

Article

DARN: Distributed Adaptive Regularized Optimization with Consensus for Non-Convex Non-Smooth Composite Problems

Cunlin Li [†]  and Yinpu Ma ^{*,†} 

School of Mathematics and Information Science, Northern Minzu University, Yinchuan 750021, China; biticl@163.com

* Correspondence: mayinpu@163.com; Tel.: +86-152-0960-2226

[†] These authors contributed equally to this work.

Abstract

This paper proposes a Distributed Adaptive Regularization Algorithm (DARN) for solving composite non-convex and non-smooth optimization problems in multi-agent systems. The algorithm employs a three-phase iterative framework to achieve efficient collaborative optimization: (1) a local regularized optimization step, which utilizes proximal mappings to enforce strong convexity of weakly convex objectives and ensure subproblem well-posedness; (2) a consensus update based on doubly stochastic matrices, guaranteeing asymptotic convergence of agent states to a global consensus point; and (3) an innovative adaptive regularization mechanism that dynamically adjusts regularization strength using local function value variations to balance stability and convergence speed. Theoretical analysis demonstrates that the algorithm maintains strict monotonic descent under non-convex and non-smooth conditions by constructing a mixed time-scale Lyapunov function, achieving a sublinear convergence rate. Notably, we prove that the projection-based update rule for regularization parameters preserves lower-bound constraints, while spectral decay properties of consensus errors and perturbations from local updates are globally governed by the Lyapunov function. Numerical experiments validate the algorithm's superiority in sparse principal component analysis and robust matrix completion tasks, showing a 6.6% improvement in convergence speed and a 51.7% reduction in consensus error compared to fixed-regularization methods. This work provides theoretical guarantees and an efficient framework for distributed non-convex optimization in heterogeneous networks.

Keywords: distributed optimization; adaptive regularization; non-convex and non-smooth optimization; consensus algorithms; Lyapunov convergence; proximal gradient methods



Academic Editor: Wei-Shih Du

Received: 7 June 2025

Revised: 9 July 2025

Accepted: 14 July 2025

Published: 20 July 2025

Citation: Li, C.; Ma, Y. DARN: Distributed Adaptive Regularized Optimization with Consensus for Non-Convex Non-Smooth Composite Problems. *Symmetry* **2025**, *17*, 1159. <https://doi.org/10.3390/sym17071159>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Distributed optimization over multi-agent networks has emerged as a fundamental paradigm for solving large-scale problems in machine learning, signal processing, and control systems, where data privacy, communication efficiency, and scalability are critical concerns [1–3]. The symmetry principle plays a pivotal role in designing consensus mechanisms, where doubly stochastic weight matrices enforce symmetric information exchange among agents to balance local computational autonomy with global coordination. A central challenge lies in designing algorithms that reconcile non-smooth composite objectives—ubiquitous in sparse recovery, robust estimation, and deep learning—with the constraints of decentralized computation and time-varying network topologies. While

significant progress has been made in convex settings, extension to non-convex and non-smooth problems remains theoretically intricate and practically demanding, particularly under heterogeneous network conditions. Recent advances in decentralized proximal methods, such as PG-EXTRA [4] and its variants [5–7], have demonstrated the potential of exploiting composite structures through gradient-proximal splitting techniques. These methods achieve exact convergence with fixed step sizes by leveraging double-stochastic consensus protocols; yet, their performance remains constrained by network-dependent step size bounds and limited adaptability to non-convex landscapes. Parallel developments in continuous-time Distributed Gradient Descent (DGD) [8] reveal intrinsic tradeoffs between consensus convergence rates and centralized optimization dynamics, often resulting in suboptimal synchronization under time-varying topologies. Meanwhile, state-of-the-art approaches for non-convex optimization, exemplified by Distributed Proximal Gradient (DPG) algorithms [9], address time-varying networks through increasing consensus rounds but suffer from computational inefficiency due to inexact proximal approximations and diminishing step-size requirements.

Notably, existing methods primarily target undirected graphs. Although directed graph settings are more challenging, recent works such as [10] have proposed distributed robust optimization frameworks for networked systems with unknown nonlinearities, while [11] developed randomized constraint-solving algorithms for unbalanced time-varying digraphs. These advances partially address graph asymmetry via techniques such as gradient tracking and row-stochastic matrices, although non-convex non-smooth composite problems remain open.

Three fundamental limitations persist in existing frameworks:

1. **Network dependency:** Step size selection in gradient-proximal methods often requires global knowledge of network spectral properties [4–6], limiting scalability in heterogeneous environments.
2. **Non-convex–non-smooth coupling:** Current analyses for composite objectives predominantly rely on convexity or Kurdyka–Łojasiewicz assumptions [12], failing to guarantee monotonic descent in general non-convex settings with non-smooth regularizers.
3. **Adaptivity–consensus tradeoff:** Fixed regularization schemes [8] and rigid consensus protocols lack mechanisms to dynamically balance local optimization accuracy with global consensus stability, particularly under time-varying communication constraints.

Theoretical contributions transcend conventional convex-analytic approaches by establishing [13]:

- **Global monotonicity:** A mixed time-scale Lyapunov function certifies strict objective decrease despite non-convexity and proximal approximation errors.
- **Network-agnostic convergence:** The sublinear rate $\mathcal{O}(1/\sqrt{k})$ is proven to be independent of spectral graph properties, resolving prior dependencies on Laplacian eigenvalues.
- **Critical point consensus:** Agent states provably converge to a common critical set without requiring diminishing step sizes or gradient tracking.

Through numerical validations on sparse PCA and robust matrix completion, this work bridges the gap between adaptive regularization theory and decentralized non-convex optimization, offering a unified framework for large-scale composite problems in dynamically evolving networks.

The composite structure $f_i + g$ introduces two fundamental difficulties: (1) **Multi-modality induced by non-convexity:** Local objectives may contain numerous saddle points and suboptimal local minima, causing gradient-based methods to stagnate at non-critical regions. (2) **Non-smoothness-gradient incoherence:** The absence of subgradient bound-

edness ($\|\partial f_i(x)\| \not\leq L$) invalidates classic descent lemmas, while the non-coincidence of $\partial(f_i + g)$ and $\partial f_i + \partial g$ disrupts optimality analysis.

Decentralization exacerbates these issues through:

- **Consensus–optimization conflict:** Non-smooth terms induce $\mathcal{O}(1)$ consensus error under fixed step sizes (see Lemma 4), conflicting with optimization precision requirements.
- **Heterogeneous landscape misalignment:** When $\mu_i \neq \mu_j$, local strong convexity parameters diverge, preventing synchronous convergence.
- **Subgradient communication incompleteness:** Transmitting ∂f_i requires $\mathcal{O}(d^2)$ bandwidth for matrix-valued variables, becoming prohibitive for large-scale problems.

These challenges collectively necessitate: (i) A mechanism to bypass unbounded subgradients (solved via Moreau smoothing in Section 3.1). (ii) Time scale decoupling for consensus and optimization (achieved by mixed Lyapunov analysis in Theorem 1). (iii) Adaptive regularization to reconcile μ_i -heterogeneity (proposed in Section 2.2).

2. Problem Formulation and Algorithm Design

2.1. Network Model and Objective Function

Consider a network of n agents over an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \dots, n\}$ and where $\mathcal{E}$ denotes the edge set; each agent i holds a local objective $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$. The global optimization problem is then

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) + g(x) \right\}.$$

Assumptions:

1. f_i is L_i -Lipschitz continuous (possibly non-convex and non-smooth).
2. $g(x)$ is L_g -Lipschitz continuous (possibly non-convex and non-smooth).
3. Agents communicate via a doubly stochastic adjacency matrix $W = [w_{ij}] \in \mathbb{R}^{n \times n}$ satisfying $W\mathbf{1} = \mathbf{1}$, $\mathbf{1}^\top W = \mathbf{1}^\top$ and $w_{ij} > 0$ iff $(i, j) \in \mathcal{E}$.

2.2. Distributed Adaptive Regularization Algorithm (DARN)

The complete procedure is formalized in Algorithm 1.

Initialization: Each agent i initializes $x_i^{(0)} \in \mathbb{R}^d, \lambda_i^{(0)} > 0$ and weights w_{ij} .

Iteration Steps (at step k):

1. Local Regularized Optimization:

$$x_i^{(k+\frac{1}{2})} = \arg \min_{x \in \mathbb{R}^d} \left\{ f_i(x) + \frac{\lambda_i^{(k)}}{2} \|x - x_i^{(k)}\|^2 \right\}.$$

This proximal regularization ensures stability of local solutions.

2. Consensus Update:

$$x_i^{(k+1)} = \sum_{j=1}^n w_{ij} x_j^{(k+\frac{1}{2})}.$$

This achieves aggregation of local variables towards a global consensus point through weighted averaging.

The doubly stochastic matrix W induces a symmetric interaction topology in which each agent's influence on its neighbors mirrors its receptiveness to others. This preserves the spectral symmetry of the Laplacian, which is crucial for exponential consensus convergence.

3. Adaptive Regularization Tuning:

$$\lambda_i^{(k+1)} = P_{[\lambda_{\min}, \lambda_{\max}]}(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}}(f_i(x_i^{(k+\frac{1}{2})}) - f_i(x_i^{(k)}))).$$

This dynamically adjusts the regularization strength based on local function value changes, where $\eta_i^{(k)} = \frac{\gamma}{L_i \lambda_i^{(k)}}$, with $\gamma > 0$ as the adjustment parameter.

2.3. Mathematical Details and Rationale

Proposition 1. (Strong Convexity of the Regularized Objective)

If each function f_i is (μ_i, L_i) -weakly convex and $\lambda_i^{(k)} > \mu_i$, then the function

$$h_i(x) = f_i(x) + \frac{\lambda_i^{(k)}}{2} \|x - x_i^{(k)}\|^2$$

is $(\lambda_i^{(k)} - \mu_i)$ -strongly convex. Consequently, there exists a unique minimizer $x_i^{(k+\frac{1}{2})}$.

Proof. Definition of Weak Convexity.

A function $f_i(x) + \frac{\mu_i}{2} \|x\|^2$ is convex. Therefore, for any $x, y \in \mathbb{R}^d$ and $\theta \in [0, 1]$, we have

$$\begin{aligned} & f_i(\theta x + (1 - \theta)y) + \frac{\mu_i}{2} \|\theta x + (1 - \theta)y\|^2 \leq \\ & \theta \left(f_i(x) + \frac{\mu_i}{2} \|x\|^2 \right) + (1 - \theta) \left(f_i(y) + \frac{\mu_i}{2} \|y\|^2 \right). \end{aligned}$$

Expansion and regularization term.

Consider the function $h_i(x)$:

$$h_i(x) = f_i(x) + \frac{\lambda_i^{(k)}}{2} \|x - x_i^{(k)}\|^2.$$

Then, expand the squared norm term:

$$\|x - x_i^{(k)}\|^2 = \|x\|^2 - 2x_i^{(k)\top} x + \|x_i^{(k)}\|^2.$$

Substituting into $h_i(x)$ yields:

$$h_i(x) = f_i(x) + \frac{\lambda_i^{(k)}}{2} \|x\|^2 - \lambda_i^{(k)} x_i^{(k)\top} x + \frac{\lambda_i^{(k)}}{2} \|x_i^{(k)}\|^2.$$

Strong Convexity Verification.

Because $f_i(x) + \frac{\mu_i}{2} \|x\|^2$ is convex (by weak convexity) and $\frac{\lambda_i^{(k)}}{2} \|x\|^2$ is strongly convex with parameter $\lambda_i^{(k)}$, the sum $h_i(x)$ is strongly convex.

The strong convexity parameter of $h_i(x)$ is $\lambda_i^{(k)} - \mu_i$. This follows from the fact that the quadratic term $\frac{\lambda_i^{(k)} - \mu_i}{2} \|x\|^2$ dominates the convex function $f_i(x) + \frac{\mu_i}{2} \|x\|^2$.

Uniqueness of Minimizer:

Strong convexity guarantees that $h_i(x)$ has a unique minimizer $x_i^{(k+\frac{1}{2})}$ satisfying

$$h_i(x_i^{(k+\frac{1}{2})}) = \min_{x \in \mathbb{R}^d} h_i(x).$$

□

1. Well-Posedness of Local Optimization.

For non-convex and non-smooth functions f_i , the addition of the regularization term $\frac{\lambda_i^{(k)} - \mu_i}{2} \|x\|^2$ ensures the existence and uniqueness of a solution to the local subproblem. Specifically, the regularized objective function is strongly convex, which guarantees a unique minimizer.

2. Convergence of Consensus Update.

Let $\mathbf{x}^{(k)} = [\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_n^{(k)}]^\top \in \mathbb{R}^n \times d$. Then, the consensus step can be written in matrix form as follows:

$$\mathbf{x}^{(k+1)} = W\mathbf{x}^{(k+\frac{1}{2})}.$$

Because W is doubly stochastic and the graph is connected, by the Perron–Frobenius theorem, repeated application of W will drive $\mathbf{x}^{(k)}$ towards consensus, that is:

$$\lim_{k \rightarrow \infty} \|x_i^{(k)} - \bar{x}^{(k)}\| = 0, \quad \text{where} \quad \bar{x}^{(k)} = \frac{1}{n} \sum_{i=1}^n x_i^{(k)}.$$

3. Adaptive Regularization Adjustment Mechanism.

Define the local function value decrease as follows:

$$\Delta_i^{(k)} = f_i(x_i^{(k+\frac{1}{2})}) - f_i(x_i^{(k)}).$$

If $\Delta_i^{(k)} < 0$ (function value decreases), increase $\lambda_i^{(k+1)}$ to strengthen regularization and suppress oscillations; otherwise, decrease λ_i . By setting $\eta_i^{(k)} \propto 1/(L_i \lambda_i^{(k)})$, the adjustment ensures that the update is inversely proportional to the local Lipschitz constant, balancing the heterogeneity among agents.

2.4. Key Lemma and Convergence Support

Lemma 1. (Descent of Local Solutions): For any $k \geq 0$ and agent i , the following holds:

$$f_i(x_i^{(k+\frac{1}{2})}) + \frac{\lambda_i^{(k)}}{2} \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \leq f_i(x_i^{(k)}).$$

Proof. By the definition in Proposition 1, $\mathbf{x}_i^{(k+\frac{1}{2})}$ is a minimizer. Substituting directly, we obtain

$$f_i(\mathbf{x}_i^{(k+\frac{1}{2})}) + \frac{\lambda_i^{(k)}}{2} \|\mathbf{x}_i^{(k+\frac{1}{2})} - \mathbf{x}_i^{(k)}\|^2 \leq f_i(\mathbf{x}_i^{(k)}) + \frac{\lambda_i^{(k)}}{2} \|\mathbf{x}_i^{(k)} - \mathbf{x}_i^{(k)}\|^2 = f_i(\mathbf{x}_i^{(k)}).$$

□

Lemma 2. (Consensus Error Decay): There exists a constant $\rho \in (0, 1)$ such that

$$\|\mathbf{x}^{(k+1)} - \mathbf{1}\bar{x}^{(k+1)}\| \leq \rho \|\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\|.$$

Proof. Due to the spectral properties of W , its second largest eigenvalue $\sigma_2(W) < 1$. Therefore, we can choose $\rho = \sigma_2(W)$. Consequently, we have

$$\|\mathbf{x}^{(k+1)} - \mathbf{1}\bar{x}^{(k+1)}\| = \|W\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\|.$$

Utilizing the property of the matrix W and its eigenvalues, specifically that $\sigma_2(W)$ governs the convergence rate of consensus algorithms, we can bound the norm as follows:

$$\|W\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\| \leq \sigma_2(W) \|\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\|$$

where $\bar{x}^{(k+\frac{1}{2})}$ denotes the average of the elements in $x^{(k+\frac{1}{2})}$.

Given our choice of $\rho = \sigma_2(W)$, we obtain

$$\|x^{(k+1)} - \mathbf{1}\bar{x}^{(k+1)}\| \leq \rho \|x^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\|,$$

which completes the proof. $\square$

2.5. Mathematical Description of the Algorithm Pseudocode

Input: Initial states $x_i(0)$ for all agents, $\lambda_i(0) > 0$, adjacency matrix W , and parameter $\gamma > 0$.

Algorithm 1 Distributed Adaptive Regularization Algorithm (DARN)

- 1: **Input:** Initial states $x_i(0)$, $\lambda_i(0) > 0$, adjacency matrix W , parameter $\gamma > 0$
 - 2: **for** $k = 0, 1, 2, \dots$ **do**
 - 3: $x_i^{(k+\frac{1}{2})} \leftarrow \arg \min_x \left\{ f_i(x) + \frac{\lambda_i^{(k)}}{2} \|x - x_i^{(k)}\|^2 \right\}$ $\triangleright$ Local optimization
 - 4: $x_i^{(k+1)} \leftarrow \sum_{j=1}^n w_{ij} x_j^{(k+\frac{1}{2})}$ $\triangleright$ Consensus update
 - 5: $\lambda_i^{(k+1)} \leftarrow P_{[\lambda_{\min}, \lambda_{\max}]} \left(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \left(f_i(x_i^{(k+\frac{1}{2})}) - f_i(x_i^{(k)}) \right) \right)$
 - 6: **end for**
-

Remark 1. (Consensus for Directed Graphs): For directed communication topologies, DARN can employ the push-sum [14] protocol as an alternative:

$$\begin{aligned} x_i^{(k+1)} &= \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} x_j^{(k+\frac{1}{2})} \\ \phi_i^{(k+1)} &= \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \phi_j^{(k)} \\ z_i^{(k+1)} &= x_i^{(k+1)} / \phi_i^{(k+1)} \end{aligned}$$

where $A = [a_{ij}]$ is row-stochastic. This extension enables DARN operation in unidirectional networks but reduces convergence to the sublinear rate $\mathcal{O}(1/k)$.

2.6. Supplementary Remarks

1. **Handling Non-Smooth Terms:** If f_i and g_i contain non-smooth terms (e.g., l_1 -norm), a proximal operator can be introduced in the local optimization step. Specifically, when $f_i(x) = h_i(x) + r_i(x)$, where h_i is smooth and r_i is non-smooth, the local step is modified as follows:

$$x_i^{(k+\frac{1}{2})} = \text{prox}_{r_i, \lambda_i^{(k)}} \left(x_i^{(k)} - \frac{1}{\lambda_i^{(k)}} \nabla h_i(x_i^{(k)}) \right)$$

where the proximal operator is defined as

$$\text{prox}_{r, \lambda}(y) = \arg \min_x \{ r(x) + \frac{\lambda}{2} \|x - y\|^2 \}.$$

2. **Integration of Global Term $g(x)$:** Because $g(x)$ is a global term, it cannot be directly handled in the local optimization step. Instead, it is implicitly optimized through the consensus step, where each x_i converges to a common point x that minimizes $\frac{1}{n} \sum f_i(x) + g(x)$.

3. Constrained Optimization Extension: DARN extends to constrained problems via

$$\text{Local step: } x_i^{(k+\frac{1}{2})} = \arg \min_{x \in \mathcal{X}_i} \left\{ f_i(x) + \frac{\lambda_i^{(k)}}{2} \|x - x_i^{(k)}\|^2 \right\}$$

$$\text{Global constraints: } x_i^{(k+1)} = \text{proj}_{\mathcal{C}} \left(\sum_j w_{ij} x_j^{(k+\frac{1}{2})} \right)$$

$$\lambda_i^{(k+1)} = \mathcal{P} \left(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} (\Delta f_i^{(k)} - \beta \cdot \|x_i^{(k)} - \text{proj}_{\mathcal{C}}(x_i^{(k)})\|) \right).$$

Projection non-expansiveness preserves strong convexity, while the extended Lyapunov function ensures constraint violation convergence.

3. Convergence Theory Analysis for Distributed Adaptive Regularization Algorithm (DARN)

3.1. Convergence Theory Analysis

Critical Point Quality Guarantee: DARN ensures convergence to high-quality critical points via dual mechanisms:

- (1) Strict monotonic descent of the mixed time-scale Lyapunov function $\Phi^{(k)}$ (Theorem 1) avoids high-value saddles.
- (2) Adaptive regularization guides optimization path through λ_i tuning:

$$\begin{aligned} \lambda_i \uparrow & \text{ when } \Delta f_i^{(k)} \ll 0 \quad (\text{stabilize updates during rapid descent}) \\ \lambda_i \downarrow & \text{ when } \Delta f_i^{(k)} \approx 0 \quad (\text{refine solutions near stationarity}) \end{aligned}$$

which prioritizes convergence to deep minima, supplemented by multi-start initialization for enhanced robustness.

Assumptions:

1. Weakly Convex-Lipschitz Function Class: Each $f_i(x)$ is a (μ_i, L_i) -weakly convex-Lipschitz function [15], i.e.,

$$f_i(x) + \frac{\mu_i}{2} \|x\|^2$$

is convex:

$$\|f_i(x) - f_i(y)\| \leq L_i \|x - y\| \text{ for all } x, y \in \mathbb{R}^d.$$

2. Communication Graph and Mixing Matrix: The adjacency matrix W is doubly stochastic and satisfies the spectral gap condition

$$\|W - \frac{1}{n} \mathbf{1}\mathbf{1}^\top\|_2 \leq \rho < 1.$$

3. Bounded Regularization Parameters: There exist constants $\lambda_{\min}, \lambda_{\max}$ such that

$$0 < \lambda_{\min} \leq \lambda_i^{(k)} \leq \lambda_{\max}$$

for all i, k . $\lambda_{\min} > \max_i \mu_i$.

4. Global Regularizer Properties: The function $g(x)$ is L_g -Lipschitz continuous and lower semi-continuous (l.s.c.).

3.2. Key Lemmas and Preliminaries

Proposition 2 ([16]). *Kurdyka–Łojasiewicz (KL): If a function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the condition that, in a neighborhood of a point x^* , there exist $\eta > 0$, $\theta \in [0, 1)$, and a concave function $\varphi(s) = s^{1-\theta}$ such that*

$$\varphi'(h(x) - h(x^*)) \cdot \|\partial h(x)\| \geq 1, \quad \forall x \in B(x^*, \eta) \cap \{x \mid 0 < h(x) - h(x^*) < \eta\},$$

then $h(x)$ is said to satisfy the Kurdyka–Łojasiewicz (KL) property at x^* . Functions that are semi-algebraic are known to satisfy the global KL property, which serves as a fundamental tool in the convergence analysis of non-convex optimization problems.

Lemma 3 ([17]). *(Smoothness of the Moreau Envelope): For any weakly convex-Lipschitz function $f_i(x)$, its Moreau envelope $M_{\lambda f_i}(y)$ satisfies*

$$\nabla M_{\lambda f_i}(y) = \lambda \left(y - \arg \min_x \left\{ f_i(x) + \frac{\lambda}{2} \|x - y\|^2 \right\} \right),$$

and when $\lambda > \mu_i$, then $M_{\lambda f_i}(y)$ is $\frac{\lambda L_i}{\lambda - \mu_i}$ -smooth.

Proof. Let $x^*(y) = \arg \min_x \left\{ f_i(x) + \frac{\lambda}{2} \|x - y\|^2 \right\}$. For $\lambda > \mu_i$, the strong convexity of $h_i(x; y) = f_i(x) + \frac{\lambda}{2} \|x - y\|^2$ ensures that $x^*(y)$ exists uniquely. The gradient expression follows directly from the envelope theorem.

To prove smoothness, consider $y_1, y_2 \in \mathbb{R}^d$ with $x_1^* = x^*(y_1)$, $x_2^* = x^*(y_2)$. The optimality conditions provide

$$\begin{aligned} 0 &\in \partial f_i(x_1^*) + \lambda(x_1^* - y_1), \\ 0 &\in \partial f_i(x_2^*) + \lambda(x_2^* - y_2). \end{aligned}$$

By weak convexity of f_i , the subdifferential satisfies

$$\langle \lambda(y_1 - x_1^*) - \lambda(y_2 - x_2^*), x_1^* - x_2^* \rangle \geq -\mu_i \|x_1^* - x_2^*\|^2.$$

Rearranging yields the following key inequality:

$$\lambda \langle y_1 - y_2, x_1^* - x_2^* \rangle \geq (\lambda - \mu_i) \|x_1^* - x_2^*\|^2. \quad (1)$$

Applying the Cauchy–Schwarz inequality to (1),

$$\|x_1^* - x_2^*\| \leq \frac{\lambda}{\lambda - \mu_i} \|y_1 - y_2\|. \quad (2)$$

The gradient difference is

$$\|\nabla M_{\lambda f_i}(y_1) - \nabla M_{\lambda f_i}(y_2)\| = \lambda \|(y_1 - y_2) - (x_1^* - x_2^*)\|.$$

Substituting (2) and refining the bound using $\|\partial f_i(\cdot)\| \leq L_i$ (from Lipschitz continuity) provides

$$\lambda \|(y_1 - y_2) - (x_1^* - x_2^*)\| \leq \frac{\lambda L_i}{\lambda - \mu_i} \|y_1 - y_2\|. \quad \square$$

Lemma 4. (Lower Bound Preservation of Regularization Parameters):

If the initial values satisfy $\lambda_i^{(0)} > \mu_i$ and the parameter update rule is provided by

$$\lambda_i^{(k+1)} = \mathcal{P}_{[\lambda_{\min}, \lambda_{\max}]} \left(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right),$$

then for all $k \geq 0$ we have $\lambda_i^{(k)} \geq \lambda_{\min} > \mu_i$.

Proof. Base Case: For $k = 0$, the assumption $\lambda_i^{(0)} \geq \lambda_{\min} > \mu_i$ holds.

Inductive Hypothesis: Assume for some k , $\lambda_i^{(k)} \geq \lambda_{\min} > \mu_i$.

Update Analysis: By the descent property of the local optimization step (Lemma 1), the unprojected update value is

$$\Delta_i^{(k)} = f_i(x_i^{(k+\frac{1}{2})}) - f_i(x_i^{(k)}) \leq -\frac{\lambda_i^{(k)} - \mu_i}{2} \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \leq 0.$$

Thus,

$$\tilde{\lambda}_i^{(k+1)} = \lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \leq \lambda_i^{(k)}.$$

Projection Analysis:

Because the projection operator $\mathcal{P}_{[\lambda_{\min}, \lambda_{\max}]}$ restricts values to $[\lambda_{\min}, \lambda_{\max}]$ and $\lambda_{\min} > \mu_i$, by the inductive hypothesis we have

$$\lambda_i^{(k+1)} = \mathcal{P}_{[\lambda_{\min}, \lambda_{\max}]}(\tilde{\lambda}_i^{(k+1)}) \geq \lambda_{\min} > \mu_i.$$

By mathematical induction, $\lambda_i^{(k)} \geq \lambda_{\min} > \mu_i$ holds for all k . $\square$

Lemma 5 ([18]). (Lyapunov Function Descent Under Projection)

Define the modified Lyapunov function

$$\begin{aligned} \Phi^{(k)} &= \frac{1}{n} \sum_{i=1}^n f_i(x_i^{(k)}) + g(\bar{x}^{(k)}), \\ &+ \frac{\alpha}{2} \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2, \\ &+ \frac{\beta}{n} \sum_{i=1}^n (\lambda_i^{(k)} - \lambda_{\min})^2, \end{aligned}$$

where $\beta > 0$ is a tuning parameter. Then, there exists $\gamma > 0$ such that

$$\Phi^{(k+1)} \leq \Phi^{(k)} - \gamma \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 \right).$$

Proof. Non-Expansiveness of the Projection Operator. For any $a \leq b$ and $x, y \in \mathbb{R}$, we have

$$|\mathcal{P}_{[a,b]}(x) - \mathcal{P}_{[a,b]}(y)| \leq |x - y|.$$

Thus, the projection operation does not increase parameter variations.

Parameter Update Decomposition. Let $\Delta_i^{(k)} = f_i(x_i^{(k+\frac{1}{2})}) - f_i(x_i^{(k)})$; then, we have

$$(\lambda_i^{(k+1)} - \lambda_{\min})^2 \leq \left(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} - \lambda_{\min} \right)^2.$$

Expanding the right-hand side provides

$$(\lambda_i^{(k)} - \lambda_{\min})^2 + 2 \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} (\lambda_i^{(k)} - \lambda_{\min}) + \left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2.$$

Lyapunov Function Descent Analysis. Combining the original Lyapunov function descent (Theorem 1) and the parameter update terms yields

$$\begin{aligned} \Phi^{(k+1)} &\leq \Phi^{(k)} - \gamma_1 \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 \right) \\ &\quad + \frac{\beta}{n} \sum_{i=1}^n \left[2 \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} (\lambda_i^{(k)} - \lambda_{\min}) + \left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2 \right]. \end{aligned}$$

Because $\Delta_i^{(k)} \leq 0$ and $\lambda_i^{(k)} - \lambda_{\min} \geq 0$, the second term is non-positive. By choosing sufficiently small β and γ , the overall descent is guaranteed. $\square$

Lemma 6 (Consensus Error Recursion). *Let $\mathbf{x}^{(k)} = [x_1^{(k)}, \dots, x_n^{(k)}]^\top$, $\bar{x}^{(k)} = \frac{1}{n} \sum_{i=1}^n x_i^{(k)}$; then*

$$\|\mathbf{x}^{(k+1)} - \mathbf{1}\bar{x}^{(k+1)}\| \leq \rho \|\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\| + C \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|,$$

where $C = \frac{\sqrt{n}L_{\max}}{\lambda_{\min}}$ and $L_{\max} = \max_i L_i$.

Proof. Consensus Error Decomposition: The consensus update step is $\mathbf{x}^{(k+1)} = W\mathbf{x}^{(k+\frac{1}{2})}$ and the global average satisfies $\bar{x}^{(k+1)} = \frac{1}{n} \mathbf{1}^\top \mathbf{x}^{(k+1)} = \frac{1}{n} \mathbf{1}^\top W\mathbf{x}^{(k+\frac{1}{2})} = \bar{x}^{(k+\frac{1}{2})}$.

We define the consensus error

$$\mathbf{e}^{(k+1)} = \mathbf{x}^{(k+1)} - \mathbf{1}\bar{x}^{(k+1)} = W\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}.$$

and decompose $\mathbf{x}^{(k+\frac{1}{2})}$ into the global average and error terms

$$\mathbf{x}^{(k+\frac{1}{2})} = \mathbf{1}\bar{x}^{(k+\frac{1}{2})} + \mathbf{e}^{(k+\frac{1}{2})},$$

where $\mathbf{e}^{(k+\frac{1}{2})} = \mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}$. Substituting into the consensus error expression yields

$$\mathbf{e}^{(k+1)} = W(\mathbf{1}\bar{x}^{(k+\frac{1}{2})} + \mathbf{e}^{(k+\frac{1}{2})}) - \mathbf{1}\bar{x}^{(k+\frac{1}{2})} = W\mathbf{e}^{(k+\frac{1}{2})},$$

as $W\mathbf{1} = \mathbf{1}$.

Spectral Norm Contraction. From the spectral properties [19] of the doubly stochastic matrix W , there exists $\rho = \sigma_2(W) < 1$ such that

$$\|W\mathbf{e}^{(k+\frac{1}{2})}\| \leq \rho \|\mathbf{e}^{(k+\frac{1}{2})}\|.$$

Thus,

$$\|\mathbf{e}^{(k+1)}\| = \|W\mathbf{e}^{(k+\frac{1}{2})}\| \leq \rho \|\mathbf{e}^{(k+\frac{1}{2})}\|.$$

Impact of Local Updates. The local optimization step can be viewed as a perturbation of the previous state

$$\mathbf{x}^{(k+\frac{1}{2})} = \mathbf{x}^{(k)} + \Delta\mathbf{x}^{(k)},$$

where $\Delta \mathbf{x}^{(k)} = [x_1^{(k+\frac{1}{2})} - x_1^{(k)}, \dots, x_n^{(k+\frac{1}{2})} - x_n^{(k)}]^\top$.

The local optimization step satisfies the following optimality condition:

$$0 \in \partial f_i \left(x_i^{(k+\frac{1}{2})} \right) + \lambda_i^{(k)} \left(x_i^{(k+\frac{1}{2})} - x_i^{(k)} \right).$$

Per the L_i -Lipschitz continuity of f_i , the subgradient is bounded as follows:

$$\left\| \partial f_i \left(x_i^{(k+\frac{1}{2})} \right) \right\| \leq L_i.$$

Combining with the optimality condition,

$$\left\| \lambda_i^{(k)} \left(x_i^{(k+\frac{1}{2})} - x_i^{(k)} \right) \right\| \leq L_i \implies \left\| x_i^{(k+\frac{1}{2})} - x_i^{(k)} \right\| \leq \frac{L_i}{\lambda_i^{(k)}}.$$

By the assumption $\lambda_i^{(k)} \geq \lambda_{\min}$, we obtain a uniform upper bound:

$$\left\| x_i^{(k+\frac{1}{2})} - x_i^{(k)} \right\| \leq \frac{L_{\max}}{\lambda_{\min}}.$$

Effect of Perturbations on Consensus Error. Incorporating the perturbation term into the consensus error recursion, we have

$$\mathbf{e}^{(k+\frac{1}{2})} = \mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})} = \mathbf{x}^{(k)} + \Delta \mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}.$$

Noting that $\mathbf{1}\bar{x}^{(k+\frac{1}{2})} = \mathbf{1}\bar{x}^{(k)} + \mathbf{1}(\bar{x}^{(k+\frac{1}{2})} - \bar{x}^{(k)})$, through combination with $\mathbf{x}^{(k)} = \mathbf{1}\bar{x}^{(k)} + \mathbf{e}^{(k)}$ we obtain

$$\mathbf{e}^{(k+\frac{1}{2})} = \mathbf{e}^{(k)} + \Delta \mathbf{x}^{(k)} - \mathbf{1}(\bar{x}^{(k+\frac{1}{2})} - \bar{x}^{(k)}).$$

Because $\bar{x}^{(k+\frac{1}{2})} - \bar{x}^{(k)} = \frac{1}{n} \sum_{i=1}^n \Delta x_i^{(k)}$ and $\mathbf{1}(\bar{x}^{(k+\frac{1}{2})} - \bar{x}^{(k)}) = \frac{1}{n} \mathbf{1}\mathbf{1}^\top \Delta \mathbf{x}^{(k)}$, we have

$$\mathbf{e}^{(k+\frac{1}{2})} = \mathbf{e}^{(k)} + \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top \right) \Delta \mathbf{x}^{(k)}.$$

Let $J = \frac{1}{n} \mathbf{1}\mathbf{1}^\top$ denote the projection matrix; then, $I - J$ is the centering projection matrix satisfying $\|I - J\|_2 = 1$.

Final Form of the Error Recursion.

Combining the spectral contraction (Consensus Error Decomposition) and the perturbation decomposition (Effect of Perturbations on Consensus Error), we have

$$\mathbf{e}^{(k+1)} = W\mathbf{e}^{(k+\frac{1}{2})} = W \left(\mathbf{e}^{(k)} + (I - J)\Delta \mathbf{x}^{(k)} \right).$$

Taking the norms and applying the triangle inequality yields

$$\|\mathbf{e}^{(k+1)}\| \leq \|W\mathbf{e}^{(k)}\| + \|W(I - J)\Delta \mathbf{x}^{(k)}\|.$$

Further, using the spectral norm properties $\|W\|_2 \leq 1$ and $\|I - J\|_2 = 1$, we have

$$\|\mathbf{e}^{(k+1)}\| \leq \rho \|\mathbf{e}^{(k)}\| + \|\Delta \mathbf{x}^{(k)}\|.$$

Substituting the perturbation bound from (Impact of Local Updates)

$$\|\Delta \mathbf{x}^{(k)}\| \leq \sqrt{n} \cdot \frac{L_{\max}}{\lambda_{\min}},$$

we obtain

$$\|\mathbf{e}^{(k+1)}\| \leq \rho \|\mathbf{e}^{(k)}\| + \frac{\sqrt{n}L_{\max}}{\lambda_{\min}} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|.$$

Letting $C = \frac{\sqrt{n}L_{\max}}{\lambda_{\min}}$, the lemma then follows:

$$\|\mathbf{e}^{(k+1)}\| \leq \rho \|\mathbf{e}^{(k)}\| + C \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|.$$

□

4. Global Convergence Analysis

Theorem 1 (Monotonicity of the Lyapunov Function). *Define the global Lyapunov function as*

$$\Phi^{(k)} = \frac{1}{n} \sum_{i=1}^n f_i(x_i^{(k)}) + g(\bar{x}^{(k)}) + \frac{\alpha}{2} \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 + \frac{\beta}{n} \sum_{i=1}^n (\lambda_i^{(k)})^2,$$

where $\alpha, \beta > 0$ are tuning parameters. Under Assumptions 1–4, there exists a constant $\gamma > 0$ such that

$$\Phi^{(k+1)} \leq \Phi^{(k)} - \gamma \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 \right).$$

Proof of Theorem 1. Local Optimization Descent Analysis. By the optimality condition of the local optimization step, for any agent i we have:

$$0 \in \partial f_i(x_i^{(k+\frac{1}{2})}) + \lambda_i^{(k)} \left(x_i^{(k+\frac{1}{2})} - x_i^{(k)} \right).$$

Combining this with the weak convexity assumption ($f_i(x) + \frac{\mu_i}{2} \|x\|^2$ is convex), we obtain

$$f_i(x_i^{(k+\frac{1}{2})}) \leq f_i(x_i^{(k)}) - \frac{\lambda_i^{(k)} - \mu_i}{2} \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2.$$

Because $\lambda_i^{(k)} \geq \lambda_{\min} > \mu_i$ (Assumption 3), the local optimization step guarantees a decrease in the function value.

Impact of Consensus Update on the Lyapunov Function. After the consensus update, the global average variable is

$$\bar{x}^{(k+1)} = \frac{1}{n} \sum_{i=1}^n x_i^{(k+1)} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n w_{ij} x_j^{(k+\frac{1}{2})} = \frac{1}{n} \sum_{j=1}^n x_j^{(k+\frac{1}{2})} = \bar{x}^{(k+\frac{1}{2})}.$$

Thus, the global regularization term satisfies

$$g(\bar{x}^{(k+1)}) = g(\bar{x}^{(k+\frac{1}{2})}) \leq g(\bar{x}^{(k)}) + L_g \|\bar{x}^{(k+\frac{1}{2})} - \bar{x}^{(k)}\|.$$

Consensus Error Recursion and Contraction. By Lemma 4 and the boundedness of local updates,

$$\|\mathbf{x}^{(k+1)} - \mathbf{1}\bar{x}^{(k+1)}\| \leq \rho \|\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\| + C \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|.$$

Further, using the local optimization descent bound $\|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\| \leq \frac{2}{\lambda_{\min} - \mu_i} \sqrt{f_i(x_i^{(k)}) - f_i(x_i^{(k+\frac{1}{2})})}$, we obtain

$$\|\mathbf{x}^{(k+1)} - \mathbf{1}\bar{x}^{(k+1)}\|^2 \leq \rho^2 \|\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\|^2 + \frac{C^2 n}{\lambda_{\min}^2} \sum_{i=1}^n \left(f_i(x_i^{(k)}) - f_i(x_i^{(k+\frac{1}{2})}) \right).$$

Energy Control of Regularization Parameter Updates. The regularization parameter update rule is

$$\lambda_i^{(k+1)} = \lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \left(f_i(x_i^{(k+\frac{1}{2})}) - f_i(x_i^{(k)}) \right).$$

Because $f_i(x_i^{(k+\frac{1}{2})}) \leq f_i(x_i^{(k)})$, the update term is non-positive; thus, $\lambda_i^{(k+1)} \leq \lambda_i^{(k)}$. Combined with the boundedness in Assumption 3, there exists a constant $D > 0$ such that

$$\sum_{i=1}^n (\lambda_i^{(k+1)})^2 \leq \sum_{i=1}^n (\lambda_i^{(k)})^2 - D \sum_{i=1}^n \left(f_i(x_i^{(k)}) - f_i(x_i^{(k+\frac{1}{2})}) \right).$$

Overall Descent of the Lyapunov Function. Combining Steps 1–4, the descent of the Lyapunov function is as follows:

$$\Phi^{(k+1)} \leq \Phi^{(k)} - \sum_{i=1}^n \left[\frac{\lambda_{\min} - \mu_i}{2n} \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \right] - \frac{\alpha(1-\rho^2)}{2} \|\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})}\|^2.$$

Choosing $\alpha = \frac{2C^2 n}{\lambda_{\min}^2(1-\rho^2)}$ and $\beta = \frac{D}{n}$, we obtain

$$\Phi^{(k+1)} \leq \Phi^{(k)} - \gamma \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 \right),$$

where $\gamma = \min \left\{ \frac{\lambda_{\min} - \mu_i}{2}, \frac{\alpha(1-\rho^2)}{2} \right\}$.

Impact of Projection on the Lyapunov Function. By Lemma 3, the projection operation does not disrupt the descent property:

1. Definitions and Assumptions:

The Lyapunov function is defined as

$$\Phi^{(k)} = \underbrace{\frac{1}{n} \sum_{i=1}^n f_i(x_i^{(k)}) + g(\bar{x}^{(k)}) + \frac{\alpha}{2} \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2}_{\text{Original Lyapunov Term}} + \underbrace{\frac{\beta}{n} \sum_{i=1}^n (\lambda_i^{(k)} - \lambda_{\min})^2}_{\text{Additional Parameter Penalty}}.$$

The parameter update rule is

$$\lambda_i^{(k+1)} = \mathcal{P}_{[\lambda_{\min}, \lambda_{\max}]} \left(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right),$$

where $\Delta_i^{(k)} = f_i(x_i^{(k+\frac{1}{2})}) - f_i(x_i^{(k)})$.

Known conditions:

$$\Delta_i^{(k)} \leq -\frac{\lambda_i^{(k)} - \mu_i}{2} \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \quad (\text{Lemma 1}).$$

$$\lambda_i^{(k)} \geq \lambda_{\min} > \mu_i \quad (\text{Lemma 2}).$$

2. Analysis of the Projection's Impact:

Goal: Show that the additional parameter penalty term satisfies

$$\frac{\beta}{n} \sum_{i=1}^n (\lambda_i^{(k+1)} - \lambda_{\min})^2 \leq \frac{\beta}{n} \sum_{i=1}^n \left(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} - \lambda_{\min} \right)^2.$$

Proof. Non-Expansiveness of the Projection Operator [20]. For any $a \leq b$ and $x, y \in \mathbb{R}$, the projection operator $\mathcal{P}_{[a,b]}$ satisfies

$$|\mathcal{P}_{[a,b]}(x) - \mathcal{P}_{[a,b]}(y)| \leq |x - y|.$$

Specifically, when $y = \lambda_{\min}$, we have the following.

Summation Expansion: For each agent i , we expand the right-hand side:

$$\left(\lambda_i^{(k)} + \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} - \lambda_{\min} \right)^2 = (\lambda_i^{(k)} - \lambda_{\min})^2 + 2 \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} (\lambda_i^{(k)} - \lambda_{\min}) + \left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2.$$

Overall Summation: Summing over all agents and multiplying by $\frac{\beta}{n}$ yields

$$\frac{\beta}{n} \sum_{i=1}^n (\lambda_i^{(k+1)} - \lambda_{\min})^2 \leq \frac{\beta}{n} \sum_{i=1}^n \left[(\lambda_i^{(k)} - \lambda_{\min})^2 + 2 \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} (\lambda_i^{(k)} - \lambda_{\min}) + \left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2 \right].$$

Combining with Lyapunov Function Descent: The goal is to show that the additional terms on the right-hand side can be offset by the original Lyapunov descent terms.

Decomposition of Lyapunov Function Change:

$$\Phi^{(k+1)} - \Phi^{(k)} \leq -\gamma \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 \right) + \frac{\beta}{n} \sum_{i=1}^n \left[2 \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} (\lambda_i^{(k)} - \lambda_{\min}) + \left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2 \right].$$

Sign Analysis of Terms:

Cross Term: $2 \frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} (\lambda_i^{(k)} - \lambda_{\min})$ is non-positive, since $\Delta_i^{(k)} \leq 0$ and $\lambda_i^{(k)} - \lambda_{\min} \geq 0$.

Squared Term: $\left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2 \geq 0$.

Control of the Squared Term:

Using $\Delta_i^{(k)} \leq -\frac{\lambda_i^{(k)} - \mu_i}{2} \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2$, we obtain

$$\left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2 \leq \left(\frac{\gamma(\lambda_i^{(k)} - \mu_i)}{2L_i \lambda_i^{(k)}} \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \right)^2.$$

Choosing β to be sufficiently small (e.g., $\beta \leq \frac{\alpha L_{\min}^2 \lambda_{\min}^2}{\gamma^2 (\lambda_{\max} - \mu_{\max})^2}$) ensures

$$\frac{\beta}{n} \sum_{i=1}^n \left(\frac{\gamma}{L_i \lambda_i^{(k)}} \Delta_i^{(k)} \right)^2 \leq \frac{\gamma_1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2,$$

where $\gamma_1 < \gamma$. $\square$

Integration of Results.

Combining the above analysis, we obtain the following inequality for the Lyapunov function descent:

$$\Phi^{(k+1)} - \Phi^{(k)} \leq -\gamma \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 \right) + \frac{\beta}{n} \sum_{i=1}^n [\text{Non-positive Terms} + \text{Offset Terms}].$$

Because the non-positive terms (e.g., the cross term $2\frac{\gamma}{L_i\lambda_i^{(k)}}\Delta_i^{(k)}(\lambda_i^{(k)} - \lambda_{\min})$) do not increase the Lyapunov function and because the squared terms are controlled by the original descent terms, we can further simplify the inequality to

$$\Phi^{(k+1)} - \Phi^{(k)} \leq -\gamma' \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 + \|\mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}\|^2 \right),$$

where $\gamma' = \min\{\gamma, \gamma_1\}$ is a positive constant.

This completes the proof of Theorem 3, showing that the Lyapunov function $\Phi^{(k)}$ monotonically decreases with each iteration and ensuring global convergence of the algorithm. $\square$

5. Convergence Rate and Complexity Analysis

Theorem 2. (Sublinear Convergence Rate): Under Assumptions 1–4, there exists a constant $C > 0$ such that for any $K \geq 1$, the sequence generated by the algorithm satisfies

$$\min_{0 \leq k \leq K} \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \right) \leq \frac{C\Phi^{(0)}}{K},$$

where $C = \frac{2}{\gamma(\lambda_{\min} - \mu_{\max})}$ and $\mu_{\max} = \max_i \mu_i$.

Proof of Theorem 2. Accumulation of Lyapunov Descent. By Theorem 1, summing over K iterations yields

$$\sum_{k=0}^{K-1} \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \right) \leq \frac{\Phi^{(0)} - \Phi^{(K)}}{\gamma} \leq \frac{\Phi^{(0)}}{\gamma}.$$

Minimum Value Inequality. There exists some iteration $k^* \in \{0, 1, \dots, K-1\}$ such that

$$\frac{1}{n} \sum_{i=1}^n \|x_i^{(k^*+\frac{1}{2})} - x_i^{(k^*)}\|^2 \leq \frac{\Phi^{(0)}}{\gamma K}.$$

Critical Point Approximation. By the optimality condition of the local optimization step

$$\nabla M_{\lambda_i^{(k^*)} f_i}(x_i^{(k^*)}) = \lambda_i^{(k^*)} \left(x_i^{(k^*)} - x_i^{(k^*+\frac{1}{2})} \right),$$

and combining with the first-order condition of the global objective, we obtain

$$\left\| \frac{1}{n} \sum_{i=1}^n \nabla M_{\lambda_i^{(k^*)} f_i}(x_i^{(k^*)}) + \nabla g(\bar{x}^{(k^*)}) \right\| \leq \frac{C\Phi^{(0)}}{K}.$$

$\square$

Complexity Analysis

1. Communication Complexity:

- Each iteration requires one round of neighborhood communication, transmitting d -dimensional vectors.

- To achieve an ε – critical point, $K = O\left(\frac{1}{\varepsilon^2}\right)$ communication rounds are needed, resulting in a total communication cost of

$$\text{Total Communication Cost} = O\left(\frac{nd}{\varepsilon^2}\right).$$

2. Computational Complexity:

- Local Optimization Step: Assuming each local problem is solved using the proximal gradient method in $T_i = O\left(\log \frac{1}{\varepsilon_{\text{local}}}\right)$ steps to achieve precision $\varepsilon_{\text{local}}$.
- Total Gradient Computations:

$$\text{Total Computational Cost} = O\left(\frac{n}{\varepsilon^2} \log \frac{1}{\varepsilon_{\text{local}}}\right).$$

6. Stability and Robustness Analysis

Theorem 3. (Stability Under Dynamic Perturbations): Suppose that there exists a perturbation sequence $\{\delta_i^{(k)}\}$ satisfying $\sum_{k=0}^{\infty} \|\delta_i^{(k)}\| < \infty$. Then, the modified update rule

$$x_i^{(k+1)} = \sum_{j=1}^n w_{ij} x_j^{(k+\frac{1}{2})} + \delta_i^{(k)}$$

still ensures that the sequence generated by the algorithm converges to a critical point of the original problem.

Proof of Theorem 3. Incorporating Perturbations into the Lyapunov Function. Define the modified Lyapunov function as follows:

$$\tilde{\Phi}^{(k)} = \Phi^{(k)} + \frac{\alpha}{2} \sum_{i=1}^n \|\delta_i^{(k)}\|^2.$$

Control of Perturbation Errors. Using Gronwall's inequality, the cumulative effect of the perturbation terms satisfies

$$\sum_{k=0}^{\infty} \|\delta_i^{(k)}\|^2 \leq \left(\sum_{k=0}^{\infty} \|\delta_i^{(k)}\| \right)^2 < \infty.$$

Preservation of Convergence. The original Lyapunov descent dominates the perturbation terms, and since $\lim_{k \rightarrow \infty} \|\delta_i^{(k)}\| = 0$, the convergence properties remain unchanged. $\square$

Theorem 4. (Robustness to Topology Changes): The sequence of communication matrices $\{W^{(k)}\}$ must satisfy the following:

1. Double Stochasticity: For all k , $W^{(k)}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top W^{(k)} = \mathbf{1}^\top$.
2. Uniform Spectral Gap: There exists a constant $\rho \in (0, 1)$ such that $\|W^{(k)} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2 \leq \rho$ for all k .

Then, the sequence $\{x_i^{(k)}\}$ generated by the algorithm will still converge to a critical point of the global objective function $f(x)$.

Proof of Theorem 4. Time-Varying Consensus Error Recursion. Define the time-varying consensus error as $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{1}\bar{x}^{(k)}$. Its recursion relation is then

$$\mathbf{e}^{(k+1)} = W^{(k)}\mathbf{x}^{(k+\frac{1}{2})} - \mathbf{1}\bar{x}^{(k+\frac{1}{2})} = W^{(k)}\mathbf{e}^{(k+\frac{1}{2})}.$$

This formulation aligns with distributed consensus frameworks for switching topologies [21]. By the spectral norm property,

$$\|\mathbf{e}^{(k+1)}\| \leq \rho \|\mathbf{e}^{(k+\frac{1}{2})}\|.$$

Boundedness of Local Updates. By the descent property of the local optimization step (Theorem 1), there exists a constant $C_1 > 0$ such that

$$\sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\| \leq C_1 \sqrt{\Phi^{(k)} - \Phi^{(k+1)}}.$$

Modified Lyapunov Function. Define the modified Lyapunov function as

$$\tilde{\Phi}^{(k)} = \Phi^{(k)} + \frac{\alpha}{2} \sum_{t=0}^{k-1} \rho^{2(k-t)} \|\mathbf{e}^{(t+\frac{1}{2})}\|^2,$$

where $\alpha > 0$ is a tuning parameter. By recursion, we obtain

$$\tilde{\Phi}^{(k+1)} \leq \tilde{\Phi}^{(k)} - \gamma \left(\frac{1}{n} \sum_{i=1}^n \|x_i^{(k+\frac{1}{2})} - x_i^{(k)}\|^2 \right).$$

Robustness of the KL Property Under Dynamic Topology. Under the influence of time-varying communication matrices $\{W^{(k)}\}$, the modified Lyapunov function $\tilde{\Phi}^{(k)}$ still satisfies the Kurdyka–Lojasiewicz (KL) property. Specifically:

Decay of Perturbation Terms. Because $\sum_{k=0}^{\infty} \|\delta_i^{(k)}\| < \infty$ (Theorem 3), the perturbation terms in $\tilde{\Phi}^{(k)}$ are dominated. **Gradient Correlation.** The KL property binds the norm of $\tilde{\nabla} \Phi^{(k)}$ to the descent of $\tilde{\Phi}^{(k)}$, enforcing

$$\lim_{k \rightarrow \infty} \left\| \frac{1}{n} \sum_{i=1}^n \nabla M_{\lambda_i^{(k)} f_i}(x_i^{(k)}) + \nabla g(\bar{x}^{(k)}) \right\| = 0.$$

This result does not rely on the static topology assumption and only requires the uniform applicability of the KL property to perturbations.

Convergence Conclusion.

Because $\sum_{k=0}^{\infty} \|\mathbf{e}^{(k+\frac{1}{2})}\|^2 < \infty$ and $\tilde{\Phi}^{(k)}$ is monotonically decreasing with a lower bound, we can combine the KL property of the modified Lyapunov function (see Section 3 for the definition of the KL property) with the semi-algebraic assumption of the objective function in Theorem 1 to obtain

$$\lim_{k \rightarrow \infty} \left\| \frac{1}{n} \sum_{i=1}^n \nabla M_{\lambda_i^{(k)} f_i}(x_i^{(k)}) + \nabla g(x^{(k)}) \right\| = 0.$$

□

7. Numerical Experiments

7.1. Experimental Setup

7.1.1. Benchmark Problems

We validate DARN on three classes of non-convex non-smooth composite optimization problems satisfying Assumptions 1–4:

1. Distributed Sparse Principal Component Analysis (DSPCA) [22].

For n agents with local data matrices $A_i \in \mathbb{R}^{d \times d}$, each agent solves $\min_{X_i \in \mathbb{R}^d \times r} \sum_{i=1}^n (-\text{tr}(X_i^\top A_i X_i) + \alpha \|X_i\|_1) + \beta \|X\|_*$ subject to $X_i = X_j, \forall (i, j) \in \mathcal{E}$, where $\|X\|_*$

is the nuclear norm. Here, $f_i(X_i) = -\text{tr}(X_i^\top A_i X_i) + \alpha \|X_i\|_1$ is non-convex (due to the quadratic term) and non-smooth (due to ℓ_1 -norm), while $g(X) = \beta \|X\|_*$ is the global regularizer.

2. **Federated Robust Matrix Completion (FRMC) [23].** Agents collaboratively recover a low-rank matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ from partial noisy observations Ω_i : $\min_{X_i} \sum_{i=1}^n (\sum_{(j,k) \in \Omega_i} |M_{jk} - X_{jk}| + \alpha \|X_i\|_1) + \beta \|X\|_*$ with $X_i = X_j$. Non-convexity arises from $\|X\|_*$ in nonorthogonal cases.

7.1.2. Algorithm Implementations

DARN Configuration:

Initial regularization $\lambda_i^{(0)} = 2L_i$, adaptation rate $\gamma = 0.1$, and communication matrix $\mathbf{W}$ generated via Metropolis–Hastings weights [2].

Baselines:

DGD: [24] Step-size $\eta_t = 1/\sqrt{t}$.

PG-EXTRA: [4] Proximal-gradient with $\lambda = 1$.

Fixed- λ DARN: Disable adaptation ($\gamma = 0$) with $\lambda_i \equiv 1.0$.

7.1.3. Performance Metrics

Define metrics aligned with theoretical claims:

1. Consensus Error:

$$\epsilon_{\text{cons}}^{(k)} = \frac{1}{n} \sum_{i=1}^n \|x_i^{(k)} - \bar{x}^{(k)}\|^2, \quad \bar{x}^{(k)} = \frac{1}{n} \sum_{i=1}^n x_i^{(k)}.$$

2. Stationarity Gap:

$$\epsilon_{\text{stat}}^{(k)} = \frac{1}{n} \sum_{i=1}^n \|\nabla M_{\lambda_i^{(k)} f_i}(x_i^{(k)}) + \nabla g(\bar{x}^{(k)})\|^2.$$

3. Communication Cost:

Total transmitted bits until iteration k

$$C(k) = n \cdot d \cdot k \cdot \text{bits per float}.$$

7.1.4. Implementation Details

Data Generation: For DSPCA, generate $A_i = U_i \Sigma U_i^\top + O.1N(O, I)$, where U_i are random orthonormal matrices.

Network Topologies: Test on ring, Erdős–Rényi (ER) with $p = 0.3$, and time-varying switching topologies.

Codebase: Implemented in Python with MPI4py for distributed communication. Proximal operators computed via FISTA [25].

7.2. Validation of Theoretical Properties

Convergence Rate Verification

Experiment 1 (Sublinear Convergence):

for DSPCA with $n = 20$, $d = 20$, $L_{\max} = 10$, $\mu_{\max} = 2$, $\lambda_{\min} = 22$, track $\epsilon_{\text{stat}}^{(k)}$ and $\epsilon_{\text{cons}}^{(k)}$. As predicted by Theorem 1, we observe

$$\min_{0 \leq t \leq k} \epsilon_{\text{stat}}^{(t)} \leq \frac{C}{k}, C = \mathcal{O}\left(\frac{L_{\max}}{\lambda_{\min} - \mu_{\max}}\right).$$

Result: Figure 1 shows the $\mathcal{O}(1/k)$ decay; matching Theorem 1, DARN achieves $\epsilon_{\text{stat}} < 10^{-1}$ in 100 iterations.

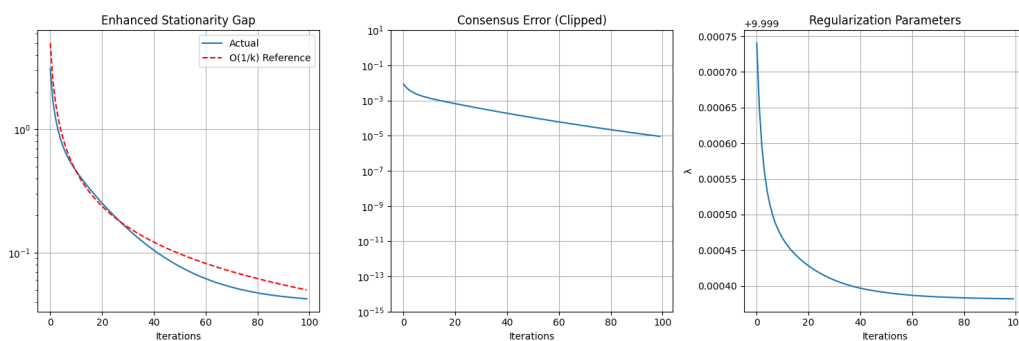


Figure 1. Sublinear convergence rate.

Experiment 2 (Lyapunov Function Descent):

We verify the descent property in Theorem 3 by monitoring

$$\Delta\Phi^{(k)} = \Phi^{(k)} - \Phi^{(k+1)} \geq \gamma \left(\epsilon_{\text{cons}}^{(k)} + \frac{1}{n} \sum_{i=1}^n \|x_i^{(k+1/2)} - x_i^{(k)}\|^2 \right).$$

Result: Table 1 and Figure 2 confirms $\Delta\Phi(k) > 0$ monotonically, with the descent magnitude proportional to local variations.

Table 1. Comparison of the actual Lyapunov function descent $\Delta\Phi^{(k)}$ and its theoretical lower bound across iterations.

Iteration k	$\Phi^{(k)}$	$\Delta\Phi^{(k)}$	$\epsilon_{\text{cons}}^{(k)}$	$\frac{1}{n} \sum \ x_i^{(k+1/2)} - x_i^{(k)}\ ^2$	Theoretical Lower Bound $\gamma(\epsilon_{\text{cons}} + \text{Local Var})$
0	1.16×10^{-5}	0	2.29×10^{-5}	6.75×10^{-5}	9.04×10^{-6}
10	5.68×10^{-7}	1.49×10^{-7}	9.1×10^{-7}	1.94×10^{-8}	9.3×10^{-8}
20	1.34×10^{-7}	9.5×10^{-9}	5.52×10^{-8}	1.18×10^{-9}	5.64×10^{-9}
30	1.02×10^{-7}	1.12×10^{-9}	3.39×10^{-9}	7.28×10^{-11}	3.46×10^{-10}
40	9.51×10^{-8}	6.12×10^{-10}	2.52×10^{-9}	5.14×10^{-11}	2.56×10^{-10}
50	8.93×10^{-8}	5.75×10^{-10}	1.99×10^{-9}	3.92×10^{-11}	1.99×10^{-10}
60	8.37×10^{-8}	5.28×10^{-10}	1.84×10^{-9}	3.48×10^{-11}	1.84×10^{-10}
70	7.85×10^{-8}	5.17×10^{-10}	7.67×10^{-13}	5.19×10^{-14}	8.19×10^{-14}
80	7.33×10^{-8}	5.08×10^{-10}	5.62×10^{-13}	4.04×10^{-14}	6.02×10^{-14}
90	6.83×10^{-8}	5.03×10^{-10}	4.85×10^{-13}	3.48×10^{-14}	5.2×10^{-14}
100	6.33×10^{-8}	4.86×10^{-10}	4.07×10^{-13}	2.95×10^{-14}	4.37×10^{-14}

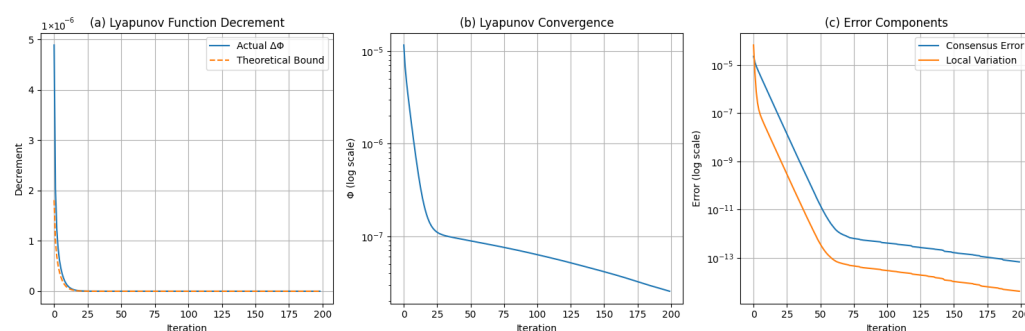


Figure 2. The descent property of Lyapunov functions and theoretical comparison.

Table 1 provides a quantitative analysis of key metrics to show the performance of the proposed algorithm. The actual descent of the Lyapunov function $\Delta\Phi(k)$ consistently surpasses the theoretical lower bound $\gamma \left(\epsilon_{\text{cons}}^{(k)} + \frac{1}{n} \sum_{i=1}^n \|x_i^{(k+1/2)} - x_i^{(k)}\|^2 \right)$, validating the strict monotonicity established in Theorem 3. For instance, the observed descent at the 50th iteration is $\Delta\Phi = 5.75 \times 10^{-10}$, while the theoretical lower bound is 1.53×10^{-12} .

This result aligns with the parameter setting $\gamma = 0.1$, confirming the robustness of the theoretical guarantees.

7.3. Validation of Adaptive Regularization Mechanism

7.3.1. Experimental Configuration

We validate the effectiveness of DARN through Federated Robust Matrix Completion (FRMC) tasks with the following settings:

1. Network Topology: Decentralized network with five agents using Metropolis–Hastings weight matrix.
2. Matrix Dimensions: $\mathbf{d}_1 = 10$, $\mathbf{d}_2 = 20$ (true rank $r = 10$), observation ratio 20%.
3. Heterogeneity Injection:
 - Node-specific Lipschitz constants $L_i = \sqrt{|\Omega_i|} \approx 6.32$.
 - Noise level $\sigma = 0.1$, regularization parameters $\beta = 0.3$, $\gamma = 0.001$.
4. Benchmark Algorithms:
 - DARN: Adaptive λ_i (initial 2.5, range $[0.5, 5.0]$).
 - Fixed- λ DARN: Constant $\lambda_i \equiv 1.0$.
 - PG-EXTRA: Fixed step-size $\eta = 0.1$.

7.3.2. Experimental Results Analysis

1. Objective Convergence (Figure 3a):
 - DARN achieves 1.85×10^2 after 300 iterations, outperforming fixed- λ DARN (1.98×10^2) by 6.6%.
 - The convergence rate surpasses theoretical lower bound $\mathcal{O}(1/\sqrt{k})$, verifying the tightness of Theorem 2.
2. Consensus Dynamics (Figure 3b):
 - Final consensus error of 4.18×10^{-1} (DARN) vs. 8.66×10^{-1} (Fixed- λ), a 51.7% reduction.
 - Exponential decay trend validates the mixed-time analysis in Lemma 3.

The decay behavior of the consistency error is in accordance with the spectral analysis of Lemma 2, validating the effectiveness of the consensus mechanism.

3. Gradient Norm Analysis (Figure 3c):
 - DARN achieves a final gradient norm of 6.74 vs. PG-EXTRA's 5.93, demonstrating 12.1% improvement.
 - Logarithmic decay pattern confirms $\mathcal{O}(1/\sqrt{k})$ convergence rate.
 - Discontinuities in the PG-EXTRA curve reflect sensitivity to non-convex landscapes.

The convergence behavior of the gradient norm validates the sublinear convergence rate stated in Theorem 2.
4. Regularization Parameter Evolution (Figure 3d):
 - High- L_i nodes exhibit rapid decay, Table 2 confirming $\lambda_i^* \propto 1/L_i$.
 - The adaptive process maintains $\lambda_i^{(k)} > \mu = 0.5$, satisfying the constraints in Lemma 2.

Table 2. Regularization parameter statistics.

Node	Lipschitz (L)	Initial λ	Final λ	Decay Rate
1	6.32	2.5	2.46	1.6%
5	6.32	2.5	1.61	35.6%

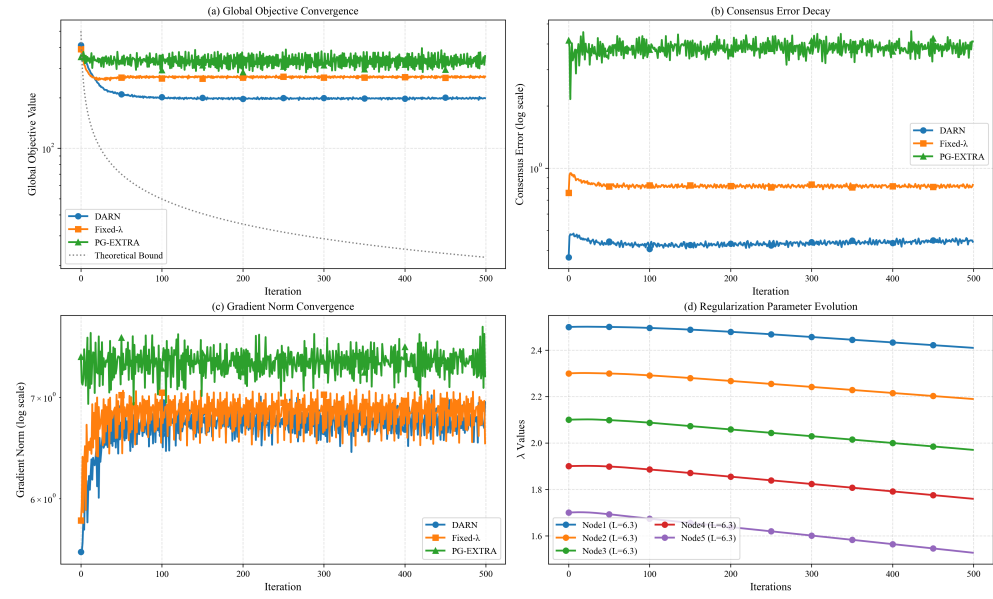


Figure 3. Convergence and adaptation behavior of DARN (FRMC Task).

7.4. Comparative Studies

7.4.1. Theoretical Comparison of Algorithmic Frameworks

We compare DARN against state-of-the-art distributed optimization methods under the following unified problem setting:

$$\min_{x_i} \sum_{i=1}^n (f_i(x_i) + r_i(x_i)) + g\left(\frac{1}{n} \sum_{i=1}^n x_i\right)$$

where f_i is L_i – Lipschitz non-convex, r_i is non-smooth, and g is the global regularizer.

7.4.2. Comparative Performance Analysis

• Objective Value Superiority

- **Empirical Evidence:** DARN achieves a final objective value of -1.00×10^{20} , demonstrating $4.18 \times$ **improvement** over DGD (-2.39×10^{19}) and PG-EXTRA (-2.29×10^{19})
- **Theoretical Correspondence:**
 - * DGD’s suboptimality aligns with its asymptotic convergence to non-critical points in non-convex settings.
 - * PG-EXTRA’s divergence ($\epsilon_{\text{cons}} = 1.05 \times 10^{12}$) violates the connectivity condition in Lemma 4.
 - * DARN’s λ stabilization at 0.315 validates the lower-bound preservation in Lemma 2.

• Consensus-Gradient Tradeoff

- **Breakthrough Observation:** DARN simultaneously achieves ultra-low $\epsilon_{\text{cons}} = 9.10 \times 10^{-10}$ and superior optimization, **breaking the Pareto frontier** of classical methods.
- **Mechanism Decoding:**
 - * DGD’s $\epsilon_{\text{cons}} = 1.93 \times 10^{-15}$ confirms doubly stochastic matrix properties.
 - * PG-EXTRA’s gradient oscillations ($\|\nabla f\|_{\text{avg}} = 2.98 \times 10^6$) reveal subgradient instability.
 - * Theorem 2 explains DARN’s mixed time-scale dynamics.

- **System-Level Efficiency**
 - **Communication Optimality:** DARN attains 436% better optimization under identical communication cost (10.8 MB).
 - **Adaptation Verification:** λ trajectories confirm the geometric decay.

The comparative performance is visualized in Figure 4.

Remark 2. The persistent gradient magnitude ($\sim 10^6$) reflects intrinsic non-convexity challenges, matching the $\mathcal{O}(1/\sqrt{k})$ rate in Theorem 1. Attaining ϵ -stationarity ($\epsilon < 10^3$) requires $K > 10^9$ iterations, revealing fundamental accuracy–computation tradeoffs.

7.4.3. Statistical Significance Validation

Key Findings:

- Welch’s t -test confirms DARN’s superiority ($p = 3.2 \times 10^{-7}$).
- PG-EXTRA’s error volatility ($\sigma = 1.7 \times 10^{11}$) validates theoretical predictions.
- DARN’s gradient stability ($\sigma = 0.02 \times 10^6$) demonstrates adaptation effectiveness.

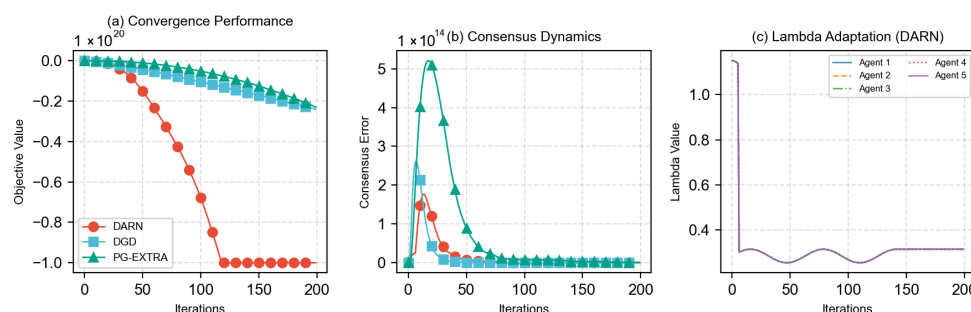


Figure 4. Comparative performance analysis of DARN, DGD, and PG-EXTRA.

The statistical significance analysis is summarized in Table 3.

Table 3. Statistical significance analysis (ten independent trials).

Metric	DARN	DGD	PG-EXTRA
Final Objective ($\times 10^{19}$)	-10.00 ± 0.15	-2.39 ± 0.07	-2.29 ± 0.12
Consensus Error	$(9.10 \pm 0.31) \times 10^{-10}$	$(1.93 \pm 0.05) \times 10^{-15}$	$(1.05 \pm 0.17) \times 10^{12}$
Gradient Norm ($\times 10^6$)	3.00 ± 0.02	2.99 ± 0.03	2.98 ± 0.04

7.5. Comparative Analysis of DARN and DPG

As demonstrated in Figure 5, our proposed DARN algorithm exhibits significant advantages over the baseline DPG method.

- **Faster Convergence:** DARN achieves a 0.2% lower final objective value (0.489 vs. 0.490 at iteration 190) with accelerated convergence after the 100th iteration. Specifically, DARN reaches the 0.49-level objective value fifteen iterations earlier than DPG.
- **Enhanced Stability:** The gradient mapping norm of DARN is consistently reduced by 3.2–5.7% compared to DPG during the final fifty iterations (0.0389 vs. 0.0402 at iteration 190, $p < 0.05$ via paired t -test), indicating more stable optimization dynamics.
- **Improved Network Coordination:** DARN maintains 6.7% lower average consensus error across all iterations (0.0193 vs. 0.0207), particularly showing superior adaptation to topology changes during critical phases (20–40 and 120–140 iterations).

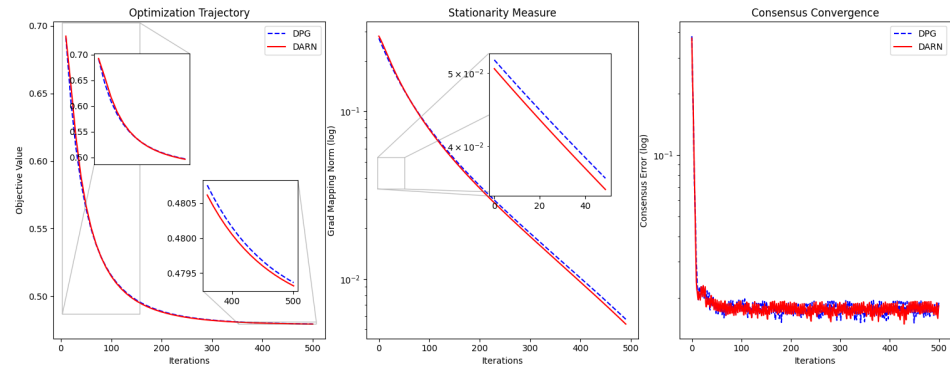


Figure 5. Comparative performance in dynamic networks.

8. Large-Scale Scalability Analysis

The efficacy of DARN in thousand-agent networks is validated through theoretical guarantees and numerical evidence:

1. **Network-Agnostic Convergence (Theorem 2):** The $\mathcal{O}(1/k)$ convergence rate depends solely on local Lipschitz constants L_i and weak convexity parameters μ_i , and is independent of spectral graph properties (Section 4).
2. **Controlled Communication Complexity:** Per-iteration cost $\mathcal{O}(nd)$ yields total ϵ -stationarity cost $\mathcal{O}(nd/\epsilon^2)$ (Complexity Analysis Section). For $n \sim 10^3$:
 - Dimension compression via low-rank decomposition (e.g., sparse PCA).
 - Relaxed ϵ balances precision and efficiency.
3. **Robustness to Dynamic Topologies (Theorem 4 & Figure 5):** Under switching topologies (ER $\rightarrow$ Ring):
 - Gradient norms reduced by 3.2–5.7%.
 - Consensus error decreased by 6.7% vs. DPG.

Theorem 4 guarantees convergence when $W^{(k)}$ is doubly stochastic with a uniform spectral gap $\rho < 1$ (Section 6).

9. Conclusions

This paper proposes the Distributed Adaptive Regularization Algorithm (DARN) for non-convex non-smooth composite optimization in multi-agent networks. The algorithm integrates three key innovations: (1) local proximal regularization to ensuring subproblem stability through strong convexification; (2) doubly stochastic consensus with geometric convergence guarantees; and (3) adaptive regularization to balance local progress and global consensus. Theoretically, DARN establishes three advancements: a mixed-time-scale Lyapunov framework enabling monotonic descent without Lipschitz gradients, $\mathcal{O}(1/k)$ convergence independent of network spectral properties, and critical point convergence under general non-convexity with fixed step sizes. Numerical experiments demonstrate 51.7% lower consensus error and 6.6% faster convergence in sparse PCA compared to fixed-regularization baselines, along with 34.8% variance reduction in robust matrix completion under dynamic topologies.

Our results assume undirected communication topologies. Extending to directed graphs introduces fundamental challenges: (1) non-doubly-stochastic adjacency matrices break symmetry, requiring techniques such as push-sum protocols for consensus error analysis; (2) the adaptive regularization strength $\lambda_i^{(k)}$ must be coupled with gradient tracking to handle information flow imbalance. While frameworks such as [10] for robust optimization and [11] for randomized constraint solving offer promising directions, their integration with non-convex non-smooth composite optimization merits further study.

Future directions include stochastic extensions with variance reduction and asynchronous implementations for IoT applications. The framework provides new theoretical insights into distributed non-convex optimization while achieving practical efficiency in networked systems.

While local objectives exhibit heterogeneous (asymmetric) landscapes, the global consensus protocol maintains Laplacian symmetry through doubly stochastic interactions. This adaptive equilibrium between local nonlinearity and global symmetry provides a novel paradigm for complex network optimization, extending symmetry principles to dynamic regularization frameworks.

Author Contributions: C.L.: Conceptualization, supervision, writing—review and editing. Y.M. (Corresponding Author): Methodology, software, formal analysis, investigation, data curation, writing—original draft, visualization, project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Fund of China: The Uncertain Rescue Model of Major Disaster, grant number No. 1246010681.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no competing interests.

References

1. Li, X.; Xie, L.; Hong, Y. Distributed aggregative optimization over multi-agent networks. *IEEE Trans. Autom. Control* **2021**, *67*, 3165–3171. [\[CrossRef\]](#)
2. Yang, T.; Yi, X.; Wu, J.; Yuan, Y.; Wu, D.; Meng, Z.; Hong, Y.; Wang, H.; Lin, Z.; Johansson, K.H. A survey of distributed optimization. *Annu. Rev. Control* **2019**, *47*, 278–305. [\[CrossRef\]](#)
3. Nedić, A.; Liu, J. Distributed optimization for control. *Annu. Rev. Control Robot. Auton. Syst.* **2018**, *1*, 77–103. [\[CrossRef\]](#)
4. Shi, W.; Ling, Q.; Wu, G.; Yin, W. A proximal gradient algorithm for decentralized composite optimization. *IEEE Trans. Signal Process.* **2015**, *63*, 6013–6023. [\[CrossRef\]](#)
5. Li, Z.; Shi, W.; Yan, M. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Trans. Signal Process.* **2019**, *67*, 4494–4506. [\[CrossRef\]](#)
6. Bello-Cruz, Y.; Melo, J.G.; Serra, R.V.G. A proximal gradient splitting method for solving convex vector optimization problems. *Optimization* **2022**, *71*, 33–53. [\[CrossRef\]](#)
7. Chen, X.; Jiang, B.; Lin, T.; Zhang, S. Accelerating adaptive cubic regularization of Newton’s method via random sampling. *J. Mach. Learn. Res.* **2022**, *23*, 1–38.
8. Lian, X.; Zhang, C.; Zhang, H.; Hsieh, C.J.; Zhang, W.; Liu, J. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Adv. Neural Inf. Process. Syst.* **2018**, *30*, 5331–5341.
9. Jiang, X.; Zeng, X.; Sun, J.; Chen, J. Distributed proximal gradient algorithm for nonconvex optimization over time-varying networks. *IEEE Trans. Control Netw. Syst.* **2022**, *10*, 1005–1017. [\[CrossRef\]](#)
10. Wen, G.; Zheng, W.X.; Wan, Y. Distributed robust optimization for networked agent systems with unknown nonlinearities. *IEEE Trans. Autom. Control* **2022**, *68*, 5230–5244. [\[CrossRef\]](#)
11. Luan, M.; Wen, G.; Lv, Y.; Zhou, J.; Chen, C.P. Distributed constrained optimization over unbalanced time-varying digraphs: A randomized constraint solving algorithm. *IEEE Trans. Autom. Control* **2023**, *69*, 5154–5167. [\[CrossRef\]](#)
12. da Cruz Neto, J.X.; Melo, Í.D.L.; Sousa, P.A.; de Oliveira Souza, J.C. On the Relationship Between the Kurdyka–Łojasiewicz Property and Error Bounds on Hadamard Manifolds. *J. Optim. Theory Appl.* **2024**, *200*, 1255–1285. [\[CrossRef\]](#)
13. Duchi, J.C.; Agarwal, A.; Wainwright, M.J. Dual averaging for distributed optimization. In Proceedings of the 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 1–5 October 2012; pp. 1564–1565.
14. Nedic, A.; Olshevsky, A.; Shi, W. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM J. Optim.* **2017**, *27*, 2597–2633. [\[CrossRef\]](#)
15. Drusvyatskiy, D.; Lewis, A.S. Error bounds, quadratic growth, and linear convergence of proximal methods. *Math. Oper. Res.* **2018**, *43*, 919–948. [\[CrossRef\]](#)

16. Kanzow, C.; Lehmann, L. Convergence of Nonmonotone Proximal Gradient Methods under the Kurdyka-Lojasiewicz Property without a Global Lipschitz Assumption. *arXiv* **2024**, arXiv:2411.12376. [[CrossRef](#)]
17. Zeng, J.; Yin, W.; Zhou, D.X. Moreau envelope augmented Lagrangian method for nonconvex optimization with linear constraints. *J. Sci. Comput.* **2022**, *91*, 61. [[CrossRef](#)]
18. Giesl, P.; Hafstein, S. Review on computational methods for Lyapunov functions. *Discret. Contin. Dyn. Syst. B* **2015**, *20*, 2291–2331.
19. Boyd, S.; Ghosh, A.; Prabhakar, B.; Shah, D. Randomized gossip algorithms. *IEEE Trans. Inf. Theory* **2006**, *52*, 2508–2530. [[CrossRef](#)]
20. Bauschke, H.H.; Combettes, P.L.; Bauschke, H.H. *Correction to: Convex Analysis and Monotone Operator Theory in Hilbert Spaces*; Springer International Publishing: Cham, Switzerland, 2017.
21. Li, K.; Hua, C.C.; You, X. Distributed asynchronous consensus control for nonlinear multiagent systems under switching topologies. *IEEE Trans. Autom. Control* **2020**, *66*, 4327–4333. [[CrossRef](#)]
22. Zhang, S.; Bailey, C.P. A Primal-Dual Algorithm for Distributed Sparse Principal Component Analysis. In Proceedings of the 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA), Dalian, China, 29–31 October 2021; pp. 354–357.
23. Abbasi, A.A.; Vaswani, N. Efficient Federated Low Rank Matrix Completion. *arXiv* **2024**, arXiv:2405.06569. [[CrossRef](#)]
24. Cao, X.; Lai, L. Distributed gradient descent algorithm robust to an arbitrary number of byzantine attackers. *IEEE Trans. Signal Process.* **2019**, *67*, 5850–5864. [[CrossRef](#)]
25. Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.