Achieving Linear Convergence with Parameter-Free Algorithms in Decentralized Optimization

Ilya Kuruzov
Innopolis University
kuruzov.ia@phystech.edu.

Gesualdo Scutari
Purdue University
gscutari@purdue.edu.

Alexander Gasnikov
Innopolis University
gasnikov@yandex.ru

Abstract

This paper addresses the minimization of the sum of strongly convex, smooth functions over a network of agents without a centralized server. Existing decentralized algorithms require knowledge of functions and network parameters, such as the Lipschitz constant of the global gradient and/or network connectivity, for hyperparameter tuning. Agents usually cannot access this information, leading to conservative selections and slow convergence or divergence. This paper introduces a decentralized algorithm that eliminates the need for specific parameter tuning. Our approach employs an operator splitting technique with a novel variable metric, enabling a local backtracking line-search to adaptively select the stepsize without global information or extensive communications. This results in favorable convergence guarantees and dependence on optimization and network parameters compared to existing nonadaptive methods. Notably, our method is the first *adaptive* decentralized algorithm that achieves linear convergence for strongly convex, smooth objectives. Preliminary numerical experiments support our theoretical findings, demonstrating superior performance in convergence speed and scalability.

1 Introduction

We study optimization across a network of $m > 1$ agents, modeled as an undirected, static graph, possibly with no centralized server. The agents cooperatively solve the following problem:

$$\min_{x\in\mathbb{R}^n} \sum_{i=1}^{m} f_i(x), \tag{P}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ is the loss function of agent i, assumed to be strongly convex and smooth (i.e., with gradient being Lipschitz continuous), and accessible only to agent i.

This formulation applies to various fields, particularly emphasizing decentralized machine learning problems where datasets are produced and collected at different locations. Traditionally, statistical and computational methods in this domain have relied on a centralized paradigm, aggregating computational resources at a single, central location. However, this approach is increasingly unsuitable for modern applications with many machines, leading to server congestion, inefficient communication, and high energy consumption [27, 23]. This has motivated the surge of learning algorithms that target *decentralized* networks with *no servers*, a.k.a. *mesh* networks, which is the setting of this paper.

Decentralized convex optimization has a long history, with numerous algorithms applicable to Problem (P); recent tutorials include [34, 41, 6, 33, 45]. **Lack of adaptivity:** These methods share the hurdle of relying sensibly on the tuning of hyperparameters, such as the stepsize (a.k.a. learning rate), for both theoretical and practical convergence. Existing theories ensure convergence under generally conservative bounds on the stepsize, which depend on parameters like the Lipschitz constant of the global gradient, the spectral gap of the graph adjacency matrix, or other topological properties. Acquiring such information is challenging in practice, due to physical or privacy limitations and computational/communication constraints. This often leads to manual tuning, which is not only tedious but also results in less predictable, problem-dependent, and non-reproducible performance.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3042

Parameter-free centralized methods: On the the hand, significant progress has been made in the *centralized* setting to automate the selection of the stepsize across various optimization and learning problem classes. **(i)** Traditional approaches in optimization–such as line-search methods [36], Barzilai-Borwein's stepsize [3], and Polyak's stepsize [37]–have been supplemented by recent adaptive stepsize rules based on estimates of local curvature [30] and subsequent techniques [31, 19, 20, 22, 50]. **(ii)** In the ML community, adaptive gradient methods such as AdaGrad [11], Adam [18], AMSGrad [40], NSGD-M [9], and variants [25, 44, 29] have gained significant attention for training large-scale learning models. These methods apply to stochastic, nonconvex optimization problems. **(iii)** Further advancements extend adaptivity to stochastic/online convex optimization problems, e.g., [5, 15].

Distributed adaptive methods: While variant of these centralized algorithms have been adapted to federated architectures (server-client systems), e.g., in [39, 24, 8], their application to mesh networks is *not feasible*. In federated learning, a central server aggregates local model updates, a process integral to its hierarchical structure. However, mesh networks, which lack a centralized coordinating node, do not support such a direct aggregation of large-scale vectors. Recent attempts to implement some form of stepsize adaptivity for *stochastic (non)convex/online* optimization problems over mesh networks are [32, 7, 21]. These methods generally achieve adaptivity by properly normalizing agents' gradients using past information. However, with the exception of [21], they rely on the strong assumption that the (population) losses are *globally* Lipschitz continuous (i.e., their gradients are bounded). In fact, Lipschitz continuity in convex optimization readily unlocks parameter-free convergence by using stepsize tuning of $\mathcal{O}(1/\sqrt{k})$ (here, k is the iteration index). Moreover, [32, 7] still require knowledge of some optimization parameters for the stepsize tuning, to guarantee convergence.

Attempts to introduce adaptivity in decentralized optimization for solving (P) have been explored in [12, 14, 13]. These methods bring the Barzilai-Borwein (BB)'s stepsize strategy into gradient tracking algorithms [43, 28, 35, 48]. **(i)** However, convergence of these algorithms *is not guaranteed* under the proposed BB strategy, unless the stepsizes *remain uniformly bounded from below and above* throughout the algorithm's trajectory–a condition the BB rule *does not inherently satisfy* in decentralized settings. Furthermore, these bounds for the stepsizes are typically unknown to the agents, as they depend on the strong convexity and smoothness constants of all agents' losses. Even with such knowledge, enforcing these conservative bounds contradicts the principle of adaptivity by potentially negating the advantages of a variable stepsize strategy that adapts based on local loss curvature, producing stepsize values significantly larger than theoretical thresholds used in nonadaptive methods. **(ii)** Additionally, to ensure contraction of the iterates, studies such as [13, 14] require multiple rounds of communications per iteration (gradient evaluation)–this demands the knowledge of network and optimization parameters at the agents' sides, making practical implementation unfeasible. **(iii)** None of these studies offer expressions of convergent rates for the explored algorithms, leaving it unclear whether the BB stepsize rule can provably outperform nonadaptive methods. **(iv)** Lastly, the methods discussed employ the traditional BB rule, which is only proven in centralized settings to produce convergence methods when minimizing *quadratic* losses. Simulations in [12, 13] are in fact performed only on quadratic functions.

Open questions and challenges: To our knowledge, no deterministic, parameter-free decentralized algorithms exist that solve Problem (P) over mesh networks, particularly achieving linear convergence when agents' functions are strongly convex and smooth. The current decentralized adaptive stochastic methods [32, 7, 21] discussed earlier do not adequately bridge this gap. Tailored for stochastic environments, these methods merely ensure that cumulative consensus errors along the iterations remain bounded, *not necessarily decreasing*. This typically involves either diminishing stepsizes or adjustments based on the final horizon to manage the bias-variance trade-off. These strategies fall short in deterministic scenarios like Problem (P), failing to ensure convergence to *exact* solutions, and achieve faster $\mathcal{O}(1/k)$ convergence rates in convex cases or *linear* rates in strongly convex scenarios.

Major contributions: This paper addresses this open problem. Our contributions are the following:

1. *A new parameter-free decentralized algorithm:* We propose a decentralized algorithm that eliminates the need for specific tuning of the stepsize. Our approach leverages a Forward-Backward operator splitting technique combined with a novel variable metric, enabling a local backtracking line-search procedure to adaptively select the stepsize at each iteration without requiring global information on optimization and network parameters or extensive communications. We are not aware of any other provable decentralized line-search methods over mesh networks.

https://doi.org/10.52202/079017-3042

Designing decentralized line-search procedures that are well-defined (terminating in a finite number of steps), locally implementable, and ensure algorithm convergence through satisfactory descent on an appropriate merit function presents significant challenges. A major issue is that line-search procedures merely based on the local curvature of agents' functions often fail to ensure convergence, producing *excessively large*, heterogeneous stepsizes that, e.g., poorly connected networks cannot support. This necessitates the identification of line-search *directions* and *surrogate functions* that encapsulate *both* optimization and network influences, aspects that have not yet formalized. Our design guidelines (cf., Sec. 3) are of independent interest; hopefully they will provide valuable insights for the development of other decentralized adaptive schemes, such as those based on alternative operator splittings.

2. Convergence guarantees: We have established linear convergence for the proposed decentralized adaptive method. Our analysis identifies critical quantities that capture the interplay between optimization conditions and network topology, directly influencing the convergence rates. Specifically: (a) In "well-connected" networks, the convergence rate exhibits a *separation property*: the overall rate is dictated by the slower of either the centralized gradient algorithm solving the same problem or a consensus algorithm run on the same mesh network. (b) Conversely, in "poorly" connected networks, the separation property vanishes, and the convergence rates are adversely affected by network degradation terms, still exhibiting a linear dependence on the condition number of the optimization loss. **(ii)** Unlike many existing distributed optimization frameworks, the optimization parameters in our rate expressions–such as smoothness and strong convexity constants–are localized to the *convex hull* of the traveled iterates. This localization arises from our adaptive stepsize strategy, which employs a line-search procedure tailored to local geometries, yielding more favorable dependencies on optimization parameters and thus enhanced convergence guarantees. **(iii)** Numerical experiments demonstrate superior performance of the proposed adaptive algorithm in convergence speed and scalability compared to existing non-adaptive methods.

1.1 Notation and paper organization

Capital letters denote matrices. Bold capital letters represent matrices where each row is an agent's variable, e.g., $\mathbf{X} = [x_1, \ldots, x_m]^\top$. For such matrices, the i-th row is denoted by the corresponding lowercase letter with the subscript i; e.g., for $\mathbf{X}$, we write x_i (as column vector). Let $\mathbb{S}^m$, $\mathbb{S}^m_+$, and $\mathbb{S}^m_{++}$ be the set of $m \times m$ (real) symmetric, symmetric positive semidefinite, and symmetric positive definite matrices, respectively; $A^\dagger$ denotes the Moore-Penrose pseudoinverse of A. The eigenvalues of $W \in \mathbb{S}^m$ are ordered in nonincreasing order, and denoted by $\lambda_1(W) \geq \cdots \geq \lambda_m(W)$. For two operators A and B of appropriate size, $(A \circ B)(\bullet)$ stands for $A(B(\bullet))$. We denote: $[m] = \{1, \ldots, m\}$, for any integer $m \geq 1$; $[x]_+ := \max(x, 0)$, $x \in \mathbb{R}$; $1_m \in \mathbb{R}^m$ is the vector of all ones; I_m (resp. 0_m) is the $m \times m$ identity (resp. the $m \times m$ zero) matrix; the information on the dimension is omitted when not necessary; $\texttt{null}(A)$ (resp. $\texttt{span}(A)$) is the nullspace (resp. range space) of the matrix A. Let $\langle X, Y \rangle := \texttt{tr}(X^\top Y)$, for any X and Y of suitable size ($\texttt{tr}(\bullet)$) is the trace operator; and $\|X\|_M := \sqrt{\langle MX, X \rangle}$, for any symmetric, positive definite M and X of suitable dimensions. We still use $\|X\|_M$ when M is positive semidefinite and $X \in \texttt{span}(M)$.

2 Problem Setup

We investigate Problem (P) over a network of $[m]$ agents, modeled as an undirected, static, connected graph $\mathcal{G} = ([m], \mathcal{E})$, where $(i, j) \in \mathcal{E}$ if there is communication link (edge) between i and j. For each agent i, we define by $\mathcal{N}_i := \{j : \,|\, (i, j) \in \mathcal{E}, \text{ for some } i \in [m]\} \cup \{i\}$ the set of immediate neighbors of agent i (including agent i itself).

Assumption 1. **(i)** *Each function f_i in* (P) *is L-smooth and μ-strong convex on $\mathbb{R}^n$, for some $L \in (0, \infty)$ and $\mu \in (0, \infty)$; and* **(ii)** *each agent i has access only to its own function f_i.*

The following matrices are commonly utilized in the design of gossip-based algorithms.

Definition 2 (Gossip matrices). *Let $\mathcal{W}_\mathcal{G}$ denote the set of matrices $\widetilde{W} = [\widetilde{W}_{ij}]_{i,j=1}^m$ that satisfy the following properties:* **(i)** *(compliance with $\mathcal{G}$) $\widetilde{W}_{ij} > 0$ if $(i, j) \in \mathcal{E}$; otherwise $\widetilde{W}_{ij} = 0$. Furthermore, $\widetilde{W}_{ii} > 0$, for all $i \in [m]$; and* **(ii)** *(doubly stochastic) $\widetilde{W} \in \mathbb{S}^m$ and $\widetilde{W} 1_m = 1_m$.*

These matrices are standard in the literature on decentralized optimization algorithms, and several instances have been employed in practice; see [34, 41, 33] for some representative examples. Notice that for any $\widetilde{W} \in \mathcal{W}_{\mathcal{G}}$ (assuming $\mathcal{G}$ connected) it hold: **(i)** (null space condition) $\texttt{null}(I_m - W) = \texttt{span}(1_m)$; and **(ii)** (eigen-spectrum distribution) $2I \succeq \widetilde{W} + I \succ 0_m$.

3 Algorithm Design

Our approach to solving Problem (P) involves a saddle-point reformulation tackled via a variable metric operator splitting, implementable across the graph $\mathcal{G}$. The innovative aspect of the proposed method lies in the selection of the variable metric that, coupled with a Forward Backward Splitting (FBS), enable adaptive stepsize selections through a decentralized line-search procedures.

Introducing local copies $x_i \in \mathbb{R}^d$ of the shared variable x (the i-th one is controlled by agent i), and the stack matrix $\mathbf{X} := [x_1, \ldots, x_m]^\top \in \mathbb{R}^{m \times n}$, let us consider the following auxiliary problem:

$$\min_{\mathbf{X} \in \mathbb{R}^{m \times n}} \left[F(K\mathbf{X}) := \sum_{i=1}^{m} f_i([K\mathbf{X}]_i) \right], \quad \text{s.t.} \;\; \text{Ł}\,\mathbf{X} = 0. \tag{P$'$}$$

Here, Ł and K are $m \times m$ matrices that meet the following criteria: **(c1)** $\text{Ł} \in \mathbb{S}^m$ and $\texttt{null}(\text{Ł}) = \texttt{span}(1_m)$; **(c2)** $K \in \mathbb{S}^m_{++}$ and $\texttt{null}(I - K) = \texttt{span}(1_m)$; and **(c3)** Ł and K commute. Conditions (c1) and (c2) ensure that (P) and (P′) are equivalent. Specifically, any solution $\mathbf{X}^\star$ of (P′) has the form of $\mathbf{X}^\star = 1_m (x^\star)^\top$, where $x^\star$ solves (P), and vice versa. While not essential, condition (c3) is postulated to simplify the algorithm derivation.

Primal-dual optimality for (P′) reads, with $\mathbf{Y}$ being the dual-variable associated with the constraints,

$$(A + B)\left(\begin{bmatrix}\mathbf{X}^\star \\ \mathbf{Y}^\star\end{bmatrix}\right) = 0, \quad \text{where} \quad A := \begin{bmatrix} K \circ \nabla F \circ K & 0 \\ 0 & 0 \end{bmatrix} \text{ and } B := \begin{bmatrix} 0 & \text{Ł} \\ -\text{Ł} & 0 \end{bmatrix}.$$

Given $\mathbf{X}^k, \mathbf{Y}^k$ at iteration k, the update $\mathbf{X}^{k+1}, \mathbf{Y}^{k+1}$ via FBS with metric $C \in \mathbb{S}^{2m}_{++}$ reads [4]

$$(C + B)\left(\begin{bmatrix}\mathbf{X}^{k+1} \\ \mathbf{Y}^{k+1}\end{bmatrix}\right) = (C - A)\left(\begin{bmatrix}\mathbf{X}^{k} \\ \mathbf{Y}^{k}\end{bmatrix}\right). \tag{1}$$

Monotone operator theory [4] ensures convergence of (1) under the following conditions:

(c4) B is a monotone operator, $C \in \mathbb{S}^{2m}_{++}$, and **(c5)** $I - C^{-1/2} A C^{-1/2}$ is an averaged operator.

Condition (c4) is satisfied by construction; (c5) can be enforced through a suitable selection of $C \in \mathbb{S}^{2m}_{++}$ while leveraging the co-coercivity of A (implied by Assumption 1). Denoting by $\alpha > 0$ the stepsize employed in the algorithm, we seek for C with the following structure:

$$C = \begin{bmatrix} \alpha^{-1} C_1 & 0 \\ 0 & C_2 \end{bmatrix}, \quad \text{with} \quad C_1, C_2 \in \mathbb{S}^m_{++}$$

to be determined. We proceed solving (1). Taking $(C + B)^{-1}$, we have

$$\begin{aligned} \mathbf{X}^{k+1} &= (I)\,(\mathbf{X}^k) - \alpha\left((II)\,(\mathbf{X}^k) + (III)\,(\mathbf{Y}^k)\right), \\ \mathbf{Y}^{k+1} &= (IV)\,(\mathbf{Y}^k) + (V)\,(\mathbf{X}^k), \end{aligned} \tag{2}$$

where

$$\begin{aligned} (I) &:= I_m - \alpha \cdot C_1^{-1}\,\text{Ł}\,\left(C_2 + \alpha \cdot \text{Ł}\,C_1^{-1}\,\text{Ł}\right)^{-1}\,\text{Ł}, \\ (II) &:= (I)\,C_1^{-1}\,K\,\nabla F \circ K, \\ (III) &:= C_1^{-1}\,\text{Ł}\,\left(C_2 + \alpha \cdot \text{Ł}\,C_1^{-1}\,\text{Ł}\right)^{-1}\,C_2, \\ (IV) &:= \left(C_2 + \alpha \cdot \text{Ł}\,C_1^{-1}\,\text{Ł}\right)^{-1}\,C_2, \\ (V) &:= \left(C_2 + \alpha \cdot \text{Ł}\,C_1^{-1}\,\text{Ł}\right)^{-1}\,\text{Ł}\left(I - \alpha \cdot C_1^{-1} K\,\nabla F \circ K\right). \end{aligned} \tag{3}$$

In addition to satisfying (c5), $C_1, C_2 \in \mathbb{S}^m_{++}$ must be strategically chosen to facilitate the design of a decentralized line-search procedure for α. We propose the following guiding principles:

(c6) The range of admissible stepsize values α ensuring convergence–hence satisfying (c5)–should be independent of the network parameters; and

(c7) the operators (I), (II), and (III) in (2) should be independent of α.

At a high level, (c6) aims to decouple the line-search mechanism from network-dependent constraints. By doing so, it ensures that performing the line-search from the agents' sides requires no mid-process communications during backtracking, relying solely on local computations. Meanwhile, (c7) facilitates the identification of $-((II)(\mathbf{X}^k) + (III)(\mathbf{Y}^k))$ in (2) as a potential line-search direction. This direction must be paired with an appropriate surrogate function, which we will define shortly.

Among several potential selections, in this paper, we consider the following for C_1 and C_2:

$$C_1 = K \quad \text{and} \quad C_2 = \alpha K^{-1}\left(c^{-1}\, I - \L^2\right), \text{ with } c < 1/2, \tag{4}$$

which satisfy all the specified requirements. Using (4) and (c3), the operators in (3) simplify to

$$(I) = I_m - c{\cdot}\L^2,\ (II) = (I)\nabla F{\circ}K,\ (III) = \L K^{-1},\ (IV) = (I),\ (V) = \frac{c}{\alpha}{\cdot}K\,\L\,(I - \alpha\nabla F{\circ}K).$$

Notice that (I), (II), and (III) are independent of the stepsize. Substituting the above expressions in (2) and introducing $\mathbf{D}^k := K^{-1}\L\mathbf{Y}^k$, the algorithm can be rewritten as

$$\begin{aligned}
\mathbf{X}^{k+1} &= (I - c\L^2)\,\mathbf{X}^k - \alpha\cdot(I - c\L^2)\left(\mathbf{D}^k + \nabla F(K\mathbf{X}^k)\right),\\
\mathbf{D}^{k+1} &= (I - c\L^2)\,\mathbf{D}^k + \frac{c}{\alpha}\cdot\L^2\left(\mathbf{X}^k - \alpha\nabla F(K\mathbf{X}^k)\right).
\end{aligned}$$

To make the above updates compliant with the graph $\mathcal{G}$ while satisfying (c1)-(c3), we set $\L^2 = (I - \widetilde{W})$, with $\widetilde{W} \in \mathcal{W}_{\mathcal{G}}$, and $K = I - c\L^2$, where $c \in (0, 1/2)$ is a free universal constant. Introducing $W := (1-c)I_m + c\widetilde{W} \in \mathcal{W}_{\mathcal{G}}$, the final decentralized algorithm can be rewritten as

$$\begin{aligned}
\mathbf{X}^{k+1/2} &= W\,\mathbf{X}^k, \quad \mathbf{D}^{k+1/2} = W\,\left(\mathbf{D}^k + \nabla F(\mathbf{X}^{k+1/2})\right),\\
\mathbf{X}^{k+1} &= \mathbf{X}^{k+1/2} - \alpha\cdot\mathbf{D}^{k+1/2},\\
\mathbf{D}^{k+1} &= \mathbf{D}^{k+1/2} + \frac{1}{\alpha}\cdot\left(\mathbf{X}^k - \mathbf{X}^{k+1/2} - \alpha\nabla F(\mathbf{X}^{k+1/2})\right).
\end{aligned} \tag{5}$$

Finally, it can be verified that (c6) is met if $(\sqrt{\alpha}K^{-1/2}) \circ \nabla F \circ (\sqrt{\alpha}K^{-1/2})$ is nonexpansive, which holds if $\alpha < 1/L$, being independent on the network parameters. Next, we introduce a line-search procedure that enables the use of an adaptive stepsize α rather than a constant one satisfying the above more conservative bound.

Decentralized backtracking: It is not difficult to check that $-\mathbf{D}^{k+1/2}$ is a descent direction of $F^k(\mathbf{X}) := F(\mathbf{X}) + \langle\mathbf{D}^k, \mathbf{X}\rangle$ at $\mathbf{X}^{k+1/2}$. This naturally suggests the following backtracking procedure for α: at iteration k, find the largest $\alpha^k > 0$ such that

$$F^k(\mathbf{X}^{k+1}) \leq F^k(\mathbf{X}^{k+1/2}) + \left\langle \nabla F^k(\mathbf{X}^{k+1/2}), \mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\right\rangle + \frac{\delta}{2\alpha^k}\|\mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\|^2, \tag{6}$$

where $\delta \in (0, 1]$ is a tuning parameter. However, this condition would require a communication round for each backtracking step. To reduce the communication burden, we introduce a local stepsize for each agent i, denoted by α_i^k, determined by a backtracking line-search on the local function $f_i^k(x) := f_i(x) + \langle d_i^k, x\rangle$. Specifically, each α_i^k is the largest positive value satisfying

$$f_i^k(x_i^{k+1}) \leq f_i^k(x_i^{k+1/2}) + \left\langle \nabla f_i^k(x_i^{k+1/2}), x_i^{k+1} - x_i^{k+1/2}\right\rangle + \frac{\delta}{2\alpha_i^k}\|x_i^{k+1} - x_i^{k+1/2}\|^2. \tag{7}$$

Clearly $\alpha^k = \min_{i\in[m]} \alpha_i^k$ also satisfies (6). Noticing that f_i^k has the same smooth (and strong convexity) constant(s) of f_i, one can replace f_i^k in (7) with f_i. The proposed decentralized algorithm is summarized in Algorithm 1, with the backtracking line-search procedure detailed in Algorithm 2.

3.1 Discussion

Several comments are in order.

On the proposed algorithm: We emphasize that selecting $K \neq I_m$ in (P′) marks a significant departure from the commonly used saddle-point reformulations of Problem (P), where $K = I_m$, e.g., [46, 34, 33, 1]. Choosing $K \neq I_m$, in conjunction with the novel variable metric C in the FBS as

Algorithm 1

Data: (i) Initialization $\mathbf{X}^0 \in \mathbb{R}^{m\times n}$ and $\mathbf{D}^0 = 0$; (ii) initial value $\alpha_{-1} \in (0,\infty)$; (iii) Backtracking parameters $\delta \in (0,1]0$; (iv) nondecreasing sequence $\{\gamma^k\}_k \subseteq [1,\infty)$ (v) Gossip matrix $W := (1-c)I_m + c\widetilde{W}$, with $\widetilde{W} \in \mathcal{W}_{\mathcal{G}}$, and $c \in (0,1/2]$. Set the iteration index $k = 0$.

```
(S.1) Communication step: Agents updates primal and dual variables via gossiping:
    $\mathbf{X}^{k+1/2} = W\,\mathbf{X}^k \quad \text{and} \quad \mathbf{D}^{k+1/2} = W\left(\mathbf{D}^k + \nabla F(\mathbf{X}^{k+1/2})\right);$
(S.2) Decentralized line-search: Each agent updates $\alpha_i^k$ according to
    $\alpha_i^k = \texttt{Backtracking}\left(\alpha^{k-1}, f_i, x_i^{k+1/2}, -d_i^{k+1/2}, \gamma^k, \delta\right);$
(S.3) Global min-consensus:
    $\alpha^k = \min_{i\in[m]} \alpha_i^k;$
(S.4) Local updates of the primal and dual variables:
    $\mathbf{X}^{k+1} = \mathbf{X}^{k+1/2} - \alpha^k \cdot \mathbf{D}^{k+1/2},$
    $\mathbf{D}^{k+1} = \mathbf{D}^{k+1/2} + \frac{1}{\alpha^k} \cdot \left(\mathbf{X}^k - \mathbf{X}^{k+1/2} - \alpha^k \nabla F(\mathbf{X}^{k+1/2})\right).$
(S.5) If a termination criterion is not met, $k \leftarrow k+1$ and go to step (S.1).
```

Algorithm 2 Backtracking(α, f, x, d, γ, δ)

```
1: $\alpha^+ := \gamma\alpha;$
2: $x^+ := x + \alpha^+ d$; set $t = 1$;
3: while $f(x^+) > f(x) + \langle \nabla f(x), x^+ - x\rangle + \frac{\delta}{2\alpha^+}\|x^+ - x\|^2$ do
4:     $\alpha^+ \leftarrow (1/2)\alpha^+;$
5:     $x^+ := x + \alpha^+ d;$
6:     $t \leftarrow t + 1;$
   return $\alpha^+$.
```

specified in (4), is critical to obtain a valid line-search procedure that is also implementable across the network. For instance, popular decentralized algorithms such as EXTRA [42] and NIDS [26] can be interpreted as FBS with suitable metrics associated with the primal-dual reformulation of (P) as (P′) but with $K = I_m$. However, these schemes do not facilitate any suitable line-search, as no stepsize-independent descent direction can be identified in their updates. Hopefully, our approach will provide principled guidelines for the design of other parameter-free decentralized algorithms, stemming from alternative decentralized formulations of (P) and their corresponding operator splittings.

On the backtracking: The following Lemma shows that the line-search procedure in Algorithm 2 is well-defined, as long as the function f therein is locally smooth.

Lemma 3. *Let f in Algorithm 2 be any L_f-smooth and μ_f-strongly convex function on the segment $[x, x+\gamma\alpha d]$, where $L_f \in (0,\infty)$, $\mu_f \in [0,\infty)$, and $\gamma \in [1,\infty)$. The following hold for Algorithm 2:*

(i) The returned α^+ satisfies

$$\min\left(\gamma\alpha, \frac{\delta}{2L_f}\right) \le \alpha^+ \le \min\left(\gamma\alpha, \frac{\delta}{\mu_f}\right) \le \infty. \tag{8}$$

Therefore, the backtracking procedure terminates in at most $\max\left(1, \lceil \log_2 \frac{2L_i\gamma\alpha}{\delta}\rceil\right)$ *t-steps;*

(ii) For any α^+ returned by Algorithm 2, any $\bar{\alpha}^+ \in (0,\alpha^+]$ also satisfies the backtracking condition.

Notice that the last statement of the lemma guarantees that the each $\alpha^k = \min_{i\in[m]} \alpha_i^k$ satisfies the descent property (6) on the global loss F^k, as each α_i^k meets the local condition (7).

The sequence $\{\gamma^k\}_{k=1}^\infty$ used in line 1 of the backtracking algorithm, with each $\gamma^k \ge 1$, is introduced to favor nonmonotone, and thus potentially larger, stepsize values between two consecutive line-search calls. Any sequence satisfying $\gamma^k \downarrow 1$ and $\prod_{k=1}^{\infty} \gamma^k = \infty$, is advisable. In our experiments, we found the following rule quite effective: $\gamma^k = \big((k+\beta_1)/(k+1)\big)^{\beta_2}$, for some $\beta_2 > 0$ and $\beta_1 \geq 1$. One can also opt for $\gamma^k = 1$, for all k, thus eliminating this extra parameter, if simplicity is desired.

On the min-consensus: Step `(S.3)` involves a min-consensus across the network to establish a common stepsize, $\alpha^k = \min_{i\in[m]} \alpha_i^k$, among the agents. This procedure is easily implemented in federated systems, where a server node facilitates information exchange between clients. Interestingly, this min-consensus protocol is also well-suited to current wireless mesh network technologies. Modern networks support multi-interface communications, including WiFi and LoRa (Low-Range) [17, 2, 16]. WiFi allows high-speed, short-range communications, supporting a mesh topology where nodes transmit large data volumes to immediate neighbors. Conversely, LoRa facilitates long-range but low-rate communications, ideal for communication flooding that reaches all network nodes in a single hop but transmits minimal information. Therefore, in multi-interface networks, the proposed algorithm operates by transmitting vector variables in Steps `(S.1)` via WiFi, while LoRa is used for the min-consensus in Step `(S.3)`. Furthermore, the values α_i^k's can be quantized to their nearest lower values using a few bits before transmission. Based on Lemma 3(ii), this quantization ensures that the descent condition (6) is still met with the resultant min quantized stepsize. This approach renders the extra communication cost for implementing the global min-consensus step negligible.

For networks where LoRa technology cannot be used, Sec. 5 proposes a variant of Algorithm 1 wherein the global min-consensus step `(S.3)` is replaced by a local min-consensus procedure.

4 Convergence Results

We begin introducing a quantity of interest that helps identifying different operational regimes of the proposed algorithm. Let $(\mathbf{X}^\star, \mathbf{D}^\star)$ be a fixed point of Algorithm 1 (whose existence is ensured by Assumption 1), and let $\{(\mathbf{X}^k, \mathbf{D}^k)\}$ be the iterates generated by Algorithm 1. Define

$$r^k = \frac{\sqrt{\frac{1}{(\alpha^k)^2}\|\mathbf{X}^k\|^2_{c(I-\widetilde{W})} + \left\| c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\right\|_M^2}}{\max\left(\frac{1}{\alpha^k}\|\mathbf{X}^k - \mathbf{X}^\star\|, \|\mathbf{D}^k - \mathbf{D}^\star\|_M\right)}, \text{ with } M := c^{-1}(I-\widetilde{W})^\dagger - I. \tag{9}$$

The following comments are in order. **(i)** Both $\mathbf{D}^k$ and $\mathbf{D}^\star$ lie in the $\mathtt{span}(I-\widetilde{W}) = \mathtt{span}(I-W)$, for all k, and M is positive defined on this span. Consequently, $\|\mathbf{D}^k - \mathbf{D}^\star\|_M > 0$ for all $\mathbf{D}^k \neq \mathbf{D}^\star$, and $\|\mathbf{D}^k - \mathbf{D}^\star\|_M = 0$ if and only if $\mathbf{D}^k = \mathbf{D}^\star$. **(ii)** Under Assumption 1, $\mathbf{X}^\star = 1(x^\star)^\top$, where $x^\star$ is the solution of (P). **(iii)** The quantity r^k reflects the algorithm's convergence progress through the evolution of the dual variables and consensus error. Rewriting the update of the dual variables as $\mathbf{D}^{k+1} = \mathbf{D}^k + \frac{c}{\alpha^k}(I-\widetilde{W})\mathbf{X}^k - c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)$, we observe that small values of $\|\frac{c}{\alpha^k}(I-\widetilde{W})\mathbf{X}^k - c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\|$ compared to $\|\mathbf{D}^k - \mathbf{D}^\star\|$ and $\|\mathbf{X}^k - \mathbf{X}^\star\|$–hence small r^k values–indicate slow convergence improvements (see Lemma 8 in the appendix).

We remark that r^k need not be known by the agents; it is instrumental only for the analysis and posterior assessment of algorithm performance.

Linear convergence is established below via contraction of the following merit function

$$V^k := \left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^{k-1})^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2. \tag{10}$$

Theorem 4. *Given Problem* (P) *under Assumption 1, let* $\{(\mathbf{X}^k, \mathbf{D}^k)\}$ *be the iterates generated by Algorithm 1. Then, the following holds*

$$V^{k+1} \leq \max\left(\frac{\alpha^k}{\alpha^{k-1}}, 1\right)^2 \left(1 - \min\left(\rho_1^k, \delta(r^k)^2\right)\right) V^k, \text{ where } \rho_1^k := \frac{\mu^k\alpha^k}{2}(1 - c(1-\lambda_m(\widetilde{W})))^2 < 1. \tag{11}$$

If $r^k < \sqrt{2}/4$, *then*

$$V^{k+2} \leq \max\left(\frac{\alpha^{k+1}}{\alpha^k}, 1\right)^2 \max\left(\frac{\alpha^k}{\alpha^{k-1}}, 1\right)^2 \left(1 - \rho_2^k\right) V^k, \tag{12}$$

where

$$\rho_2^k := \frac{(1-c(1-\lambda_m(\widetilde{W})))^2}{128(\gamma^k)^2\max(1, \lambda_{\max}(M))} \min\left(\mu^{k+1}\alpha^{k+1}, \mu^k\alpha^k, \frac{1}{L^k\alpha^k}\right) < 1.$$

Here μ^k (resp. μ^{k+1}) and L^k are the strong convexity and smoothness constants of (each) f_i along the segment $[x_i^{k+1/2}, x^\star]$ (resp. $[x_i^{k+1+1/2}, x^\star]$), respectively.

The theorem establishes linear convergence of Algorithm 1. As $\max(1, (\alpha^k/\alpha^{k-1})^2)$ is bounded away from zero and uniformly upper bounded (with value depending on the sequence $\{\gamma^k\}$)–see Lemma 3–the convergence rate is predominantly determined by $\{\rho_1^k\}$, $\{\rho_2^k\}$, and $\{r^k\}$. Notice that, in the setting of the theorem, each $\rho_1^k, \rho_2^k \in [0, 1)$. Intriguingly, the algorithm exhibits different operational regimes based on the range of values r^k takes along the traveled trajectory. At the high level, if r^k remains "large", faster convergence can be guaranteed, as certified by (11); otherwise V^k decreases every two consecutive iterations (see (12)), yielding to a slower convergence. The number of iterations required to reach a desired termination accuracy is given next.

Corollary 4.1. *Instate the setting of Theorem 4, with now $\{\gamma^k\}$ being chosen such that $\gamma^k \leq \big((k+\beta_1)/(k+1)\big)^{\beta_2}$, for all k and some $\beta_1 \geq 1, \beta_2 > 0$. Then*

$$\left\|\mathbf{X}^{k+1} - \mathbf{X}^\star\right\|^2 + \frac{1}{4L^2}\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M^2 \leq \varepsilon,$$

for all $k \geq N_\varepsilon$, where N_ε is given as follows:

If $r^k \geq r_{\texttt{low}} := \frac{1}{\sqrt{2}} \min\left(\frac{1}{2}, \frac{1}{\sqrt{\lambda_{\max}(M)}}\right)$, *for all* k,

$$N_\varepsilon = \mathcal{O}\left(\frac{1}{\delta}\max\left(\frac{1}{c(1-\lambda_2(\widetilde{W})}, \frac{\kappa}{(1-c(1-\lambda_m(\widetilde{W})))^2}\right)\log(V^0/\varepsilon)\right); \tag{13}$$

otherwise,

$$N_\varepsilon = \mathcal{O}\left(\frac{1}{\delta}\frac{\kappa}{(1-c(1-\lambda_m(\widetilde{W})))^2 c(1-\lambda_2(\widetilde{W}))}\log(V^0/\varepsilon)\right). \tag{14}$$

Here κ is the condition number of each f_i restricted to the convex hull of $\{x^\star, \{x_i^k, x_i^{k+1/2}\}_{k=0}^{N_\varepsilon}\}$, and $\mathcal{O}$ hides the dependence on β_1 and β_2.

Corollary 4.1 identifies the following two different operational regimes of the algorithm, resulting in difference performance based upon the network connectivity and optimization condition number.

(1) Strong connectivity regime: when $r^k \geq r_{\texttt{low}}$, for all k, a fact that numerically has been observed for 'relatively good' network connectivity, the convergence rate exhibits a separation in the dependence on the network and optimization parameters. Since $1 - c(1 - \lambda_m(\widetilde{W})) > 1 - 2c$, it follows that, when $c(1 - \lambda_2(\widetilde{W})) \geq (1 - 2c)/\sqrt{\kappa}$, N_ε reduces to $\mathcal{O}(\kappa)$ (omitting the dependence on ε), which matches the complexity of the centralized gradient algorithm. This suggests scenarios where the optimization problem is harder than a consensus problem over the same network, resulting in the bottleneck between the two. Conversely, when the condition number κ is large relative to the network connectivity $1 - \lambda_2(\widetilde{W})$, the rate is determined by that of the consensus algorithm running on the same network, that is, $\mathcal{O}((1 - \lambda_2(\widetilde{W}))^{-1})$. The above rate separation property mirrors that of certain *nonadaptive* primal-dual decentralized schemes including NEXT [10], AugDGM [47], Exact Diffusion [49] (with rate expression as improved in [46]), NIDS [26], and ABC [46].

(2) Worst-case regime: This regime reflects the algorithm's worst-case performance, typically registered in "weakly" connected networks: the convergence rate reads $\mathcal{O}(\kappa/(1 - \lambda_2(\widetilde{W})))$, where optimization and network parameters are now mixed. This rate aligns with those of *nonadaptive* decentralized gradient-tracking schemes, such as DGing [35], SONATA [43] (subject to sufficiently small network connectivity), and [38].

In summary, the convergence rate of Algorithm 1 resembles in the form that of existing nonadaptive decentralized methods, but offers more favorable dependence on the condition number than that typically found in those algorithms. Specifically, the condition number in (13) and (14) is the *local* condition number, defined on the convex hull of the trajectory, which is generally much smaller than the *global* condition number governing decentralized algorithms in the literature. This demonstrates the algorithm's capability to adapt to the local geometry of the optimization problem.

5 From Global to Local Min-Consensus

This section extends Algorithm 1 to settings where the global min-consensus procedure in `(S.3)` is not implementable. For these cases, we propose to replace such a step with a *local* min-consensus procedure. The new algorithm is formally described in Algorithm 3 and briefly commented next.

In step `(S.3)`, each agent now computes its stepsize taking the minimum values among those of their immediate neighbors only (including itself). This produces possibly different stepsizes α_i^k for each agent, collected in the diagonal matrix $\Lambda^k = \texttt{diag}(\alpha_1^k \ldots \alpha_m^k)$. Because of that, in order to still guarantee $\mathbf{D}^k \in \texttt{span}(I - \widetilde{W})$–a key property for the convergence of the algorithm–we slightly modified the updates of the dual variable in `(S.4)`, compared with the same step in Algorithm 1. Specifically, the updating direction of the dual variable as in Algorithm 1, $(\alpha^k)^{-1}(\mathbf{X}^k - \mathbf{X}^{k+1/2} - \alpha^k \nabla F(\mathbf{X}^{k+1/2}))$, is replaced in Algorithm 3 by $(\Lambda^k)^{-1}\mathbf{X}^k - \mathbf{X}_\Lambda^{k+1/2} - \nabla F(\mathbf{X}^{k+1/2})$, where $\mathbf{X}_\Lambda^{k+1/2} = W\,(\Lambda^k)^{-1}\mathbf{X}^k$. Notice that, if all the stepsizes are equal, the update `(S.4)` in Algorithm 3 reduced to that in Algorithm 1. Finally, we point out that the computation of $\mathbf{X}_\Lambda^{k+1/2}$ requires only the extra communication of neighboring stepsizes (thus scalar) values, which has a negligible cost.

Algorithm 3

Data: (i) Initialization $\mathbf{X}^0 \in \mathbb{R}^{m\times n}$ and $\mathbf{D}^0 = 0$; (ii) initial value $\alpha_{-1} \in (0, \infty)$; (iii) Backtracking parameters $\delta \in (0, 1]$; (iv) nondecreasing sequence $\{\gamma^k\}_k \subseteq [1, \infty)$ (v) Gossip matrix $W := (1-c)I_m + c\widetilde{W}$, with $\widetilde{W} \in \mathcal{W}_{\mathcal{G}}$, and $c \in (0, 1/2]$. Set the iteration index $k = 0$.

```
(S.1) Communication step: Agents updates primal and dual variables via gossiping:
      $$\mathbf{X}^{k+1/2} = W\,\mathbf{X}^k \quad \text{and} \quad \mathbf{D}^{k+1/2} = W\left(\mathbf{D}^k + \nabla F(\mathbf{X}^{k+1/2})\right);$$
(S.2) Decentralized line-search: Each agent updates $\overline{\alpha}_i^k$ according to
      $$\overline{\alpha}_i^k = \texttt{Backtracking}\left(\alpha^{k-1}, f_i, x_i^{k+1/2}, d_i^{k+1/2}, \gamma^k, \delta\right);$$
(S.3) Local min-consensus:
      $$\alpha_i^k = \min_{j\in\mathcal{N}_i} \overline{\alpha}_j^k, \quad \forall i \in [m];$$
      Define $\Lambda^k = \texttt{diag}(\alpha_1^k \ldots \alpha_m^k)$;
(S.4) Local updates of the primal and dual variables:
      $$\mathbf{X}^{k+1} = \mathbf{X}^{k+1/2} - \Lambda^k \cdot \mathbf{D}^{k+1/2}, \quad \mathbf{X}_\Lambda^{k+1/2} = W\,(\Lambda^k)^{-1}\mathbf{X}^k,$$
      $$\mathbf{D}^{k+1} = \mathbf{D}^{k+1/2} + (\Lambda^k)^{-1}\mathbf{X}^k - \mathbf{X}_\Lambda^{k+1/2} - \nabla F(\mathbf{X}^{k+1/2}).$$
(S.5) If a termination criterion is not met, $k \leftarrow k+1$ and go to step (S.1).
```

Convergence of Algorithm 3 is established in the following theorem.

Theorem 5. *Instate assumptions in Theorem 4, applied now to Algorithm 3, with $\{\gamma^k\}$ being chosen such that $\gamma^k \le \big((k+\beta_1)/(k+1)\big)^{\beta_2}$, for all k and some $\beta_1 \ge 1, \beta_2 > 0$. Further, suppose there exists a constant $R > 0$ such that $V^k \le R$, for all k. Then*

$$\min_{j\in[1,N+1]} \left\|\mathbf{X}^j - \mathbf{X}^\star\right\|^2 + \frac{1}{4L^2}\|\mathbf{D}^j - \mathbf{D}^\star\|_M^2 \le \varepsilon,$$

with

$$N = \mathcal{O}\left(\max\left(\log d_{\mathcal{G}} + \log N_\varepsilon, \log \alpha_0 L\right) \max(N_\varepsilon, d_{\mathcal{G}})\right),$$

where N_ε is defined as in Corollary 4.1 (replacing therein V_0 with R).

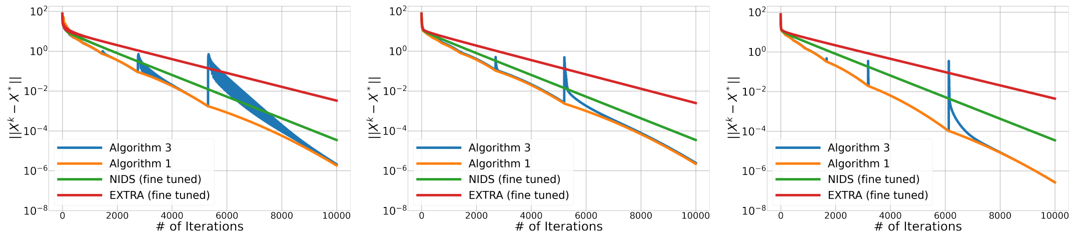
Interestingly, Theorem 5 states that the degradation of the convergence rate when a local-min consensus is used instead of the global one is mild. Specifically, up to log factors, the total number of iterations to ε-optimality depends on $d_{\mathcal{G}}$ (the diameter of the graph $\mathcal{G}$), if $d_{\mathcal{G}} > N_\varepsilon$. This result is somehow expected, as min-consensus requires a number of iterations proportional to $d_{\mathcal{G}}$ to propagate through the entire network. However, monotonicity in the decrease of the primal and dual errors can no longer be guaranteed when min-consensus is employed.

6 Numerical Results

This section presents some preliminary numerical results. We compare Algorithm 1 and Algorithm 3 with EXTRA [42] and NIDS [26] on a ridge regression problem using synthetic data. Further experiments are presented in the appendix. All experiments are run on Acer Swift 5 SF514-55TA-56B6, with processor Intel(R) Core(TM) i5-8250U @ CPU 1.60GHz, 1800 MHz.

Ridge regression: It is an instance of (P), with $f_i(x) = \|A_i x_i - b_i\|^2 + \sigma\|x_i\|_2^2$, where we set $A_i \in \mathbb{R}^{20\times 300}, b_i \in \mathbb{R}^{20}$, and $\sigma = 0.1$.The elements of A_i, b_i are independently sampled from the standard normal distribution; the regularization is set to $\sigma = 0.1$. We simulated a network of $m = 20$ agents, and the following three different graph topologies, reflecting varying connectivity levels: **(i)** $\mathcal{G}_1$: Graph-path with $m-1$ edges and diameter $m-1$, i.e., $\mathcal{G} = \{[m], \{(i, i+1)\}_{i=1}^{m-1}\}$; **(ii)** $\mathcal{G}_2$: Erdős–Rényi graph, sparsely connected; and **(iii)** $\mathcal{G}_3$: Erdős–Rényi graph, well-connected.

Results are summarized in Fig. 1 and Fig. 2. For EXTRA and NIDS we use a grid-search tuning, chosen to achieve the best practical performance. Algorithm 1 and Algorithm 3 are simulated under the following choice of the line-search parameters satisfying Corollary 4.1: $\gamma^k = (k+2)/(k+1)$, $\delta = 1$. For all the algorithms we used the Metropolis-Hastings weight matrix $W \in \mathcal{G}_{\mathcal{W}}$ [34].

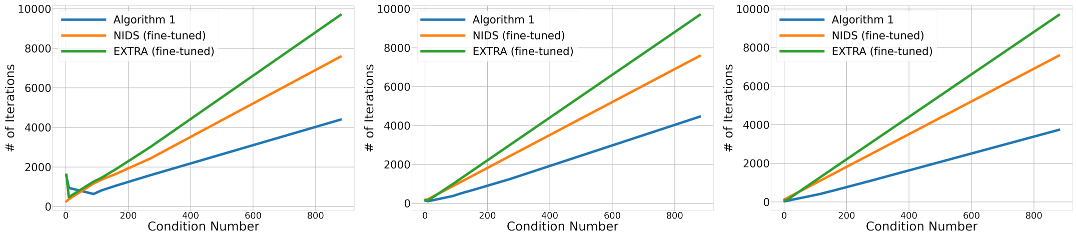


(a) Line Graph

(b) Erdős-Rényi Graph, $p = 0.1$

(c) Erdős-Rényi Graph, $p = 0.5$

Figure 1: **Ridge regression** on different graphs: (1a) Line graph; (1b) Erdős-Rényi Graph with edge activation probability $p = 0.1$; (1c) Erdős-Rényi Graph with edge activation probability $p = 0.5$



(a) Line Graph

(b) Erdős-Rényi Graph, $p = 0.1$

(c) Erdős-Rényi Graph, $p = 0.5$

Figure 2: **Ridge regression**: Number of iterations N for $\|\mathbf{X}^N - \mathbf{X}^\star\| \leq 10^{-5}$ versus the condition number of agents’ looses on different graphs; (2a) Line graph; (2b) Erdős-Rényi Graph with edge activation probability $p = 0.1$; (2c) Erdős-Rényi Graph with edge activation probability $p = 0.5$

The figures demonstrate that the proposed method consistently outperforms both EXTRA and NIDS, even when using the local min-consensus strategy, with a significant gap emerging as the condition number increases. This performance is particularly noteworthy given that Algorithm 1 and 3 operate effectively without requiring tedious tuning or global knowledge of the optimization and network parameters. Notably, Algorithms 1 and 3 exhibit different convergence behaviors: as predicted by Theorem 5, local min-consensus results in nonmonotonic error dynamics $\|\mathbf{X}^k - \mathbf{X}^\star\|^2$. However, the practical convergence speed remains largely unaffected compared to the global min-consensus.

Acknowledgment

The work of I. Kuruzov has been supported by the Grant No. 70-2021-00143 01.11.2021, IGK 000000D730324P540002.

References

[1] S. A. Alghunaim, E. K. Ryu, K. Yuan, and A. H. Sayed. Decentralized proximal gradient algorithms with linear convergence rates. *IEEE Transactions on Automatic Control*, 66(6):2787–2794, June 2021.

[2] A. Askhedkar, B. Chaudhari, and M. Zennaro. *Hardware and software platforms for low-power wide-area networks*, page 397–407. Elsevier, 2020.

[3] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[4] H. H. Bauschke and P.L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, New York, NY, 2011.

[5] Y. Carmon and O. Hinder. Making sgd parameter-free. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 2360–2389. PMLR, 02–05 Jul 2022.

[6] T. Chang, M. Hong, H. Wai, X. Zhang, and S. Lu. Distributed learning in the nonconvex world: From batch data to streaming and beyond. *IEEE Signal Processing Magazine*, 37(3):26–38, 2020.

[7] X. Chen, B. Karimi, W. Zhao, and P. Li. On the convergence of decentralized adaptive gradient methods. In *Asian Conference on Machine Learning*, pages 217–232. PMLR, 2023.

[8] X. Chen, X. Li, and P. Li. Toward communication efficient adaptive gradient method. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, page 119–128, Virtual Event USA, October 2020. ACM.

[9] A. Cutkosky and H. Mehta. Momentum improves normalized SGD. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 13–18 Jul 2020.

[10] P. Di Lorenzo and G. Scutari. NEXT: In-network nonconvex optimization. *IEEE Trans. Signal Inf. Process. Netw.*, 2(2):120–136, June 2016.

[11] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 257–265, 2011.

[12] Iyanuoluwa E. and Chinwendu E. Q-linear convergence of distributed optimization with barzilai-borwein step sizes. In *58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8, 2022.

[13] J. Gao, XW. Liu, YH. Dai, HUang Y., and P. Yang. Achieving geometric convergence for distributed optimization with barzilai-borwein step sizes. *Sci. China Inf. Sci.*, 65:149–204, 2022.

[14] J. Hu, X. Chen, L. Zheng, L. Zhang, and H. Li. (rectified version) the barzilai–borwein method for distributed optimization over unbalanced directed networks. *arXiv:2305.11469v3*, 2024.

[15] M. Ivgi, O. Hinder, and Y. Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14465–14499. PMLR, 23–29 Jul 2023.

[16] T. Janssen, N. BniLam, M. Aernouts, R. Berkvens, and M. Weyn. Lora 2.4 ghz communication link and range. *Sensors*, 20(16):4366, August 2020.

[17] D.H. Kim, J.Y. Lim, and J.D. Kim. Low-power, long-range, high-data transmission using wi-fi and lora. In *2016 6th International Conference on IT Convergence and Security (ICITCS)*, page 1–3, Prague, Czech Republic, September 2016. IEEE.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

https://doi.org/10.52202/079017-3042

[19] P. Latafat, A. Themelis, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *arXiv preprint arXiv:2301.04431*, 2023.

[20] P. Latafat, A. Themelis, and P. Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. *arXiv preprint arXiv:2311.18431*, 2023.

[21] J. Li, X. Chen, S. Ma, and M. Hong. Problem-parameter-free decentralized nonconvex stochastic optimization. *arXiv preprint arXiv:2402.08821*, 2024.

[22] T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. *arXiv preprint arXiv:2310.10082*, 2023.

[23] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[24] X. Li, B. Karimi, and P. Li. On distributed adaptive optimization with gradient compression. In *International Conference on Learning Representations (ICLR)*, 2022.

[25] X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTAT)*. PMLR, 2019.

[26] Z. Li, W. Shi, and M. 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.

[27] X. Lian, C. Zhang, H. Zhang, C. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 5330–5340, 2017.

[28] P. Di Lorenzo and G. Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

[29] L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, Louisiana, May 2019.

[30] Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. In *International Conference on Machine Learning*, 2019.

[31] Y. Malitsky and K. Mishchenko. Adaptive proximal gradient method for convex optimization. *arXiv preprint arXiv:2308.02261*, 2024.

[32] P. Nazari, D.A. Tarzanagh, and G. Michailidis. Dadam: A consensus-based distributed adaptive gradient method for online optimization. *IEEE Transactions on Signal Processing*, 70:6065–6079, 2022.

[33] A. Nedic. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.

[34] A. Nedić, A. Olshevsky, and M. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106:953–976, 2018.

[35] A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27:2597–2633, July 2016.

[36] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 2 edition, 2006.

[37] B.T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

[38] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, Sept 2018.

[39] S. Reddi, Z. Burr Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konev$\mathbf{v}cn$, S. Kumar, and B. McMahan. Adaptive federated optimization. In *International Conference on Learning Representations (ICLR)*, 2021.

[40] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018.

[41] A. H. Sayed. Adaptation, learning, and optimization over networks. *Foundations and Trends in Machine Learning*, 7:311–801, January 2014.

[42] W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM J. on Optimization*, 25(2):944–966, November 2015.

[43] Y. Sun, G. Scutari, and A. Daneshmand. Distributed optimization based on gradient-tracking revisited: Enhancing convergence rate via surrogation. *SIAM J. on Optimization*, 32:354–385, 2022.

[44] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *The Journal of Machine Learning Research*, 21:1–30, 2020.

[45] R. Xin, S. Pu, A. Nedic, and U. A. Khan. A general framework for decentralized optimization with first-order methods. *Proceedings of the IEEE*, 108(11):1869–1889, November 2020.

[46] J. Xu, Y. Tian, Y. Sun, and G. Scutari. Distributed algorithms for composite optimization: Unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69:3555–3570, 2021.

[47] J. Xu, S. Zhu, Y.-C. Soh, and L. Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *Proceedings of the 54th IEEE Conference on Decision and Control*, pages 2055–2060, 2015.

[48] J. Xu, S. Zhu, Y. Chai Soh, and L. Xie. Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Trans. Automat. Contr.*, 63(2):434–448, 2017.

[49] K. Yuan, B. Ying, X. Zhao, and A. H. Sayed. Exact diffusion for distributed optimization and learning–part i: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723, 2018.

[50] D. Zhou, S. Ma, and J. Yang. Adabb: Adaptive barzilai-borwein method for convex optimization. *arXiv preprint arXiv:2401.08024*, 2024.

Appendix

A Proof of Lemma 3

The proof of the lemma is quite standard, and it is reported here for the sake of completeness.

(i) Smoothness of f implies that Algorithm 2 terminates when $\alpha^+ \leq \delta/L_f$. Therefore, it must be $\alpha^+ \geq \min(\delta/2L_f, \gamma\alpha)$. Furthermore, it follows from the strong convexity of f that $\delta/(2\alpha^+) \geq \mu/2$; hence, $\alpha^+ \leq \min(\delta/\mu, \gamma\alpha)$. This proves (8).

Further, by the lower bound above, one infers that the backtracking procedure terminates when $\alpha^+ \leq \frac{\delta}{L_i}$. Noting that $\alpha^+ = 2^{-t+1}\gamma$, we deduce that $t = \left\lfloor \log_2 \frac{2L_i\gamma\alpha}{\delta} \right\rfloor$ interations suffice.

(ii) Let $\phi(\alpha) := f(x + \alpha d)$. Notice that ϕ is convex and $\phi'(0) = \langle \nabla f(x), d\rangle$. The termination condition in Algorithm 2 can be equivalently rewritten in terms of ϕ as

$$\phi(\alpha^+) \leq \phi(0) + \phi'(0)\,\alpha^+ + \alpha^+ \frac{\delta}{2}\|d\|^2. \tag{15}$$

Given $\lambda \in [0, 1]$, let $\bar{\alpha} = \lambda\alpha^+$. Invoking convexity of ϕ, we can write

$$\begin{aligned}\phi(\bar{\alpha}) &= \phi\left(\lambda\alpha^+ + (1-\lambda)0\right) \leq \lambda\phi(\alpha^+) + (1-\lambda)\phi(0) \\ &\overset{(15)}{\leq} \phi(0) + \phi'(0)\,(\lambda\alpha^+) + (\lambda\alpha^+)\frac{\delta}{2}\|d\|^2,\end{aligned}$$

which completes the proof. □

B Proof of Theorem 4

We begin establishing the dynamics of V^k defined in (10) along two consecutive updates.

Lemma 6. *The following holds along the update* $(\mathbf{D}^k, \mathbf{X}^k) \to (\mathbf{D}^{k+1}, \mathbf{X}^{k+1})$*:*

$$\begin{aligned}V^{k+1} = &\left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ &- \|\mathbf{X}^k - \mathbf{X}^{k+1}\|^2 - (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^{k+1}\|_M^2 \\ &+ 2\alpha^k\left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^\star - \mathbf{X}^{k+1}\right\rangle.\end{aligned} \tag{16}$$

Proof. See Appendix E.1. □

Using the properties of the backtracking procedure (Lemma 3) and leveraging strong convexity and smoothness of F, the inner product in (16) can be bounded as follows.

Lemma 7. *The following holds:*

$$\begin{aligned}\left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^\star - \mathbf{X}^{k+1}\right\rangle \leq &\frac{\delta}{2\alpha^k}\|\mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\|^2 \\ &- \max\left(\frac{\mu^k}{2}\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\|^2, \frac{1}{2L^k}\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^*\|^2\right),\end{aligned}$$

where μ^k (resp. L^k) is the strong convexity (resp. smoothness) constant of each f_i along the segment $[x_i^{k+1/2}, x^\star]$.

Proof. See Appendix E.2. □

Combining Lemma 6 and Lemma 7, after some algebraic manipulation, we obtain the following.

Lemma 8. *In the setting above, it holds*

$$\begin{aligned}V^{k+1} \leq &\left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ &- \max\left(\mu^k\alpha^k\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\|^2, \frac{\alpha^k}{L^k}\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^*\|^2\right) \\ &- \delta\left(\|\mathbf{X}^k\|^2_{c(I-\widetilde{W})} + (\alpha^k)^2\left\|c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\right\|_M^2\right).\end{aligned} \tag{17}$$

Proof. See Appendix E.3. □

Lemma 8 suggests the path for the rest of the analysis: the decrease of V^{k+1} relies on the values of the terms

$$\|\mathbf{X}^k\|^2_{c(I-\widetilde{W})} + (\alpha^k)^2 \left\| c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\right\|^2_M$$

and

$$\max\left(\mu^k\alpha^k\|\mathbf{X}^k-\mathbf{X}^\star\|^2, \frac{\alpha^k}{L}\|\nabla F(\mathbf{X}^{k+1/2})+\mathbf{D}^*\|^2\right)$$

relative to the the primal and dual optimality gaps $\|\mathbf{X}^k-\mathbf{X}^\star\|^2$ and $(\alpha^k)^2\|\mathbf{D}^k-\mathbf{D}^\star\|^2_M$, respectively. At the high-level, one can say that "higher" values of such quantities relative to $\|\mathbf{X}^k-\mathbf{X}^\star\|^2$ and $(\alpha^k)^2\|\mathbf{D}^k-\mathbf{D}^\star\|^2_M$, determine larger decrease of V^{k+1}.

The above argument can be formally recorded by the following quantities:

$$r^k := \frac{\sqrt{\frac{1}{(\alpha^k)^2}\|\mathbf{X}^k\|^2_{c(I-\widetilde{W})} + \left\| c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\right\|^2_M}}{\max\left(\frac{1}{\alpha^k}\|\mathbf{X}^k-\mathbf{X}^\star\|, \|\mathbf{D}^k-\mathbf{D}^\star\|_M\right)}, \tag{18}$$

and

$$g^k := \frac{\max\left(\frac{1}{\alpha^k}\|\mathbf{X}^k-\mathbf{X}^\star\|, \|\nabla F(\mathbf{X}^{k+1/2})+\mathbf{D}^*\|\right)}{\|\mathbf{D}^k-\mathbf{D}^\star\|_M}. \tag{19}$$

Using r^k and g^k in (17), the next lemma establishes contraction of V^{k+1}, with a contraction factor depending in particular on such quantities.

Lemma 9. *The following holds*

$$V^{k+1} \le \max\left(\frac{\alpha^k}{\alpha^{k-1}}, 1\right)^2 \left(1-\min\left(\rho_1^k, \zeta^k\right)\right) V^k, \tag{20}$$

where

$$\rho_1^k := \frac{\mu^k\alpha^k}{2}(1-c(1-\lambda_m(\widetilde{W})))^2 < 1$$

and

$$\zeta^k := \max\left(\delta(r^k)^2, (g^k)^2 \min\left(\frac{\mu^k\alpha^k(1-c(1-\lambda_m(\widetilde{W})))^2}{2}, \frac{1}{2L^k\alpha^k}\right)\right).$$

Proof. See Appendix E.4. □

The final expression (11) in Theorem 4 follows easily from $\zeta^k \ge \delta(r^k)^2$.

The above result ensures a "sufficient" descent of V^{k+1} when r^k (or g^k) is large enough. However, the contraction factor in (20) becomes vacuous for arbitrarily small values of r^k (or g^k).

Next, we examine the unfavorable case where both r^k *and* g^k are "small", leading to the proof of the decay of V^{k+1} as stated in (12) of Theorem 4. We build on the following key property of the sequence g^k in this scenario: under low r^k values, if g^k is "small", the subsequent value g^{k+1} cannot become arbitrarily small.

Lemma 10. *Suppose*

$$r^k < \frac{1}{\sqrt{2}} \quad \textit{and} \quad g^k \le \min\left(\frac{1-r^k\sqrt{2}}{2\sqrt{\lambda_{\max}(M)}}, 1\right). \tag{21}$$

Then,

$$\frac{1}{(\alpha^k)^2}\frac{\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2}{\|\mathbf{D}^{k+1}-\mathbf{D}^\star\|^2_M} \ge \frac{1}{\lambda_{\max}(M)}\left(1-\frac{2g^k\sqrt{\lambda_{\max}(M)}}{1-r^k\sqrt{2}}\right)^2. \tag{22}$$

Proof. See Appendix E.5. □

We infer from Lemma 10 that

$$g^{k+1} \overset{(19)}{\geq} \frac{1}{\alpha^{k+1}} \frac{\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|}{\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M}$$
$$\overset{(22)}{\geq} \frac{1}{\gamma^k \sqrt{\lambda_{\max}(M)}} \left(1 - \frac{2g^k \sqrt{\lambda_{\max}(M)}}{1 - r^k \sqrt{2}}\right),$$

where we used $\alpha^k/\alpha^{k+1} \geq 1/\gamma^k$ (due to $\alpha^{k+1} \leq \gamma^k \alpha^k$, see Step 1 of Algorithm 2). Notice that (i) the term in the parenthesis will be around one for small enough values of g^k and r^k; and (ii) the sequence $\{\gamma^k\}$ is chosen being eventually uniformly lower bounded. Therefore, the above bound implies that r^k and g^k cannot both progressively diminish along the iterates. Consequently, in the unfavorable scenario described by (21), V^{k+1} still decreases, albeit over two consecutive iterations. This outcome is formalized in the following lemma.

Lemma 11. *Suppose condition* (21) *holds. Then,*

$$V^{k+2} \leq \max\left(\frac{\alpha^{k+1}}{\alpha^k}, 1\right)^2 \max\left(\frac{\alpha^k}{\alpha^{k-1}}, 1\right)^2 \left(1 - \hat{\rho}_2^k\right) V^k, \tag{23}$$

where

$$\hat{\rho}_2^k := \frac{\mu^{k+1}\alpha^{k+1}\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2}{2(\gamma^k)^2 \max(\lambda_{\max}(M), 1)} \left(1 - \frac{2g^k\sqrt{\lambda_{\max}(M)}}{1 - r^k\sqrt{2}}\right)^2 < 1.$$

Here, μ^{k+1} is the strong convexity constants of (each) f_i along the segment $[x_i^{(k+1)+1/2}, x^\star]$.

Proof. See Appendix E.6. □

The final convergence result as stated in (12) is obtained using Lemma 9 and Lemma 11, with the variable g^k, absorbed, as outlined next. We strengthen condition on r^k in (21) by $r^k \leq \sqrt{2}/4$. We consider two cases for the value of g^k, in the above scenario. Specifically, **(Case 1)** g^k is bounded away from zero, implying ζ^k in (20) to be so. Therefore, $\left(1 - \min\left(\rho_1^k, \zeta^k\right)\right) < 1$. This will be sufficient to ensure enough descent for V^{k+1}, though in two consecutive iterations. **(Case 2):** g^k may be arbitrarily small; in this case, $\hat{\rho}_2^k$ in (23) remains bounded away from zero, ensuring contraction though from V^k to V^{k+2}. More formally we have the following.

- **Case 1:** Consider (20), under (21) strengthened by $r^k \leq \sqrt{2}/4$. If

$$g^k \geq \min\left(\frac{1}{8\sqrt{\lambda_{\max}(M)}}, 1\right), \tag{24}$$

then

$$\min(\rho_1^k, \zeta^k) \geq \frac{\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2}{16\max(1, 8\lambda_{\max}(M))} \min\left(\mu^k\alpha^k, \frac{1}{\alpha^k L^k}\right).$$

Using lower bound above in (20), yields

$$V^{k+2} \leq \max\left(\frac{\alpha^{k+1}}{\alpha^k}, 1\right)^2 V^{k+1} \leq \max\left(\frac{\alpha^{k+1}}{\alpha^k}, 1\right)^2 \max\left(\frac{\alpha^k}{\alpha^{k-1}}, 1\right)^2 \left(1 - \hat{\rho}_1^k\right) V^k,$$

with

$$\hat{\rho}_1^k = \frac{\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2}{16\max(1, 8\lambda_{\max}(M))} \min\left(\mu^k\alpha^k, \frac{1}{\alpha^k L^k}\right). \tag{25}$$

Notice that if the interval of admissible values for g^k, as specified by (21) and (24) is empty, this case does not apply.

- **Case 2:** Consider (23) under (21) strengthened by $r^k \leq \sqrt{2}/4$. If

$$0 < g^k \leq \min\left(1, \frac{1}{8\sqrt{\lambda_{\max}(M)}}\right),$$

https://doi.org/10.52202/079017-3042

then

$$1 - \frac{2g^k\sqrt{\lambda_{\max}(M)}}{1 - r^k\sqrt{2}} > \frac{1}{2},$$

where we used $r^k \leq \sqrt{2}/4$. This implies

$$\hat{\rho}_2^k \geq \frac{\mu^{k+1}\alpha^{k+1}\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2}{8(\gamma^k)^2 \max(\lambda_{\max}(M), 1)}.$$

Using this lower bound in (23), yields

$$V^{k+2} \leq \max\left(\frac{\alpha^{k+1}}{\alpha^k}, 1\right)^2 \max\left(\frac{\alpha^k}{\alpha^{k-1}}, 1\right)^2 \left(1 - \hat{\rho}_3^k\right) V^k,$$

with

$$\hat{\rho}_3^k = \frac{\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2}{8\max(\lambda_{\max}(M), 1)} \frac{\mu^{k+1}\alpha^{k+1}}{(\gamma^k)^2}. \tag{26}$$

Combining Case 1 and Case 2 above–taking the minimum between (25) and (26) and using the fact that $\gamma^k \geq 1$, leads to the desired decay of V^{k+1} as in (12).

This completes the proof of Theorem 4. □

C Proof of Corollary 4.1

Let us consider the first $N > 0$ iterations of Algorithm 1. Let us denote by L and μ the constants of smoothness and strong convexity of each f_i restricted to the convex hull of $\{x^\star, \{x_i^k, x_i^{k+1/2}\}_{k=0}^N\}$. We proceed lower bounding ρ_1^k and ρ_2^k given in Theorem 4. We will use the following facts:

$$\alpha^k \geq \delta/(2L), \quad \alpha^k < \delta/\mu, \quad \text{and} \quad \frac{\alpha^{k+1}}{\alpha^k} \leq \gamma^k,$$

due to Lemma 7, and given $\lambda_{\max}(M) = (c(1 - \lambda_2(\widetilde{W}))^{-1} - 1$,

$$\frac{1}{\max(\lambda_{\max}(M), 1)} \geq c(1 - \lambda_2(\widetilde{W})) \quad \text{and} \quad \frac{1}{\lambda_{\max}(M)} \geq c(1 - \lambda_2(\widetilde{W})).$$

We can bound ρ_1^k and ρ_2^k as

$$\rho_1^k \geq \delta\frac{\mu}{4L}(1 - c(1 - \lambda_m(\widetilde{W})))^2 \quad \text{and} \quad \rho_2^k \geq \delta\frac{\mu}{L}\frac{\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2 c(1 - \lambda_2(\widetilde{W}))}{256(\gamma^k)^2}, \quad \forall k \leq N. \tag{27}$$

Using (27), we can simplify the rate decay of V^N in Theorem 4 as follows.

- **Case 1:** Suppose

$$r^k \geq r_{\texttt{low}} := \frac{1}{\sqrt{2}}\min\left(\frac{1}{2}, \frac{1}{\sqrt{\lambda_{\max}(M)}}\right), \quad \forall k \geq N. \tag{28}$$

Substituting the lower bounds of ρ_1^k and r^k in (11), we obtain the following simplified convergence rate:

$$V^N \leq \left(\prod_{k=0}^{N-1}\gamma^k\right)^2 \left(1 - \frac{\delta}{8}\min\left(\frac{\mu}{L}\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2, c(1 - \lambda_2(\widetilde{W}))\right)\right)^N V^0.$$

- **Case 2:** Condition (28) does not hold. For the values of k such that $r^k \leq \frac{1}{\sqrt{2}}\min\left(\frac{1}{2}, \frac{1}{\sqrt{\lambda_{\max}(M)}}\right) \leq \frac{\sqrt{2}}{4}$, we can use (12). Substituting threin the lower bound for ρ_2 and $\gamma^k = ((k + \beta_1)/(k+1))^{\beta_2} \geq \beta_1^{\beta_2}$, yields

$$V^{k+2} \leq \left(\gamma^k\gamma^{k+1}\right)^2 \left(1 - \delta\frac{(1 - c(1 - \lambda_m(\widetilde{W})))^2 c(1 - \lambda_2(\widetilde{W}))}{256\beta_1^{2\beta_2}}\frac{\mu}{L}\right) V^k. \tag{29}$$

On the other hand, for k such that $r^k \geq \frac{1}{\sqrt{2}} \min\left(\frac{1}{2}, \frac{1}{\sqrt{\lambda_{\max}(M)}}\right)$, using (11) on two consecutive iterations, we have

$$
\begin{aligned}
V^{k+2} \leq & (\gamma^{k+1})^2 V^{k+1} \\
& \leq \left(\gamma^k \gamma^{k+1}\right)^2 \left(1 - \frac{\delta}{8} \min\left(\frac{\mu}{L}\left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2, c(1 - \lambda_2(\widetilde{W}))\right)\right) V^k \\
& \leq \left(\gamma^k \gamma^{k+1}\right)^2 \left(1 - \delta \frac{(1 - c(1 - \lambda_m(\widetilde{W})))^2 c(1 - \lambda_2(\widetilde{W}))}{256 \beta_1^{2\beta_2}} \frac{\mu}{L}\right) V^k. \qquad (30)
\end{aligned}
$$

Therefore, in either situations of Case 2, one can ensure contraction after two consecutive iterations by a factor

$$
\delta \frac{(1 - c(1 - \lambda_m(\widetilde{W})))^2 c(1 - \lambda_2(\widetilde{W}))}{256 \beta_1^{2\beta_2}} \frac{\mu}{L} < 1.
$$

Using $V^{k+1} \leq (\gamma^k)^2 V^k$, we can merge (29) and (30) as follows:

$$
\begin{aligned}
V^N & \leq \left(\prod_{k=0}^{N-1} \gamma^k\right)^2 \left(1 - \delta \frac{(1 - c(1 - \lambda_m(\widetilde{W})))^2 c(1 - \lambda_2(\widetilde{W}))}{256 \beta_1^{2\beta_2}} \frac{\mu}{L}\right)^{\lfloor N/2 \rfloor} V^0 \\
& \leq \left(\prod_{k=0}^{N-1} \gamma^k\right)^2 \left(1 - \delta \frac{(1 - c(1 - \lambda_m(\widetilde{W})))^2 c(1 - \lambda_2(\widetilde{W}))}{256 \beta_1^{2\beta_2}} \frac{\mu}{L}\right)^{(N-1)/2} V^0 \\
& \leq \left(\prod_{k=0}^{N-1} \gamma^k\right)^2 \left(1 - \delta \frac{(1 - c(1 - \lambda_m(\widetilde{W})))^2 c(1 - \lambda_2(\widetilde{W}))}{512 \beta_1^{2\beta_2}} \frac{\mu}{L}\right)^{N-1} V^0.
\end{aligned}
$$

- **Case 1 + Case 2:** We can combine the rate expressions derived in the two cases above as follows:

$$
V^N \leq \left(\prod_{k=0}^{N-1} \gamma^k\right)^2 (1 - \rho)^{N-1} V^0,
$$

where

$$
\rho = \begin{cases} \frac{\delta}{8} \min\left(\frac{\mu}{4L}(1 - c(1 - \lambda_m(\widetilde{W})))^2, c(1 - \lambda_2(\widetilde{W}))\right), & \text{if } r^k \geq \frac{1}{\sqrt{2}} \min\left(\frac{1}{2}, \frac{1}{\sqrt{\lambda_{\max}(M)}}\right) \text{ for all } k; \\ \delta \frac{(1 - c(1 - \lambda_m(\widetilde{W})))^2 c(1 - \lambda_2(\widetilde{W}))}{512 (\gamma^k)^2} \frac{\mu}{L}, & \text{else.} \end{cases}
$$

Notice that $\rho \in (0, 1)$.

Finally, we can obtain the desired asymptotic convergent rate noting that the growth of $\prod_k \gamma^k$ is dominated by the geometric decay of the contraction factor. This is formalized next.

Lemma 12. *Let $\gamma^k = ((k + \beta_1)/(k + 1))^{\beta_2}$ with $\beta_1 \geq 1, \beta_2 \geq 0$. Then the following holds:*

$$
\prod_{k=0}^{N-1} \gamma^k \leq \beta_1^{\beta_2(\lceil \beta_1 \rceil + 1)} N^{\beta_2(\beta_1 - 1)}. \qquad (31)
$$

Furthermore, for any given $\rho \in (0, 1)$, we have

$$
\left(\prod_{k=0}^{N-1} \gamma^k\right)^2 (1 - \rho)^{N-1} \leq (1 - \rho/2)^{N-1}, \qquad (32)
$$

for all

$$
N \geq N_0 := \frac{4}{\rho} \max\left(2\beta_2(\lceil \beta_1 \rceil + 1) \ln \beta_1 + \ln(2), 4\beta_2 \beta_1 \ln \frac{8\beta_1 \beta_2}{\rho}\right).
$$

Proof. See Appendix E.7 □

Inequality (32) provides the asymptotic rate expression, as stated in the corollary where $\mathcal{O}$ hides the dependence on β_1 and β_2. □

D Proof of Theorem 5

We begin noticing that if the stepsizes in Algorithm 3 are identical across agents, Algorithm 3 reduces to Algorithm 1. For the iterates where this happens, one can rely on the convergence guarantees established for Algorithm 1. Specifically, we have the following result, whose proof is straightforward.

Lemma 13. *Suppose that exists some $k_\varepsilon \geq 1$ such that $\alpha_1^k = \cdots = \alpha_m^k$, for $k = k_\varepsilon, \ldots, k_\varepsilon + N_\varepsilon$. Then one can guarantee*

$$\left\|\mathbf{X}^{k_\varepsilon+N_\varepsilon} - \mathbf{X}^\star\right\|^2 + \frac{1}{4L^2}\|\mathbf{D}^{K_\varepsilon+N_\varepsilon} - \mathbf{D}^\star\|_M^2 \leq \varepsilon,$$

where N_ε is defined as in Corollary 4.1 (replacing therein V^0 with V^{K_ε}).

The remainder of the proof focuses on characterizing the properties of certain key events identified as detrimental for the local-min consensus algorithm to achieve convergence. We will demonstrate that the occurrence of such events within N consecutive iterations is of the order of $\log N$, indicating that these are sporadic events relative to the total of N iterations.

Given α_i^k's, as defined in Step `(S.3)` of Algorithm 3, let us denote their minimum across *all* agents at time k as

$$\alpha_{\min}^k := \min_{i\in[m]} \alpha_i^k.$$

If the backtracking loop (steps 3-6 in Algorithm 2) is not activated at iteration $k-1$ in *any* of the agents' local line searches, then all output stepsizes α_i^k will increase by the same factor γ^k, that is, $\alpha_i^k = \gamma^k \alpha_i^{k-1}$; hence, does $\alpha_{\min}^k$. On the other hand, if all stepsizes are consensual at iteration $k-1$ and the backtracking procedure at *some* of the agents' side enters its steps- 3-6, a "desynchronization" of the stepsizes occurs. This event can be detected by the condition

$$\alpha_{\min}^k < \gamma^k \alpha_{\min}^{k-1}.$$

When the stepsize at time $k-1$ are not consensual, the condition above identifies increases in stepsize disagreements from iteration $k-1$ to k, measured by

$$\max_{j\in[m]} \alpha_j^k - \alpha_{\min}^k > \gamma^k \left(\max_{j\in[m]} \alpha_j^{k-1} - \alpha_{\min}^{k-1}\right).$$

This motivates the definition of the following index set: given $N = 1, 2, \ldots,$ let

$$\mathcal{I}_N = \left\{k \in [N] : \alpha_{\min}^k < \gamma^k \alpha_{\min}^{k-1}\right\}.$$

The key properties of interest of this set are summarized below.

Lemma 14. *For any given $N = 1, 2, \ldots,$ the following statements hold:*

1. *If $\alpha_i^k = \alpha_{\min}^k$, for all $i \in [m]$, and $k+1 \notin \mathcal{I}_N$, then $\alpha_i^{k+1} = \alpha_{\min}^{k+1}$, for all $i \in [m]$;*
2. *If $k \in \mathcal{I}_N$, $k < N - d_\mathcal{G}$, and $k+1, \ldots k+d_\mathcal{G} \notin \mathcal{I}_N$, then $\alpha_i^{k+d_\mathcal{G}} = \alpha_{\min}^{k+d_\mathcal{G}}$, for all $i \in [m]$;*
3. *$|\mathcal{I}_N| \leq \max\left(\ln \alpha_0 L + \ln \prod_{k=0}^{N-1} \gamma^k + \ln \frac{2}{\delta}, 0\right)$.*

Proof. See Appendix F.1. □

In words, the first statement confirms that consensus on the stepsizes is maintained if none of the local backtracking procedures are triggered. The second statement ensures that, the local-min consensus algorithm requires at most $d_\mathcal{G}$ iterations to converge, from any initialization, provided that during those iterations no backtracking events alter the minimum stepsize across agents. Lastly, the third assertion provides a limit on the maximum number of detrimental events that can occur during the N iterations under consideration. If this number is small relative to N, a fact that will be proved shortly, one we can find (multiple) window(s) of consecutive iterations wherein the stepsizes remain consensual across all agents. Within these windows, Lemma 13 can be applied, to establish convergence. This idea is formalized next.

Lemma 15. *Suppose $V^k \leq R$. Then,*

$$\min_{j\in[1,N+1]} \left\|\mathbf{X}^j - \mathbf{X}^\star\right\|^2 + \frac{1}{4L^2}\|\mathbf{D}^j - \mathbf{D}^\star\|_M^2 \leq \varepsilon, \tag{33}$$

if

$$\frac{N}{N_\varepsilon + d_\mathcal{G}} > |\mathcal{I}_N| + 1. \tag{34}$$

Here, N_ε is defined as in Corollary 4.1 (replacing therein V_0 with R).

Proof. See Appendix F.2. □

To finalize our proof, let us simplify an upper bound of $|\mathcal{I}_N|$, when $\gamma^k = \left(\frac{k+\beta_1}{k+1}\right)^{\beta_2}$. For the sake of simplicity, we consider the case $\ln N \geq 1$. Invoking Lemma 12, we have

$$\begin{aligned}\ln\frac{2}{\delta} + \ln\prod_{k=0}^{N-1}\gamma^k &\overset{(31)}{\leq} \ln\frac{2}{\delta} + \beta_2(\lfloor\beta_1\rfloor+1)\ln\beta_1 + \beta_2(\beta_1-1)\ln N\\ &\leq \underbrace{\left(\beta_2(\lfloor\beta_1\rfloor+1)\ln\beta_1 + \beta_2(\beta_1-1) + \ln\ln\frac{2}{\delta}\right)}_{\xi:=}\ln N\\ &= \xi\ln N,\end{aligned}$$

where the constant ξ depends on the algorithm parameters. Therefore, one can guarantee (33), under the following condition

$$\frac{N}{\xi\ln N + \ln\alpha_0 L} \geq N_\varepsilon + d_\mathcal{G}. \tag{35}$$

A sufficient condition for (35) is

$$\frac{N}{\max\left(\ln N, \ln\alpha_0 L\right)} \geq 2\max(\xi,1)\max(N_\varepsilon, d_\mathcal{G}).$$

Let N^* be the smallest iteration for which the above inequality holds. Then,

$$N^* = \mathcal{O}(\max\left[\log d_\mathcal{G} + \log N_\varepsilon, \log\alpha_0 L\right]\max(N_\varepsilon, d_\mathcal{G})).$$

□

E Proof of the Intermediate Results in Appendix B

E.1 Proof of Lemma 6

Let us rewrite V^{k+1} in terms of $\|\mathbf{X}^k - \mathbf{X}^\star\|^2$ and $(\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2$, to link it back to V^k:

$$\begin{aligned}V^{k+1} =& \|\mathbf{X}^k - \mathbf{X}^\star\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2\\ &- \|\mathbf{X}^k - \mathbf{X}^{k+1}\|^2 - (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^{k+1}\|_M^2\\ &- 2\underbrace{\left\langle\mathbf{X}^{k+1} - \mathbf{X}^\star, \mathbf{X}^k - \mathbf{X}^{k+1}\right\rangle}_{\texttt{term I}} - 2\underbrace{(\alpha^k)^2\left\langle\mathbf{D}^{k+1} - \mathbf{D}^\star, \mathbf{D}^k - \mathbf{D}^{k+1}\right\rangle_M}_{\texttt{term II}}.\end{aligned} \tag{36}$$

where the equality follows from $\|a\|^2 = \|b\|^2 - \|a-b\|^2 - 2\langle a, b-a\rangle$.

Notice that the negative terms on the RHS of (36) will contribute to the decrease of V^{k+1}. We are thus left to deal with `term I` and `term II`. The idea is to bound them so that they can overall being controlled by the backtracking inequality.

Let us proceed bounding `term I` and `term II` using the algorithm dynamics. We have the following:

$$\begin{aligned}\texttt{term I} =& \left\langle (\mathbf{X}^k - \alpha^k \mathbf{D}^{k+1}) - \mathbf{X}^\star, \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle - \alpha^k \left\langle \nabla F(\mathbf{X}^{k+1/2}), \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle \\ =& \left\langle \mathbf{X}^k - \mathbf{X}^\star, \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle + \alpha^k \left\langle \mathbf{D}^{k+1} - \mathbf{D}^\star, \mathbf{X}^{k+1} - \mathbf{X}^k \right\rangle \\ & - \alpha^k \left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle ,\end{aligned} \tag{37}$$

where in the second equality we used $\nabla F(\mathbf{X}^\star) = -\mathbf{D}^\star$. Note that the last term in the expression above can be controlled through the backtracking procedure. Hence, we proceed bounding `term II` to "cancel" out the other terms on the RHS of (37). Specifically,

$$\begin{aligned}\texttt{term II} =& -\alpha^k \left\langle \mathbf{D}^{k+1} - \mathbf{D}^\star, \alpha^k c^{-1} \left(I - \widetilde{W} \right)^\dagger (\mathbf{D}^{k+1} - \mathbf{D}^k) + \alpha^k \mathbf{D}^k - \alpha^k \mathbf{D}^{k+1} \right\rangle \\ =& -\alpha^k \left\langle \mathbf{D}^{k+1} - \mathbf{D}^\star, \mathbf{X}^k - \alpha^k \nabla F(\mathbf{X}^{k+1/2}) - \alpha^k \mathbf{D}^{k+1} - \mathbf{X}^\star \right\rangle \\ =& -\alpha^k \left\langle \mathbf{D}^{k+1} - \mathbf{D}^\star, \mathbf{X}^{k+1} - \mathbf{X}^\star \right\rangle ,\end{aligned} \tag{38}$$

where we used the update of $\mathbf{X}^{k+1}$ and the facts $\mathbf{D}^{k+1} - \mathbf{D}^\star \in \texttt{span}(I - \widetilde{W})$ and $\mathbf{X}^\star \in \texttt{null}(I - \widetilde{W})$. Summing up (37) and (38), we obtain

$$\begin{aligned}\texttt{term I} + \texttt{term II} =& \left\langle \mathbf{X}^k - \mathbf{X}^\star, \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle + \alpha^k \left\langle \mathbf{D}^{k+1} - \mathbf{D}^\star, \mathbf{X}^{k+1} - \mathbf{X}^k \right\rangle \\ & - \alpha^k \left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle \\ & - \alpha^k \left\langle \mathbf{D}^{k+1} - \mathbf{D}^\star, \mathbf{X}^{k+1} - \mathbf{X}^\star \right\rangle \\ =& \left\langle \mathbf{X}^k - \mathbf{X}^\star, \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle - \alpha^k \left\langle \mathbf{D}^{k+1} - \mathbf{D}^\star, \mathbf{X}^k - \mathbf{X}^\star \right\rangle \\ & - \alpha^k \left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle \\ =& \alpha^k \left\langle \mathbf{X}^k - \mathbf{X}^\star, \frac{1}{\alpha^k} (\mathbf{X}^k - \mathbf{X}^{k+1}) - \mathbf{D}^{k+1} + \mathbf{D}^\star \right\rangle \\ & - \alpha^k \left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle \\ =& \alpha^k \left\langle \mathbf{X}^k - \mathbf{X}^\star, \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star) \right\rangle \\ & - \alpha^k \left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^k - \mathbf{X}^{k+1} \right\rangle \\ =& -\alpha^k \left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^\star - \mathbf{X}^{k+1} \right\rangle .\end{aligned} \tag{39}$$

The statement of the Lemma follows readily substituting (39) in (36). □

E.2 Proof of Lemma 7

We preliminary notice that, in view of Lemma 3, the backtracking inequality (6) on F^k holds with $\alpha^k = \min_{i\in[m]} \alpha_i^k$, where each α_i^k is the outcome of the backtracking procedure on the local f_i^k. Since F and F^k have the same curvature, it follows that (6) holds also on F, that is,

$$F(\mathbf{X}^{k+1}) \leq F(\mathbf{X}^{k+1/2}) + \left\langle \nabla F(\mathbf{X}^{k+1/2}), \mathbf{X}^{k+1} - \mathbf{X}^{k+1/2} \right\rangle + \frac{\delta}{2\alpha^k} \|\mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\|^2. \tag{40}$$

We proceed now bounding $\left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^\star - \mathbf{X}^{k+1} \right\rangle$ building on (40). To do so, we decompose the inner product as follows

$$\begin{aligned}& \left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^\star - \mathbf{X}^{k+1} \right\rangle \\ =& \underbrace{\left\langle \nabla F(\mathbf{X}^{k+1/2}), \mathbf{X}^\star - \mathbf{X}^{k+1/2} \right\rangle}_{\texttt{term I}} \underbrace{- \left\langle \nabla F(\mathbf{X}^{k+1/2}), \mathbf{X}^{k+1} - \mathbf{X}^{k+1/2} \right\rangle}_{\texttt{term II}} + \left\langle \nabla F(\mathbf{X}^\star), \mathbf{X}^{k+1} - \mathbf{X}^\star \right\rangle\end{aligned} \tag{41}$$

We bound `term I` invoking strong convexity and co-coercivity of F while we use (40) to bound `term II`. Specifically,

$$\texttt{term I} \leq F(\mathbf{X}^\star) - F(\mathbf{X}^{k+1/2}) + \begin{cases} -\dfrac{\mu^k}{2}\left\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\right\|^2, & \text{(by strong convexity)} \\ -\dfrac{1}{2L^k}\left\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^\star\right\|^2, & \text{(by co-coercivity)}, \end{cases}$$

where we also used $\nabla F(\mathbf{X}^\star) = -\mathbf{D}^\star$. Therefore

$$\texttt{term I} \leq F(\mathbf{X}^\star) - F(\mathbf{X}^{k+1/2}) - \max\left(\frac{\mu^k}{2}\left\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\right\|^2, \frac{1}{2L^k}\left\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^\star\right\|^2\right). \tag{42}$$

Using (40), `term II` can be bounded as

$$\texttt{term II} \leq F(\mathbf{X}^{k+1/2}) - F(\mathbf{X}^{k+1}) + \frac{\delta}{2\alpha^k}\left\|\mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\right\|^2. \tag{43}$$

Using (42) and (43) in (41), yields

$$\begin{aligned} &\left\langle \nabla F(\mathbf{X}^{k+1/2}) - \nabla F(\mathbf{X}^\star), \mathbf{X}^\star - \mathbf{X}^{k+1}\right\rangle \\ &\leq \frac{\delta}{2\alpha^k}\left\|\mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\right\|^2 - \max\left(\frac{\mu^k}{2}\left\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\right\|^2, \frac{1}{2L^k}\left\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^\star\right\|^2\right) \\ &+ \underbrace{F(\mathbf{X}^\star) + \langle \nabla F(\mathbf{X}^\star), \mathbf{X}^{k+1} - \mathbf{X}^\star\rangle - F(\mathbf{X}^{k+1})}_{\leq 0}. \end{aligned}$$

This completes the proof. □

E.3 Proof of Lemma 8

Combining Lemma 6 and Lemma 7, we can write

$$\begin{aligned} V^{k+1} \leq &\left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ &- \max\left(\frac{\mu^k}{2}\|\mathbf{X}_k^{k+1/2} - \mathbf{X}^\star\|^2, \frac{1}{2L^k}\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^*\|^2\right) \\ &- \underbrace{\|\mathbf{X}^k - \mathbf{X}^{k+1}\|^2}_{\texttt{term I}} - \underbrace{(\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^{k+1}\|_M^2}_{\texttt{term II}} + \delta\underbrace{\|\mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\|^2}_{\texttt{term III}}. \end{aligned}$$

Next, we demonstrate that the sum of the last three terms contributes to the decrease of V^{k+1}. Using the definition of $\mathbf{X}^{k+1}$ and $\mathbf{X}^{k+1/2}$, we can bound `term III` as follows:

$$\begin{aligned} \texttt{term III} &= \left\|\left(\mathbf{X}^{k+1} - \mathbf{X}^k\right) - \left(\mathbf{X}^{k+1/2} - \mathbf{X}^k\right)\right\|^2 \\ &= -\left\|\mathbf{X}^{k+1/2} - \mathbf{X}^k\right\|^2 - 2\left\langle \mathbf{X}^{k+1/2} - \mathbf{X}^k, \mathbf{X}^{k+1} - \mathbf{X}^{k+1/2}\right\rangle + \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|^2 \\ &= -\left\|c(I - \widetilde{W})\mathbf{X}^k\right\|^2 - 2\alpha^k\left\langle \mathbf{D}^{k+1/2}, c(I - \widetilde{W})\mathbf{X}^k\right\rangle + \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|^2 \\ &= -\left\|c(I - \widetilde{W})\mathbf{X}^k\right\|^2 - 2\alpha^k\left\langle \left(I - c(I - \widetilde{W})\right)\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right), c(I - \widetilde{W})\mathbf{X}^k\right\rangle \\ &\quad + \|\mathbf{X}^k - \mathbf{X}^{k+1}\|^2. \end{aligned}$$

Proceeding with $\|\mathbf{D}^k - \mathbf{D}^{k+1}\|_M^2$, we have

$$\begin{aligned} \texttt{term II} &= (\alpha^k)^2\left\|-c(I - \widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right) + \frac{c}{\alpha^k}(I - \widetilde{W})\mathbf{X}^k\right\|_M^2 \\ &= -\left\|c(I - \widetilde{W})\mathbf{X}^k\right\|^2 + \|\mathbf{X}^k\|_{c(I - \widetilde{W})}^2 \\ &\quad - 2\alpha^k\left\langle \left(I - c(I - \widetilde{W})\right)\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right), c(I - \widetilde{W})\mathbf{X}^k\right\rangle \\ &\quad + (\alpha^k)^2\left\|c(I - \widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\right\|_M^2, \end{aligned}$$

where the second equality follows from the definition of $M = c^{-1}(I - \widetilde{W})^{\dagger} - I$.

Combining the three terms yields

$$\begin{aligned} & -\texttt{term I} - \texttt{term II} + \delta\texttt{term III} \\ =&\delta(-\texttt{term I} - \texttt{term II} + \texttt{term III}) - (1-\delta)(\texttt{term I} + \texttt{term II}) \\ \leq&\delta(-\texttt{term I} - \texttt{term II} + \texttt{term III}) \\ =& -\delta\|\mathbf{X}^k\|^2_{c(I-\widetilde{W})} - \delta(\alpha^k)^2 \left\|c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\right\|_M^2, \end{aligned}$$

which proves the statement of the lemma. □

E.4 Proof of Lemma 9

The proof involves further bounding the RHS of (17) in Lemma 8, appropriately in terms of r^k and g^k. To achieve this, we construct two alternative bounds of the RHS of (17), as discussed below.

The first bound will reveal the dependence on r^k, and it is based on using in (17)

$$\max\left(\mu^k\alpha^k\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\|^2, \frac{\alpha^k}{L^k}\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^*\|^2\right) \geq \frac{\mu^k\alpha^k}{2}\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\|^2$$

along with

$$\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\|^2 \geq \left(1 - c(1 - \lambda_m(\widetilde{W}))\right)^2 \|\mathbf{X}^k - \mathbf{X}^\star\|^2. \tag{44}$$

We obtain

$$\begin{aligned} V^{k+1} \leq & \left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ & - \frac{\mu^k\alpha^k}{2}(1 - c(1 - \lambda_m(\widetilde{W})))^2\|\mathbf{X}^k - \mathbf{X}^\star\|^2 \\ & - \delta\left\|\mathbf{X}^k\right\|^2_{c(I-\widetilde{W})} - \delta(\alpha^k)^2\left\|c(I-\widetilde{W})\left(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k\right)\right\|_M^2 \\ \overset{(18)}{\leq} & \left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ & - \frac{\mu^k\alpha^k}{2}(1 - c(1 - \lambda_m(\widetilde{W})))^2\|\mathbf{X}^k - \mathbf{X}^\star\|^2 \\ & - \delta(r^k)^2(\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^*\|_M^2 \\ = & \left(1 - \frac{\mu^k\alpha^k}{2}(1 - c(1 - \lambda_m(\widetilde{W})))^2\right)\|\mathbf{X}^k - \mathbf{X}^\star\|^2 + \left(1 - \delta(r^k)^2\right)(\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2. \end{aligned}$$

The second bound of the RHS of (17) aims to obtain an explicit dependence on g^k. This is done by just neglecting the last two negative terms in the RHS of (17):

$$\begin{aligned} V^{k+1} \leq & \left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ & - \max\left(\mu^k\alpha^k\|\mathbf{X}^{k+1/2} - \mathbf{X}^\star\|^2, \frac{\alpha^k}{L^k}\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^*\|^2\right) \\ \overset{(44)}{\leq} & \left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ & - \max\left(\mu^k\alpha^k(1 - c(1 - \lambda_m(\widetilde{W})))^2\|\mathbf{X}^k - \mathbf{X}^\star\|^2, \frac{\alpha^k}{L^k}\|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^*\|^2\right) \\ \leq & \left\|\mathbf{X}^k - \mathbf{X}^\star\right\|^2 + (\alpha^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 - \frac{\mu^k\alpha^k}{2}(1 - c(1 - \lambda_m(\widetilde{W})))^2\|\mathbf{X}^k - \mathbf{X}^\star\|^2 \\ & - \min\left(\frac{\mu^k(\alpha^k)^3(1 - c(1 - \lambda_m(\widetilde{W})))^2}{2}, \frac{\alpha^k}{2L^k}\right)\max\left(\frac{1}{(\alpha^k)^2}\|\mathbf{X}^k - \mathbf{X}^\star\|^2, \|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^*\|^2\right) \end{aligned}$$

$$= \left(1 - \frac{\mu^k \alpha^k}{2}(1 - c(1 - \lambda_m(\widetilde{W})))^2\right) \|\mathbf{X}^k - \mathbf{X}^\star\|^2$$
$$+ \left(1 - (g^k)^2 \min\left(\frac{\mu^k \alpha^k (1 - c(1 - \lambda_m(\widetilde{W})))^2}{2}, \frac{1}{2L^k \alpha^k}\right)\right) (\alpha^k)^2 \|\mathbf{D}^k - \mathbf{D}^\star\|_M^2.$$

The final result follows combining the above two bounds while using the definition of ρ_1^k and ζ^k. □

E.5 Proof of Lemma 10

Invoking the update of the primal variable in the form

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \alpha^k(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^\star) - \alpha^k(\mathbf{D}^{k+1} - \mathbf{D}^\star),$$

where we used $\mathbf{D}^\star + \nabla F(\mathbf{X}^\star) = 0$, and the definition of g^k, we can write

$$\begin{aligned} \frac{1}{\alpha^k}\|\mathbf{X}^{k+1} - \mathbf{X}^*\| \geq & \|\mathbf{D}^{k+1} - \mathbf{D}^\star\| - \frac{1}{\alpha^k}\|\mathbf{X}^k - \mathbf{X}^*\| - \|\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^\star\| \\ \geq & \|\mathbf{D}^{k+1} - \mathbf{D}^\star\| - 2g^k\|\mathbf{D}^k - \mathbf{D}^\star\|_M \\ \geq & \frac{1}{\sqrt{\lambda_{\max}(M)}}\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M - 2(g^k)\|\mathbf{D}^k - \mathbf{D}^\star\|_M. \end{aligned} \tag{45}$$

Let us proceed lower bounding $\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M$. Using the update of the dual variable, in the form

$$\mathbf{D}^{k+1} = \mathbf{D}^k + \frac{1}{\alpha^k}c(I - \widetilde{W})\mathbf{X}^k - c(I - \widetilde{W})(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k),$$

we obtain

$$\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M \geq \|\mathbf{D}^k - \mathbf{D}^\star\|_M - \frac{1}{\alpha^k}\|c(I - \widetilde{W})\mathbf{X}^k\|_M - \|c(I - \widetilde{W})(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k)\|_M. \tag{46}$$

Using

$$\|c(I - \widetilde{W})\mathbf{X}^k\|_M \leq \|\mathbf{X}^k\|_{c(I - \widetilde{W})},$$

the following holds for the last two terms on the RHS of (46):

$$\begin{aligned} & \frac{1}{(\alpha^k)^2}\|\mathbf{X}^k\|^2_{c(I-\widetilde{W})} + \|c(I - \widetilde{W})(\nabla F(\mathbf{X}^{k+1/2}) + \mathbf{D}^k)\|_M^2 \\ & = (r^k)^2 \max\left(\frac{1}{(\alpha^k)^2}\|\mathbf{X}^k - \mathbf{X}^\star\|^2, \|\mathbf{D}^k - \mathbf{D}^\star\|_M^2\right) \\ & \leq (r^k)^2 \max((g^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2, \|\mathbf{D}^k - \mathbf{D}^\star\|_M^2) \\ & = (r^k)^2 \max((g^k)^2, 1)\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2 \\ & = (r^k)^2\|\mathbf{D}^k - \mathbf{D}^\star\|_M^2, \end{aligned}$$

where the last equality follows from $g^k \leq 1$ (as postulated in (21)).

Finally, using $\sqrt{2}\sqrt{a^2 + b^2} \geq a + b$, we deduce

$$\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M \geq (1 - \sqrt{2}r^k)\|\mathbf{D}^k - \mathbf{D}^\star\|_M.$$

Substituting the above inequality in (45) yields the desired result. □

E.6 Proof of Lemma 11

Using (22) and (44) in (17) (while neglecting therein the last two negative terms on the RHS), yields

$$
\begin{aligned}
V^{k+2} &\le \left\|\mathbf{X}^{k+1} - \mathbf{X}^\star\right\|^2 + (\alpha^{k+1})^2\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M^2 - \mu^{k+1}\alpha^{k+1}\|\mathbf{X}^{(k+1)+1/2} - \mathbf{X}^\star\|^2\\
&\overset{(44)}{\le} \left\|\mathbf{X}^{k+1} - \mathbf{X}^\star\right\|^2 + (\alpha^{k+1})^2\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M^2 - \mu^{k+1}\alpha^{k+1}(1 - c(1 - \lambda_m(\widetilde{W})))^2\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2\\
&\overset{(22)}{\le} (1 - \mu^{k+1}\alpha^{k+1}(1 - c(1 - \lambda_m(\widetilde{W}))^2)/2)\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2 + (\alpha^{k+1})^2\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M^2\\
&\quad - \frac{\mu^{k+1}\alpha^{k+1}}{2}(1 - c(1 - \lambda_m(\widetilde{W})))^2\frac{1}{\lambda_{\max}(M)}\left(1 - \frac{2g^k\sqrt{\lambda_{\max}(M)}}{1 - r^k\sqrt{2}}\right)^2(\alpha^k)^2\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M^2\\
&\le (1 - \mu^{k+1}\alpha^{k+1}(1 - c(1 - \lambda_m(\widetilde{W}))^2)/2)\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2\\
&\quad + \left(1 - \frac{\mu^{k+1}\alpha^{k+1}}{2(\gamma^k)^2}(1 - c(1 - \lambda_m(\widetilde{W})))^2\frac{1}{\lambda_{\max}(M)}\left(1 - \frac{2g^k\sqrt{\lambda_{\max}(M)}}{1 - r^k\sqrt{2}}\right)^2\right)\\
&\quad\times(\alpha^{k+1})^2\|\mathbf{D}^{k+1} - \mathbf{D}^\star\|_M^2,
\end{aligned}
$$

where the last inequality follows from $\alpha^k/\alpha^{k+1} \ge 1/\gamma^k$.

We deduce

$$
V^{k+2} \le \max\left(\frac{\alpha^{k+1}}{\alpha^k}, 1\right)^2\left(1 - \frac{\mu^{k+1}\alpha^{k+1}(1 - c(1 - \lambda_m(\widetilde{W})))^2}{2(\gamma^k)^2\max(\lambda_{\max}(M), 1)}\frac{1}{(\gamma^k)^2}\left(1 - \frac{2g^k\sqrt{\lambda_{\max}(M)}}{1 - r^k\sqrt{2}}\right)^2\right)V^{k+1}.
$$

The final statement of the lemma follows from the above inequality and

$$
V^{k+1} \le \max\left(\frac{\alpha^k}{\alpha^{k-1}}, 1\right)^2 V^k,
$$

due to (20) and $\rho_1^k < 1$. □

E.7 Proof of Lemma 12

Let us consider the case $N \ge \lceil\beta_1\rceil + 2$. Using $\gamma^k = \left(\frac{k+\beta_1}{k+1}\right)^{\beta_2}$, we can bound the product of γ^k's as

$$
\begin{aligned}
\ln\prod_{k=0}^{N-1}\gamma^k &= \beta_2\sum_{k=0}^{N-1}\ln\frac{k+\beta_1}{k+1}\\
&= \beta_2\sum_{k=0}^{\lceil\beta_1\rceil}\ln\frac{k+\beta_1}{k+1} + \beta_2\sum_{k=\lceil\beta_1\rceil+1}^{N-1}\ln\left(1 + \frac{\beta_1 - 1}{k+1}\right)\\
&\le \beta_2(\lceil\beta_1\rceil + 1)\ln\beta_1 + \beta_2(\beta_1 - 1)\sum_{k=\lceil\beta_1\rceil+1}^{N-1}\frac{1}{k+1}\\
&\le \beta_2(\lceil\beta_1\rceil + 1)\ln\beta_1 + \beta_2(\beta_1 - 1)\sum_{k=1}^{N-1}\frac{1}{k+1}\\
&\le \beta_2(\lceil\beta_1\rceil + 1)\ln\beta_1 + \beta_2(\beta_1 - 1)\ln N,
\end{aligned}
\tag{47}
$$

which proves (31). Notice that the above bound holds also if $N \le \lceil\beta_1\rceil + 2$.

Let us determine now N_0 such that (32) holds. Condition (32) is met if the following inequality holds

$$
\ln\left(\prod_{k=0}^{N-1}\gamma^k\right)^2 + (N-1)\ln(1-\rho/2) \le 0,
$$

where we used that $1-\rho \leq (1-\rho/2)^2$ for $\rho \in (0,1)$. Bounding the LHS yields

$$\begin{aligned}\ln\left(\prod_{k=0}^{N-1}\gamma^k\right)^2 + (N-1)\ln(1-\rho/2) &\overset{(47)}{\leq} 2\beta_2(\lceil\beta_1\rceil+1)\ln\beta_1 + 2\beta_2(\beta_1-1)\ln N + (N-1)\ln(1-\rho/2)\\ &\leq 2\beta_2(\lceil\beta_1\rceil+1)\ln\beta_1 - \ln(1-\rho/2) + 2\beta_2(\beta_1-1)\ln N - \frac{\rho}{2}N\\ &\leq 2\beta_2(\lceil\beta_1\rceil+1)\ln\beta_1 + \ln 2 + 2\beta_2(\beta_1-1)\ln N - \frac{\rho}{2}N.\end{aligned}$$

It follows that (32) holds if

$$N \geq \frac{4}{\rho}(2\beta_2(\lceil\beta_1\rceil+1)\ln\beta_1 + \ln 2)$$

and

$$N \geq \frac{8\beta_2(\beta_1-1)}{\rho}\ln N.$$

A sufficient condition for the last inequality to hold is

$$N \geq \frac{16\beta_2\beta_1}{\rho}\ln\frac{8\beta_2\beta_1}{\rho}.$$

This completes the proof. □

F Proof of the Intermediate Results in Appendix D

F.1 Proof of Lemma 14

1. This assertion comes readily from the definition of $\mathcal{I}_N$ and the backtracking procedure.

2. Let $\mathcal{N}_i(k)$ be the set of neighbors of agent i that are at most $k \geq 1$ hops away from agent i, including agent i itself. For notational consistency, $\mathcal{N}_i(1) = \mathcal{N}_i \cup \{i\}$.Using $\mathcal{N}_i(k)$, we can rewrite the local-min consensus step of each agent i at iteration k as

$$\alpha_i^k = \min_{j\in\mathcal{N}_i(1)} \overline{\alpha}_j^k,$$

where $\overline{\alpha}_j^k$ is the stepsizes produced by the line-search of agent j at iteration k.

Let $\overline{k} > k$ be an iteration such that $k+1,\ldots,\overline{k} \notin \mathcal{I}_N$. It must be

$$\alpha_{\text{min}}^t = \gamma^t\alpha_{\text{min}}^{t-1}, \quad \forall t = k+1,\ldots,\bar{k}.$$

In particular, this implies

$$\overline{\alpha}_i^t = \gamma^t \min_{j\in\mathcal{N}_i(1)} \overline{\alpha}_j^{t-1}, \quad \forall t = k+1,\ldots,\bar{k},$$

which leads to

$$\alpha_i^{\overline{k}} = \left(\prod_{l=k+1}^{\overline{k}}\gamma^l\right)\min_{j\in N_i(1+\overline{k}-k)} \overline{\alpha}_j^k.$$

Finally, taking $\overline{k} = k + d_{\mathcal{G}}$ and noting $N_i(d_{\mathcal{G}}) = [m]$, we proved the second statement of the lemma.

3. From the backtracking line-search and the definition of $\mathcal{I}_N$ it follows that

$$\alpha_{\text{min}}^k \begin{cases} = \gamma^k\,\alpha_{\text{min}}^{k-1}, & \text{if } k \notin \mathcal{I}_N;\\ \leq \dfrac{\gamma^k}{2}\,\alpha_{\text{min}}^{k-1}, & \text{if } k \in \mathcal{I}_N.\end{cases}$$

Applying the above relation iteratively, yields

$$\alpha_{\text{min}}^N \leq \alpha^0 2^{-|\mathcal{I}_N|}\prod_{k=0}^{N-1}\gamma^k.$$

https://doi.org/10.52202/079017-3042

At the same time, it follows from Lemma 3 that

$$\alpha_{\min}^{N} \geq \min\left(\frac{\delta}{2L}, \gamma^{k}\alpha^{0}\right).$$

Combining the lower and upper bounds above, yields the desired result

$$|\mathcal{I}_N| \leq \max\left(\ln \alpha_0 L + \ln \prod_{k=0}^{N-1} \gamma^k + \ln \frac{2}{\delta}, 0\right).$$

□

F.2 Proof of Lemma 15

From (34), it follows

$$\left\lfloor \frac{N}{N_\varepsilon + d_{\mathcal{G}}} \right\rfloor > |\mathcal{I}_N|.$$

Then, according to the Dirichlet's principle there exists two iteration indices k_1 and k_2 such that

1. $\forall k \in [k_1, k_2] \Rightarrow k \notin \mathcal{I}_N$; and
2. $k_2 - k_1 \geq N_\varepsilon + d_{\mathcal{G}}$.

Invoking Lemma 14.(1) and Lemma 14.(2), it follows that all agents' stepsizes reach consensus after $k_1 + d_{\mathcal{G}}$ iterations and remain consensual for the subsequent N_ε iterations. One can then invoke Lemma 13, and conclude

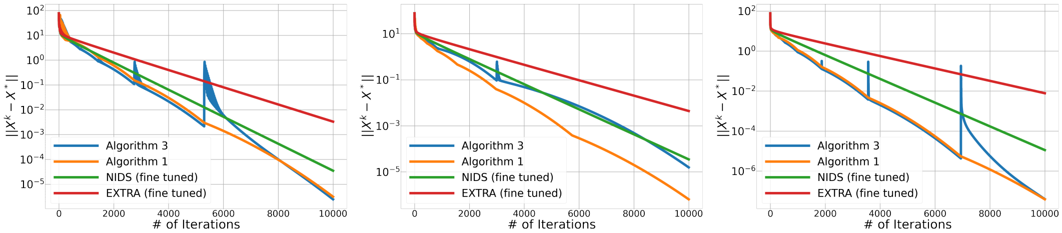
$$\left\|\mathbf{X}^{k_2} - \mathbf{X}^{\star}\right\|^2 + \frac{1}{4L^2}\|\mathbf{D}^{k_2} - \mathbf{D}^{\star}\|_M^2 \leq \varepsilon.$$

This concludes the proof. □

G Additional Numerical Results

This section presents additional experiments for the Ridge Regression problem, introduced in Section 6. Here, we consider additional graph topologies, namely: 1) Ring Graphs; 2) Random Regular Graphs with degree 3; and 3) Random Regular Graphs with degree 10. The rest of the setup (including algorithms' tuning) is the same of that described in Section 6.

The experiments are summarized in Fig. 3. The findings corroborate the conclusions presented in Sec. 6: both Algorithm 1 and Algorithm 3 outperform EXTRA and NIDS, which were finely tuned for rapid practical convergence. Quite interestingly, the performance of the proposed methods appears to be less affected by network topology and depends primarily on network connectivity.



(a) Ring Graph

(b) Regular graph, degree 3

(c) Regular graph, degree 10

Figure 3: **Ridge regression** on different **regular graphs**: (3a) Ring graph; (3b) Random Regular Graph, with degree 3; (3c) Random Regular Graph with degree 10.

https://doi.org/10.52202/079017-3042

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract gives accurate presentation of our result. Part Major contributions of Introduction contains full description of our work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: The main limitaion of proposed procedure is min-consensus. The technology for its implementation is carefully discribed in part 3.1.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-3042

Answer: [Yes]

Justification: Main assumptions and definitions are presented in Section 2. All main theoretical results presented in Section 4 with all required assumptions. Proofs are placed in Appendix A-F because of their large size.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All setup for numerical experiments are described in Section 5. It is enough to reproduce all experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/079017-3042

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: code in the form of an attached archive.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our paper demonstrates performance of optimization algorithm. Because of that, we do not need test some models. But Section 5 contains full information about our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Numerical experiments demonstrate performance of optimization algorithm on a given problems. Besides, our algorithm is deterministic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-3042

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Information is given at the end of Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Authors are familiar with NeurIPS Code of Ethics and paper conform it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There are different methods of distributed optimization. The paper propose new method of distributed optimization that has no additional societal impact as the authors think.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/079017-3042

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The proposed method does not require safeguard.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Numerical experiments use one of datasets from LIBSVM. Authors cite corresponding work of owners (see reference [6] in Section 5 and References)

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification:contains contains README file with sufficient description.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

https://doi.org/10.52202/079017-3042

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3042