A Line-search-free Method for Adaptive Decentralized Optimization

Xiaokai Chen, Ilya Kuruzov, and Gesualdo Scutari

***Abstract*— We study decentralized optimization over networks where agents cooperatively minimize a smooth (strongly) convex sum of local losses while communicating only with immediate neighbors. Prevailing decentralized methods require either centralized knowledge of global problem and network parameters for stepsize tuning–hence impractical, or costly per-iteration line-searches that demand access to local function values. We propose line-search-free, fully decentralized algorithms in which each agent adapts its stepsize using only past local iterates and gradients–with no extra function evaluations and no global tuning. The key technical ingredient is a new Lyapunov function, from which a natural adaptive stepsize rule emerges: at each iteration, each agent selects the largest stepsize that guarantees descent, based solely on a local curvature estimate built from successive gradients. The proposed algorithms enjoy strong theoretical guarantees: sublinear convergence rates for merely convex objectives and linear rates under strong convexity. Numerical experiments on standard benchmarks show consistent improvements over the state of the art, both adaptive and non-adaptive.**

***Index Terms*—Adaptive stepsize, decentralized optimization, line-search-free, networks.**

I. Introduction

We consider the following convex optimization problem

$$\min_{x\in\mathbb{R}^d} f(x), \quad f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x), \tag{P}$$

over a network of m agents, where f_i is the cost function accessible only to agent $i \in [m] := \{1, \cdots, m\}$, assumed to be (strongly) convex and *locally* smooth. All agents are embedded in a communication network modeled as a fixed, undirected and connected graph.

Decentralized optimization in the form (P) arises in several scientific and engineering areas, including distributed machine learning, multiple-agent control and coordination, and sensor network information processing. A large body of methods exists for (P); representative methods include gradient tracking [1]–[3] and primal–dual schemes [4], [5]; see [6], [7] for some overviews. As in centralized optimization, convergence of decentralized methods hinges on an appropriate stepsize policy–typically a *single* stepsize shared across agents–with either **constant** or **diminishing** rules.

Under **constant** stepsize rules, existing theory imposes upper bounds on the stepsize values that depend on global problem/network constants–e.g., Lipschitz constants of ∇f_i, the spectral gap of the gossip matrix, and related topological descriptors. These quantities are rarely known a priori in practice. Consequently, despite theoretical guidance, practitioners often fall back on manual tuning of a single, common stepsize for each problem/network instance, which can be slow and brittle because it ignores the heterogeneous geometry of the loss landscape along the trajectory. **Diminishing stepsizes** have been adopted to partially sidestep unavailable global problem/network-dependent constants, e.g., [8]–[10]. This, however, comes at the cost of significantly degraded rates (e.g., sublinear iteration complexity even under strong convexity) and sensitivity to schedule design.

This has recently motivated growing interest in *adaptive* methods that adjust stepsizes based on loss curvature information while still guaranteeing global convergence. Below we summarize the most relevant literature on the subject.

A. Literature review

The literature for adaptive **centralized** optimization has grown rapidly in the past few years. Several stepsize selection rules based on local estimates of the loss curvature have been proposed, ranging from the classical line-search [11], Barzilai-Borwein [12], [13], and Polyak [14] rules, to more recent schemes leveraging past gradient and iterate information [15]–[17], as well as methods designed for stochastic and large-scale settings such as AdaGrad [18], Adam [19], AMSGrad [20], and their variants [21], [22]. However, all these methods cannot be implemented over mesh networks without a centralized node.

In contrast, the literature for adaptive **decentralized** algorithms is fairly scarce. Direct adaptations of centralized methods [17], [23] to decentralized settings exhibit significant practical limitations: the stepsize selection either leads to extensive communications (to verify line-search based terminations) per iteration or requires global scalar information not available at the agents' side (see [24] for details). Other approaches [25], [26], despite being adaptive, still require *problem/network-dependent* parameter knowledge for stepsize selection. In [27], a decentralized Polyak stepsize rule is proposed; however, its $\mathcal{O}(1/\sqrt{mk})$ (k is the iteration index) rate matches that of non-adaptive methods, so the adaptivity yields no rate improvement, and slow convergence in the strongly convex case. Furthermore, it requires agents' losses to be *globally Lipschitz* continuous (i.e., ∇f_i uniformly bounded), which excludes a vast class of problems. The proposals [28]–[30] focus on the stochastic instance of problem (P); they explore adaptivity through stochastic gradient normalization, achieving convergence rate of $\mathcal{O}(1/\sqrt{k})$, which are quite unfavorable in determinist setting considered in this paper.

Advancements have been made in [24], [31]–[34]. They proposed decentralized algorithms that adaptively select stepsizes via local backtracking line-searches combined with global/local min-consensus protocols, achieving robust and

Chen and Scutari are with the School of Industrial Engineering, Purdue University; emails:<chen4373,gscutari>@purdue.edu.

fast convergence guarantees. These represent the current state of the art; however, the line-search procedures introduce extra per-iteration function evaluations–hence computational complexity–and the algorithms are not implementable when objective function values are unavailable.

To summarize, no existing adaptive decentralized method provides robust convergence guarantees while avoiding the computational overhead of line-searches. This paper addresses this gap; our main contributions are as follows.

B. Main contributions

1. Algorithm design: Line-search-free adaptive algorithms. We propose two decentralized algorithms (Algorithm 1 and 2) for (P) that adapt each agent's stepsize without line-searches and without knowledge of any global problem or network constant. Algorithm 1 relies on a lightweight global scalar averaging protocol, while Algorithm 2 requires only local (neighbor-to-neighbor) scalar exchanges, making it implementable on any connected network.

2. Key design principle. The algorithms are derived from a strengthened convergence analysis of the Condat-Vũ splitting [35], [36], which is of independent interest. We propose a novel augmented Lyapunov function that *incorporates trajectory information*–consecutive iterate differences and primal Lagrangian gaps. Unlike standard Lyapunov functions used in the literature, whose descent conditions couple the stepsize to both smoothness and network spectral quantities–not computable by the agents, ours yields a descent condition that depends only on local curvature estimates built from past gradients and iterates. Each agent can then maximize its local stepsize subject to this computable descent condition, obtaining adaptivity as a byproduct of the convergence mechanism itself rather than as a separate heuristic layer.

3. Strong convergence guarantees: For merely convex losses f_i, the algorithms are proved to reach an ε-solution (measured by a suitable metric) of Problem (P) in $\mathcal{O}(\widetilde{L}/\varepsilon)$ iterations (communications), where $\widetilde{L}$ is the Lipschitz constant of ∇f_i's *restricted to the convex hull* of the trajectory generated by the algorithm. For strongly convex f_i's, linear convergence is certified with communication complexity in the order of $\mathcal{O}\left(\frac{\widetilde{\kappa}}{1-\lambda_2(W)}\log(1/\varepsilon)\right)$, where $\lambda_2(W)$ is the second largest eigenvalue of the gossip matrix W and $\widetilde{\kappa}$ is the condition number of f_i's *restricted to the convex hull* of the iterates. Quite remarkably, both rates outperform those of existing nonadaptive algorithms, since $\widetilde{L}$ and $\widetilde{\kappa}$ can be significantly smaller than their global counterparts. Our numerical results validate these theoretical findings.

II. Algorithm Design and Convergence Analysis

We study (P) under the following standard assumptions.

Assumption 1 (Objective function)**.** *Each function $f_i : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, locally smooth, and μ-strongly convex, with $\mu \in [0,\infty)$. If $\mu = 0$, f admits a minimizer over $\mathbb{R}^d$.*

Unlike most of decentralized optimization literature (see, e.g., [6], [7] and references therein), this work does *not*require *global* smoothness of the agents' losses. This permits covering a broader class of practical problems, including Poisson linear inverse problem [37] and Poisson Regression [38], to name a few.

Assumption 2 (Communication graph)**.** *The graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$ is undirected and connected, where $\mathcal{V} := [m] := \{1,\ldots,m\}$ is the set of agents and $(i,j) \in \mathcal{E}$ if and only if there is a communication link between agents i and j.*

We denote by $\mathbb{S}_+^m$ (resp. $\mathbb{S}_{++}^m$) the set of $m \times m$ real symmetric positive semidefinite (resp. definite) matrices.

We begin rewriting (P) in an equivalent lifted form. Introducing local copies $x_i \in \mathbb{R}^d$ of the variable $x \in \mathbb{R}^d$—the i-th controlled by agent i—the row-wise stack matrix $\mathbf{X} := [x_1, x_2, \ldots, x_m]^\top \in \mathbb{R}^{m\times d}$, and the lifted functions

$$F(\mathbf{X}) := \sum_{i=1}^m f_i(x_i), \quad G(\mathbf{X}) := \delta_{\{0\}}(\mathbf{X}),$$

where $\delta_{\{0\}} : \mathbb{R}^{m\times d} \to \mathbb{R}\cup\{\infty\}$ is the indicator function of $\{0\}$, Problem (P) can be equivalently written as:

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times d}} F(\mathbf{X}) + G(\text{Ł}\mathbf{X}), \tag{P$'$}$$

where $\text{Ł} \in \mathbb{S}_+^m$ is an $m\times m$ matrix satisfying $\texttt{null}(\text{Ł}) = \texttt{span}(\mathbf{1}_m)$. This condition ensures that any $\mathbf{X}$ satisfying $\text{Ł}\mathbf{X} = \mathbf{0}$ is consensual, i.e., $\mathbf{X} = \mathbf{1}_m x^\top$ for some $x \in \mathbb{R}^d$. A standard choice for Ł in decentralized algorithms is $\text{Ł} = (I - \widetilde{W})^{1/2}$ for some $\widetilde{W} \in \mathcal{W}_\mathcal{G}$, where $\mathcal{W}_\mathcal{G}$ is the class of gossip matrices defined next.

Assumption 3 (Gossip matrices)**.** *$\mathcal{W}_\mathcal{G}$ denotes the set of symmetric, doubly stochastic matrices $\widetilde{W} := (\widetilde{w}_{i,j})_{i,j=1}^m$ that are compliant with $\mathcal{G}$, i.e., $\widetilde{w}_{ii} > 0$ for all $i \in [m]$; $\widetilde{w}_{ij} > 0$ for all $(i,j)\in\mathcal{E}$; and $\widetilde{w}_{ij} = 0$ otherwise.*

Several instances of such matrices have been employed in practice; see, e.g., [6], [7], [39]. Throughout the paper we number the eigenvalues of matrix $W \in \mathbb{S}^m$ in nonincreasing order, i.e., $\lambda_1(W) \geq \cdots \geq \lambda_m(W)$.

A. Revisiting convergence of the Condat-Vũ splitting method

We start from the Condat-Vũ splitting [35], [36]—a classical primal-dual method for problems in the form of (P$'$). Given parameters $\alpha,\gamma,\sigma > 0$ (to be properly determined) and an initialization $\mathbf{X}^0, \mathbf{X}^{-1}, \mathbf{Y}^0 \in \mathbb{R}^{m\times d}$, the algorithm reads: for $k = 0,1,2,\ldots$,

$$\begin{aligned}\mathbf{Y}^{k+1} &= \texttt{prox}_{\sigma\alpha G^*}(\mathbf{Y}^k + \sigma\alpha\text{Ł}((1+\gamma)\mathbf{X}^k - \gamma\mathbf{X}^{k-1}));\\ \mathbf{X}^{k+1} &= \mathbf{X}^k - \alpha(\nabla F(\mathbf{X}^k) + \text{Ł}\mathbf{Y}^{k+1}),\end{aligned} \tag{1}$$

where $\texttt{prox}_{\sigma\alpha G^*}(\mathbf{X}) := \arg\min_\mathbf{Y} G^*(\mathbf{Y}) + \frac{1}{2\sigma\alpha}\|\mathbf{X}-\mathbf{Y}\|^2$ denotes the proximal operator of the conjugate function G^* of G (applied row-wise), scaled by $\sigma\alpha$.

Using algorithm (1) with the existing convergence machinery [40] presents two fundamental obstacles.

(i) *The classical stepsize condition is impractical for adaptive methods.* Setting $\text{Ł} = (I - \widetilde{W})^{1/2}$ for some $\widetilde{W} \in \mathcal{W}_\mathcal{G}$, the classical convergence analysis [40]—which requires the stronger assumption that F is globally L-smooth (hence violating Assumption 1)—guarantees that $\{\mathbf{X}^k\}_{k\geq 0}$ generated by (1) converges to a solution $\mathbf{X}^*$ of (P′), provided

$$\alpha L/2 + \sigma\alpha^2\|I - \widetilde{W}\| < 1. \tag{2}$$

This condition suffers from several limitations. First, it couples the stepsize α and the tuning parameter $\sigma > 0$ with the network-dependent quantity $\|I - \widetilde{W}\|$, which is typically unavailable to individual agents. One could upper bound $\|I - \widetilde{W}\| \leq 2$ and substitute into (2); however, this yields a highly conservative stepsize that leads to slow convergence in practice [24]. Second, the dependence on the *global* smoothness constant L (unknown to the agents) is restrictive, as L can be significantly larger than the *local* smoothness along the trajectory of the algorithm. Third, (2) involves the auxiliary parameter σ, whose tuning interacts nontrivially with α: making α adaptive while σ is fixed (or vice versa) proves problematic, and existing attempts along this route exhibit high sensitivity of the algorithm performance to the choice of σ [24].

(ii) *Algorithm* (1) *is not directly implementable in a decentralized setting.* The updates in (1) involve the linear operator $\text{Ł} = (I-\widetilde{W})^{1/2}$, whose application to the X-and Y-variables requires global coordination and cannot be carried out via local communication among neighbors.

Next, we address both limitations above: **(i)** we develop a new convergence analysis for (1) based on a novel Lyapunov function, yielding a stepsize condition that is *independent* of the network parameters (Sec. II-B); **(ii)** we make (1) implementable in a decentralized fashion and use the network-independent analysis to guide the adaptive selection of the stepsize, tailored to the local curvature of the agents' losses, thereby getting rid of the dependence on the much more conservative global smoothness constant L (Sec. II-C).

B. Network-independent convergence analysis

Let $\mathcal{L}(\mathbf{X}, \mathbf{Y})$ be the Lagrangian function of Problem (P′):

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}) = F(\mathbf{X}) + \langle \text{Ł}\mathbf{X}, \mathbf{Y}\rangle - G^*(\mathbf{Y}); \tag{3}$$

and, for any given saddle point $(\mathbf{X}^*, \mathbf{Y}^*)$ of (3), define the primal gap at $\mathbf{X}$ as:

$$\begin{aligned}\Delta\mathcal{L}_*^k &:= \mathcal{L}(\mathbf{X}^k, \mathbf{Y}^*) - \mathcal{L}(\mathbf{X}^*, \mathbf{Y}^*)\\ &= F(\mathbf{X}^k) - F(\mathbf{X}^*) + \langle \text{Ł}\mathbf{Y}^*, \mathbf{X}^k - \mathbf{X}^*\rangle \geq 0,\end{aligned} \tag{4}$$

where the inequality follows from the convexity of F.

Our enhanced convergence analysis of algorithm (1) hinges on the following novel Lyapunov function evaluated along the iterates $\{(\mathbf{X}^k, \mathbf{Y}^k)\}$ of (1): for any given saddle point $(\mathbf{X}^*, \mathbf{Y}^*)$ of (3), let

$$\begin{aligned}V_*^k &:= \|\mathbf{X}^k - \mathbf{X}^*\|^2 + \frac{1}{\sigma}\|\mathbf{Y}^k - \mathbf{Y}^*\|^2\\ &+ \frac{1}{2}\|\mathbf{X}^k - \mathbf{X}^{k-1}\|^2 + 2\gamma\alpha\Delta\mathcal{L}_\star(\mathbf{X}^{k-1}),\end{aligned} \tag{5}$$

where $\alpha, \gamma, \sigma > 0$ are parameters to be properly chosen. Note that $V_\star^k$ is continuous in its arguments and $V_\star^k \geq 0$, with equality if and only if $(\mathbf{X}^k, \mathbf{Y}^k) = (\mathbf{X}^{k-1}, \mathbf{Y}^{k-1}) = (\mathbf{X}^*, \mathbf{Y}^*)$; hence, it is a valid measure of optimality.

The following result provides conditions for the monotone decrease of $V_\star^k$ along $\{(\mathbf{X}^k, \mathbf{Y}^k)\}$.

Lemma 1. *Let* $\{(\mathbf{X}^k, \mathbf{Y}^k)\}$ *be the iterates generated by* (1) *under Assumptions 1-3, with the additional requirement that* F *is* L*-smooth globally, and the following tuning:* $\widetilde{W} \in \mathcal{W}_\mathcal{G} \cap \mathbb{S}^m_{++}$, $\sigma > 0$, $\gamma = 1$, *and*

$$\alpha \leq \frac{1}{\sqrt{L^2 + 2\sigma} + L}. \tag{6}$$

Then, for any given saddle point $(\mathbf{X}^*, \mathbf{Y}^*)$ *of* (3), *it holds*

$$V_\star^{k+1} \leq V_\star^k, \quad \forall k = 0, 1, \ldots. \tag{7}$$

The requirement that $\widetilde{W} \succ 0$ can be readily met by using the shifted matrix $W = (1-c)I + c\widetilde{W} \in \mathcal{W}_\mathcal{G}$, for some $c \in (0, 1/2)$ and $\widetilde{W} \in \mathcal{W}_\mathcal{G}$ [6], [7], [39].

Crucially, the stepsize condition (6) depends only on the smoothness constant L and the free parameter σ, with no appearance of the network quantity $\|I - \widetilde{W}\|$. This decoupling is the key insight that enables the design of adaptive stepsizes based on local curvature alone.

C. Decentralization of (1) and adaptive stepsize selection

We now address limitation (ii) of Section II-A. To make the algorithm (1) implementable in a decentralized setting, using $\text{Ł} = (I - \widetilde{W})^{1/2}$ with $\widetilde{W} \in \mathcal{W}_\mathcal{G}$, we introduce the change of variable $\mathbf{D}^k = \text{Ł}\mathbf{Y}^k$, with $\mathbf{D}^0 = \text{Ł}\mathbf{Y}^0 = \mathbf{0}$. Since $G = \delta_{\{0\}}$, we have $G^*(\mathbf{Y}) = 0$ for all $\mathbf{Y}$, and the proximal step on $\mathbf{Y}$ in (1) reduces to a linear update. Using these facts, one can verify that (1) is equivalent to

$$\begin{aligned}\mathbf{D}^{k+1} &= \mathbf{D}^k + \sigma\alpha(I - \widetilde{W})((1+\gamma)\mathbf{X}^k - \gamma\mathbf{X}^{k-1});\\ \mathbf{X}^{k+1} &= \mathbf{X}^k - \alpha(\nabla F(\mathbf{X}^k) + \mathbf{D}^{k+1}).\end{aligned} \tag{8}$$

Algorithm (8) is now implementable in a decentralized setting; furthermore, unlike existing adaptive decentralized methods [24], [31]–[34], it requires only one round of communication per iteration.

Building on the network-independent analysis of Lemma 1, we now drop the global L-smoothness assumption and select the stepsize adaptively. We apply a Lyapunov function analogous to $V_\star^k$ to the decentralized update (8), replacing the fixed parameters α, σ, γ with non-stationary sequences $\{\alpha^k\}_{k\geq 0}$, $\{\sigma^k\}_{k\geq 0}$, and $\{\gamma^k\}_{k\geq 0}$ that are adapted along the iterates. Specifically, given $\mathbf{X}^{-1}, \mathbf{X}^0 \in \mathbb{R}^{m\times d}$, $\mathbf{D}^0 = \mathbf{0}$, and $\widetilde{W} \in \mathcal{W}_\mathcal{G} \cap \mathbb{S}^m_{++}$, we have:

$$\mathbf{D}^{k+1} = \mathbf{D}^k + \sigma^k\alpha^k(I - W)((1+\gamma^k)\mathbf{X}^k - \gamma^k\mathbf{X}^{k-1}); \tag{9a}$$

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \alpha^k(\nabla F(\mathbf{X}^k) + \mathbf{D}^{k+1}), \tag{9b}$$

for $k = 0, 1, 2, \ldots,$. Recalling that $\mathbf{D}^k = \text{Ł}\mathbf{Y}^k$, the update of the auxiliary variable $\{\mathbf{Y}^k\}_{k\geq 0}$ follows from (9a):

$$\mathbf{Y}^{k+1} = \mathbf{Y}^k + \sigma^k\alpha^k((1+\gamma^k)\text{Ł}\mathbf{X}^k - \gamma^k\text{Ł}\mathbf{X}^{k-1}), \tag{10}$$

for $k = 0, 1, 2, \cdots$, and $\mathbf{Y}^0 = \mathbf{0}$.

The Lyapunov function $V_\star^k$ in (5), adjusted to the non-stationary sequences $\{\alpha^k\}_{k\geq 0}$, $\{\sigma^k\}_{k\geq 0}$, and $\{\gamma^k\}_{k\geq 0}$, becomes (with a slight abuse of notation):

$$\begin{aligned} V_\star^k &:= \|\mathbf{X}^k - \mathbf{X}^*\|^2 + \frac{1}{\sigma^k}\|\mathbf{Y}^k - \mathbf{Y}^*\|^2 \\ &+ \frac{1}{2}\|\mathbf{X}^k - \mathbf{X}^{k-1}\|^2 + 2\gamma^k\alpha^k\Delta\mathcal{L}_\star^{k-1} \geq 0, \end{aligned} \tag{11}$$

where $\mathbf{X}^k, \mathbf{X}^{k-1}$ and $\mathbf{Y}^k$ are given by (9b) and (10), respectively. The counterpart of Lemma 1 reads as follows.

Lemma 2. *Let $\{(\mathbf{X}^k, \mathbf{D}^k)\}$ be the iterates generated by* (9) *and $\{\mathbf{Y}^k\}$ the corresponding auxiliary variables defined in* (10)*, under Assumptions 1-3 and the following tuning: $W \in \mathcal{W}_\mathcal{G} \cap \mathbb{S}_{++}^m$, $\gamma^0 = 1$, and $\gamma^k = \alpha^k/\alpha^{k-1}$ for $k \geq 1$. Then, for any given saddle point $(\mathbf{X}^*, \mathbf{Y}^*)$ of* (3) *and any $\zeta^k > 0$, it holds*

$$\begin{aligned} &V_\star^{k+1} \\ \leq& V_\star^k - \left(\frac{1}{2} - \alpha^k L^k - \alpha^k\zeta^k\right)\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2 \\ &- \left(\frac{1}{2} - \alpha^k L^k\right)\|\mathbf{X}^k - \mathbf{X}^{k-1}\|^2 \\ &- \left(\frac{1}{\sigma^k} - \frac{\alpha^k}{\zeta^k}\right)\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|^2 - \frac{\sigma^{k+1} - \sigma^k}{\sigma^k\sigma^{k+1}}\|\mathbf{Y}^{k+1} - \mathbf{Y}^*\|^2 \\ &- [(2 + 2\gamma^k)\alpha^k - 2\gamma^{k+1}\alpha^{k+1}]\Delta\mathcal{L}_\star^k, \quad \forall k \geq 0, \end{aligned} \tag{12}$$

where L^k is a proxy for the local Lipschitz constant of ∇f_i's, defined as

$$L^k := \frac{\|\nabla F(\mathbf{X}) - \nabla F(\mathbf{X}^{k-1})\|}{\|\mathbf{X}^k - \mathbf{X}^{k-1}\|}. \tag{13}$$

To guarantee descent of the Lyapunov function, i.e., $V_\star^{k+1} \leq V_\star^k$, it suffices to require that all the coefficients multiplying $\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2$, $\|\mathbf{X}^k - \mathbf{X}^{k-1}\|^2$, $\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|^2$, $\|\mathbf{Y}^{k+1} - \mathbf{Y}^*\|^2$, and $\Delta\mathcal{L}_\star^k$ in (12) are non-negative. This is ensured by selecting, for a proper $\zeta^k > 0$, α^k and $\{\sigma^k\}_{k\geq 0}$ such that the following hold for all $k \geq 0$ (note that (14a) simultaneously enforces non-negativity of the coefficients of $\|\mathbf{X}^{k+1} - \mathbf{X}^k\|^2$ and $\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\|^2$; furthermore, since $\zeta^k > 0$, the coefficient of $\|\mathbf{X}^k - \mathbf{X}^{k-1}\|^2$, namely $1/2 - \alpha^k L^k$, is also automatically non-negative):

$$\alpha^k \leq \min\left\{\frac{1}{2(L^k + \zeta^k)}, \frac{\zeta^k}{\sigma^k}\right\}; \tag{14a}$$

$$(2 + 2\gamma^k)\alpha^k - 2\gamma^{k+1}\alpha^{k+1} \geq 0; \tag{14b}$$

$$1/\sigma^k - 1/\sigma^{k+1} \geq 0. \tag{14c}$$

Optimizing (14a) over $\zeta^k > 0$ (i.e., equating the two terms in the minimum), the largest admissible stepsize satisfying (14a) is $\alpha^k = 1/(\sqrt{(L^k)^2 + 2\sigma^k} + L^k)$. Introducing slack parameters $c_1, c_2 \in (0, 1]$ in (14a) and (14b), respectively, and an additional stepsize-control policy $\pi^k : \mathbb{R} \to \mathbb{R}$ to add extra control on the stepsize selection policy, conditions (14) are satisfied by the following selection rule:

$$\begin{aligned} \alpha^k = \min\Bigg\{&\frac{1}{\sqrt{(L^k)^2 + 2\sigma^k/c_1} + L^k}, \\ &\sqrt{1 + c_2\gamma^{k-1}}\alpha^{k-1}, \pi^k(\alpha^{k-1})\Bigg\}. \end{aligned} \tag{15}$$

The stepsize-growth policy functions $\{\pi^k\}_{k\geq 0}$ impose additional restrictions on possible increases of α^k when such control is needed for convergence. When agents' losses are merely convex, this extra restriction is unnecessary and the term can be omitted, or one may simply take any admissible policy with $\pi^k(x) \geq x$ (see Assumption 6). Our analysis shows that specific choices of π^k (as in Assumption 5) are needed in the strongly convex setting to ensue *linear* convergence with desirable rate scaling.

The resulting line-search-free adaptive decentralized algorithm is summarized in Algorithm 1. Notice that the updates of the primal and dual variables (Step 3) require only one communication round/iteration within neighboring nodes.

Algorithm 1 **A**daptive **D**ecentralized **O**ptimization Algorithm with **L**ine-search-**F**ree Stepsize Selection (ADOLF)

Data: $\mathbf{X}^{-1}, \mathbf{X}^0 \in \mathbb{R}^{m\times d}$, $\mathbf{D}^0 = 0$; $\alpha^0, \sigma^k > 0$, $\forall k \geq 0$, $c_1, c_2 \in (0, 1]$, $\gamma^0 = 1$; stepsize control policy $\{\pi^k\}$; $W = (1 - c)I + c\widetilde{W}$, $c \in (0, 1/2)$ and $\widetilde{W} \in \mathcal{W}_\mathcal{G}$.

1: `(S.0)` Initialize:

$$\mathbf{D}^1 = \mathbf{D}^0 + \sigma^0\alpha^0(I - W)((1 + \gamma^0)\mathbf{X}^0 - \gamma^0\mathbf{X}^{-1});$$
$$\mathbf{X}^1 = \mathbf{X}^0 - \alpha^0(\nabla F(\mathbf{X}^0) + \mathbf{D}^1);$$

2: **for** $k = 1, 2, \cdots$ **do**

3: `(S.1) Global scalar average:` Calculate

$$L^k = \sqrt{\frac{\sum_{i\in[m]}\|\nabla f_i(x_i^k) - \nabla f_i(x_i^{k-1})\|^2}{\sum_{i\in[m]}\|x_i^k - x_i^{k-1}\|^2}};$$

4: `(S.2) Stepsize selection:` Choose α^k according to (15) and take $\gamma^k = \alpha^k/\alpha^{k-1}$;

5: `(S.3) Dual and primal update:`

$$\mathbf{D}^{k+1} = \mathbf{D}^k + \sigma^k\alpha^k(I - W)((1 + \gamma^k)\mathbf{X}^k - \gamma^k\mathbf{X}^{k-1});$$
$$\mathbf{X}^{k+1} = \mathbf{X}^k - \alpha^k(\nabla F(\mathbf{X}^k) + \mathbf{D}^{k+1}).$$

On the global scalar average. Computing L^k in `(S.1)` requires a global scalar average across the network. In modern wireless mesh networks this can be realized over a low-rate, long-range interface such as LoRa [41] (broadcasting a single scalar per node in one hop), while vector transmissions use a high-rate interface such as WiFi. Alternatively, the global scalar average can be replaced by a global-max consensus step at the cost of a slightly smaller stepsize. A fully local variant that removes any global communication requirement is presented in Section III.

D. Convergence guarantees

This subsection provides the convergence results for Algorithm 1. We postulate the following assumption.

Assumption 4. *The sequence $\{\sigma^k\}_{k\geq 0}$ is monotonically nondecreasing and $0 < \sigma^k \leq \overline{\sigma}$, for some $0 < \overline{\sigma} < \infty$ and all $k \geq 0$. The stepsize-growth policy functions $\{\pi^k\}$, $\pi^k : \mathbb{R} \to \mathbb{R}$, satisfy $\pi^k(x) \geq x$, for all $x \geq 0$.*

1) Convex f_i's$-$Sublinear convergence rate: For any given saddle point $(\mathbf{X}^*, \mathbf{Y}^*)$ of (3), let

$$\mathcal{M}_*(\mathbf{X}) := F(\mathbf{X}) - F(\mathbf{X}^*) + \langle \text{Ł}\mathbf{Y}^*, \mathbf{X} - \mathbf{X}^* \rangle + \|\text{Ł}\mathbf{X}\|^2 \geq 0.$$

Notice that $\mathcal{M}_*(\mathbf{X}) = 0$ iff $\text{Ł}\mathbf{X} = 0$ and $F(\mathbf{X}) = F(\mathbf{X}^*)$, meaning that $\mathbf{X}$ is a solution of Problem (P$'$); furthermore, $\mathcal{M}_*(\bullet)$ is continuous; hence, it is a valid measure of optimality. Finally, define the ergodic sequence:

$$\overline{\mathbf{X}}^k := \frac{1}{\theta^k} \sum_{t=0}^{k-1} \gamma^t \mathbf{X}^t, \quad \text{with} \quad \theta^k = \sum_{t=0}^{k-1} \gamma^t, \quad , \forall k \geq 1.$$

Sublinear convergence of Algorithm 1 is summarized next.

Theorem 1 (sublinear rate)**.** *Let $\{(\mathbf{X}^k, \mathbf{D}^k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1 under Assumptions 1-4, with $\mu = 0$ and $c_1 \in (0,1)$; and let $\{\mathbf{Y}^k\}_{k\geq 0}$ be the auxiliary sequence defined in* (10)*. Then, the following hold:*

(a) `sequence convergence`*: Sequences $\{\mathbf{X}^k\}$, $\{\mathbf{D}^k\}$,$\{\mathbf{Y}^k\}$ globally converge:*

$$\mathbf{X}^k \to \mathbf{X}^\star, \quad \mathbf{Y}^k \to \mathbf{Y}^\star, \quad \mathbf{D}^k \to \text{Ł}\mathbf{Y}^\star, \tag{16}$$

where $(\mathbf{X}^\star, \mathbf{Y}^\star)$ is a saddle points of (3)*;*

(b) `sublinear rate`*: Setting $c_2 \in (0,1)$, it holds that*

$$\mathcal{M}_\star(\overline{\mathbf{X}}^k) \leq \varepsilon, \quad \textit{for all } k \geq N_\varepsilon := \mathcal{O}\left(\widetilde{L}/\varepsilon\right), \tag{17}$$

where $\widetilde{L}$ is the Lipschitz constant of ∇F restricted to the convex hull of the compact set $\cup_{t=0}^\infty [\mathbf{X}^{t-1}, \mathbf{X}^t]$.

Theorem 1 certifies *sequence* convergence of Algorithm 1 sublinear convergence rate of the merit function $\mathcal{M}_\star$, under only *local* smoothness of the f_i's. This is a major departure from the existing nonadaptive decentralized methods (see, e.g., [6], [7] and references therein), which are guaranteed to converge only under *global* smoothness of f_i's. Furthermore, they achieve a $\mathcal{O}(L/\varepsilon)$ complexity whereas Algorithm 1's rate only depends on the *restricted* Lipschitz constant $\widetilde{L}$, which can be significantly smaller than L. This certifies faster convergence, under weaker assumptions. Our rate improves also upon adaptive decentralized methods [27]–[30], and matches the order of adaptive line-search-based methods [24], [31]–[34] while being line-search free.

2) Strongly convex f_i's$-$Linear convergence: We establish now linear convergence of Algorithm 1 when the agents' losses are locally strongly convex ($\mu > 0$).

We find the following choice of $\{\sigma^k\}_{k\geq 0}$ and $\{\pi^k(\bullet)\}$ convenient to achieve the desired rate scaling.

Assumption 5. *The sequence $\{\sigma^k\}_{k\geq 0}$ is chosen as $\sigma^k = \sigma/(\alpha^k)^2$, for some $\sigma \in (0, c_1/2)$ and all $k \geq 0$. The functions $\pi^k : \mathbb{R} \to \mathbb{R}$ are selected as $\pi^k(x) \leq ((k + \beta_1)/(k+1))^{\beta_2} x$, for some $\beta_1, \beta_2 > 0$ and any $k = 0, 1, \ldots$.*

Remark 1. *It's easy to show that, under Assumption 5,* (15) *can be equivalently rewritten as*

$$\alpha^k = \min\left\{\frac{1/2 - \sigma/c_1}{L^k}, \sqrt{1 + c_2\gamma^{k-1}}\alpha^{k-1}, \pi^k(\alpha^{k-1})\right\}, \tag{18}$$

which can be computed using only local information.

Theorem 2 (linear convergence)**.** *Let $\{(\mathbf{X}^k, \mathbf{D}^k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1 under Assumptions 1-3 and 5, with $\mu > 0$ and $c_1, c_2 \in (0,1)$; and let $\{\mathbf{Y}^k\}_{k\geq 0}$ be the auxiliary sequence defined in* (10)*. Then,*

(a) `global sequence convergence`*:*

$$\mathbf{X}^k \to \mathbf{X}^\star, \quad \mathbf{Y}^k \to \mathbf{Y}^\star, \quad \mathbf{D}^k \to \text{Ł}\mathbf{Y}^\star, \tag{19}$$

where $(\mathbf{X}^\star, \mathbf{Y}^\star)$ is the saddle points $(\mathbf{X}^\star, \mathbf{Y}^\star)$ of (3)*;*

(b) `linear rate`*: It holds that*

$$\|\mathbf{X}^k - \mathbf{X}^*\|^2 \leq \varepsilon, \quad \forall k \geq N_\varepsilon := \mathcal{O}\left(\frac{\widetilde{\kappa}}{1 - \lambda_2(W)} \log(1/\varepsilon)\right), \tag{20}$$

where $\widetilde{\kappa} := \widetilde{L}/\widetilde{\mu}$, and $\widetilde{\mu}$ is the strong convexity parameter of F restricted to the the convex hull of the compact set $\cup_{t=0}^\infty [\mathbf{X}^{t-1}, \mathbf{X}^t]$.

As in the convex case, the linear rate in (20) compares favorably with that of nonadaptive methods; it depends only on the *restricted* condition number $\widetilde{\kappa}$, generally smaller than the global one, $\kappa = L/\mu$. It also matches the rates of adaptive line-search methods [24], [31]–[33] while avoiding line-search steps, thereby reducing the computation cost.

III. Beyond global scalar averages

In this section, we introduce a fully local variant of Algorithm 1, requiring neighbor-to-neighbor communications also for the scalar stepsize information.

For each agent $i \in [m]$ and iteration $k \geq 0$, let α_i^k, σ_i^k, and γ_i^k denote the local counterparts of α^k, σ^k, and γ^k in Algorithm 1. To simplify the presentation, we further denote by Λ^k, Σ^k and Γ^k the diagonal matrices of $\{\alpha_i^k\}_{i\in[m]}$, $\{\sigma_i^k\}_{i\in[m]}$ and $\{\gamma_i^k\}_{i\in[m]}$ respectively. With those local values, the primal dual variable are now written as

$$\mathbf{D}^{k+1} = \mathbf{D}^k + (I - W)\Sigma^k\Lambda^k((I + \Gamma^k)\mathbf{X}^k - \Gamma^k\mathbf{X}^{k-1}); \tag{21a}$$

$$\mathbf{X}^{k+1} = \mathbf{X}^k - \Lambda^k(\nabla F(\mathbf{X}^k) + \mathbf{D}^{k+1}). \tag{21b}$$

Supported by the convergence analysis in Sec. II, ideally we would like to select the stepsize according to (15); which however is not compatible to neighboring communications. Next we put forth a procedure that satisfies (15) almost all iterations while requiring only local communication.

At iteration $k \geq 1$, agent i first computes the local curvature estimate (the local counterpart of (13))

$$L_i^k := \frac{\|\nabla f_i(x_i^k) - \nabla f_i(x_i^{k-1})\|}{\|x_i^k - x_i^{k-1}\|},$$

with the convention $L_i^k = 0$ if $x_i^k = x_i^{k-1}$, and forms the curvature-based candidate

$$\hat{\alpha}_i^k := \left(\sqrt{(L_i^k)^2 + 2\sigma_i^k/c_1} + L_i^k\right)^{-1}. \tag{22}$$

Mimicking (15), one would set $\alpha_i^k = \min_{i\in[m]} \widetilde{\alpha}_i^k$ and $\gamma_i^k = \alpha_i^k/\alpha_i^{k-1}$, with

$$\widetilde{\alpha}_i^k \leq \min\left\{\hat{\alpha}_i^k, \sqrt{1 + c_2\gamma_i^{k-1}}\alpha_i^{k-1}, \pi^k(\alpha_i^{k-1})\right\}. \tag{23}$$

However, this would call for a global min-consensus step. As proxy, we replace the global min therein with a local one and adjust the definition of γ^k as follows: at each $k \geq 0$, given $\widetilde{\alpha}_i^k$ defined in (23), let

$$\alpha_i^k = \min_{j \in \mathcal{N}_i} \widetilde{\alpha}_j^k \quad \text{and} \quad \gamma_i^k = \frac{\widetilde{\alpha}_i^k}{\alpha_i^{k-1}}, \quad i \in [m]. \tag{24}$$

This choice is implementable by neighboring communications. Furthermore, under suitable $\{\pi^k\}$ and σ_i^k's, as stated in Assumption 6 below, (24) achieves *global* consensus in *finite* time, still preserving adaptivity in the consensual stepsize value. Lemma 3 below captures this property.

Assumption 6. *For each $i \in [m]$, the sequence $\{\sigma^{k,i}\}_{k \geq 0}$ is monotonically nondecreasing and $0 < \sigma^{k,i} \leq \overline{\sigma}$, for some $0 < \overline{\sigma} < \infty$ and all $i \in [m]$ and $k \geq 0$. The stepsize policy correction sequence $\{\pi^k\}$, $\pi^k : \mathbb{R} \to \mathbb{R}$, satisfies*

$$\pi^k(x) \leq x + \delta^k, \quad ,\forall k \geq 0, \tag{25}$$

where $\{\delta^k\}$ is any summable, nonnegative sequence.

Lemma 3. *Let $\{(\mathbf{X}^k, \mathbf{D}^k)\}$ be the sequence of iterates generated by* (21) *under Assumption 6, with $\{\Lambda^k\}$, $\{\Gamma^k\}$ selected by* (24)*. If the set $\mathcal{K} := \{k : \exists i \in [m], \hat{\alpha}_i^k \leq \pi^k(\alpha_i^{k-1})\}$ is finite and $\{\mathbf{X}^k\}$ is bounded, then there exists $K > 0$ such that, for any $k \geq K$, it holds*

$$\Lambda^k = \alpha^k I \quad \textit{and} \quad \Gamma^k = \gamma^k I, \quad \textit{with} \quad \gamma^k = \frac{\alpha^k}{\alpha^{k-1}},$$

and α^k satifying (15)*, with $\sigma^k := \min_{i \in [m]} \sigma_i^k$.*

It remains to ensure $\hat{\alpha}_i^k \leq \pi^k(\alpha_i^{k-1})$ occurs finitely often. To this end, with $\eta \in (0,1)$, we adopt the strategy to enforce a sufficient decrease of α_i^k, when $\hat{\alpha}_i^k \leq \pi^k(\alpha_i^{k-1})$ by taking

$$\widetilde{\alpha}_i^k := \begin{cases} \min\{\eta \alpha_i^{k-1}, \hat{\alpha}_i^k\}, & \text{if } \hat{\alpha}_i^k \leq \pi^k(\alpha_i^{k-1}); \\ \min\left\{\pi^k(\alpha_i^{k-1}), \sqrt{1 + c_2 \gamma_i^{k-1}}\, \alpha_i^{k-1}\right\}, & \text{else.} \end{cases} \tag{26}$$

In fact, $\widetilde{\alpha}_i^k$ in (26) satisfies (23) and certifies Lemma 4.

Lemma 4. *Instate Lemma 3, but with $\{\widetilde{\alpha}_i^k\}$ chosen according to* (26)*. Then $\{\alpha_i^k\}$ is bounded from below and $|\mathcal{K}| < \infty$.*

The algorithm, based on the above stepsize selection, is summarized in Algorithm 2.

Theorem 3 summarizes sublinear convergence of Algorithm 2 when the f_i's are merely convex, where we defined the following ergodic sequences: given $K \geq 0$, let

$$\overline{\mathbf{X}}_K^k := \frac{1}{\theta_K^k} \sum_{t=K}^{k-1} \gamma^t \mathbf{X}^t, \quad \theta_K^k = \sum_{t=K}^{k-1} \gamma^t, \quad k \geq K+1. \tag{27}$$

Theorem 3 (sublinear rate)**.** *Let $\{(\mathbf{X}^k, \mathbf{D}^k)\}$ be the sequence of iterates generated by Algorithm 2 under Assumption 1-3, and 6, with $\mu = 0$ and $c_1 \in (0,1)$; further assume that $\{\mathbf{X}^k\}$ is bounded. Then, the following hold:*

(a) *global sequence convergence:*

$$\mathbf{X}^k \to \mathbf{X}^\infty, \quad \mathbf{D}^k \to Ł\mathbf{Y}^\infty := \mathbf{D}^\infty, \tag{28}$$

where $(\mathbf{X}^\infty, \mathbf{Y}^\infty)$ is a saddle points $(\mathbf{X}^\star, \mathbf{Y}^\star)$ of (3)*;*

(b) *sublinear rate: Setting $c_2 \in (0,1)$, $\exists K > 0$ such that*

$$\mathcal{M}_*(\overline{\mathbf{X}}_K^k) \leq \varepsilon, \quad \textit{for all } k \geq N_\varepsilon := \mathcal{O}\left(K + \widetilde{L}_K/\varepsilon\right), \tag{29}$$

where $\widetilde{L}_K$ is the Lipschitz constant of ∇F over the convex hull of compact set $\cup_{t=K}^\infty [\mathbf{X}^{t-1}, \mathbf{X}^t]$.

Algorithm 2 ADOLF-local

Data: $\mathbf{X}^{-1}, \mathbf{X}^0 \in \mathbb{R}^{m \times d}$, $\mathbf{D}^0 = 0$; $\Lambda^0, \Sigma^k \succ 0, \forall k \geq 0, \Gamma^0 = I$; $c_1, c_2 \in (0,1]$ $\eta \in (0,1)$; stepsize control policy $\{\pi^k\}$; $W = (1-c)I + cW$, $c \in (0, 1/2)$.

`(S.0)` Initialize:

$$\mathbf{D}^1 = \mathbf{D}^0 + (I - W)\Sigma^0 \Lambda^0((I + \Gamma^0)\mathbf{X}^0 - \Gamma^0 \mathbf{X}^{-1});$$
$$\mathbf{X}^1 = \mathbf{X}^0 + \Lambda^0(\nabla F(\mathbf{X}^0) + \mathbf{D}^1);$$

for $k = 1, 2, \cdots$ **do**

  `(S.2) Local stepsize selection`: Select $\widetilde{\alpha}_i^k$ according to (26).

  `(S.3) Local min-consensus`: Select α_i^k and γ_i^k according to (24).

  `(S.4) Dual and Primal update`: Update $\mathbf{D}^{k+1}$ and $\mathbf{X}^{k+1}$ according to (21).

Similar to Algorithm 1, Algorithm 2 achieves linear convergence with a favorable rate under another choice of $\{\Sigma^k\}$ when $\mu > 0$ (f_i's strongly convex).

Assumption 7. *The sequence $\{\Sigma^k\}_{k \geq 0}$ is selected such that $\sigma_i^k = \sigma/(\alpha_i^k)^2$ for some $\sigma \in (0, c_1/2)$ for any $k \geq 0$ and $\{\pi^k\}$ is selected according to* (25)*.*

Theorem 4 (linear convergence)**.** *Let $\{(\mathbf{X}^k, \mathbf{D}^k)\}_{k \geq 0}$ be the sequence generated by Algorithm 2 under Assumptions 1-3, and 7, with $\mu > 0$ and and $c_1, c_2 \in (0,1)$. Let $(\mathbf{X}^*, \mathbf{Y}^*)$ be the unique saddle point of* (3)*, then if the sequence of iterates $\{\mathbf{X}^k\}$ is bounded, the following hold:*

(a) *Global sequence convergence:*

$$\mathbf{X}^k \to \mathbf{X}^*, \quad \mathbf{D}^k \to Ł\mathbf{Y}^* := \mathbf{D}^*, \tag{30}$$

(b) *Linear rate: There exists a finite $K > 0$ such that $\|\mathbf{X}^k - \mathbf{X}^*\|^2 \leq \varepsilon$, for all $k \geq N_\varepsilon$, with*

$$N_\varepsilon = \mathcal{O}\left(K + \frac{\widetilde{\kappa}_K}{1 - \lambda_2(W)} \log(1/\varepsilon)\right), \tag{31}$$

where $\widetilde{\kappa}_K := \widetilde{L}_K/\widetilde{\mu}_K$ and $\widetilde{\mu}_K$ is the strong convexity parameter of F restricted to the the convex hull of the compact set $\cup_{t=K}^\infty [\mathbf{X}^{t-1}, \mathbf{X}^t]$.

A comparison between Algorithms 1 and 2 shows the price paid for removing the global scalar average: In the local variant, agents may initially use different stepsizes, inducing a *nonmonotone* decrease of the merit function from the first iteration. This mismatch is only transient, that is, there exists a finite index K such that, for all $k \geq K$, the local method uses a common scalar stepsize (still changing), i.e., $\Lambda^k = \alpha^k I$ and $\Gamma^k = \gamma^k I$. The offset K in Theorems 3 and 4 accounts exactly for this initial nonhomogeneous phase.

After iteration K, the monotonically decrease of the merit function is restored.

The boundedness assumption in Theorems 3 and 4 is due to the fact that one only assumed local smoothness, rather than the global one, commonly used in decentralized optimization. Under local smoothness, curvature constants are uniformly bounded (hence the stepsizes) only on bounded regions; boundedness of the generated trajectory ensures then that the restricted constants $\widetilde{L}$ and $\widetilde{\kappa}$ are uniformly bounded. This type of assumption is standard in the analyses under local Lipschitz gradients. It is also not a tuning requirement: the algorithm does not need to know any bound on the iterates. If the standard global-smoothness assumption is postulated instead, as in most existing decentralized methods, our analysis apply readily just the global smoothness constants, and *the boundedness of the iterates guaranteed*; hence Theorems 3 and 4 will be unconditional.

IV. Numerical Results

This section presents some preliminary numerical results, comparing Algorithms 1 (ADOLF) and 2 (ADOLF-local) with the following benchmarks: EXTRA [4] (example of nonadaptive algorithm) and the recent adaptive proposals DATOS and DATOS-local [33]. EXTRA requires full knowledge of network and optimization parameters; we select the stepsize based on a grid-search to identify the choice ensuring fastest practical convergence. DATOS and DATOS-local perform a local line-search to select the stepsize, followed by a global and local-min consensus (communication) step, respectively. In ADOLF-local, we chose $\{\pi^k\}$ such that $\pi^k(x) = x + \frac{6}{\pi^2} \cdot \frac{1}{k^2}$, and $\eta = 0.9$. No $\{\pi^k\}$ is used in ADOLF for the case-study in Sec IV-A whereas we chose $\pi^k(x) = [(k+10)/(k+1)]x$ for ADOLF applied to the ridge regression problem in Sec IV-B.

We simulate line graphs and Erdos-Renyi graphs with edge-probability $p = 0.1$ and $p = 0.9$; there are $m = 20$ agents. The gossip weights are the Metropolis-Hasting [42].

A. Logistic regression

Consider the decentralized logistic regression problem, which is an instance of (P), with

$$f_i(x) = \frac{1}{n}\sum_{j=1}^{n} \log(1 + \exp(-b_{ij} \cdot \langle x, a_{ij}\rangle)),$$

where $a_{ij} \in \mathbb{R}^d$ and $b_{ij} \in \{-1, 1\}$. Data $\{(a_{ij}, b_{ij})\}_{j=1}^n$ is owned by agent i. We use the MNIST dataset. The feature dimension is $d = 784$.

Figure 1 plots the optimality gap $(1/m)\sum_{i=1}^m f(x_i) - f^*$ versus the number of communications, achieved by all the algorihtms, where u is the objective function in (P) and u^* is its minimal value. The figures show that both proposed methods consistently outperform the EXTRA algorithm with fixed stepsize, with the advantage that do not require any user's intervention for the tuning of the stepsize. Moreover, our methods perform comparably with DATOS an DATOS-local, with the advantage that no extra line-search process.

B. Ridge regression

We also report experiments solving a ridge regression problem, a strongly convex instance of (P), with

$$f_i(x) = \frac{1}{n}\|A_i x - b_i\|^2 + \frac{\gamma^i}{2}\|x\|^2, \tag{32}$$

where $(A_i, b_i) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ are the data owned by agent i. The elements of A_i, b_i are independently sampled from the standard normal distribution. Here, $n = 20$ and $d = 500$. We set $\gamma^i = 0.1 + (i-1) \times 0.1$, so that the smoothness parameter L_i are different among each local function.

Figure 2 plots the optimality gap measured by $\|\mathbf{X}^k - \mathbf{X}^*\|^2$ versus the number of communications. Similar to the convex scenario, both proposed algorithms clearly outperform existing decentralized benchmarks.

References

[1] P. Di Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.

[2] Y. Sun, G. Scutari, and A. Daneshmand, "Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation," *SIAM Journal on Optimization*, vol. 32, no. 2, pp. 354–385, 2022.

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

[4] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[5] 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*, vol. 67, no. 3, pp. 708–723, 2018.

[6] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.

[7] A. H. Sayed *et al.*, "Adaptation, learning, and optimization over networks," *Foundations and Trends® in Machine Learning*, vol. 7, no. 4-5, pp. 311–801, 2014.

[8] A. I.-A. Chen, "Fast distributed first-order methods,". Ph.D. thesis, Massachusetts Institute of Technology, 2012.

[9] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2014.

[10] W. Choi and J. Kim, "On the convergence of decentralized gradient descent with diminishing stepsize, revisited," *arXiv preprint arXiv:2203.09079*, 2022.

[11] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.

[12] J. Barzilai and J. M. Borwein, "Two-point step size gradient methods," *IMA journal of numerical analysis*, vol. 8, no. 1, pp. 141–148, 1988.

[13] D. Zhou, S. Ma, and J. Yang, "Adabb: Adaptive barzilai-borwein method for convex optimization," *Mathematics of Operations Research*, 2025.

[14] B. T. Polyak, "Minimization of unsmooth functionals," *USSR Computational Mathematics and Mathematical Physics*, vol. 9, no. 3, pp. 14–29, 1969.

[15] Y. Malitsky and K. Mishchenko, "Adaptive gradient descent without descent," *arXiv preprint arXiv:1910.09529*, 2019.

[16] Y. Malitsky and K. Mishchenko, "Adaptive proximal gradient method for convex optimization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 100670–100697, 2024.

[17] Y. Malitsky and T. Pock, "A first-order primal-dual algorithm with linesearch," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 411–432, 2018.

[18] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization.," *Journal of machine learning research*, vol. 12, no. 7, 2011.

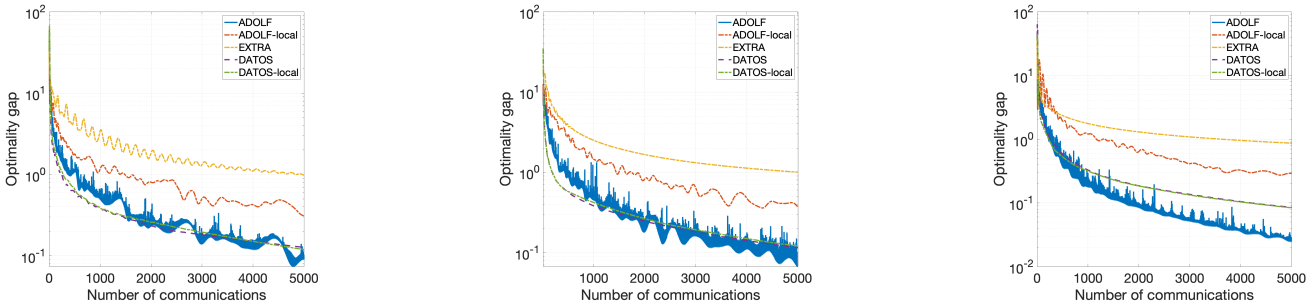

Fig. 1. Logistic regression: $\frac{1}{m}\sum_{i=1}^{m} f(x_i) - f(x^*)$ v.s. # communications. Comparison of EXTRA, DATOS, DATOS-local, ADOLF and ADOLF-local on a line graph (left) and Erdos-Renyi graphs with different edge-probability: $p = 0.1$ (middle); and $p = 0.9$ (right).

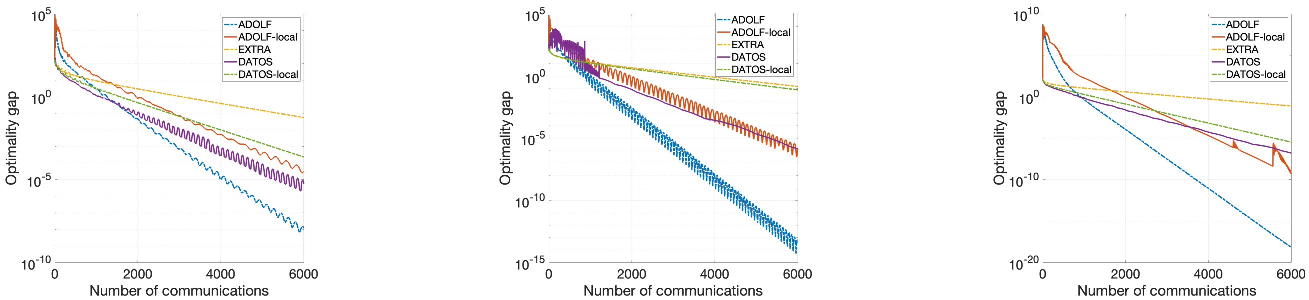

Fig. 2. Linear regression with ℓ_2 regularization: $\|\mathbf{X}^k - \mathbf{X}^*\|^2$ v.s. # communications. Comparison of EXTRA, DATOS, DATOS-local, ADOLF, and ADOLF-local on a line graph (left) and Erdos-Renyi graphs with different edge-probability: $p = 0.1$ (middle); and $p = 0.9$ (right).

[19] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[20] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," *arXiv preprint arXiv:1904.09237*, 2019.

[21] X. Li and F. Orabona, "On the convergence of stochastic gradient descent with adaptive stepsizes," in *The 22nd international conference on artificial intelligence and statistics*, pp. 983–992, PMLR, 2019.

[22] R. Ward, X. Wu, and L. Bottou, "Adagrad stepsizes: Sharp convergence over nonconvex landscapes," *Journal of Machine Learning Research*, vol. 21, no. 219, pp. 1–30, 2020.

[23] P. Latafat, A. Themelis, L. Stella, and P. Patrinos, "Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient," *Mathematical Programming*, pp. 1–39, 2024.

[24] X. Chen, I. Kuruzov, G. Scutari, and A. Gasnikov, "A parameter-free decentralized algorithm for composite convex optimization," *arXiv preprint arXiv:2508.01466*, 2025.

[25] F. Atenas, M. N. Dao, and M. K. Tam, "A distributed proximal splitting method with linesearch for locally lipschitz gradients," *arXiv preprint arXiv:2410.15583*, 2024.

[26] R. Aldana-López, A. Macchelli, G. Notarstefano, R. Aragüés, and C. Sagüés, "Towards parameter-free distributed optimization: a port-hamiltonian approach," *arXiv preprint arXiv:2404.13529*, 2024.

[27] C. Ouyang, Y. Xiong, J. Xu, K. You, and Y. Shi, "Adaptive polyak stepsize with level-value adjustment for distributed optimization," *arXiv preprint arXiv:2603.09097*, 2026.

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

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

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

[31] I. Kuruzov, X. Chen, G. Scutari, and A. Gasnikov, "Adaptive stepsize selection in decentralized convex optimization," *arXiv preprint arXiv:2507.23725*, 2025.

[32] I. Kuruzov, G. Scutari, and A. Gasnikov, "Achieving linear convergence with parameter-free algorithms in decentralized optimization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 96011–96044, 2024.

[33] X. Chen, I. Kuruzov, and G. Scutari, "Adaptive decentralized composite optimization via three-operator splitting," *arXiv preprint arXiv:2602.17545*, 2026.

[34] Q. Xu, N. S. Aybat, and M. Gurbuzbalaban, "An accelerated primal dual algorithm with backtracking for decentralized constrained optimization," *Journal of Nonlinear and Variational Analysis*, 2025.

[35] L. Condat, "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, 2013.

[36] B. C. Vũ, "A variable metric extension of the forward–backward–forward algorithm for monotone operators," *Numerical Functional Analysis and Optimization*, vol. 34, no. 9, pp. 1050–1065, 2013.

[37] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 330–348, 2017.

[38] F. Atenas, M. N. Dao, and M. K. Tam, "A distributed proximal splitting method with linesearch for locally lipschitz gradients," *Journal of Global Optimization*, pp. 1–32, 2025.

[39] A. Nedic, "Distributed gradient methods for convex machine learning problems in networks: Distributed optimization," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 92–101, 2020.

[40] E. K. Ryu and W. Yin, *Large-scale convex optimization: algorithms & analyses via monotone operators*. Cambridge University Press, 2022.

[41] T. Janssen, N. BniLam, M. Aernouts, R. Berkvens, and M. Weyn, "Lora 2.4 ghz communication link and range," *Sensors*, vol. 20, no. 16, p. 4366, 2020.

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