystem is totally governed by the high-frequency subsystem, which is a closed-loop Lur'e system with $G(z)$ and $\tilde{\Delta}_{\boldsymbol{f}}$, as shown in Figure 5.

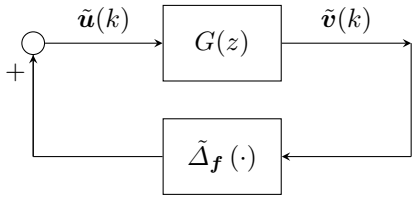


Figure 5: The high-frequency subsystem.

If the closed-loop Lur'e system is γ-stable, then the state variables converge to zero at the rate γ Khalil & Grizzle (2002). Therefore the problem (P1) becomes to find the minimal γ so that the closed-loop Lure system is γ-stable for all $\lambda_i \in [\alpha, \beta]$, and $\tilde{\Delta}_{\boldsymbol{f}}$ induced by $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$. Note that $\tilde{\Delta}_{\boldsymbol{f}}$ is sector-bounded by $[\mu, L]$. By loop transformation shown in Figure 2(b), we obtain an equivalent feedback system with $H(z)$ and Δ_{loop}, where $H(z) = \left(I - \frac{\mu+L}{2}G(z)\right)^{-1} G(z)$.

A.2 A Lower Bound on the Convergence Rate

Next, the proof of the main results is divided into two steps. First we derive the optimal worst-case convergence rate over the function subset $Q_{\mu,L}$. Since $Q_{\mu,L}$ is a subset of $\mathcal{F}_{\mu,L}$, we know that $\gamma^* \geqslant \gamma_Q^*$. Then we use the small gain theorem to show that this rate γ_Q^* can be achieved for the whole set $\mathcal{F}_{\mu,L}$.

Consider a subset of quadratic objective functions $Q_{\mu,L} \subseteq \mathcal{F}_{\mu,L}$, defined as follows

$$Q_{\mu,L} \triangleq \left\{ \boldsymbol{f} : \tilde{\Delta}_{\boldsymbol{f}} = qI_{Nd}, q \in [\mu, L] \right\}.$$

For $\boldsymbol{f} \in Q_{\mu,L}$, the Lur'e system turns out to be a linear system, of which the stability can be easily analyzed. We can obtain the exact worst-case convergence rate γ defined over $Q_{\mu,L}$ and $\{\mathcal{G}\}_{[\alpha,\beta]}$.

Lemma 5 *Consider the Lur'e system shown in Figure 2(b). If $\boldsymbol{f} \in Q_{\mu,L}$, we have the transfer matrix $H(z)$ and the whole closed-loop system are given by $H(z) = \mathrm{diag}\{H_i(z)\} \otimes I_d$ and $\Phi(z) = \mathrm{diag}\{\Phi_i(z)\} \otimes I_d$, $i = 1, \cdots, N$, respectively, where*

$$H_1(z) = \frac{-\eta}{z - 1 + \frac{\mu+L}{2}\eta}, \tag{17}$$

$$H_i(z) = \frac{-\eta(z-1)(1-\varepsilon\lambda_i)^2}{(z-1+\varepsilon\lambda_i)^2 + \frac{\mu+L}{2}\eta(z-1)(1-\varepsilon\lambda_i)^2}, \tag{18}$$

$$\Phi_1(z) = -\frac{\eta}{z-1+q\eta}, \tag{19}$$

$$\Phi_i(z) = \frac{-\eta(z-1)(1-\varepsilon\lambda_i)^2}{(z-1+\varepsilon\lambda_i)^2 + q\eta(z-1)(1-\varepsilon\lambda_i)^2}, \tag{20}$$

for $i = 2, \cdots, N$.

Proof. *Lemma 5 can be easily derived by the following computation*

$$H(z) = \left(I - \frac{\mu+L}{2}G(z)\right)^{-1} G(z),$$
$$\Phi(z) = (I - qG(z))^{-1} G(z).$$

Finding the optimal worst-case convergence rate of DIGing-ATC over $Q_{\mu,L}$ and $\{\mathcal{G}\}_{[\alpha,\beta]}$ is equivalent to the following minimization problem (P2)

$$\begin{aligned} &\min_{\eta>0,\varepsilon>0} \ \gamma \\ s.t.\ &\Phi_1(\gamma z) \text{ is stable for any } q \in [\mu, L], \\ &\Phi_i(\gamma z) \text{ is stable for any } q \in [\mu, L] \text{ and } \lambda_i \in [\alpha, \beta], i = 2, \cdots, N. \end{aligned}$$

By bilinear transform $z = \frac{s+1}{s-1}$, we know that $\Phi_i(\gamma z)$ is stable if and only if the poles of $\Phi_i(\gamma s)$ lie in the left half plane. The denominator polynomial of $\Phi_i(\gamma s)$, $i = 2, \cdots, N$ can be derived directly as follows

$$\begin{aligned} &\left[(\gamma - 1 + \varepsilon\lambda_i)^2 - q\eta(1-\gamma)(1-\varepsilon\lambda_i)^2\right] s^2 + 2\left[\gamma^2 - (1-\varepsilon\lambda_i)^2 + q\eta(1-\varepsilon\lambda_i)^2\right] s \\ &+ (\gamma + 1 - \varepsilon\lambda_i)^2 - q\eta(\gamma+1)(1-\varepsilon\lambda_i)^2. \end{aligned} \tag{21}$$

It follows from the Routh's Stability Criterion Nise (2019) that the roots of (21) lie in left half plane if and only if all the coefficients in (21) are non-negative. That is,

$$\begin{cases} (\gamma - 1 + \varepsilon\lambda)^2 - q\eta(1-\varepsilon\lambda)^2(1-\gamma) \geqslant 0 \\ \gamma^2 - (1-\varepsilon\lambda)^2 + q\eta(1-\varepsilon\lambda)^2 \geqslant 0 \\ (\gamma + 1 - \varepsilon\lambda)^2 - q\eta(1-\varepsilon\lambda)^2(1+\gamma) \geqslant 0 \end{cases}, \tag{22}$$

for all $\lambda \in [\alpha, \beta]$ and $q \in [\mu, L]$.

Since $L\eta \geqslant q\eta \geqslant \mu\eta$, we have (22) holds for all $q \in [\mu, L]$ is equivalent to the following inequalities hold

$$\begin{cases} (\gamma - 1 + \varepsilon\lambda)^2 - L\eta(1-\varepsilon\lambda)^2(1-\gamma) \geqslant 0 \\ \gamma^2 - (1-\varepsilon\lambda)^2 + \mu\eta(1-\varepsilon\lambda)^2 \geqslant 0 \\ (\gamma + 1 - \varepsilon\lambda)^2 - L\eta(1-\varepsilon\lambda)^2(1+\gamma) \geqslant 0 \end{cases}.$$

Note that $\Phi_1(\gamma z; q)$ is stable for all $q \in [\mu, L]$ if and only if $1 - \mu\eta \leqslant \gamma$ and $L\eta - 1 \leqslant \gamma$. The optimization problem in (P2) becomes

$$\min_{\eta>0,\varepsilon>0} \gamma \tag{23}$$

$s.t.$

$$\begin{cases} L\eta \leqslant 1+\gamma, & \text{(24a)} \\ \mu\eta \geqslant 1-\gamma, & \text{(24b)} \\ (\gamma - 1 + \varepsilon\lambda)^2 - L\eta(1-\varepsilon\lambda)^2(1-\gamma) \geqslant 0, & \text{(24c)} \\ (\gamma + 1 - \varepsilon\lambda)^2 - L\eta(1-\varepsilon\lambda)^2(1+\gamma) \geqslant 0, & \text{(24d)} \\ \gamma^2 - (1-\varepsilon\lambda)^2 + \mu\eta(1-\varepsilon\lambda)^2 \geqslant 0, & \text{(24e)} \end{cases}$$

for all $\lambda \in [\alpha, \beta]$. Note that (24a)-(24e) is copied from (11a)-(11e) for readability.

Proof of Lemma 1.

Proof. *First we prove that*

$$1 - \gamma - \varepsilon\alpha \leqslant 0 \text{ and } \varepsilon\beta - 1 - \gamma \leqslant 0. \tag{25}$$

It follows from (24b) and (24c) that

$$\begin{aligned} (1-\gamma-\varepsilon\lambda)^2 &\geqslant L\eta(1-\varepsilon\lambda)^2(1-\gamma) \geqslant \frac{L}{\mu}(1-\gamma)^2(1-\varepsilon\lambda)^2 \\ &= \frac{L}{\mu}(1-\gamma-\varepsilon\lambda+\varepsilon\lambda\gamma)^2 \geqslant (1-\gamma-\varepsilon\lambda+\varepsilon\lambda\gamma)^2. \end{aligned} \tag{26}$$

Since $\varepsilon\lambda\gamma > 0$, (26) implies that $1 - \gamma - \varepsilon\lambda \leqslant 0$ for any $\lambda \in [\alpha, \beta]$, which is equivalent to $1 - \gamma - \varepsilon\alpha \leqslant 0$.

Adding (24d) and (24e) yields that

$$2\gamma^2 + 2\gamma(1-\varepsilon\lambda) + (1-\varepsilon\lambda)^2[\mu\eta - (1+\gamma)L\eta] \geqslant 0.$$

Therefore $2\gamma(1+\gamma-\varepsilon\lambda) \geqslant (1-\varepsilon\lambda)^2(L\eta - \mu\eta + \gamma L\eta) \geqslant 0$ for any $\lambda \in [\alpha, \beta]$.

Obviously, $1 + \gamma - \varepsilon\lambda \geqslant 0$ for any $\lambda \in [\alpha, \beta]$ is equivalent to $\varepsilon\beta - 1 - \gamma \leqslant 0$.

It follows from $1 - \gamma - \varepsilon\lambda \leqslant 0$ and $1 + \gamma - \varepsilon\lambda \geqslant 0$ that $\gamma^2 - (1-\varepsilon\lambda)^2 = (\gamma + 1 - \varepsilon\lambda)(\gamma - 1 + \varepsilon\lambda) \geqslant 0$. Hence, (24c) and (24d) hold imply that (24e) holds.

Therefore, $\min_{\eta,\varepsilon} \gamma$ subject to (24a)-(24e) is equivalent to $\min_{\eta,\varepsilon} \gamma$ subject to (24a)-(24b) and inequalities

$$\begin{cases} \gamma - 1 + \varepsilon\lambda \geqslant \sqrt{L\eta(1-\gamma)}\,|\varepsilon\lambda - 1|, & \text{(27a)} \\ 1 + \gamma - \varepsilon\lambda \geqslant \sqrt{L\eta(1+\gamma)}\,|\varepsilon\lambda - 1|, & \text{(27b)} \end{cases}$$

for all $\lambda \in [\alpha, \beta]$.

Next we show that (27a)-(27b) hold for any $\lambda \in [\alpha, \beta]$ if and only if (12a)-(12b) hold. Note that the "only if" part trivially holds.

Suppose that (12a)-(12b) hold. If $\varepsilon\lambda - 1 \geqslant 0$, then $\varepsilon\beta - 1 \geqslant \varepsilon\lambda - 1 \geqslant 0$. We have $\frac{\gamma}{\varepsilon\lambda-1} - 1 \geqslant \frac{\gamma}{\varepsilon\beta-1} - 1 \geqslant \sqrt{L\eta(1+\gamma)}$, which implies (27b) holds. Since $\frac{\gamma}{\varepsilon\lambda-1} + 1 \geqslant \frac{\gamma}{\varepsilon\lambda-1} - 1 \geqslant \sqrt{L\eta(1+\gamma)} > \sqrt{L\eta(1-\gamma)}$, we obtain (27a) holds.

If $\varepsilon\lambda - 1 < 0$, then $0 < 1 - \varepsilon\lambda \leqslant 1 - \varepsilon\alpha$. We have $\frac{\gamma}{|\varepsilon\lambda-1|} - 1 \geqslant \frac{\gamma}{1-\varepsilon\alpha} - 1 \geqslant \sqrt{L\eta(1-\gamma)}$, which implies (27a) holds. Since $\frac{\gamma}{|\varepsilon\lambda-1|} + 1 \geqslant \frac{\gamma}{1-\varepsilon\alpha} + 1 \geqslant \sqrt{L\eta(1-\gamma)} + 2 > 1 + \gamma \overset{(24a)}{\geqslant} \sqrt{L\eta(1+\gamma)}$, we obtain that (27b) holds. The above derivation use the fact that $0 < \gamma < 1$.

Therefore, if (12a)-(12b) hold, then (27a)-(27b) hold for $\lambda \in [\alpha, \beta]$.

A.3 Proof of Necessary Condition of Theorem 1

The optimal solution of minimizing γ subject to (24a)-(24b) is $\gamma^* = \frac{\varrho-1}{\varrho+1}$ and $\eta^* = \frac{2}{L+\mu}$. Recall that the optimal worst-case convergence rate of CGD is $\gamma_c = \frac{\varrho-1}{\varrho+1}$. Therefore the algorithm DIGing-ATC achieves the rate of γ_c only if there exists ε such that (27a)-(27b) hold for $\gamma = \gamma_c$ and $\eta = \eta^*$, $\lambda \in [\alpha, \beta]$. Substituting γ_c and η^* into (27a) and (27b), we have

$$\gamma_c + \varepsilon\lambda - 1 \geqslant |\varepsilon\lambda - 1| \sqrt{\frac{L}{\mu}} (1 - \gamma_c), \tag{28a}$$

$$\gamma_c + 1 - \varepsilon\lambda \geqslant |\varepsilon\lambda - 1| (1 + \gamma_c). \tag{28b}$$

If $\varepsilon\lambda - 1 \geqslant 0$, then (28b) can be written as

$$\gamma_c - (\varepsilon\lambda - 1) \geqslant (\varepsilon\lambda - 1)(1 + \gamma_c), \tag{29}$$

which yields

$$\varepsilon\lambda - 1 \leqslant \frac{\gamma_c}{2 + \gamma_c} = \frac{\varrho - 1}{3\varrho + 1}. \tag{30}$$

Note that (29) implies (28a) since

$$\gamma_c + \varepsilon\lambda - 1 \geqslant \gamma_c \geqslant \gamma_c + 1 - \varepsilon\lambda \overset{from\ (28b)}{\geqslant} (\varepsilon\lambda - 1)(1 + \gamma_c)$$

$$= (\varepsilon\lambda - 1) \frac{2\varrho}{\varrho + 1} \geqslant (\varepsilon\lambda - 1) \frac{2\sqrt{\varrho}}{\varrho + 1} = (\varepsilon\lambda - 1) \sqrt{\frac{L}{\mu}} (1 - \gamma_c).$$

If $\varepsilon\lambda - 1 \leqslant 0$, then (28a) can be written as

$$\gamma_c + \varepsilon\lambda - 1 \geqslant (1 - \varepsilon\lambda) \frac{2\sqrt{\varrho}}{\varrho + 1},$$

which yields

$$1 - \varepsilon\lambda \leqslant \frac{1 + \varrho}{\left(1 + \sqrt{\varrho}\right)^2} \gamma_c = \frac{\sqrt{\varrho} - 1}{\sqrt{\varrho} + 1} < 1.$$

(28b) holds trivially since $\gamma_c + 1 - \varepsilon\lambda \geqslant (1 - \varepsilon\lambda)\gamma_c + 1 - \varepsilon\lambda = (1 - \varepsilon\lambda)(1 + \gamma_c)$.

Therefore, there exists ε such that (27a)-(27b) hold for $\gamma = \gamma_c$, $\eta = \eta^*$ and $\lambda \in [\alpha, \beta]$ only if

$$-\frac{\sqrt{\varrho} - 1}{\sqrt{\varrho} + 1} \leqslant \varepsilon\alpha - 1 \leqslant \varepsilon\beta - 1 \leqslant \frac{\varrho - 1}{3\varrho + 1}.$$

That is

$$\frac{1}{\alpha} \frac{2}{\sqrt{\varrho} + 1} \leqslant \varepsilon \leqslant \frac{1}{\beta} \frac{4\varrho}{3\varrho + 1}. \tag{31}$$

It follows from (31) that

$$\frac{\beta}{\alpha} \leqslant \frac{2\varrho\left(\sqrt{\varrho} + 1\right)}{3\varrho + 1}.$$

Hence, (P1) has the solution of $\gamma^* = \gamma_c$ only if $\frac{\beta}{\alpha} \leqslant \frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1}$.

The "if" part will be proved in Section A.7.

A.4 Derivation of the Lower Bound in Theorem 2

It follows from (12a)-(12b) that

$$\varepsilon\alpha \geqslant \frac{\sqrt{L\eta(1-\gamma)} + 1 - \gamma}{\sqrt{L\eta(1-\gamma)} + 1} = 1 - \frac{\gamma}{\sqrt{L\eta(1-\gamma)} + 1}, \tag{32a}$$

$$\varepsilon\beta \leqslant \frac{\sqrt{L\eta(1+\gamma)} + 1 + \gamma}{\sqrt{L\eta(1+\gamma)} + 1} = 1 + \frac{\gamma}{\sqrt{L\eta(1+\gamma)} + 1}. \tag{32b}$$

Using (24b), we have $\sqrt{L\eta} \geqslant \sqrt{\frac{L}{\mu}(1-\gamma)}$. Hence,

$$\varepsilon\alpha \geqslant 1 - \frac{\gamma}{\sqrt{\varrho}(1-\gamma)+1}, \tag{33a}$$

$$\varepsilon\beta \leqslant 1 + \frac{\gamma}{\sqrt{\varrho(1-\gamma^2)}+1}. \tag{33b}$$

(33a)-(33b) yield that

$$\frac{1}{\alpha}\left(1 - \frac{\gamma}{\sqrt{\varrho}(1-\gamma)+1}\right) \leqslant \varepsilon \leqslant \frac{1}{\beta}\left(1 + \frac{\gamma}{\sqrt{\varrho(1-\gamma^2)}+1}\right). \tag{34}$$

As $\gamma \in \left(\frac{\varrho-1}{\varrho+1}, 1\right)$ decreases, the left-hand side of inequality (34) increases monotonically and the right-hand side decreases monotonically. Therefore, the minimum γ^* satisfying (34) is given by the root of the following equation

$$\frac{\beta}{\alpha}\left(1 - \frac{\gamma}{\sqrt{\varrho}(1-\gamma)+1}\right) = 1 + \frac{\gamma}{\sqrt{\varrho(1-\gamma^2)}+1}. \tag{35}$$

A bisection method can be easily employed to find γ^* by using the properties of $\gamma^* \in \left[\frac{\varrho-1}{\varrho+1}, 1\right)$ and the monotonicity of both sides in (35).

Since $\min\limits_{s.t.(24a),(24b),(32a),(32b)} \gamma \geqslant \min\limits_{s.t.(24b),(32a),(32b)} \gamma$, we know that γ^* is a lower bound of the solution of (P2). That is, $\gamma_{P2} \geqslant \gamma^*$.

Once γ^* is computed, we obtain that $\eta^* = \frac{1-\gamma^*}{\mu}$ and ε^* is given by $\varepsilon^* = \frac{1}{\beta}g_1(\gamma^*) = \frac{1}{\beta}g_2(\gamma^*)$.

A.5 Derivation of the Lower Bound in Theorem 3

It follows from (11b) and (12a) that

$$\sqrt{\frac{1-\gamma}{\mu}} \leqslant \sqrt{\eta} \leqslant \frac{\gamma - 1 + \varepsilon\alpha}{\sqrt{L(1-\gamma)}(1-\varepsilon\alpha)}. \tag{36}$$

There exists η such that (36) holds if and only if $\sqrt{\frac{L}{\mu}}(1-\gamma)(1-\varepsilon\alpha) \leqslant \gamma - (1-\varepsilon\alpha)$. Hence $\gamma \geqslant \frac{(1-\varepsilon\alpha)\left(\sqrt{\varrho}+1\right)}{(1-\varepsilon\alpha)\sqrt{\varrho}+1} = \frac{\sqrt{\varrho}+1}{\sqrt{\varrho}+\frac{1}{1-\varepsilon\alpha}}$.

For simplicity, denote $\tau = \varepsilon\beta - 1$. It follows from (11b) and (12b) that

$$\sqrt{\frac{1-\gamma}{\mu}} \leqslant \sqrt{\eta} \leqslant \frac{\gamma - \tau}{\sqrt{\eta(1+\gamma)}\tau}. \tag{37}$$

There exists η such that (37) holds if and only if $\sqrt{\frac{L}{\mu}(1-\gamma^2)}\tau \leqslant \gamma - \tau$. That is

$$\varrho\left(1-\gamma^2\right)\tau^2 \leqslant \gamma^2 - 2\tau\gamma + \tau^2. \tag{38}$$

(38) can be written as $\left(\varrho\tau^2+1\right)\gamma^2 - 2\tau\gamma + \tau^2 - \varrho\tau^2 \geqslant 0$. Since $0 < \gamma < 1$, we have $\gamma \geqslant \gamma_2(\tau)$, where $\gamma_2(\cdot)$ is defined in (7). Since $\frac{\beta}{\alpha} > \frac{2\varrho\left(\sqrt{\varrho}+1\right)}{3\varrho+1}$, we have $\frac{4\varrho}{\beta(3\varrho+1)} < \frac{2}{\alpha\left(\sqrt{\varrho}+1\right)}$.

When $\varepsilon = \frac{4\varrho}{\beta(3\varrho+1)}$, we have $1-\varepsilon\alpha = 1 - \frac{4\varrho\alpha}{\beta(3\varrho+1)} = b_1$. We can verify that $\gamma_1^* = \gamma_1(b_1)$, $\eta_1^* = \frac{1-\gamma_1^*}{\mu}$ and ε satisfies the inequalities of (11a),(11b),(12a),(12b). $\gamma_1(b_1)$ is a suboptimal convergence rate.

When $\varepsilon = \frac{1}{\alpha}\frac{2}{\sqrt{\varrho}+1}$, we have $\varepsilon\beta - 1 = \frac{\beta}{\alpha}\frac{2}{\sqrt{\varrho}+1} - 1 = b_2$. We can verify that $\gamma_2^* = \gamma_2(b_2)$, $\eta_2^* = \frac{1-\gamma_2^*}{\mu}$ and ε satisfies the inequalities of (11a),(11b),(12a),(12b). $\gamma_2(b_2)$ is a suboptimal convergence rate. Similarly, we can verify that $\gamma_3^*, \eta_2^*, \varepsilon_3^*$ also satisfy the inequalities of (11a),(11b),(12a),(12b). $\gamma_3(b_3)$ is a suboptimal convergence rate.

A.6 Necessary Condition and Derivation of the Lower Bound in Theorem 4

Since the original DIGing-ATC is a special case of $\varepsilon = 1$ and W is a doubly stochastic matrix, we have $0 < \alpha < \beta < 2$. Substitute $\varepsilon = 1$ into (11a),(11b),(12a),(12b), we have

$$\alpha \geqslant 1 - \frac{\gamma}{\sqrt{\varrho}\,(1-\gamma)+1}, \tag{39a}$$

$$\beta \leqslant 1 + \frac{\gamma}{\sqrt{\varrho}\,(1-\gamma^2)+1}. \tag{39b}$$

Therefore, the convergence rate of CGD can be achieved if and only if both inequalities hold. Substituting $\gamma_c = \frac{\varrho-1}{\varrho+1}$ into (39a) and (39b) yields that $\beta \leqslant \frac{4\varrho}{3\varrho+1}$, $\alpha \geqslant \frac{\sqrt{\varrho}+1}{2}$. Following the proof of Theorem 2, we obtain that $\gamma^* = \max\{\gamma_1(1-\alpha), \gamma_2(\beta-1)\}$, where $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ are defined in (7).

A.7 Verifying the Attainability of the Lower Bound via Small Gain Theorem

Next we show that the Lur'e system shown in Figure 5, with $G(z)$ and parameters η^*, ε^*, is γ^*-stable for all $\tilde{\Delta}_{\boldsymbol{f}}$, $\boldsymbol{f} \in \mathcal{F}_{\mu,L}$, by using the small gain theorem. Note that $\|\Delta_{loop}\|_\infty = \left\|\tilde{\Delta}_{\boldsymbol{f}} - \frac{L+\mu}{2}\right\|_\infty \leqslant \frac{L-\mu}{2}$ follows immediately from $\tilde{\Delta}_{\boldsymbol{f}}$ being sector bounded by $[\mu, L]$.

Substituting γ^*, ε^* and η^* into (17)-(18), we have

$$H_1(z) = \frac{-\eta^*}{z - 1 + \frac{\mu+L}{2}\eta^*}, \tag{40}$$

$$H_i(z) = \frac{-\eta^*(z-1)(1-\varepsilon^*\lambda_i)^2}{(z-1+\varepsilon^*\lambda_i)^2 + \frac{\mu+L}{2}\eta^*(z-1)(1-\varepsilon^*\lambda_i)^2},\ i = 2, \cdots, N. \tag{41}$$

The following lemma gives the norm of $H(\gamma z)$.

Lemma 6 *For $H(z)$ defined in Lemma 5, we have*

$$\|H(\gamma^* z)\|_\infty = \max_{\substack{i\in\{1,\cdots,N\},\theta\in[0,2\pi),\\ \lambda_2,\cdots,\lambda_N\in[\alpha,\beta]}} \left|H_i\left(\gamma^* e^{j\theta}\right)\right| = \frac{2}{L-\mu}.$$

Proof. *By direct computation, we have*

$$\|H_1(\gamma^* z)\|_\infty = \max_\theta \left|\frac{\eta^*}{\gamma^* e^{j\theta} - 1 + \frac{\mu+L}{2}\eta^*}\right| = \frac{\eta^*}{\gamma^* - 1 + \frac{\mu+L}{2}\eta^*} = \frac{2}{L-\mu}. \tag{42}$$

The last equality comes from $\eta^* = \frac{1-\gamma^*}{\mu}$.

$$H_i(\gamma^* z) = \frac{-\eta^*(\gamma^* z - 1)(1-\varepsilon^*\lambda_i)^2}{(\gamma^* z - 1 + \varepsilon^*\lambda_i)^2 + \frac{\mu+L}{2}\eta^*(\gamma^* z - 1)(1-\varepsilon^*\lambda_i)^2},\ i = 2, \cdots, N.$$

For notation simplicity, we omit the '$*$' in γ^*, ε^* and η^* from here till the end of the proof without causing confusion.

Write $H_i(\gamma z)$, $i = 2, \cdots, N$ as follows

$$H_i(\gamma z) = \frac{-\frac{2}{\mu+L}}{1 + \frac{2}{\mu+L}\frac{1}{\eta(\gamma z-1)}\frac{(\gamma z-1+\varepsilon\lambda_i)^2}{(1-\varepsilon\lambda_i)^2}}.$$

Denote

$$h(z,\lambda) = 1 + \frac{2}{\mu+L}\frac{1}{\eta(\gamma z-1)}\frac{(\gamma z-1+\varepsilon\lambda)^2}{(1-\varepsilon\lambda)^2}, \lambda \in [\alpha,\beta].$$

Obviously,

$$\max_{\substack{\theta\in[0,2\pi)\\ \lambda_i\in[\alpha,\beta]}} \left|H_i\left(\gamma e^{j\theta}\right)\right| = \frac{2}{\mu+L}\Bigg/ \min_{\substack{\theta\in[0,2\pi)\\ \lambda\in[\alpha,\beta]}} \left|h\left(e^{j\theta},\lambda\right)\right|. \tag{43}$$

Obviously,

$$h\left(e^{j\theta},\lambda\right) = \mathrm{Re}\left\{h\left(e^{j\theta},\lambda\right)\right\} + j\mathrm{Im}\left\{h\left(e^{j\theta},\lambda\right)\right\},$$

where

$$\mathrm{Re}\left\{h\left(e^{j\theta},\lambda\right)\right\} = 1+\frac{2}{(\mu+L)\eta(1-\varepsilon\lambda)^2}\left[2\varepsilon\lambda+\frac{(\gamma\cos\theta-1)\left(\gamma^2-2\gamma\cos\theta+1+\varepsilon^2\lambda^2\right)}{\gamma^2-2\gamma\cos\theta+1}\right],$$

$$\mathrm{Im}\left\{h\left(e^{j\theta},\lambda\right)\right\} = \frac{2}{(\mu+L)\eta(1-\varepsilon\lambda)^2}\cdot\frac{\gamma\sin\theta\left(\gamma^2-2\gamma\cos\theta+1-\varepsilon^2\lambda^2\right)}{\gamma^2-2\gamma\cos\theta+1}.$$

Note that

$$\left|h\left(e^{j\theta},\lambda\right)\right|^2 = \mathrm{Re}^2\left\{h\left(e^{j\theta},\lambda\right)\right\} + \mathrm{Im}^2\left\{h\left(e^{j\theta},\lambda\right)\right\}. \tag{44}$$

Recall that $\eta=\frac{1-\gamma}{\mu}$, $\varepsilon\alpha=\frac{\sqrt{\frac{L}{\mu}}(1-\gamma)+(1-\gamma)}{\sqrt{\frac{L}{\mu}}(1-\gamma)+1}$, $\varepsilon\beta=\frac{1+\gamma+\sqrt{\frac{L}{\mu}(1-\gamma^2)}}{1+\sqrt{\frac{L}{\mu}(1-\gamma^2)}}$. *Since* $-\gamma<\varepsilon\alpha-1\leqslant\varepsilon\lambda-1\leqslant\varepsilon\beta-1<\gamma$, *we can show that*

$$\min_{\substack{\lambda\in[\alpha,\beta]\\ \theta\in[0,2\pi)}} \left|h\left(e^{j\theta},\lambda\right)\right| = \min\left\{\left|h\left(e^{j0},\alpha\right)\right|,\left|h\left(e^{j\pi},\beta\right)\right|\right\}.$$

Therefore, we have

$$h\left(e^{j0},\alpha\right) = 1+\frac{2}{\mu+L}\frac{1}{\eta\left(\gamma-1\right)}\frac{L}{\mu}\left(1-\gamma\right)^2 = \frac{\mu-L}{L+\mu},$$
$$h\left(e^{j\pi},\beta\right) = 1-\frac{2}{\mu+L}\frac{1}{\eta\left(\gamma+1\right)}\frac{L}{\mu}\left(1-\gamma^2\right) = \frac{\mu-L}{L+\mu}.$$

Hence, $\left|h\left(e^{j0},\alpha\right)\right| = \left|h\left(e^{j\pi},\beta\right)\right| = \frac{L-\mu}{L+\mu}$, *and*

$$\min_{\substack{\lambda\in[\alpha,\beta]\\ \theta\in[0,\pi]}} \left|h\left(e^{j\theta},\lambda\right)\right| = \frac{L-\mu}{L+\mu}. \tag{45}$$

Substituting (45) to (43) yields

$$\max_{\substack{\theta\in[0,2\pi)\\ \lambda_i\in[\alpha,\beta]}} \left|H_i\left(\gamma e^{j\theta}\right)\right| = \frac{2}{\mu+L}\Bigg/\frac{L-\mu}{L+\mu} = \frac{2}{L-\mu}. \tag{46}$$

The proof is completed by combining (42) and (46).

Hence, $\|H\left(\gamma^* z\right)\|_\infty \|\Delta_{loop}\|_\infty \leqslant 1$. The γ^*-stability of the Lur'e system follows from the small gain theorem.

B Sector Bound of $\tilde{\Delta}(\cdot)$ and Boundedness of the Operator Δ_{loop}

For $\boldsymbol{f}(\cdot)$ being μ-strongly convex and L-smooth, Theorem 2.1.11 in Nesterov et al. (2018) yields that

$$\left\langle \nabla\boldsymbol{f}\left(\tilde{\boldsymbol{x}}\left(k+1\right)\right)-\nabla\boldsymbol{f}\left(\tilde{\boldsymbol{x}}\left(k\right)\right), \tilde{\boldsymbol{x}}\left(k+1\right)-\tilde{\boldsymbol{x}}\left(k\right)\right\rangle \geqslant$$
$$\frac{\mu L}{\mu+L}\left\|\tilde{\boldsymbol{x}}\left(k+1\right)-\tilde{\boldsymbol{x}}\left(k\right)\right\|_2^2+\frac{1}{\mu+L}\left\|\nabla\boldsymbol{f}\left(\tilde{\boldsymbol{x}}\left(k+1\right)\right)-\nabla\boldsymbol{f}\left(\tilde{\boldsymbol{x}}\left(k\right)\right)\right\|_2^2,$$

which is obtained by the fact that unitary transformation does not change the inner-product of vectors. Recall that $\tilde{\boldsymbol{u}}(k) := \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k)) - \nabla \boldsymbol{f}(\tilde{\boldsymbol{x}}(k+1))$, $\tilde{\boldsymbol{v}}(k) := \tilde{\boldsymbol{x}}(k) - \tilde{\boldsymbol{x}}(k+1)$, we have

$$\|\tilde{\boldsymbol{u}}(k)\|_2^2 - (\mu + L)\langle \tilde{\boldsymbol{u}}(k), \tilde{\boldsymbol{v}}(k)\rangle + \mu L \|\tilde{\boldsymbol{v}}(k)\|_2^2 \leqslant 0.$$

It is equivalent to

$$(\tilde{\boldsymbol{u}}(k) - \mu\tilde{\boldsymbol{v}}(k))^T (\tilde{\boldsymbol{u}}(k) - L\tilde{\boldsymbol{v}}(k)) \leqslant 0. \tag{47}$$

Hence the nonlinear operator $\tilde{\Delta}(\cdot)$ is sector-bounded in $[\mu, L]$. The formula (47) can be formulated as

$$\left[\left(\tilde{\boldsymbol{u}}(k) - \frac{L+\mu}{2}\tilde{\boldsymbol{v}}(k)\right) - \frac{\mu - L}{2}\tilde{\boldsymbol{v}}(k)\right]^T \left[\left(\tilde{\boldsymbol{u}}(k) - \frac{L+\mu}{2}\tilde{\boldsymbol{v}}(k)\right) - \frac{L-\mu}{2}\tilde{\boldsymbol{v}}(k)\right] \leqslant 0. \tag{48}$$

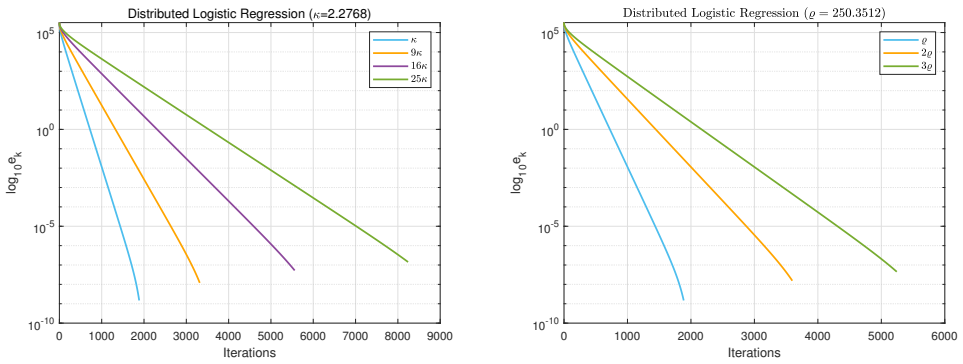
Formula (48) indicates that the nonlinear operator $\Delta_{loop}(\cdot)$ is sector-bounded in $\left[-\frac{L-\mu}{2}, \frac{L-\mu}{2}\right]$. Thus, we obtain $\sup_{\|\tilde{\boldsymbol{v}}(k)\|_2 > 0} \frac{\|\Delta_{loop}(\tilde{\boldsymbol{v}}(k))\|_2}{\|\tilde{\boldsymbol{v}}(k)\|_2} \leqslant \frac{L-\mu}{2}$, which is given in Lessard (2022).

C Additional Experiments

In this section, we present several additional experiments. Consider the logistic regression objective function (14). The first experiment is shown in Section 5.

The full $a9a$ dataset contains 32,561 samples, comprising 24,720 negative samples and 7,841 positive samples, each with 123 features. These features are normalized to unit vectors. We select a subset of the a9a dataset, using the first 7,500 samples, which include 5,696 negative samples and 1,804 positive samples.

In the second experiment, we show the robustness of the DIGing-ATC algorithm to the estimation of parameters κ and ϱ. We conduct experiments on the randomly generated graph with a Metropolis-Hastings communication matrix and $\kappa = 2.2768$. By multiplying β by 3, 4, and 5 while dividing α by 3, 4, and 5, we obtained estimates of 9κ, 16κ, and 25κ. Similarly, by multiplying L by 2 and 3, we obtained estimates of 2ϱ and 3ϱ. Results are shown in Figure 6. The results demonstrate that the algorithm exhibits strong robustness to parameter estimation.



(a) (b)

Figure 6: Estimation robustness of κ and ϱ.

In the third experiment, we show the effectiveness of ε-weight adjustment for the communication matrix. To this end, we separately compare the iteration counts of the two distributed acceleration algorithms on the original doubly stochastic communication matrix and the ε-weight-adjusted communication matrix. We use $(1-\varepsilon)I + \varepsilon W$ as the ε-weight-adjusted communication matrix, where the value of ε is computed by the optimal parameter design of DIGing-ATC in Theorem 1-2. We consider three randomly generated 50-node graphs with different connectivity properties, with $\kappa = 49.0211, 25.4885, 2.2768$. Unless otherwise stated, the step size of the acceleration algorithm is set to $\frac{0.05}{L}$ in all subsequent experiments. Results are shown in Figure 7(a), 7(b), 7(c).

In Figure 7(a) and 7(b), the results show that both the two distributed acceleration algorithms converge faster on the communication network with ε-weight adjustment, especially when the original doubly-stochastic network has a poor connectivity. In Figure 7(c), this result indicates that ε-weight adjustment may have no effect on graphs with good connectivity. Note that although the value of ε obtained from the parameter design for the DIGing-ATC algorithm may not be optimal for the acceleration algorithms, the experimental results confirm the effectiveness of adjusting the weights of the communication matrix. This motivates future research on parameter design for distributed acceleration algorithms.

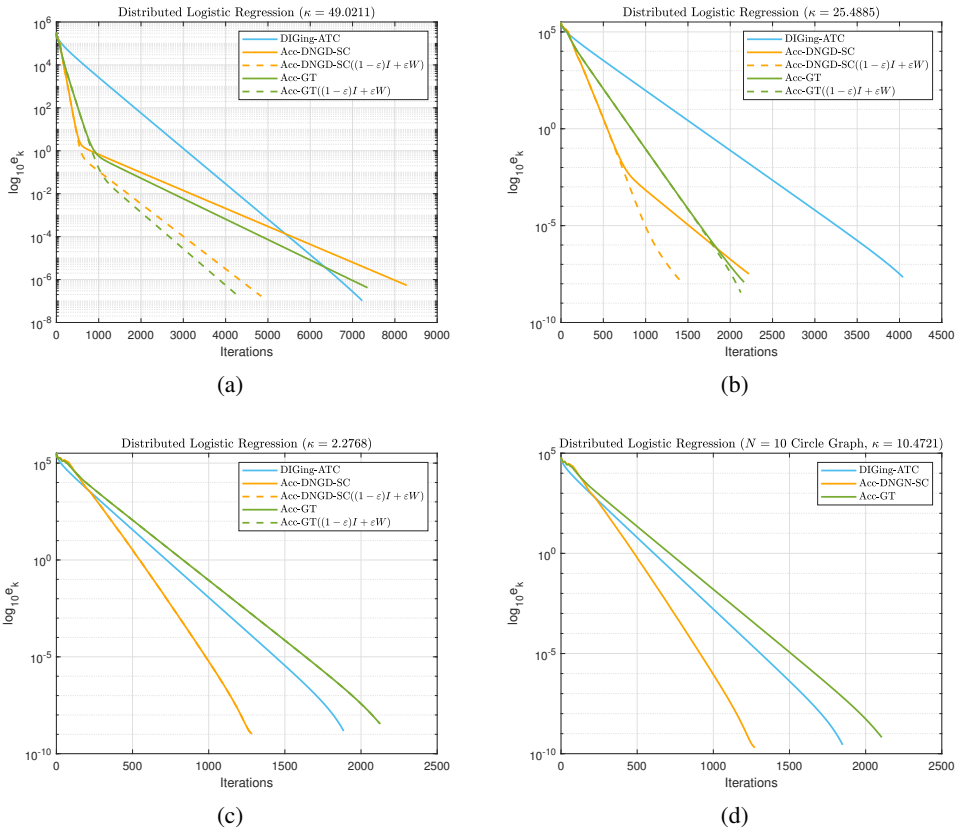


Figure 7: Results of ε-weight adjustment effectiveness (the third experiment) and ε-weight-adjusted $N = 10$ undirected circle graph (the final experiment).

In the final experiment, we demonstrate that ε-weight adjustment relaxes the requirements on communication networks: they need not be doubly stochastic, but only undirected and connected. Moreover, it is also applicable to the class of special doubly stochastic matrices that are typically excluded under the spectral radius assumption $\rho\left(W - \frac{\mathbf{1}\mathbf{1}^T}{N}\right) < 1$ Xiao & Boyd (2004). This assumption excludes undirected cycle graphs with an even number of nodes (and no self-loops), for which all diagonal entries of the adjacency matrix are zero Huang et al. (2026). To demonstrate this, we consider the simple undirected circle graph (without self-loops) of $N = 10$ nodes. All nodes have degree 2, and all edge weights are 0.5. It is doubly stochastic, i.e., all row and column sums are 1. But all diagonal entries of the communication matrix are zero. Thus, distributed algorithms based on the assumption that the spectral radius is less than 1 would diverge on such undirected circle graphs, including Acc-DNGD-SC and Acc-GT. However, we show that by using the ε-weight-adjusted communication matrix, the two acceleration algorithms could converge. Both step sizes are set as $\frac{0.05}{L}$. Results are shown in Figure 7(d). Given that undirected circle graphs are common and widely used, the results demonstrate the importance of ε-weight adjustment and the broader applicability of the algorithm across networks.