# A Decentralized Primal-Dual Framework for Non-convex Smooth Consensus Optimization

Gabriel Mancino-Ball, Yangyang Xu, and Jie Chen

***Abstract*—In this work, we introduce ADAPD, A DecentrAlized Primal-Dual algorithmic framework for solving non-convex and smooth consensus optimization problems over a network of distributed agents. The proposed framework relies on a novel problem formulation that elicits ADMM-type updates, where each agent first inexactly solves a local strongly convex subproblem with any method of its choice and then performs a neighbor communication to update a set of dual variables. We present two variants that allow for a single gradient step for the primal updates or multiple communications for the dual updates, to exploit the tradeoff between the per-iteration cost and the number of iterations. When multiple communications are performed, ADAPD can achieve theoretically optimal communication complexity results for non-convex and smooth consensus problems. Numerical experiments on several applications, including a deep-learning one, demonstrate the superiority of ADAPD over several popularly used decentralized methods.**

***Index Terms*—non-convex consensus optimization, decentralized optimization, primal-dual method, decentralized learning.**

## I. Introduction

Given a set of $N$ agents connected by an undirected network (graph) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, N\}$ denotes the set of agents and $\mathcal{E} = \{(i, j)\colon \text{agent } i \text{ is connected to agent } j\}$ denotes the set of feasible local communications among agents, consensus optimization methods solve the following problem using only local computation and local communication,

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) \triangleq \tfrac{1}{N}\textstyle\sum_{i=1}^N f_i(\mathbf{x}) \tag{1}$$

where each $f_i\colon \mathbb{R}^p \to \mathbb{R}$ is a differentiable, potentially non-convex, cost function known only to agent $i$.

Problem (1) arises naturally in various scientific and engineering applications such as distributed machine learning/federated learning [1]–[3], decentralized matrix factorization [4], network sensing and localization [5]–[7], and multi-vehicle coordination [8], to name a few. The decision variable $\mathbf{x}$ can represent the weights of a neural network [1], the location of a particular object [8], or the state of a smart grid system [6], for example. Essentially, any scenario in which data is either too large or naturally distributed fits problem (1).

### *A. Problem Formulation*

It is well known [6], [9] that if $\mathcal{G}$ is connected, the following problem is equivalent to (1):

$$\min_{\mathbf{X}} \; F(\mathbf{X}) \;\text{ subject to } \mathbf{W}\mathbf{X} = \mathbf{X} \tag{2}$$

where $\mathbf{W}$ is a *mixing matrix* [6], [10], [11] that satisfies the conditions in Assumption 1 below, and

$$\mathbf{X} \triangleq \begin{bmatrix}\mathbf{x}_1 & \ldots & \mathbf{x}_N\end{bmatrix}^\top \in \mathbb{R}^{N\times p},\; F(\mathbf{X}) \triangleq \tfrac{1}{N}\textstyle\sum_{i=1}^N f_i(\mathbf{x}_i), \tag{3}$$

is the concatenation of local decision variables and the global objective function, respectively, written in matrix notation. Here, $\mathbf{x}_i$ is agent $i$'s local copy of the global variable $\mathbf{x}$, and $\mathbf{W} \in \mathbb{R}^{N\times N}$ represents the connectivity structure of the network $\mathcal{G}$.

*Assumption 1:* The mixing matrix, $\mathbf{W} \in \mathbb{R}^{N\times N}$, satisfies,

*(i)* **(Decentralized property)** $w_{ij} > 0$ if $(i, j) \in \mathcal{E}$, otherwise $w_{ij} = 0$,

*(ii)* **(Symmetric property)** $\mathbf{W} = \mathbf{W}^\top$,

*(iii)* **(Null space property)** $\operatorname{null}(\mathbf{I} - \mathbf{W}) = \operatorname{span}\{\mathbf{e}\}$, where $\mathbf{e} \in \mathbb{R}^N$ is the vector of all ones, and

*(iv)* **(Spectral property)** the eigenvalues of $\mathbf{W}$ lie in the range $(-1, 1]$ and can be ordered as

$$-1 < \lambda_N(\mathbf{W}) \le \cdots \le \lambda_2(\mathbf{W}) < \lambda_1(\mathbf{W}) = 1.$$

Several common choices for mixing matrices are presented in [6].

- *Laplacian-based constant edge weight matrix,*

$$\mathbf{W} = \mathbf{I} - \tfrac{\mathbf{L}}{\tau} \tag{4}$$

where $\mathbf{L}$ is the Laplacian matrix of $\mathcal{G}$ and $\tau > \frac{1}{2}\lambda_1(\mathbf{L})$. Here, $\lambda_1(\mathbf{L})$ is the largest positive eigenvalue of $\mathbf{L}$. If the eigenvalues of $\mathbf{L}$ are unknown, by the Gershgorin circle theorem one can use $\tau = \max_{i\in\mathcal{V}}\{|\mathcal{N}_i|\} + \epsilon$, for some $\epsilon > 0$, where $\mathcal{N}_i \triangleq \{j\colon (i, j) \in \mathcal{E}\}$ is the set of agents that can communicate with agent $i$.

- *Metropolis constant edge weight matrix,* for some $\epsilon > 0$,

$$w_{ij} = \begin{cases} \frac{1}{\max\{|\mathcal{N}_i|,|\mathcal{N}_j|\}+\epsilon}, & (i, j) \in \mathcal{E}, \\ 0, & (i, j) \notin \mathcal{E} \text{ and } i \neq j, \\ 1 - \sum_{k\in\mathcal{V}} w_{ik}, & i = j. \end{cases} \tag{5}$$

- *Symmetric fastest distributed linear averaging matrix*, (FDLA), which is a matrix that achieves the fastest information diffusion through $\mathcal{G}$ and is obtained by solving a semidefinite program [12].

Gabriel Mancino-Ball and Yangyang Xu are with Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA (e-mail: mancig@rpi.edu; xuy21@rpi.edu).

Jie Chen is with MIT-IBM Watson AI Lab, IBM, Cambridge, MA, 02142, USA (email: chenjie@us.ibm.com).

This work was supported by the Rensselaer-IBM AI Research Collaboration, part of the IBM AI Horizons Network. Y. Xu is supported in part by NSF grants DMS-2053493 and DMS-2208394 and the ONR award N00014-22-1-2573. J. Chen is supported in part by DOE Award DE-OE0000910.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes appendices and supplemental proofs. This material is less than 200 KB in size.

Note that the constraint formulation $\mathbf{WX} = \mathbf{X}$ in (2) is not the only choice for a consensus problem. Under Assumption 1, an equivalent consensus constraint adopted by others [4], [13], [14] is $\mathbf{x}_i = \mathbf{x}_j$ for all $(i,j) \in \mathcal{E}$. This constraint is an *edge-based* constraint, whereas we consider a *vertex-based* constraint. When a primal-dual approach is designed, if $\mathcal{G}$ is dense, then a vertex-based constraint introduces fewer dual variables than an edge-based constraint. Further, an optimal $\mathbf{W}$ can be designed, given $\mathcal{G}$ [12].

A vital quantity for our analysis comes from the spectral properties of $\mathcal{G}$. We define

$$\rho \triangleq \left\|\mathbf{W} - \tfrac{1}{N}\mathbf{e}\mathbf{e}^\top\right\|_2 = \max\left\{|\lambda_2(\mathbf{W})|, |\lambda_N(\mathbf{W})|\right\} \in [0,1). \tag{6}$$

The metric in (6) is one way to measure the connectivity of $\mathcal{G}$, where $\rho \approx 0$ implies good connectivity.

Under Assumption 1, particularly $\operatorname{null}\big(\sqrt{\mathbf{I}-\mathbf{W}}\big) = \operatorname{null}\big(\mathbf{I} - \mathbf{W}\big) = \operatorname{span}\{\mathbf{e}\}$, a further equivalent reformulation to (1) is

$$\min_{\mathbf{X}} \; F(\mathbf{X}) \;\text{ subject to }\; \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X} = \mathbf{0}, \tag{7}$$

where $\mathbf{0} \in \mathbb{R}^{N\times p}$ is the matrix of all zeros. A benefit of this formulation is that the constraint $\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X} = \mathbf{0}$ can now be incorporated into a penalty term, $\frac{1}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}\right\|_F^2$, where $\eta > 0$ is a penalty parameter. The gradient associated with this term is $\frac{1}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}$, which can be computed by a single neighbor communication.

One way to solve (7) is to form the augmented Lagrangian with dual variables $\mathbf{Y} \triangleq \begin{bmatrix}\mathbf{y}_1 & \dots & \mathbf{y}_N\end{bmatrix}^\top$ and perform a primal-dual type update as in IDEAL [15]. The issue with this approach is that the communication and computation phases are inherently coupled, illustrated as follows. The classic primal-dual updates at iteration $k$ used to solve (7) are

$$\begin{aligned}\mathbf{X}^{k+1} &= \operatorname*{argmin}_{\mathbf{X}} F(\mathbf{X}) + \left\langle \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Y}^k, \mathbf{X}\right\rangle + \tfrac{1}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}\right\|_F^2, \\ \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Y}^{k+1} &= \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Y}^k + \tfrac{1}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}^{k+1}.\end{aligned} \tag{8}$$

If a first-order method is used to solve the $\mathbf{X}$ subproblem, then part of the gradient will contain $\mathbf{WX}$ *at each gradient computation*, thus for every gradient computed, one neighbor communication must be performed.

Another way to solve (7), as suggested by the Prox-PDA method [4], is to introduce an additional proximal term of the form $\frac{1}{2\eta}\left\|\mathbf{X}-\mathbf{X}^k\right\|_{\mathbf{B}^\top\mathbf{B}}^2$, where $\mathbf{B}^\top\mathbf{B} = -(\mathbf{I}-\mathbf{W}) + \mathbf{D}$ with some diagonal matrix $\mathbf{D}$. This negates the neighbor communication required in the $\mathbf{X}$ subproblem of (8), but introduces a new parameter $\mathbf{D}$ that impacts this method's numerical performance.

Interestingly, Prox-PDA with a special choice of $\mathbf{B}^\top\mathbf{B}$ recovers the distributed ADMM algorithm [16] for consensus optimization with edge-based constraint; see the *Supplemental Material* for details. Hence, part of this work serves to compare ADMM-type methods derived from using edge-based constraints versus vertex-based constraints for solving problem (1) in a decentralized manner. Our numerical findings in Section IV indicate that our below derived inexact ADMM gives better performance than distributed ADMM [16].

To remove the addition of $\mathbf{D}$ from Prox-PDA, yet still decouple the communication and computation phases of traditional primal-dual methods, we propose adding an extra variable (and constraint), leading to the following formulation:

$$\min_{\mathbf{X},\mathbf{X}_0} \; F(\mathbf{X}) \;\text{ subject to }\; \mathbf{X} = \mathbf{X}_0, \;\; \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0 = \mathbf{0}. \tag{9}$$

Governed now by two blocks of primal variables, a natural approach to solve (9) would be to use an ADMM-type update [17], [18], but as argued in Section II, the classic ADMM cannot be implemented in a decentralized manner to solve (9). Hence, we are motivated to design a method that: (i) solves (9) using only decentralized communication and local gradient computations and (ii) achieves the optimal communication complexity results established in [19].

We state the technical assumptions on $F$ below.

*Assumption 2:* The objective function $F$ in (9) satisfies:

*(i)* $F$ is $L$-smooth, i.e. there is $0 < L < \infty$ such that

$$\|\nabla F(\mathbf{X}) - \nabla F(\mathbf{Y})\|_F \le L\,\|\mathbf{X}-\mathbf{Y}\|_F\,, \;\; \forall\; \mathbf{X},\mathbf{Y} \in \mathbb{R}^{N\times p}. \tag{10}$$

*(ii)* $F$ is lower bounded, i.e. there is $\underline{f}$ such that

$$-\infty < \underline{f} \le F(\mathbf{X}), \;\; \forall\; \mathbf{X} \in \mathbb{R}^{N\times p}. \tag{11}$$

The gradient of $F$, written in matrix notation, is

$$\nabla F(\mathbf{X}) \triangleq \tfrac{1}{N}\begin{bmatrix}\nabla f_1(\mathbf{x}_1) & \dots & \nabla f_N(\mathbf{x}_N)\end{bmatrix}^\top \in \mathbb{R}^{N\times p}. \tag{12}$$

Note that the assumptions (10) and (11) are standard in non-convex optimization. If each $f_i$ is $L_i$-smooth then $L \ge \max_i L_i$ and the lower boundedness assumption is equivalent to the existence of a minimizer of $F$.

Before demonstrating a brief literature review, we state a standard definition [4], [11], [19] for stationary points of (1).

*Definition 1 ($\boldsymbol{\varepsilon}$-stationary point):* A matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ is called an $\varepsilon$-stationary point of (1) if

$$\left\|\tfrac{1}{N}\textstyle\sum_{i=1}^N \nabla f_i(\bar{\mathbf{x}})\right\|_2^2 + \left\|\mathbf{X}-\bar{\mathbf{X}}\right\|_F^2 \le \varepsilon \tag{13}$$

where $\bar{\mathbf{x}} \triangleq \frac{1}{N}\mathbf{e}^\top\mathbf{X}$ is the average vector across the $N$ rows of $\mathbf{X}$ and $\bar{\mathbf{X}} \triangleq \frac{1}{N}\mathbf{e}\mathbf{e}^\top\mathbf{X}$ is a matrix version of this same average.

### *B. Related Works*

Distributed computing dates back decades ago to the seminal work [20]. *Centralized* computing paradigms, where $\mathbf{W} = \frac{1}{N}\mathbf{e}\mathbf{e}^\top$ in (2) have been heavily studied; when each $f_i$ is convex, methods such as ADMM [17] and FedAVG [21] have theoretical convergence guarantees. The focus of this paper is on *decentralized* computing paradigms. Methods such as DGD [10] and the distributed subgradient method in [22] have been shown to have sublinear convergence in the convex differentiable and convex non-differentiable settings, respectively. When strong convexity is assumed, the NEAR-DGD [23] method improved the convergence result of DGD by allowing for multiple communications during each iteration. If $f_i$ has Lipschitz continuous gradient and is strongly convex, ADMM [6] and Acc-DNGD-SC [24] exhibit linear convergence. The EXTRA [6] method also exhibits linear convergence if the global function $f$ is restricted strongly convex[1]. SSDA [9] and the recent distributed FGM [25]

[1] A convex, differentiable function $h\colon \mathbb{R}^p \to \mathbb{R}$ is restricted strongly convex about a point $\tilde{\mathbf{x}}$ with parameter $\mu > 0$ if $\langle \nabla h(\mathbf{x}) - \nabla h(\tilde{\mathbf{x}}), \mathbf{x}-\tilde{\mathbf{x}}\rangle \ge \mu\,\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2$ for all $\mathbf{x} \in \mathbb{R}^p$.

were designed for convex problems where the gradients of the Fenchel conjugates[2] of the objective functions $f_i$ are computable, with the later providing complexity results when only approximate gradients are computable. IDEAL [15] and FlexPD [26] are recent primal-dual methods that perform many, or just a few, local neighbor communications per local primal update, respectively.

Of particular interest to us are algorithms dealing explicitly with non-convex local cost functions, e.g. neural networks. When each $f_i$ has Lipschitz continuous gradient, the celebrated DGD [27] has been shown to converge using diminishing step-sizes with a rate $O\left((1-\rho)^{-2}K^{-1}\right)$, where $\rho$ is defined in (6) and $K$ is the iteration number. As indicated in the introduction, Prox-PDA [4] is a primal-dual method that is closely related to non-convex ADMM [28] which converges at a rate $O\left(K^{-1}\right)$, but has superior numerical performance when compared to DGD. SONATA/NEXT [7], [29] is a primal method that exhibits the same convergence rate as Prox-PDA and incorporated a potentially non-smooth but convex regularizer into the objective. While SONATA is applicable to a larger class of problems, it needs to take step-sizes proportional to $N^{-1}$ for convergence; as $N \to \infty$, SONATA's performance can suffer because of this requirement. If the Chebyshev communication protocol [30] is used, SONATA additionally can achieve the $O\left((1-\rho)^{-0.5}K^{-1}\right)$ rate, but SONATA must communicate two variables for every algorithm update. Both Prox-PDA and SONATA require agents to solve a local strongly convex subproblem. Our proposed framework can achieve a convergence rate of $O\left((1-\rho)^{-0.5}K^{-1}\right)$ when multiple neighbor communications are performed; note this is optimal for the class of smooth nonconvex problems [19].

Methods that use stochastic gradients have also been heavily studied. Adapting DGD to stochastic updates yields D-PSGD [2] which is shown to have a convergence rate of $O\left(K^{-0.5}\right)$. Recent works such as $D^2$ [31], DSGT [32], and D-GET [11] make use of stochastic gradient updates mixed with a gradient tracking scheme and draw inspiration from their non-stochastic and centralized counterparts [6], [33]. $D^2$ improves the convergence of D-PSGD, but requires more restrictive assumptions on the eigenvalues of $\mathbf{W}$. The convergence rate of DSGT was shown to be $O\left(K^{-0.5}+(1-\rho)^{-3}K^{-1}\right)$ in [32] and later improved to $\tilde{O}\left(K^{-0.5}+(1-\rho^2)^{-1}K^{-1}\right)$ in [34]. D-GET is able to achieve a rate $O\left(K^{-1}\right)$ but requires a full gradient computation every few iterations; GT-SARAH [35] achieves the same rate but removes the full gradient computation. The authors in [36] develop a primal-dual method with convergence rate $O\left(K^{-0.5}\right)$, where each agent computes one local stochastic gradient per update. The recent SPPDM [37] can also achieve a stochastic $\varepsilon$-stationary point in $O\left(\varepsilon^{-1}\right)$ iterations using stochastic gradients and incorporates a potentially non-smooth but convex regularizer into the objective; SPPDM requires a mini-batch of size $\Omega\left(\varepsilon^{-1}\right)$ to achieve this rate. We remark that our framework can exhibit the optimal convergence rate when *deterministic* gradients are used, yet we include relevant decentralized stochastic methods here for sake of completeness.

Additional algorithms to consider are asynchronous algorithms that do not require a synchronous communication step and algorithms that use time-varying mixing matrices and/or mixing matrices that do not satisfy Assumption 1. Some prominent asynchronous algorithms include AD-PSGD [38], the Asynchronous Primal-Dual method in [39], APPG [40], and the asynchronous ADMM [13], [41]. Algorithms that handle different network structures from those considered here have also been considered: Push-Pull [42] handles directed graphs and DIGing [43] is a gradient tracking algorithm that works for network structures where $\mathbf{W}$ changes at every iteration. While these scenarios are certainly interesting, we focus on synchronous updates and undirected graphs.

### C. Summary of Contributions

Our main contributions are listed below:

- We motivate the novel problem formulation of (9) for solving the non-convex and smooth decentralized consensus optimization problem. We propose ADAPD, **A** **D**ecentr**A**lized **P**rimal-**D**ual algorithmic framework for solving such problem. Our framework is based on performing inexact ADMM-type updates by the augmented Lagrangian function of problem (9). Two variants to our framework: ADAPD-OG (ADAPD-**O**ne **G**radient) and ADAPD-MC (ADAPD-**M**ultiple **C**ommunications) are presented. ADAPD-OG performs a single gradient step instead of inexactly solving a local strongly convex subproblem. ADAPD-MC allows each agent to communicate multiple times with their neighbors for each update. These variants allow for agents to optimize the balance between performing local computation and local communication.
- We prove that ADAPD and ADAPD-OG converge to an $\varepsilon$-stationary point, see (13), in $O\left(L(1-\rho)^{-2}\varepsilon^{-1}\right)$ neighbor communications. When the MC variant is used, this complexity is reduced to $O\left(L(1-\rho)^{-0.5}\varepsilon^{-1}\right)$, which is optimal for smooth, non-convex consensus optimization problems [19]. For both ADAPD and ADAPD-OG, a key ingredient of our analysis is defining a Lyapunov function that decreases with every iteration.
- We compare ADAPD on several non-convex problems to state-of-the-art methods such as DGD [27], Prox-PDA [4], D-PSGD [2], DSGT [32], D-GET [11], and SPPDM [37]. Four experiments are conducted in total: two using deterministic gradients and two using stochastic gradients. In all cases, ADAPD demonstrates numerical superiority over these popularly used methods.

### D. Notation

We use bold face letters such as $\mathbf{X}$ and $\mathbf{x}$ to denote a matrix and a vector, respectively. Let $x_{ij}$ denote the element in the $i^{th}$ row and $j^{th}$ column of the matrix $\mathbf{X}$. The Frobenius norm of a matrix is denoted $\|\cdot\|_F$, while the Euclidean norm of a vector is denoted $\|\cdot\|_2$. Define the standard matrix inner product of $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N\times p}$ to be $\langle \mathbf{A}, \mathbf{B}\rangle \triangleq \sum_{i=1}^{N}\sum_{j=1}^{p} a_{ij}b_{ij}$. For a given symmetric matrix $\mathbf{U} \in \mathbb{R}^{N\times N}$, we denote $\|\mathbf{A}\|_{\mathbf{U}}^2 \triangleq \langle \mathbf{A}, \mathbf{U}\mathbf{A}\rangle$. If

[2] The Fenchel conjugate of a convex function $h\colon \mathbb{R}^p \to \mathbb{R}$ is $h^*(\mathbf{y}) \triangleq \sup_{\mathbf{x}} \langle \mathbf{x}, \mathbf{y}\rangle - h(\mathbf{x})$.

$\mathbf{U}$ is positive definite, then $\|\mathbf{A}\|_{\mathbf{U}}^2$ defines a norm. For square matrices $\mathbf{A},\mathbf{B} \in \mathbb{R}^{N\times N}$, define the matrix inequality $\mathbf{A} \preccurlyeq \mathbf{B}$ to hold if and only if $\mathbf{B}-\mathbf{A}$ is positive semi-definite.

## II. ADAPD Framework

To solve (9), we employ the augmented Lagrangian function with penalty parameter $0 < \eta < \frac{1}{L}$, which is

$$\begin{aligned}\mathcal{L}_\eta(\mathbf{X},\mathbf{X}_0;\mathbf{Y},\mathbf{Z}) &= F(\mathbf{X}) + \langle \mathbf{Y}, \mathbf{X}-\mathbf{X}_0\rangle + \tfrac{1}{2\eta}\|\mathbf{X}-\mathbf{X}_0\|_F^2 \\ &+ \left\langle \mathbf{Z}, \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0\right\rangle + \tfrac{1}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0\right\|_F^2\end{aligned} \tag{14}$$

with dual variables

$$\mathbf{Y} \triangleq \begin{bmatrix}\mathbf{y}_1 & \dots & \mathbf{y}_N\end{bmatrix}^\top,\ \mathbf{Z} \triangleq \begin{bmatrix}\mathbf{z}_1 & \dots & \mathbf{z}_N\end{bmatrix}^\top \in \mathbb{R}^{N\times p}. \tag{15}$$

The classic ADMM [17] approach for solving (9) performs the following updates using (14):

$$\mathbf{X}^{k+1} = \operatorname*{argmin}_{\mathbf{X}} \mathcal{L}_\eta(\mathbf{X},\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k) \tag{16}$$

$$\mathbf{X}_0^{k+1} = \operatorname*{argmin}_{\mathbf{X}_0} \mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0;\mathbf{Y}^k,\mathbf{Z}^k) \tag{17}$$

$$\mathbf{Y}^{k+1} = \mathbf{Y}^k + \beta_1\left(\mathbf{X}^{k+1} - \mathbf{X}_0^{k+1}\right) \tag{18}$$

$$\mathbf{Z}^{k+1} = \mathbf{Z}^k + \beta_2\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1} \tag{19}$$

where $\beta_1, \beta_2 > 0$ are the step-sizes for the dual variables.

Notice that in practice, the exact minimizer of (16) is difficult to find; thus a natural alternative to (16) would be to perform an *inexact* update to the local decision variable as in [44], [45]. This would lead to a computationally efficient way to solve the local subproblem (16) that fully utilizes local computing power without overburdening the agents.

Further, notice that the optimal solution to (17) involves solving

$$\tfrac{1}{\eta}(2\mathbf{I}-\mathbf{W})\mathbf{X}_0 = \tfrac{1}{\eta}\mathbf{X}^{k+1} + \mathbf{Y}^k - \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^k. \tag{20}$$

It should be stated that $\left(2\mathbf{I}-\mathbf{W}\right)^{-1}$ exists (by Assumption 1(iv)). However, it is not easy to solve in a decentralized setting: since $\mathbf{W}$ is not diagonal, solving (17) would involve another iterative method (e.g. Jacobi method), which may require many communications to find the exact minimizer. To remedy this, we note that (20) is a linear equation and apply a simple split for the unknown $\mathbf{X}_0$:

$$\tfrac{1}{\eta}2\mathbf{X}_0^{k+1} - \tfrac{1}{\eta}\mathbf{W}\mathbf{X}_0^k = \tfrac{1}{\eta}\mathbf{X}^{k+1} + \mathbf{Y}^k - \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^k. \tag{21}$$

Remarkably, such a rough estimate for solving the $\mathbf{X}_0$ subproblem based on the past iterate $\mathbf{X}_0^k$ still guarantees convergence, based on the following intuition. Let $\widehat{\mathbf{X}}_0^{k+1}$ be the solution of (20). Then the one gradient step in (21) replaces the unknown term $\mathbf{W}\widehat{\mathbf{X}}_0^{k+1}$ by $\mathbf{W}\mathbf{X}_0^k$. Our analysis will show that $\mathbf{X}_0^{k+1} - \mathbf{X}_0^k \to \mathbf{0}$. Hence, the one-step gradient descent update will become a close approximation to the exact update, and thus it can still guarantee convergence.

Additionally, the $\mathbf{Z}$ update in (19) cannot be implemented in a decentralized manner, as $\sqrt{\mathbf{I}-\mathbf{W}}$ in general will not preserve the underlying network topology. However, notice that if $\mathbf{Z}^0 \in \operatorname{range}(\sqrt{\mathbf{I}-\mathbf{W}})$, then $\mathbf{Z}^k \in \operatorname{range}(\sqrt{\mathbf{I}-\mathbf{W}})$, for all $k \geq 0$ from (19). Hence, we can multiply $\sqrt{\mathbf{I}-\mathbf{W}}$ to the left of all terms in (19) and obtain the equivalent update

$$\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+1} = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^k + \tfrac{1}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}. \tag{22}$$

Doing so allows us to use $\tilde{\mathbf{Z}}^k = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^k$ to simplify all relevant terms in (21) and (22).

To summarize, defining $\beta_1 = \beta_2 = \frac{1}{\eta}$, we propose the following modifications to (16)-(19):

$$\mathbf{X}^{k+1} \approx \operatorname*{argmin}_{\mathbf{X}} \mathcal{L}_\eta(\mathbf{X},\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k) \tag{23}$$

$$\mathbf{X}_0^{k+1} = \tfrac{1}{2}\left(\mathbf{W}\mathbf{X}_0^k + \mathbf{X}^{k+1} + \eta(\mathbf{Y}^k - \tilde{\mathbf{Z}}^k)\right) \tag{24}$$

$$\mathbf{Y}^{k+1} = \mathbf{Y}^k + \tfrac{1}{\eta}\left(\mathbf{X}^{k+1} - \mathbf{X}_0^{k+1}\right) \tag{25}$$

$$\tilde{\mathbf{Z}}^{k+1} = \tilde{\mathbf{Z}}^k + \tfrac{1}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}. \tag{26}$$

On the surface, there are two multiplications with $\mathbf{W}$ in (23)-(26). However, they involve the same variable $\mathbf{X}_0$ differing in only two consecutive iterations. This implies that except for the first iteration, our framework requires only one multiplication by $\mathbf{W}$ per iteration and hence only one communication among agents (for networks where multiple communications are permitted, see Section II-A).

We make two remarks on the solution of the local subproblem (23).

*Remark 1:* For $\eta < \frac{1}{L}$, the $\mathbf{X}$ update performed in (23) is accomplished by inexactly solving the following strongly convex local subproblem for all agents $i = 1, \dots, N$,

$$\min_{\mathbf{x}_i} f_i(\mathbf{x}_i) + \left\langle \mathbf{y}_i^k, \mathbf{x}_i - \mathbf{x}_{0,i}^k\right\rangle + \tfrac{1}{2\eta}\left\|\mathbf{x}_i - \mathbf{x}_{0,i}^k\right\|_2^2, \tag{27}$$

where the inexactness is quantified by the following error quantities. We require

$$\begin{gathered}\|\mathbf{r}_i^{k+1}\|_2^2 \leq \tfrac{\epsilon_{k+1}}{N}, \text{ with} \\ \mathbf{r}_i^{k+1} \triangleq \nabla f_i(\mathbf{x}_i^{k+1}) + \mathbf{y}_i^k + \tfrac{1}{\eta}(\mathbf{x}_i^{k+1} - \mathbf{x}_{0,i}^k),\ \forall k \geq 0,\end{gathered} \tag{28}$$

to hold for the local error at iteration $k$ and for the cumulative error we require,

$$\textstyle\sum_{k=1}^{\infty} \epsilon_{k+1} = O\left(1-\rho\right). \tag{29}$$

*Remark 2:* Similar to the results in [46], we can require,

$$\mathbb{E}\left[\|\mathbf{r}_i^{k+1}\|_2^2\right] \leq \tfrac{\epsilon_{k+1}}{N},\ \forall k \geq 0, \text{ and (29)} \tag{30}$$

and the theoretical results are not significantly affected. This allows for stochastic solvers to be used by each local agent.

From an agent's point of view, (23)-(26) can be summarized in Alg. 1 below.

Recall that we obtain a unique sequence $\{\mathbf{Z}^k\}_{k=1}^K$ in $\operatorname{range}\left(\sqrt{\mathbf{I}-\mathbf{W}}\right)$ from the generated $\tilde{\mathbf{Z}}$-sequence. Therefore, without causing confusion, we can use the corresponding $\mathbf{Z}$-sequence in our analysis. Notice that our framework is sufficiently flexible to allow each agent to use different local subroutines to solve (27). In networks where the computing power of the agents differs vastly (see, e.g. [1]), this flexible framework allows for agents with higher compute capabilities to fully utilize their compute power whereas agents with lower compute capabilities are not expected to utilize heavy optimization tools to solve their local subproblems.

**Algorithm 1: ADAPD** (agent view)

```
Input: $\mathbf{X}^0$, $\mathbf{X}_0^0$, $\mathbf{Y}^0$, $\tilde{\mathbf{Z}}^0 = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^0$ with
       $\mathbf{Z}^0 \in \mathrm{range}(\sqrt{\mathbf{I}-\mathbf{W}})$, $K$, $\eta > 0$, a non-increasing
       sequence $\{\epsilon_k\}_{k=1}^K$.
1 for $k = 0, \ldots, K-1$ do
2     for $i = 1, \ldots, N$ in parallel do
3         Update $\mathbf{x}_i$ until $\|\mathbf{r}_i^{k+1}\|_2^2 \le \frac{\epsilon_{k+1}}{N}$ with $\mathbf{r}_i^{k+1}$ in (28)
4         if $k = 0$ then
5             $\mathbf{x}_{0,i}^{k+1} \leftarrow \frac{1}{2}\left(\sum_{j\in\mathcal{N}_i\cup\{i\}} w_{ij}\mathbf{x}_{0,j}^k + \mathbf{x}_i^{k+1} + \eta(\mathbf{y}_i^k - \tilde{z}_i^k)\right)$
6         else
7             $\mathbf{x}_{0,i}^{k+1} \leftarrow \frac{1}{2}\left(\mathbf{x}_i^{k+1} + \mathbf{x}_{0,i}^k + \eta(\mathbf{y}_i^k - 2\tilde{\mathbf{z}}_i^k + \tilde{\mathbf{z}}_i^{k-1})\right)$
8         $\mathbf{y}_i^{k+1} \leftarrow \mathbf{y}_i^k + \frac{1}{\eta}\left(\mathbf{x}_i^{k+1} - \mathbf{x}_{0,i}^{k+1}\right)$
9         $\tilde{\mathbf{z}}_i^{k+1} \leftarrow \tilde{\mathbf{z}}_i^k + \frac{1}{\eta}(1-w_{ii})\mathbf{x}_{0,i}^{k+1} - \frac{1}{\eta}\sum_{j\in\mathcal{N}_i} w_{ij}\mathbf{x}_{0,j}^{k+1}$
```

We now describe two variants/modifications to Alg. 1 that can be employed if the computational constraints and/or the communication constraints are relaxed.

### A. Framework Variants

*1) Computation Variant:* In scenarios where agents may face a lack of computational resources to solve (27), it may be inefficient to compute $\nabla f_i(\cdot)$ many times. To remedy this, we propose ADAPD-OG (**O**ne **G**radient), which requires each agent to only compute a single gradient during every iteration. More precisely, we do the update:

$$\mathbf{X}^{k+1} = \mathbf{X}_0^k - \eta\left(\nabla F(\mathbf{X}^k) + \mathbf{Y}^k\right). \tag{31}$$

Notice that if $\widehat{\mathbf{x}}_i^{k+1}$ is the exact solution of (27), then $\widehat{\mathbf{X}}^{k+1} = \mathbf{X}_0^k - \eta\left(\nabla F(\widehat{\mathbf{X}}^{k+1}) + \mathbf{Y}^k\right)$, which is a backward step because $\nabla F(\widehat{\mathbf{X}}^{k+1})$ is unknown. The update in (31) is a forward step. Since we can show $\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_F \to 0$, the forward step will eventually be a close approximation of the backward step, and thus we can expect convergence. Alg. 2 displays the pseudocode of ADAPD-OG.

**Algorithm 2: ADAPD-OG** (agent view)

```
Input: $\mathbf{X}^0$, $\mathbf{X}_0^0$, $\mathbf{Y}^0$, $\tilde{\mathbf{Z}}^0 = \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^0$ with
       $\mathbf{Z}^0 \in \mathrm{range}(\sqrt{\mathbf{I}-\mathbf{W}})$, $K$, $\eta > 0$.
for $k = 0, \ldots, K-1$ do
    for $i = 1, \ldots, N$ in parallel do
        $\mathbf{x}_i^{k+1} \leftarrow \mathbf{x}_{0,i}^k - \eta\left(\nabla f_i(\mathbf{x}_i^k) + \mathbf{y}_i^k\right)$
        Perform lines 4 - 9 in Alg. 1 to update $\mathbf{x}_{0,i}^{k+1}$, $\mathbf{y}_i^{k+1}$,
          and $\mathbf{z}_i^{k+1}$
```

*2) Communication Variant:* For convergence, it may be practical to allow agents more than one communication during each ADAPD update. We denote the following multiple communication modification (either to Alg. 1 or Alg. 2) with appending an "-MC" (**M**ultiple **C**ommunications) to the algorithm name.

As stated in the introduction, our analysis depends on the value of $\rho$ which measures how quickly an average value can be computed in a decentralized manner. In a *centralized* computing paradigm, where each agent is allowed to communicate with all other agents either directly or via a central server, the mixing matrix $\mathbf{W}$ can be replaced by the averaging matrix $\frac{1}{N}\mathbf{e}\mathbf{e}^\top$. In this instance $\rho = 0$, which can lead to the fastest convergence for our algorithm in both theory and practice. However, by Assumption 1(i), the communication pattern of the agents is limited to performing only neighbor communications.

One straightforward modification to improve our method's dependence on $\rho$ is to replace $\mathbf{W}$ by $\mathbf{W}^R$ ($R$ denotes a power, not an iteration number here) for the $\tilde{\mathbf{Z}}$ update in (26) and the computation of $\mathbf{X}_0^1$, where $R \ge 1$ is an integer. Notice that $\mathbf{W}^R$ satisfies Assumption 1(ii)-(iv). Thus all our theoretical results hold for this MC modification. Since $\rho(\mathbf{W}^R) = \|\mathbf{W}^R - \frac{1}{N}\mathbf{e}\mathbf{e}^\top\|_2 = \rho(\mathbf{W})^R$, this MC modification can lead to a smaller $\rho$ if $R > 1$. However, if $\rho(\mathbf{W})$ is very close to *one*, $R$ needs to be very large in order to push $\rho(\mathbf{W}^R)$ to *zero*. For this case, more efficient methods have been proposed in the literature for distributed averaging [30], [47], [48]. We employ the Chebyshev accelerated method considered in [30]. The pseudocode is shown in Alg. 3. While the Chebyshev acceleration is called at iteration $k$ of ADAPD-MC or ADAPD-OG-MC, the input $\mathbf{A}^0$ will be $\mathbf{X}_0^{k+1}$.

**Algorithm 3: Chebyshev acceleration**

```
Input: $\mathbf{W}, \mathbf{A}^0, \mathbf{A}^1 = \mathbf{W}\mathbf{A}^0$, $R$.
Compute the step-sizes $\mu_0 = 1, \mu_1 = \frac{1}{\rho}$
for $r = 1, \ldots, R$ do
    $\mu_{r+1} \leftarrow \frac{2}{\rho}\mu_r - \mu_{r-1}$
    $\mathbf{A}^{r+1} \leftarrow \frac{2\mu_r}{\rho\mu_{r+1}}\mathbf{W}\mathbf{A}^r - \frac{\mu_{r-1}}{\mu_{r+1}}\mathbf{A}^{r-1}$
Output: $\mathbf{A}^R$
```

The following lemma shows that the properties in Assumption 1(ii)-(iv) still hold for the operator used in the Chebyshev acceleration and provides an explicit convergence rate for Alg. 3. For a proof, see the proof of Theorem 4 in [9] and Corollary 6.1 in [30].

*Lemma 1:* The output of Alg. 3 can be represented as $\mathbf{A}^R = \mathcal{P}(\mathbf{W}, R)\,\mathbf{A}^0$, where $\mathcal{P}(\mathbf{W}, R)$ is a degree-$R$ polynomial of $\mathbf{W}$ and satisfies Assumptions 1(ii)-(iv). Additionally, we have that $\bar{\mathbf{A}}^R = \bar{\mathbf{A}}^0 \triangleq \bar{\mathbf{A}}$ for any $R$ and

$$\left\|\mathbf{A}^R - \bar{\mathbf{A}}\right\|_F \le 2\left(1 - \sqrt{1-\rho}\right)^R \left\|\mathbf{A}^0 - \bar{\mathbf{A}}\right\|_F. \tag{32}$$

We note that employing Alg. 3 is only feasible if either: (i) the communication pattern is so sparse that consensus error is the main bottleneck for convergence, or (ii) communication cost is low relative to the computation cost, meaning that agents can communicate faster than they can compute. In practice, it is suggested that agents find a balance that distributes work evenly between communication and computation.

## III. Theoretical Guarantees

Our theoretical analysis draws from decentralized analytical methods such as [4], [27] and classical non-convex ADMM analyses, as in [18]. We first show the change in the augmented Lagrangian function value after one ADAPD iteration, i.e. (23)-(26). Then we define a Lyapunov function and use it to show convergence. A crucial quantity for our analysis is

$$\mathbf{V}_0^k \triangleq \left(\mathbf{X}_0^{k+1} - \mathbf{X}_0^k\right) - \left(\mathbf{X}_0^k - \mathbf{X}_0^{k-1}\right). \tag{33}$$

We define $\mathbf{X}_0^{-1} \triangleq \mathbf{X}_0^0$, to ensure that $\mathbf{V}_0^k$ is defined for all $k \geq 0$. In the convergence analysis of Alg. 1, we will make consistent use of the following two facts.

*Fact 1 (Peter-Paul and Young's Inequality):* For any $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times p}$, for any $\delta > 0$ and $i = 1, \ldots, m$, we have,

$$\langle \mathbf{A}, \mathbf{B} \rangle \leq \frac{\delta}{2} \|\mathbf{A}\|_F^2 + \frac{1}{2\delta} \|\mathbf{B}\|_F^2, \tag{34}$$

$$\left\|\sum_{i=1}^m \mathbf{A}_i\right\|_F^2 \leq m \sum_{i=1}^m \|\mathbf{A}_i\|_F^2. \tag{35}$$

## A. Convergence Results of ADAPD

The first step in the analysis creates an equivalence expression among the dual and primal variables. The proofs of all lemmas are located in the *Supplementary Material.*

*Lemma 2:* For all $k \geq 0$, the dual variables in (25) and (26) can be expressed as

$$\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{Z}^k = \mathbf{Y}^k - \frac{1}{\eta}\mathbf{W}(\mathbf{X}_0^k - \mathbf{X}_0^{k-1}) \tag{36}$$

$$\mathbf{Y}^k = \mathbf{R}^k - \nabla F(\mathbf{X}^k) - \frac{1}{\eta}(\mathbf{X}_0^k - \mathbf{X}_0^{k-1}) \tag{37}$$

where $\mathbf{R}^k \triangleq \begin{bmatrix}\mathbf{r}_1^k & \ldots & \mathbf{r}_N^k\end{bmatrix}^\top$ for $\mathbf{r}_i^k$ defined in (28) for all $i = 1, \ldots, N$.

Next, we characterize the change of the augmented Lagrangian function value after one ADAPD iteration.

*Lemma 3:* Let $\{(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k)\}$ be obtained from Alg. 1 or equivalently by updates (23)-(26) such that (28) holds. If $\eta < \frac{1}{2L}$, then it holds for all $k \geq 0$ that

$$\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^{k+1}) - \mathcal{L}_\eta(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k) \\
&\leq \frac{2L\eta - 1}{2\eta} \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 + \frac{\epsilon_{k+1}}{2L} - \frac{1}{2\eta} \left\|\mathbf{X}_0^{k+1} - \mathbf{X}_0^k\right\|_F^2 \\
&\quad + \eta \left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\right\|_F^2 + \eta \left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2.
\end{aligned} \tag{38}$$

Notice that the inequality in (38) does not imply the non-increasing monotonicity of the augmented Lagrangian function at the generated iterates. Below, we bound the dual variable change by the primal variable change and the $\mathbf{V}_0^k$ term. Then we establish another inequality and add it to (38) to build a non-increasing Lyapunov function.

*Lemma 4:* Under the assumptions of Lemma 3, it holds that for all $k \geq 0$,

$$\eta \left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\right\|_F^2 \leq 4L^2\eta \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 + \frac{4}{\eta} \left\|\mathbf{V}_0^k\right\|_F^2 + 8\eta\epsilon_k, \tag{39}$$

$$\eta(1-\rho) \left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2 \leq 8L^2\eta \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 + \frac{10}{\eta} \left\|\mathbf{V}_0^k\right\|_F^2 + 16\eta\epsilon_k, \tag{40}$$

where $\mathbf{V}_0^k$ is defined in (33).

*Lemma 5:* For all $k \geq 0$, the following relation holds

$$\begin{aligned}
&\frac{1}{2\eta}\left(\left\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2 + \left\|\sqrt{\mathbf{I} - \mathbf{W}}(\mathbf{X}_0^{k+1} - \mathbf{X}_0^k)\right\|_F^2\right) \\
&\quad - \frac{1}{2\eta}\left\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{X}_0^k\right\|_F^2 \\
&\quad + \frac{1}{2\eta}\left(\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2 + \left\|\mathbf{X}_0^{k+1} - \mathbf{X}_0^k\right\|_{\mathbf{W}}^2 - \left\|\mathbf{X}_0^k - \mathbf{X}_0^{k-1}\right\|_{\mathbf{W}}^2\right) \\
&\leq (L - \frac{1}{2\eta})\left\|\mathbf{X}_0^{k+1} - \mathbf{X}_0^k\right\|_F^2 + \frac{L}{2}\left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 \\
&\quad + \frac{1}{2\eta}\left\|\mathbf{X}_0^k - \mathbf{X}_0^{k-1}\right\|_F^2 - \frac{1}{2\eta}\left\|\mathbf{V}_0^k\right\|_F^2 + \frac{2}{L}\epsilon_k
\end{aligned} \tag{41}$$

where $\mathbf{V}_0^k$ is defined in (33).

Using Lemmas 4 and 5, we are ready to build a non-increasing Lyapunov function as follows.

*Lemma 6:* Let $\{(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k)\}$ be obtained from Alg. 1 or equivalently by updates (23)-(26) such that (28) holds. If $\eta < \frac{1}{2L}$, then

$$\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^{k+1}) + \frac{C}{2\eta}\left\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2 \\
&\quad + \frac{C}{2\eta}\left\|\mathbf{X}_0^{k+1} - \mathbf{X}_0^k\right\|_F^2 \\
&\leq \mathcal{L}_\eta(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k) + \frac{C}{2\eta}\left\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{X}_0^k\right\|_F^2 + \frac{C}{\eta}\left\|\mathbf{X}_0^k - \mathbf{X}_0^{k-1}\right\|_F^2 \\
&\quad + \left(\frac{(8L^2(1-\rho) + 16L^2)\eta^2 + (C+2)L(1-\rho)\eta - (1-\rho)}{2(1-\rho)\eta}\right)\left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 \\
&\quad + \left(\frac{2CL\eta - C - 1}{2\eta}\right)\left\|\mathbf{X}_0^{k+1} - \mathbf{X}_0^k\right\|_F^2 \\
&\quad + \frac{(1-\rho) + (32L + 16L(1-\rho))\eta + 4C(1-\rho)}{2L(1-\rho)}\epsilon_k.
\end{aligned} \tag{42}$$

for all $k \geq 0$, where $C \geq \frac{20 + 8(1-\rho)}{(1-\rho)^2}$ is a fixed constant.

For the rest of the analysis, we fix $C \triangleq \frac{28}{(1-\rho)^2}$ as used in Lemma 6, define the Lyapunov function:

$$\Phi^k \triangleq \mathcal{L}_\eta(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k) + \frac{C}{2\eta}\left\|\sqrt{\mathbf{I} - \mathbf{W}}\mathbf{X}_0^k\right\|_F^2 + \frac{C}{\eta}\left\|\mathbf{X}_0^k - \mathbf{X}_0^{k-1}\right\|_F^2. \tag{43}$$

We show the lower boundedness of this Lyapunov function in the following proposition and use this to obtain the convergence of Alg. 1.

*Proposition 1:* Under Assumptions 1 and 2, let $\{(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k)\}$ be obtained from Alg. 1 or equivalently by updates (23)-(26) such that (28) and (29) hold. Choose $C$ and $\eta$ such that

$$C = \frac{28}{(1-\rho)^2} \text{ and } \eta < \frac{1}{2CL}. \tag{44}$$

Then the Lyapunov function (43) is uniformly lower bounded. More specifically, for all $k \geq 0$,

$$\Phi^k \geq \underline{\phi} := \underline{f} - \frac{(1-\rho) + (32L + 16L(1-\rho))\eta + 4C(1-\rho)}{2L(1-\rho)}\sum_{k=0}^\infty \epsilon_k - 1 > -\infty, \tag{45}$$

where we take $\epsilon_0 = \epsilon_1$ and $\underline{f}$ is defined in Assumption 2.

We are now in position to prove the convergence rate results of ADAPD.

*Theorem 1:* Under the same conditions assumed in Proposition 1, it holds that

$$\begin{aligned}&\tfrac{C_1}{K}\textstyle\sum_{k=0}^{K-1}\left(\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2+\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2\right)\\&\le\tfrac{\Delta_\Phi}{K}+\tfrac{(32L+16L(1-\rho))\eta+(4C+1)(1-\rho)}{2L(1-\rho)K}\textstyle\sum_{k=0}^{K-1}\epsilon_k,\end{aligned}\tag{46}$$

where $\Delta_\Phi \triangleq \Phi^0 - \underline{\phi}$ and

$$C_1 \triangleq \tfrac{1-2CL\eta}{2\eta}.\tag{47}$$

*Theorem 2 (Convergence of ADAPD):* Under the same conditions assumed in Proposition 1, it holds

$$\begin{aligned}&\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left(\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_2^2+\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2\right)\\&\le\tfrac{((2L^2+1)C_2+C_4)\Delta_\Phi}{KC_1}+\tfrac{(192L^2+96)\eta^2}{KC_1(1-\rho)^2}\textstyle\sum_{k=0}^{K-1}\epsilon_k\\&\quad+\tfrac{((2L^2+1)C_2C_3+C_3C_4+4C_1)}{KC_1}\textstyle\sum_{k=0}^{K-1}\epsilon_k\end{aligned}\tag{48}$$

where $C_1$ is defined in (47), $C_2 \triangleq \frac{208}{(1-\rho)^2}$, $C_3 \triangleq \frac{(32L+16L(1-\rho))\eta+(4C+1)(1-\rho)}{2L(1-\rho)}$, $C_4 \triangleq \frac{16}{\eta^2}$, $\bar{\mathbf{x}}^k \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^k$, and $\bar{\mathbf{X}}^k \triangleq \frac{1}{N}\mathbf{e}\mathbf{e}^\top\mathbf{X}^k$.

*Remark 3:* Let $k_0 = \operatorname*{argmin}_{1\le k\le K}\left(\left\|\nabla f(\bar{\mathbf{x}}^k)\right\|_2^2+\left\|\mathbf{X}^{k'}-\bar{\mathbf{X}}^k\right\|_F^2\right)$. Then $\left\|\nabla f(\bar{\mathbf{x}}^{k_0})\right\|_2^2+\left\|\mathbf{X}^{k_0}-\bar{\mathbf{X}}^{k_0}\right\|_F^2=\mathcal{O}\left(\frac{1}{K}\right)$. Hence, in order to produce an $\varepsilon$-stationary point as defined in Definition 1, we need $K=\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ iterations. Furthermore, notice that all the problems in (27) are smooth and strongly convex. The steepest gradient method has linear convergence to solve such problems. Hence, to produce $\mathbf{x}_i^{k+1}$ as a $\frac{\varepsilon_{k+1}}{N}$-accurate solution of the problem in (27), it needs $\mathcal{O}\left(\log\frac{N}{\varepsilon_{k+1}}\right)$ gradient evaluations for each $i=1,\ldots,N$. Choose $\varepsilon_{k+1}=\frac{\epsilon_0}{(k+1)^\gamma}$ for all $k\ge 0$ and for some $\gamma>1$ where $\epsilon_0=\mathcal{O}(1-\rho)$. Then $\{\varepsilon_{k+1}\}$ is summable, and the total gradient evaluations to produce an $\varepsilon$-stationary point of (1) would be $\sum_{k=0}^{K-1}\mathcal{O}\left(\log N(k+1)^\gamma\right)=\mathcal{O}\left(\frac{1}{\varepsilon}\log\frac{N}{\varepsilon^\gamma}\right)$.

### B. Convergence Results of ADAPD-OG

The convergence rate results of the ADAPD-OG follow the same logic as the results for ADAPD, hence all supporting Lemmas and proofs are located in the *Supplementary Material*. Notice that (37) is no longer a valid relation when Alg. 2 is used. Instead, we have the following from (25) and (31):

$$\mathbf{Y}^k=-\nabla F(\mathbf{X}^{k-1})-\tfrac{1}{\eta}\left(\mathbf{X}_0^k-\mathbf{X}_0^{k-1}\right),\forall\, k\ge 0.\tag{49}$$

As in the analysis for ADAPD, we define $\mathbf{X}_0^{-1}\triangleq\mathbf{X}_0^0$ and further define $\mathbf{X}^{-1}\triangleq\mathbf{X}^0$. We have the following result.

*Theorem 3 (Convergence of ADAPD-OG):* Under Assumptions 1 and 2, let $\{(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)\}$ be obtained from Alg. 2 or equivalently by (31) and (24)-(26). Choose $\hat{C}$ and $\eta$ such that

$$\hat{C}\triangleq\tfrac{16}{(1-\rho)^2}\text{ and }\eta<\tfrac{1}{2\hat{C}L}.\tag{50}$$

Then, it holds

$$\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left(\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_2^2+\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2\right)\le\dfrac{\left((2L^2+1)\hat{C}_2+\hat{C}_3\right)\Delta_{\hat{\Phi}}}{\hat{C}_1K}$$

where $\hat{C}_1\triangleq\frac{L}{(1-\rho)^2}\le\frac{(1-\rho)-(\hat{C}+1)L(1-\rho)\eta-((1-\rho)+1)4L^2\eta^2}{2(1-\rho)\eta}$, $\hat{C}_2\triangleq\frac{112}{(1-\rho)^2}$, $\hat{C}_3\triangleq\frac{8}{\eta^2}$, $\Delta_{\hat{\Phi}}\triangleq\hat{\Phi}^0-\underline{f}+1$, $\bar{\mathbf{x}}^k\triangleq\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^k$, and $\bar{\mathbf{X}}^k\triangleq\frac{1}{N}\mathbf{e}\mathbf{e}^\top\mathbf{X}^k$.

*Remark 4:* Theorems 2 and 3 give the convergence results in terms of the constants $C_1, C_2, C_3$, and $C_4$ for Alg. 1 (or $\hat{C}_1, \hat{C}_2, \hat{C}_3$, and $\hat{C}_4$ for Alg. 2) which depend on $C$ ($\hat{C}$) and $\eta$, and in turn depend on $L$ and $\rho$. To make this dependency clearer, we use the $\mathcal{O}(\cdot)$ notation to give dependency only in terms of $L,\rho$, and the algorithm iteration number $K$. For Alg. 1, using (29), and for Alg. 2, we have

$$\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left(\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_2^2+\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2\right)=\mathcal{O}\left(\tfrac{L}{(1-\rho)^2K}\right).\tag{51}$$

### C. Complexity Analysis

We now give a complexity analysis for Alg.'s 1 and 2 regarding the number of primal gradient computations and neighbor communications each method must perform to find an $\varepsilon$-stationary point (see Definition 1); we refer to these quantities as the *computation* and *communication* complexities, respectively. This leads to the following corollaries, whose proofs are in the *Supplementary Materials*.

*Corollary 1 (Complexity results of ADAPD):* Under the same conditions assumed in Theorem 2, if steepest gradient descent is used to solve the subproblem (27), such that conditions (28) and (29) hold, then Alg. 1 can produce an $\varepsilon$-stationary point in respectively

$$\tilde{\mathcal{O}}\left(\tfrac{L}{(1-\rho)^2\varepsilon}\right)\text{ and }\mathcal{O}\left(\tfrac{L}{(1-\rho)^2\varepsilon}\right)\tag{52}$$

gradient computations[3] and neighbor communications.

*Corollary 2 (Complexity results of ADAPD-OG):* Under the same conditions assumed in Theorem 3, Alg. 2 can produce an $\varepsilon$-stationary point in

$$\mathcal{O}\left(\tfrac{L}{(1-\rho)^2\varepsilon}\right)\tag{53}$$

gradient computations and neighbor communications.

Corollaries 1 and 2 show that both ADAPD and ADAPD-OG depend upon the quantity $(1-\rho)^{-2}$ in terms of the number of communications required to achieve $\varepsilon$-stationarity. To improve this to the optimal communication complexity in terms of the dependence on $\rho$ (see, e.g. [9]), we have the following theorem.

*Theorem 4 (Complexity results of ADAPD-MC):* Under the same conditions assumed in Theorem 2, let $R=\lceil\frac{2}{\sqrt{1-\rho}}\rceil$ iterations of the Chebyshev acceleration Alg. 3 be performed during the line 9 update of Alg. 1. Then Alg. 1 can produce an $\varepsilon$-stationary point in $\tilde{\mathcal{O}}\left(\frac{L}{\varepsilon}\right)$ and $\mathcal{O}\left(\frac{L}{\sqrt{1-\rho\varepsilon}}\right)$ gradient computations and neighbor communications, respectively.

[3] The $\tilde{\mathcal{O}}(\cdot)$ hides a log dependency on $\varepsilon$ here.

*Theorem 5 (Complexity results of ADAPD-OG-MC):* Under the same conditions assumed in Theorem 3, let $R = \lceil \frac{2}{\sqrt{1-\rho}} \rceil$ iterations of the Chebyshev acceleration Alg. 3 be performed during the line 9 update of Alg. 2. Then Alg. 2 can produce an $\varepsilon$-stationary point in $O\left(\frac{L}{\varepsilon}\right)$ and $O\left(\frac{L}{\sqrt{1-\rho}\varepsilon}\right)$ gradient computations and neighbor communications, respectively.

## IV. Numerical Experiments

We test our proposed methods on several non-convex problems: (i) a binary classification problem using logistic regression with a non-convex regularizer, (ii) a multi-target cooperative localization problem, and (iii) two image classification problems using convolutional neural networks. The experiments serve to verify both the flexibility of our methods, as well as their numerical superiority over other decentralized optimization methods. Implementations of our methods are made available at https://github.com/RPI-OPT/ADAPD.

For experiments (i) and (ii), we compare our methods to DGD with a diminishing step-size [27] and the single gradient version of Prox-PDA, called Prox-GPDA [4]. We also ran experiments with Prox-PDA but found no advantage over using Prox-PDA versus Prox-GPDA; since Prox-GPDA only requires one gradient computation per update, we use this as a baseline. For Alg. 1, we use $\epsilon_k = \frac{\hat{\epsilon}}{(k+1)^d}$ in (28) where $\hat{\epsilon}$ and $d$ are tuned from a fixed set of values and solve each agent's local problem (27) by the FISTA [49] method. For experiment (iii), we compare to D-PSGD [2], DSGT [32], D-GET [11], a single stochastic gradient implementation of Prox-PDA [4], and SPPDM [37]. For all experiments, we fix a set of penalty parameters/step-sizes and optimize each algorithm over this set, choosing whichever penalty/step-size performs the best. For all methods besides Prox-(G)PDA and SPPDM, we use the same mixing matrix, which will be described in each subsection below. For Prox-(G)PDA, we take $\mathbf{W}$ to be the formulation as given in [4] (see equation (23) in [4] and the discussion that follows) and for SPPDM, we use the graph Laplacian as stated in their problem formulation.

### A. Non-convex Regularized Logistic Regression

We consider the non-convex decentralized binary classification problem [4], [46]. Utilizing a logistic regression formulation, the local agent cost functions are given by,

$$f_i(\mathbf{x}_i) = \frac{1}{m_i}\sum_{j=1}^{m_i}\log\left(1+\exp(-b_j\left\langle \mathbf{x}_i,\mathbf{a}_j\right\rangle)\right) + \sum_{d=1}^{D}\frac{\alpha(\mathbf{x}_i[d])^2}{1+(\mathbf{x}_i[d])^2} \quad (54)$$

where $\mathbf{x}_i[d]$ denotes the $d^{th}$ component of the vector $\mathbf{x}_i$. Given a set of data $\{(\mathbf{a}_j, b_j)\}_{j=1}^{m_i}$ for all $i = 1,\dots,N$, where $b_j \in \{-1,+1\}$ denotes a particular class label, (54) can be used to perform binary classification and the non-convex regularizer, $\sum_{d=1}^{D}\frac{\alpha(\mathbf{x}_i[d])^2}{1+(\mathbf{x}_i[d])^2}$ helps to promote sparsity on the solutions. We use the a9a dataset [50], [51] which consists of 32,561 training data points and 16,281 testing data points. Each data point $\mathbf{a}_j \in \mathbb{R}^{123}$ contains numerical features about adults from the 1994 Census database and $b_j$ indicates whether or not the adults earn more or less than \$50,000 per year. We fix $N = 50$ for this experiment and simulate agent connectivity in two ways: (i) using a ring-structured graph and (ii) using a random Erdös Rényi graph, with connection probability equal to 0.3 (i.e. each agent is connected to roughly 15 other agents).

For the ring-structured graph, we choose $\mathbf{W}$ to be

$$w_{ij} = \begin{cases} \frac{1}{2}, & i = j, \\ \frac{1}{4}, & (i,j) \in \mathcal{E} \text{ and } i \neq j, \\ 0, & \text{otherwise,} \end{cases}$$

and for the random Erdös Rényi graph, we use the Laplacian-based constant edge weight matrix from (4). We vary $\alpha \in \{0.01, 1.0\}$ to study the effect that the non-convex term has on each agent's local subproblem. For all runs, we fix the communication budget to 500 neighbor communications. Additionally, we compare ADAPD-MC and ADAPD-OG-MC to the other methods. We perform 5 iterations of the Chebyshev acceleration in Alg. 3 during every outer iteration of Alg. 1 for ADAPD-MC and 2 iterations for ADAPD-OG-MC. This means we only compute 250 gradients for ADAPD-OG-MC, to keep with the 500 communication budget. We report the $\varepsilon$-stationarity violation (13) for the following four scenarios: (i) the random Erdös Rényi graph with $\alpha = 1.0$, (ii) the random Erdös Rényi graph with $\alpha = 0.01$, (iii) the ring graph with $\alpha = 1.0$, and (iv) the ring graph with $\alpha = 0.01$.

From Figure 1, it is evident that when the communication pattern is sparse (i.e. the two rightmost plots), performing multiple communications and multiple local updates can reduce the stationarity violation faster over performing just one neighbor communication or just one local update. When the communication pattern is not too sparse (i.e. the two leftmost plots), ADAPD-OG performs significantly better and requires fewer gradients than the other methods compared here. In all cases, ADAPD and it's variants outperform DGD and Prox-GPDA.

### B. Multi-Target Cooperative Localization

Multi-target cooperative localization is a target locating problem [5]: given only a noisy distance metric, can $N$ agents locate $N_T$ common targets? Let $\{\omega_i\}_{i=1}^{N}$ be a set of locations of the agents, i.e. $\omega_i \in \mathbb{R}^2$ for all $i = 1,\dots,N$. Then the local objective function for each agent is given by

$$f_i(\mathbf{x}_i) = \frac{1}{4}\sum_{t=1}^{N_T}\left(\xi_{i,t} - \|\mathbf{x}_i[t] - \omega_i\|_2^2\right)^2 \quad (55)$$

where $\xi_{i,t}$ is a random variable that represents a noisy distance metric, and $\mathbf{x}_i = \begin{bmatrix}\mathbf{x}_i[1]^\top & \dots & \mathbf{x}_i[N_T]^\top\end{bmatrix}^\top \in \mathbb{R}^{2N_T}$ is a stacking of the vectors $\{\mathbf{x}_i[t]\}_{t=1}^{N_T}$. Note that (55) is indeed non-convex, but it is not globally $L$-smooth for any $L \geq 0$. However, we still find this problem is valuable to test our methods. Denote the true targets as $\mathbf{x}^*[t]$ for all $t = 1,\dots,N_T$; these are used to generate $\xi_{i,t}$ for all $i$ and $t$ by computing $\xi_{i,t} = \|\mathbf{x}^*[t] - \omega_i\|_2^2 + \epsilon_{i,t}$ where $\epsilon_{i,t}$ is drawn from a normal distribution with mean 0 and variance $\sigma^2 > 0$. For all of our experiments we set $\sigma^2 = 0.01$. We simulate agent connectivity by randomly generating $N = 50$ agents in $[-1,1]\times[-1,1]$ grid and creating an edge between agents if the Euclidean distance between them is less than or equal to 0.3. Each coordinate in

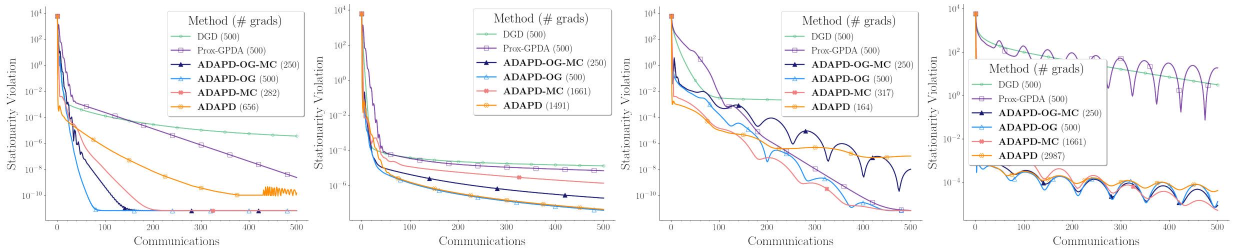

Fig. 1. Stationarity violation for the non-convex logistic regression problem (in order from left to right): random Erdös Rényi graph with $\alpha = 1.0$, random Erdös Rényi graph with $\alpha = 0.01$, ring-structured graph with $\alpha = 1.0$, and ring-structured graph with $\alpha = 0.01$.

the targets $\{\mathbf{x}^*[t]\}_{t=1}^{N_T}$ is drawn independently from a normal distribution with mean 0 and variance 0.1. Figure 2 shows the connectivity of the agents, as well as an example of target locations.

For this example, $\mathbf{W}$ is chosen to be the Laplacian-based constant edge weight matrix from (4). We randomly generate $N_T = 5$ targets and limit the communications to be 1,500 for all algorithm runs, with all methods starting from the same initial point. Since the targets are randomly generated for each experiment, we perform 10 independent trials and plot the mean results, with an associated 95% confidence interval.

Figure 2 shows that in terms of stationarity violation, ADAPD is superior to DGD and Prox-GPDA. Using only 20% more gradient computations on each agent, ADAPD is able to both solve the localization problem and find the true targets with fewer communications than the other methods. Additionally, ADAPD-OG utilizes the same number of gradient computations and neighbor communications as Prox-GPDA and DGD, but still performs better.

### C. Convolutional Neural Networks

For these experiments, we fix $N = 8$ agents and use a ring-structured graph where self-weighting and neighbor weighting is set to be $\frac{1}{3}$. We train the models on a cluster of 8 NVIDIA Tesla V100 GPUs, where each GPU represents an agent. PyTorch is used in the training of the models and OpenMPI is used to perform the neighbor communication of the neural network weights. All experiments are performed with 10 different initial starting points. We report the average results, as well as a 95% confidence interval taken over the 10 trials.

*1) MNIST:* The first Convolutional Neural Network (CNN) experiment we perform is training LeNet [52] on the MNIST dataset. We make the activation function for each layer the hyperbolic tangent function to ensure smoothness of the local objective functions. Since methods like DSGT [32] and D-GET [11] require multiple neighbor communications during each update, we instead fix the number of *epochs* for this experiment to 50 and fix the mini-batch size to 64 for all methods. We randomly generate 10 sets of initial points for the agents and report the average of all relevant metrics, as well as a 95% confidence interval. For ADAPD and ADAPD-OG, we simply replace the full gradient computation by a stochastic gradient during each local agent update. It is worth noting that neither ADAPD, nor Prox-PDA, have theoretical convergence guarantees in this experimental setting. Nonetheless, we see impressive results for this problem and thus include it. To see the effect of stochasticity here, we run the ADAPD in Alg. 1 by computing both one and two stochastic gradients during line 3. For Prox-PDA, we compute one stochastic gradient step. Similar to [2], we report the stationarity violation for all methods, as well as the training loss and testing accuracy using the average of the local agent's weights[4]. In practice, this is not feasible due to the decentralized communication pattern, however, an average model can be obtained after all local training has been done by performing many neighbor communication rounds [12]. Note that the training loss reported here is not scaled by $\frac{1}{N}$ to facilitate a fair comparison with standard CNN training methods (i.e. centralized training).

Additionally, we report the wall-clock time taken to reach and stay above 97% testing accuracy for the MNIST image classification problem in Table I. This value comes from selecting the highest whole number of testing accuracy that most methods exceed. D-GET does not achieve this accuracy in the alloted amount of epochs. The "Samples" column indicates the amount of data visited by each agent to achieve the 97% testing accuracy and the "Communications" column indicates the corresponding number of communications performed by each agent (for D-GET, these values are simply the total numbers used during training). We also include each method's highest testing accuracy in the last column.

While D-GET is able to achieve the lowest stationarity violation, the training loss and testing accuracy indicate it does not converge to a solution that solves the classification problem well. Figure 3 and Table I show that both ADAPD and ADAPD-OG outperform competitors in terms of testing accuracy, suggesting that ADAPD is able to find a solution that generalizes better than other methods. Additionally, Table I shows that ADAPD (with 2 SGD steps) and ADAPD-OG require far fewer communications to achieve a high testing accuracy. In a network setting where communication time dominates the computation time, ADAPD and its variants can outperform the competitors.

*2) CIFAR-10:* The second CNN experiment we perform is training the ALL-CNN model [53] on the CIFAR-10 dataset [54]. We add batch normalization after every ReLU activation function and perform no data augmentation prior to training. For these experiments, we fix the mini-batch size to 128 for all methods and limit the number of updates so that

[4] Similar results for both the MNIST and CIFAR-10 image classification problems are observed if the local weights are used to compute the training loss and testing accuracy.

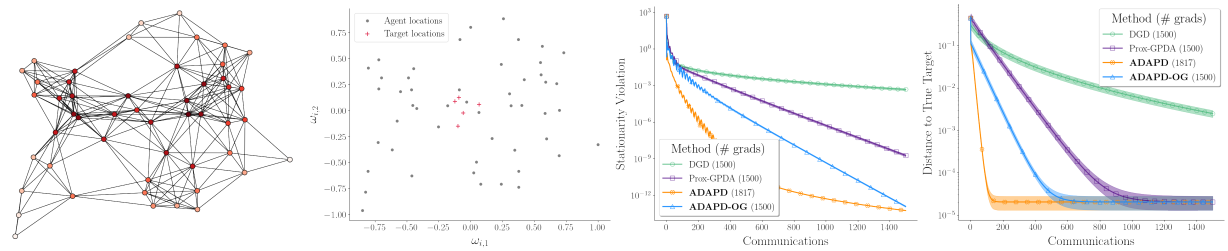

Fig. 2. In order from left to right: agent locations and their connectivity (darker colors indicate more connections), example of target locations, stationarity violation, and distance to true targets for the multi-target cooperative localization problem.

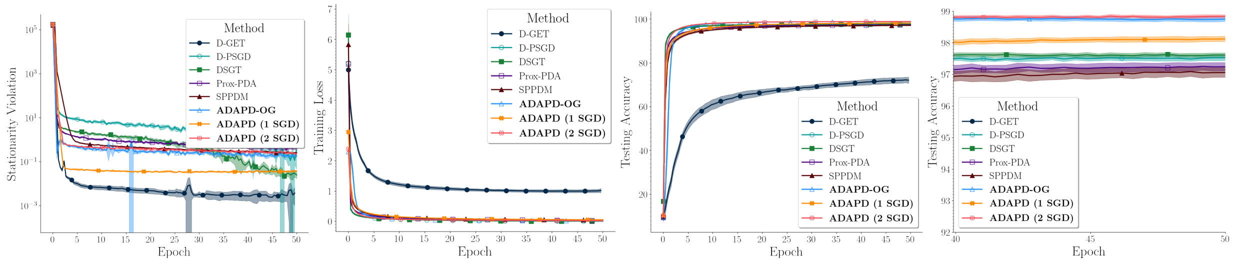

Fig. 3. In order from left to right: stationarity violation, training loss, testing accuracy, and a zoomed version of the testing accuracy over the last ten epochs for the MNIST image classification problem.

| Method | To reach 97% testing accuracy | | | Highest accuracy (%) |
|---|---|---|---|---|
| | Time (s) | Samples | Communications | |
| D-GET | ✗ | 376,524 | 4,096 | 72.17 |
| D-PSGD | 31.56 | 92,800 | 1,450 | 97.53 |
| DSGT | 26.78 | 73,600 | 2,300 | 97.64 |
| Prox-PDA | 80.99 | 227,200 | 3,550 | 97.25 |
| SPPDM | 116.73 | 326,400 | 5,100 | 97.09 |
| ADAPD-OG | **16.35** | **48,000** | 750 | 98.77 |
| ADAPD (1 SGD) | 74.2 | 121,600 | 1,900 | 98.12 |
| ADAPD (2 SGD) | 47.5 | 83,200 | **650** | **98.85** |

TABLE I
TIME TO REACH 97% TESTING ACCURACY ON THE MNIST IMAGE CLASSIFICATION PROBLEM. FINAL COLUMN REPRESENTS HIGHEST OVERALL TESTING ACCURACY. BOLD ITEMS INDICATE THE BEST VALUE.

each method runs for 500 epochs. We only use the ADAPD algorithm with 1 stochastic gradient step for these experiments, but we tune the dual step-size in (25) and (26). In Figure 4, we report the same metrics as in the MNIST experiment.

Similar to the MNIST image classification problem, we report the wall-clock time taken to reach and stay above 88% testing accuracy in Table II. In terms of stationarity, all methods besides D-GET struggle. However, ADAPD performs better than the competitors in terms of testing accuracy (see Figure 4 and Table II). Similar to the MNIST results, this suggests that ADAPD is able to find a solution to the image classification problem that generalizes better than the competitors. Additionally, ADAPD greatly saves on the number of data samples and communications necessary to achieve a high testing accuracy.

## V. Conclusion

In this work, we presented ADAPD: **A** **D**ecentr**A**lized **P**rimal-**D**ual framework for solving non-convex and smooth consensus optimization problems over a network of agents. Two variants to ADAPD are presented, the ADAPD-OG (**O**ne **G**radient) and the ADAPD-MC (**M**ultiple **C**ommunications). We demonstrated that ADAPD and ADAPD-OG achieves $O\left(L(1-\rho)^{-2}\varepsilon^{-1}\right)$ communication complexity to find an $\varepsilon$-stationary point and showed this can be reduced to $O\left(L(1-\rho)^{-0.5}\varepsilon^{-1}\right)$ when the MC variant is used; this is optimal for the class of smooth, non-convex, decentralized consensus problems considered in this work. Finally, we presented four numerical experiments that validate our claim that ADAPD outperforms other state-of-the-art decentralized methods. Future research topics would be extending the theoretical guarantees of ADAPD-OG to the stochastic case and demonstrating convergence in a time-varying/asynchronous setting of ADAPD and its variants.

## Appendix A
## Supporting Lemmas and proofs for ADAPD

*Proof* [of Proposition 1] First, it is obvious that $\mathcal{L}_\eta(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k) \leq \Phi^k$ for all $k \geq 0$ by the definition of $\Phi^k$ in (43). Second, notice

$$\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^{k+1})$$

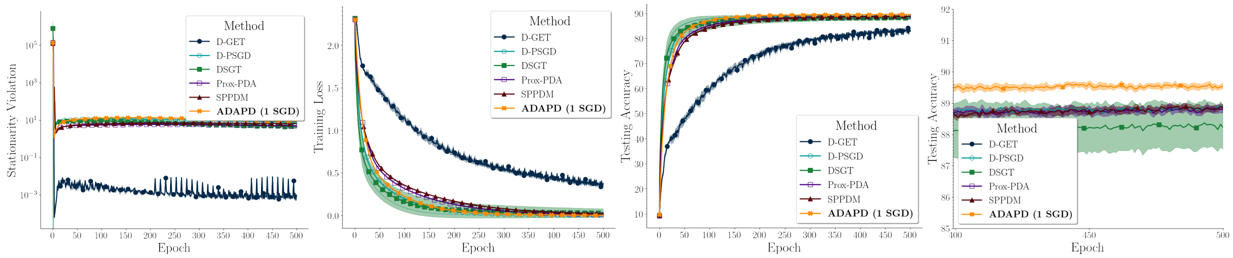

Fig. 4. In order from left to right: stationarity violation, training loss, testing accuracy, and a zoomed version of the testing accuracy over the last hundred epochs for the CIFAR-10 image classification problem.

| Method | To reach 88% testing accuracy | | | Highest accuracy (%) |
|---|---|---|---|---|
| | Time (s) | Samples | Communications | |
| D-GET | ✗ | 3,125,006 | 35,530 | 84.16 |
| D-PSGD | **651.88** | 1,011,200 | 7,900 | 88.92 |
| DSGT | 1,900.23 | 2,348,800 | 36,700 | 88.37 |
| Prox-PDA | 1,025.57 | 1,523,200 | 11,900 | 88.88 |
| SPPDM | 1,395.84 | 1,708,800 | 13,350 | 88.91 |
| ADAPD (1 SGD) | 870.11 | **806,400** | **6,300** | **89.62** |

TABLE II

TIME TO REACH 88% TESTING ACCURACY ON THE CIFAR-10 IMAGE CLASSIFICATION PROBLEM. FINAL COLUMN REPRESENTS HIGHEST OVERALL TESTING ACCURACY. BOLD ITEMS INDICATE THE BEST VALUE.

$$
\begin{aligned}
&= F(\mathbf{X}^{k+1}) + \left\langle \mathbf{Y}^{k+1}, \mathbf{X}^{k+1} - \mathbf{X}_0^{k+1} \right\rangle + \tfrac{1}{2\eta} \left\| \mathbf{X}^{k+1} - \mathbf{X}_0^{k+1} \right\|_F^2 \\
&\quad + \left\langle \mathbf{Z}^{k+1}, \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1} \right\rangle + \tfrac{1}{2\eta} \left\| \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1} \right\|_F^2 \\
&\overset{(25),(26)}{=} F(\mathbf{X}^{k+1}) + \left\langle \mathbf{Y}^{k+1}, \eta(\mathbf{Y}^{k+1} - \mathbf{Y}^k) \right\rangle + \tfrac{1}{2\eta} \left\| \mathbf{X}^{k+1} - \mathbf{X}_0^{k+1} \right\|_F^2 \\
&\quad + \left\langle \mathbf{Z}^{k+1}, \eta(\mathbf{Z}^{k+1} - \mathbf{Z}^k) \right\rangle + \tfrac{1}{2\eta} \left\| \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1} \right\|_F^2 \\
&= F(\mathbf{X}^{k+1}) + \tfrac{\eta}{2} \left( \left\|\mathbf{Y}^{k+1}\right\|_F^2 + \left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\right\|_F^2 - \left\|\mathbf{Y}^k\right\|_F^2 \right) \\
&\quad + \tfrac{1}{2\eta} \left\| \mathbf{X}^{k+1} - \mathbf{X}_0^{k+1} \right\|_F^2 + \tfrac{1}{2\eta} \left\| \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1} \right\|_F^2 \\
&\quad + \tfrac{\eta}{2} \left( \left\|\mathbf{Z}^{k+1}\right\|_F^2 + \left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2 - \left\|\mathbf{Z}^k\right\|_F^2 \right)
\end{aligned}
$$

Thus, by the definition of $\underline{f}$ in (11), we have that for any integer number $K \geq 1$,

$$
\begin{aligned}
&\sum_{k=0}^{K-1} \left( \Phi^{k+1} - \underline{f} \right) \\
&\geq \sum_{k=0}^{K-1} \left( \mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^{k+1}) - \underline{f} \right) \\
&= \sum_{k=0}^{K-1} \left( F(\mathbf{X}^{k+1}) - \underline{f} + \tfrac{\eta}{2} \left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\right\|_F^2 + \tfrac{1}{2\eta} \left\| \mathbf{X}^{k+1} - \mathbf{X}_0^{k+1} \right\|_F^2 \right) \\
&+ \sum_{k=0}^{K-1} \left( \tfrac{\eta}{2} \left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2 + \tfrac{1}{2\eta} \left\| \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1} \right\|_F^2 \right) \\
&\quad + \tfrac{\eta}{2} \left\|\mathbf{Y}^K\right\|_F^2 - \tfrac{\eta}{2} \left\|\mathbf{Y}^0\right\|_F^2 + \tfrac{\eta}{2} \left\|\mathbf{Z}^K\right\|_F^2 - \tfrac{\eta}{2} \left\|\mathbf{Z}^0\right\|_F^2 \\
&\geq -\tfrac{\eta}{2} \left\|\mathbf{Y}^0\right\|_F^2 - \tfrac{\eta}{2} \left\|\mathbf{Z}^0\right\|_F^2 \triangleq -M. \qquad (56)
\end{aligned}
$$

Thirdly, by (42) and the definition of $\Phi^k$ in (43), we have

$$
\begin{aligned}
&\Phi^{k+1} + \tfrac{(1-\rho)-(C+2)L(1-\rho)\eta-(8L^2(1-\rho)+16L^2)\eta^2}{2(1-\rho)\eta} \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 \\
&+ \left( \tfrac{1}{2\eta} - CL \right) \left\| \mathbf{X}_0^{k+1} - \mathbf{X}_0^k \right\|_F^2 \leq \Phi^k + \tfrac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)} \epsilon_k. \qquad (57)
\end{aligned}
$$

Hence, it holds from the choice of $C$ and $\eta$ that

$$
\Phi^{k+1} \leq \Phi^k + \tfrac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)} \epsilon_k. \qquad (58)
$$

Now assume that there exists $k_0 \geq 0$ such that $\Phi^{k_0} - \underline{f} < -\frac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)} \sum_{k=0}^{\infty} \epsilon_k - 1$. Then summing up (58) gives $\Phi^k - \underline{f} \leq \Phi^{k_0} - \underline{f} + \frac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)} \sum_{k=k_0}^{\infty} \epsilon_k < -1$ for all $k \geq k_0$. Hence, $\sum_{k=k_0+1}^{\infty} \left( \Phi^k - \underline{f} \right) = -\infty$, which contradicts (56). Therefore, we conclude that $\Phi^k - \underline{f} \geq -\frac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)} \sum_{k=0}^{\infty} \epsilon_k - 1$, for all $k \geq 0$ and complete the proof. □

*Proof* [of Theorem 1] Summing up (57) from $k = 0$ to $K - 1$ and dividing by $K$ results in

$$
\begin{aligned}
&\left( \tfrac{(1-\rho)-(C+2)L(1-\rho)\eta-(8L^2(1-\rho)+16L^2)\eta^2}{2(1-\rho)\eta} \right) \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 \\
&+ \tfrac{1-2CL\eta}{2\eta} \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \left\| \mathbf{X}_0^{k+1} - \mathbf{X}_0^k \right\|_F^2 \\
&\leq \tfrac{\Phi^0 - \Phi^K}{K} + \tfrac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)} \cdot \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \epsilon_k \\
&\overset{(45)}{\leq} \tfrac{\Phi^0 - \underline{\phi}}{K} + \tfrac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)} \cdot \tfrac{1}{K} \textstyle\sum_{k=0}^{K-1} \epsilon_k. \qquad (59)
\end{aligned}
$$

By the choice of $C$ and $\eta$, it holds that $\frac{1-2CL\eta}{2\eta} \leq \frac{(1-\rho)-(C+2)L(1-\rho)\eta-(8L^2(1-\rho)+16L^2)\eta^2}{2(1-\rho)\eta}$ so $C_1$ as defined in (47) is positive, and thus the inequality in (59) implies the desired result. □

*Proof* [of Theorem 2] First, we have for all $k \geq 0$ that

$$
\begin{aligned}
&\left\| \mathbf{X}^{k+1} - \bar{\mathbf{X}}^{k+1} \right\|_F^2 \\
&= \left\| \mathbf{X}^{k+1} - \mathbf{W}\mathbf{X}^{k+1} + \mathbf{W}\mathbf{X}^{k+1} - \bar{\mathbf{X}}^{k+1} \right\|_F^2 \\
&= \left\| \mathbf{X}^{k+1} - \mathbf{W}\mathbf{X}^{k+1} \right\|_F^2 + \left\| \left( \mathbf{W} - \tfrac{1}{N}\mathbf{e}\mathbf{e}^\top \right) (\mathbf{X}^{k+1} - \bar{\mathbf{X}}^{k+1}) \right\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
&+2\left\langle \mathbf{X}^{k+1}-\mathbf{W}\mathbf{X}^{k+1}, \left(\mathbf{W}-\tfrac{1}{N}\mathbf{e}\mathbf{e}^\top\right)(\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1})\right\rangle \\
\overset{(34)}{\le}& \left\|\mathbf{X}^{k+1}-\mathbf{W}\mathbf{X}^{k+1}\right\|_F^2 + \left\|\left(\mathbf{W}-\tfrac{1}{N}\mathbf{e}\mathbf{e}^\top\right)(\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1})\right\|_F^2 \\
&+\tfrac{1}{\delta}\left\|\mathbf{X}^{k+1}-\mathbf{W}\mathbf{X}^{k+1}\right\|_F^2 + \delta\left\|\left(\mathbf{W}-\tfrac{1}{N}\mathbf{e}\mathbf{e}^\top\right)(\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1})\right\|_F^2 \\
\le\,& \rho^2(1+\delta)\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2 + \left(1+\tfrac{1}{\delta}\right)\left\|\mathbf{X}^{k+1}-\mathbf{W}\mathbf{X}^{k+1}\right\|_F^2,
\end{aligned}
$$

where $\delta > 0$, and we have used $\rho = \left\|\mathbf{W}-\frac{1}{N}\mathbf{e}\mathbf{e}^\top\right\|_2$ from (6). Choosing $\delta = \frac{1-\rho}{\rho} > 0$ and simplifying the result, we obtain

$$
\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2 \le \tfrac{1}{(1-\rho)^2}\left\|\mathbf{X}^{k+1}-\mathbf{W}\mathbf{X}^{k+1}\right\|_F^2. \tag{60}
$$

Now, by (26) and the proof of Lemma 4, we have

$$
\begin{aligned}
&\tfrac{1}{\eta}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}\right\|_F^2 = \eta\left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{Z}^{k+1}-\mathbf{Z}^k)\right\|_F^2 \\
&\le 8L^2\eta\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2 + \tfrac{10}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2 + 16\eta\epsilon_k
\end{aligned} \tag{61}
$$

and by (25),

$$
\begin{aligned}
&\left\|\mathbf{X}^{k+1}-\mathbf{X}_0^{k+1}\right\|_F^2 = \eta^2\left\|\mathbf{Y}^{k+1}-\mathbf{Y}^k\right\|_F^2 \\
&\overset{(39)}{\le} 4L^2\eta^2\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2 + 4\left\|\mathbf{V}_0^k\right\|_F^2 + 8\eta^2\epsilon_k.
\end{aligned} \tag{62}
$$

Thus,

$$
\begin{aligned}
&\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2 \\
\overset{(60)}{\le}& \tfrac{1}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}^{k+1}\right\|_F^2 \\
\le\,& \tfrac{2}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\left\|(\mathbf{I}-\mathbf{W})(\mathbf{X}^{k+1}-\mathbf{X}_0^{k+1})\right\|_F^2 \\
&+\tfrac{2}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}\right\|_F^2 \\
\le\,& \tfrac{2}{(1-\rho)^2K}\left(4\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\mathbf{X}_0^{k+1}\right\|_F^2 + \textstyle\sum_{k=0}^{K-1}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}\right\|_F^2\right) \\
\overset{(61),(62)}{\le}& \tfrac{48L^2\eta^2}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2 + \tfrac{52}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{V}_0^k\right\|_F^2 \\
&+\tfrac{96\eta^2}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\epsilon_k \\
\overset{(35)}{\le}& \tfrac{48L^2\eta^2}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2 \\
&+\tfrac{208}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2 + \tfrac{96\eta^2}{(1-\rho)^2K}\textstyle\sum_{k=0}^{K-1}\epsilon_k \\
\overset{(46)}{\le}& \tfrac{1}{C_1K}\left(C_2\Delta_\Phi + \tfrac{96\eta^2+C_2C_3(1-\rho)^2}{(1-\rho)^2}\textstyle\sum_{k=0}^{K-1}\epsilon_k\right),
\end{aligned} \tag{63}
$$

where we have used the fact that $\|\mathbf{I}-\mathbf{W}\|_2 \le 2$ and the choice of $\eta$ in (50) to have $\max\left\{\frac{48L^2\eta^2}{(1-\rho)^2}, \frac{208}{(1-\rho)^2}\right\} = \frac{208}{(1-\rho)^2} \triangleq C_2$ and defined $C_3 \triangleq \frac{(32L+16L(1-\rho))\eta+(4C+1)(1-\rho)}{2L(1-\rho)}$. Furthermore, we use (36) and (37) to have

$$
\nabla F(\mathbf{X}^{k+1}) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+1} = \mathbf{R}^{k+1} - \tfrac{1}{\eta}(\mathbf{I}+\mathbf{W})[\mathbf{X}_0^{k+1}-\mathbf{X}_0^k]. \tag{64}
$$

Now, by Assumption 1(iii), we have $\mathbf{e}^\top\sqrt{\mathbf{I}-\mathbf{W}} = \mathbf{0}$. Hence

$$
\begin{aligned}
&\tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_F^2 \\
&= \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left\|\tfrac{1}{N}\mathbf{e}\mathbf{e}^\top\left(\nabla F(\bar{\mathbf{X}}^{k+1}) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+1}\right)\right\|_F^2 \\
&\le \left\|\tfrac{1}{N}\mathbf{e}\mathbf{e}^\top\right\|_2^2 \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left\|F(\bar{\mathbf{X}}^{k+1}) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+1}\right\|_F^2 \\
&\le \tfrac{2}{K}\textstyle\sum_{k=0}^{K-1}\left\|F(\mathbf{X}^{k+1}) + \sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+1}\right\|_F^2 \\
&+ \tfrac{2}{K}\textstyle\sum_{k=0}^{K-1}\left\|\nabla F(\bar{\mathbf{X}}^{k+1}) - \nabla F(\mathbf{X}^{k+1})\right\|_F^2 \\
&\overset{(64),(10)}{\le} \tfrac{2}{K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{R}^{k+1} - \tfrac{1}{\eta}(\mathbf{I}+\mathbf{W})[\mathbf{X}_0^{k+1}-\mathbf{X}_0^k]\right\|_F^2 \\
&+ \tfrac{2L^2}{K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2 \\
&\overset{(35),(28)}{\le} \tfrac{16}{K\eta^2}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2 + \tfrac{4}{K}\textstyle\sum_{k=0}^{K-1}\epsilon_k \\
&+ \tfrac{2L^2}{K}\textstyle\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2 \\
&\overset{(46),(63)}{\le} \tfrac{(2C_2L^2+C_4)\Delta_\Phi}{KC_1} \\
&+ \tfrac{192L^2\eta^2+(2C_2C_3L^2+C_3C_4+4C_1)(1-\rho)^2}{KC_1(1-\rho)^2}\textstyle\sum_{k=0}^{K-1}\epsilon_k
\end{aligned} \tag{65}
$$

where we have used $\|\mathbf{I}+\mathbf{W}\|_2 \le 2$ in the fourth inequality and defined $C_4 \triangleq \frac{16}{\eta^2}$. Finally, we have that

$$
\begin{aligned}
&\min_{1\le k'\le K}\left(\left\|\nabla f(\bar{\mathbf{x}}^{k'})\right\|_2^2 + \left\|\mathbf{X}^{k'}-\bar{\mathbf{X}}^{k'}\right\|_F^2\right) \\
&\le \tfrac{1}{K}\textstyle\sum_{k=0}^{K-1}\left(\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_2^2 + \left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2\right) \\
&\overset{(63),(65)}{\le} \tfrac{((2L^2+1)C_2+C_4)\Delta_\Phi}{KC_1} + \tfrac{(192L^2+96)\eta^2}{KC_1(1-\rho)^2}\textstyle\sum_{k=0}^{K-1}\epsilon_k \\
&+ \tfrac{((2L^2+1)C_2C_3+C_3C_4+4C_1)}{KC_1}\textstyle\sum_{k=0}^{K-1}\epsilon_k.
\end{aligned} \tag{66}
$$

We complete the proof. □

# References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (A. Singh and J. Zhu, eds.), vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, PMLR, 20–22 Apr 2017. 1, 4

[2] X. Lian, C. Zhang, H. Zhang, C.-J. 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* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, pp. 5330–5340, Curran Associates, Inc., 2017. 1, 3, 8, 9

[3] X. Liang, A. M. Javid, M. Skoglund, and S. Chatterjee, "Asynchrounous decentralized learning of a neural network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3947–3951, 2020. 1

[4] M. Hong, D. Hajinezhad, and M.-M. Zhao, "Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, (International Convention Centre, Sydney, Australia), pp. 1529–1538, PMLR, 06–11 Aug 2017. 1, 2, 3, 5, 8

[5] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, 2012. 1, 8

[6] 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, pp. 944 – 966, 2015. 1, 2, 3

[7] P. D. 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. 1, 3

[8] P. Abichandani, H. Y. Benson, and M. Kam, "Decentralized multi-vehicle path coordination under communication constraints," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2306–2313, 2011. 1

[9] K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié, "Optimal algorithms for smooth and strongly convex distributed optimization in networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, (International Convention Centre, Sydney, Australia), pp. 3027–3036, PMLR, 06–11 Aug 2017. 1, 2, 5, 7

[10] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, pp. 1835 – 1854, 2016. 1, 2

[11] H. Sun, S. Lu, and M. Hong, "Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, (Virtual), pp. 9217–9228, PMLR, 13–18 Jul 2020. 1, 2, 3, 8, 9

[12] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Systems and Control Letters*, vol. 53, no. 1, pp. 65–78, 2004. 1, 2, 9

[13] E. Wei and A. Ozdaglar, "On the o(1/k) convergence of asynchronous distributed alternating direction method of multipliers," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 551–554, 2013. 2, 3

[14] K. Wang, I. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: when experts are not enough," *Quantitative Science Studies*, vol. 1, pp. 396–413, February 2020. 2

[15] Y. Arjevani, J. Bruna, B. Can, M. Gurbuzbalaban, S. Jegelka, and H. Lin, "Ideal: Inexact decentralized accelerated augmented lagrangian method," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 20648–20659, Curran Associates, Inc., 2020. 2, 3

[16] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014. 2

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," vol. 3, no. 1, 2011. 2, 4

[18] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019. 2, 6

[19] H. Sun and M. Hong, "Distributed non-convex first-order optimization and information processing: Lower complexity bounds and rate optimal algorithms," *IEEE Transactions on Signal Processing*, vol. 67, no. 22, pp. 5912–5928, 2019. 2, 3

[20] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. USA: Prentice-Hall, Inc., 1989. 2

[21] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2020. 2

[22] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, pp. 48 – 61, 2009. 2

[23] A. S. Berahas, R. Bollapragada, N. S. Keskar, and E. Wei, "Balancing communication and computation in distributed optimization," *IEEE Transactions on Automatic Control*, vol. 64, no. 8, pp. 3141–3155, 2019. 2

[24] G. Qu and N. Li, "Accelerated distributed nesterov gradient descent," *IEEE Transactions on Automatic Control*, 2019. 2

[25] C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić, "A dual approach for optimal algorithms in distributed optimization over networks," *Optimization Methods and Software*, vol. 36, no. 1, pp. 171–210, 2021. 2

[26] F. Mansoori and E. Wei, "Flexpd: A flexible framework of first-order primal-dual algorithms for distributed optimization," *IEEE Transactions on Signal Processing*, vol. 69, pp. 3500–3512, 2021. 3

[27] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on Signal Processing*, vol. 66, pp. 2834 – 2848, 2018. 3, 5, 8

[28] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016. 3

[29] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, no. 1, pp. 497–544, 2019. 3

[30] W. Auzinger and J. M. Melenk, "Iterative solution of large linear systems," *TU Wien, Lecture Notes*, 2011. 3, 5

[31] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "$d^2$: Decentralized training over decentralized data," in *Proceedings of the 35th International Conference on Machine Learning* (J. Dy and A. Krause, eds.), vol. 80 of *Proceedings of Machine Learning Research*, (Stockholmsmässan, Stockholm Sweden), pp. 4848–4856, PMLR, 10–15 Jul 2018. 3

[32] J. Zhang and K. You, "Decentralized stochastic gradient tracking for non-convex empirical risk minimization," *arXiv preprint arXiv:1909.02712*, 2020. 3, 8, 9

[33] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "SARAH: A novel method for machine learning problems using stochastic recursive gradient," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, (International Convention Centre, Sydney, Australia), pp. 2613–2621, PMLR, 06–11 Aug 2017. 3

[34] A. Koloskova, T. Lin, and S. U. Stich, "An improved analysis of gradient tracking for decentralized machine learning," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021. 3

[35] R. Xin, U. A. Khan, and S. Kar, "Fast decentralized nonconvex finite-sum optimization with recursive variance reduction," *SIAM Journal on Optimization*, vol. 32, no. 1, pp. 1–28, 2022. 3

[36] X. Yi, S. Zhang, T. Yang, T. Chai, and K. H. Johansson, "A primal-dual sgd algorithm for distributed nonconvex optimization," *arXiv preprint arXiv:2006.03474*, 2020. 3

[37] Z. Wang, J. Zhang, T.-H. Chang, J. Li, and Z.-Q. Luo, "Distributed stochastic consensus optimization with momentum for nonconvex nonsmooth problems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4486–4501, 2021. 3, 8

[38] X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, pp. 3043–3052, 2018. 3

[39] T. Wu, K. Yuan, Q. Ling, W. Yin, and A. Sayed, "Decentralized consensus optimization with asynchrony and delays," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 4, pp. 293 – 307, 2017. 3

[40] J. Zhang and K. You, "Fully asynchronous distributed optimization with linear convergence in directed networks," *arXiv preprint arXiv:1901.08215*, 2021, 1901.08215. 3

[41] M. Hong, "A distributed, asynchronous, and incremental algorithm for nonconvex optimization: An admm approach," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 935–945, 2018. 3

[42] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push-pull gradient methods for distributed optimization in networks," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 1–16, 2021. 3

[43] A. Nedic, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, pp. 2597 – 2633, 2017. 3

[44] J. Eckstein and W. Yao, "Relative-error approximate versions of douglas–rachford splitting and special cases of the admm," *Mathematical Programming*, vol. 170, no. 2, pp. 417–444, 2018. 4

[45] S. Kumar, K. Rajawat, and D. P. Palomar, "Distributed inexact successive convex approximation admm: Analysis-part i," 2019, 1907.08969. 4

[46] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "Fedpd: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021. 4, 8

[47] J. Liu and A. S. Morse, "Accelerated linear iterations for distributed averaging," *Annual Reviews in Control*, vol. 35, no. 2, pp. 160–165, 2011. 5

[48] H. Ye, Z. Zhou, L. Luo, and T. Zhang, "Decentralized accelerated proximal gradient descent," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 18308–18317, Curran Associates, Inc., 2020. 5

[49] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 693–696, 2009. 8

[50] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 8

[51] D. Dua and C. Graff, "Uci machine learning repository," 2017. 8

[52] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. 9

[53] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2015, 1412.6806. 9

[54] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., 2009. 9

# Supplementary Material

## APPENDIX A
## ON THE EQUIVALENCE BETWEEN PROX-PDA [1] AND DISTRIBUTED ADMM [2]

Here, we show that the distributed ADMM algorithm [2], which uses edge-based constraints to enforce consensus (see (3) in [2]), can reduce to the Prox-PDA method in [1]. Under the assumption of $\mathbf{A}^\top\mathbf{A} = \mathbf{L}^-$ (the signed Laplacian of $\mathcal{G}$), Prox-PDA performs global updates:

$$\begin{aligned}\mathbf{X}^{(k+1)} &= \operatorname*{argmin}_{\mathbf{X}} \left\{ F(\mathbf{X}) - \left\langle \boldsymbol{\mu}^{(k)}, \mathbf{A}\mathbf{X} \right\rangle + \frac{\beta}{2} H(\mathbf{X}, \mathbf{X}^k; \mathbf{A}, \mathbf{B}) \right\}, \\ \boldsymbol{\mu}^{(k+1)} &= \boldsymbol{\mu}^{(k)} + \beta \mathbf{A}\mathbf{X}^{k+1},\end{aligned} \tag{A.1}$$

where $H(\mathbf{X}, \mathbf{X}^k; \mathbf{A}, \mathbf{B}) = \|\mathbf{A}\mathbf{X}\|_F^2 + \left\|\mathbf{X} - \mathbf{X}^k\right\|_{\mathbf{B}^\top\mathbf{B}}^2$. Choosing $\mathbf{B}^\top\mathbf{B} = \mathbf{L}^+$ (the unsigned Laplacian of $\mathcal{G}$), we have

$$\begin{aligned} &H(\mathbf{X}, \mathbf{X}^k; \mathbf{A}, \mathbf{B}) \\ =& \frac{\beta}{2} \sum_{i=1}^{N} \left( 2\left|\mathcal{N}_i\right| \|\mathbf{x}_i\|_2^2 - \left\langle \mathbf{x}_i, \sum_{j\in\mathcal{N}_i} \mathbf{x}_j \right\rangle \right) + \frac{\beta}{2} \left\langle \mathbf{x}^k, \mathbf{B}^\top\mathbf{B}\mathbf{X}^k \right\rangle \\ &+ \frac{\beta}{2} \sum_{i=1}^{N} \left( \left\langle \mathbf{x}_i, \sum_{j\in\mathcal{N}_i} \mathbf{x}_j \right\rangle - 2\left|\mathcal{N}_i\right| \left\langle \mathbf{x}_i, \mathbf{x}_i^k \right\rangle - 2 \left\langle \mathbf{x}_i, \sum_{j\in\mathcal{N}_i} \mathbf{x}_j^k \right\rangle \right). \end{aligned} \tag{A.2}$$

Multiplying both sides of the $\boldsymbol{\mu}$ update in (A.1) by $\mathbf{A}^\top$, letting $\boldsymbol{\alpha}^k \triangleq \mathbf{A}^\top\boldsymbol{\mu}^k \,\forall\, k$, and dropping the $\frac{\beta}{2}\left\langle \mathbf{x}^k, \mathbf{B}^\top\mathbf{B}\mathbf{X}^k \right\rangle$ term from (A.2) (as the argmin is about $\mathbf{X}$ in (A.1)) results in (10) from [2]. Hence, the two algorithms are equivalent. As a result, the distributed ADMM updates from [2] converge in the non-convex case by the convergence of Prox-PDA [1].

## APPENDIX B
## SUPPORTING LEMMAS AND PROOFS FOR ADAPD (CON'T.)

The proofs of Lemma 2 and the following three Lemmas can be found in the longer version of this work [3].

*Lemma B.1:* (Lemma 6 in [4]) If (28) is satisfied and $\eta < \frac{1}{2L}$, then

$$\begin{aligned} &\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k) - \mathcal{L}_\eta(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k) \\ &\le \tfrac{2L\eta - 1}{2\eta} \left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 + \tfrac{\epsilon_{k+1}}{2L}, \forall\, k \ge 0. \end{aligned} \tag{B.1}$$

*Lemma B.2:* The partial gradient $\nabla_{\mathbf{X}_0}\mathcal{L}_\eta(\mathbf{X}, \mathbf{X}_0; \mathbf{Y}, \mathbf{Z})$ is $\frac{3}{\eta}$-Lipschitz continuous about $\mathbf{X}_0$ for any $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. Further, for all $k \ge 0$, we have

$$\begin{aligned} &\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^k, \mathbf{Z}^k) - \mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k) \\ &\le -\tfrac{1}{2\eta} \left\|\mathbf{X}_0^{k+1} - \mathbf{X}_0^k\right\|_F^2. \end{aligned} \tag{B.2}$$

*Lemma B.3:* For all $k \ge 0$, the followings hold:

$$\begin{aligned} &\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^k) - \mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^k, \mathbf{Z}^k) \\ &= \eta \left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\right\|_F^2, \end{aligned} \tag{B.3}$$

$$\begin{aligned} &\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^{k+1}) - \mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^k) \\ &= \eta \left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2. \end{aligned} \tag{B.4}$$

*Proof* [Of Lemma 3] The inequality follows from rewriting $\mathcal{L}_\eta(\mathbf{X}^{k+1}, \mathbf{X}_0^{k+1}; \mathbf{Y}^{k+1}, \mathbf{Z}^{k+1}) - \mathcal{L}_\eta(\mathbf{X}^k, \mathbf{X}_0^k; \mathbf{Y}^k, \mathbf{Z}^k)$ as the summation of the left-hand sides of (B.1), (B.2), (B.3), and (B.4) and using those four inequalities. □

*Proof* [Of Lemma 4] To prove (39), by (37), we have $\eta\left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\right\|_F^2 = \eta\left\|\mathbf{R}^{k+1} - \mathbf{R}^k - \nabla F(\mathbf{X}^{k+1}) + \nabla F(\mathbf{X}^k) - \frac{1}{\eta}\mathbf{V}_0^k\right\|_F^2 \overset{(35)}{\le} 4\eta\left(\left\|\mathbf{R}^{k+1}\right\|_F^2 + \left\|\mathbf{R}^k\right\|_F^2 + \left\|\nabla F(\mathbf{X}^{k+1}) - \nabla F(\mathbf{X}^k)\right\|_F^2 + \frac{1}{\eta^2}\left\|\mathbf{V}_0^k\right\|_F^2\right) \overset{(28),(10)}{\le} 4L^2\eta\left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 + \frac{4}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2 + 8\eta\epsilon_k$ where in the last inequality we have further used $\epsilon_{k+1} \le \epsilon_k$ for all $k \ge 0$.
To prove (40), notice that if $\mathbf{Z}^0 \in \operatorname{range}\left(\sqrt{\mathbf{I} - \mathbf{W}}\right)$, then by (19), $\mathbf{Z}^k \in \operatorname{range}\left(\sqrt{\mathbf{I} - \mathbf{W}}\right)$ for all $k \ge 0$. Thus with $\rho_2 \triangleq 1 - \lambda_2(\mathbf{W})$, we have

$$\eta\rho_2\left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2 \le \eta\left\|\sqrt{\mathbf{I} - \mathbf{W}}(\mathbf{Z}^{k+1} - \mathbf{Z}^k)\right\|_F^2, \tag{B.5}$$

and since $1 - \rho \le \rho_2$, it further holds that,

$$\eta(1 - \rho)\left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2 \le \eta\rho_2\left\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\right\|_F^2 \tag{B.6}$$

In addition,

$$\eta\left\|\sqrt{\mathbf{I} - \mathbf{W}}(\mathbf{Z}^{k+1} - \mathbf{Z}^k)\right\|_F^2 \tag{B.7}$$

$$\overset{(36)}{=} \eta\left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k - \tfrac{1}{\eta}\mathbf{W}\mathbf{V}_0^k\right\|_F^2$$

$$\overset{(34)}{\le} 2\eta\left\|\mathbf{Y}^{k+1} - \mathbf{Y}^k\right\|_F^2 + \tfrac{2}{\eta}\left\|\mathbf{W}\mathbf{V}_0^k\right\|_F^2 \tag{B.8}$$

$$\overset{(39)}{\le} 8L^2\eta\left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 + \tfrac{8}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2 + 16\eta\epsilon_k + \tfrac{2}{\eta}\left\|\mathbf{W}\mathbf{V}_0^k\right\|_F^2$$

$$\le 8L^2\eta\left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|_F^2 + \tfrac{10}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2 + 16\eta\epsilon_k \tag{B.9}$$

where the last inequality uses Assumption 1(iv). Using (B.5) with (B.6), we complete the proof. □

*Proof* [Of Lemma 5] By (37), we have

$$\begin{aligned} &\left\langle \mathbf{Y}^{k+1} - \mathbf{Y}^k, \mathbf{X}_0^{k+1} - \mathbf{X}_0^k \right\rangle \\ &= \left\langle \mathbf{R}^{k+1} - \mathbf{R}^k - \nabla F(\mathbf{X}^{k+1}) + \nabla F(\mathbf{X}^k) - \tfrac{1}{\eta}\mathbf{V}_0^k, \mathbf{X}_0^{k+1} - \mathbf{X}_0^k \right\rangle. \end{aligned} \tag{B.10}$$

We handle the two sides of (B.10) separately. First, we have

$$\begin{aligned} &\left\langle \mathbf{Y}^{k+1} - \mathbf{Y}^k, \mathbf{X}_0^{k+1} - \mathbf{X}_0^k \right\rangle \\ &\overset{(36)}{=} \left\langle \sqrt{\mathbf{I} - \mathbf{W}}\mathbf{Z}^{k+1} - \sqrt{\mathbf{I} - \mathbf{W}}\mathbf{Z}^k + \tfrac{1}{\eta}\mathbf{W}\mathbf{V}_0^k, \mathbf{X}_0^{k+1} - \mathbf{X}_0^k \right\rangle \end{aligned}$$

$$
\begin{aligned}
&\overset{(26)}{=}\left\langle \tfrac{1}{\eta}(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}+\tfrac{1}{\eta}\mathbf{W}\mathbf{V}_0^k,\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\rangle\\
&=\tfrac{1}{2\eta}\left(\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2+\left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{X}_0^{k+1}-\mathbf{X}_0^k)\right\|_F^2\right)\\
&\quad-\tfrac{1}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^k\right\|_F^2\\
&\quad+\tfrac{1}{2\eta}\left(\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2+\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_{\mathbf{W}}^2-\left\|\mathbf{X}_0^k-\mathbf{X}_0^{k-1}\right\|_{\mathbf{W}}^2\right)
\end{aligned}
$$

where the last equality can be verified straightforwardly. Second, we have

$$
\begin{aligned}
&\left\langle \mathbf{R}^{k+1}-\mathbf{R}^k-\nabla F(\mathbf{X}^{k+1})+\nabla F(\mathbf{X}^k)-\tfrac{1}{\eta}\mathbf{V}_0^k,\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\rangle\\
&\overset{(34),(10)}{\leq}\tfrac{1}{2L}\left\|\mathbf{R}^{k+1}-\mathbf{R}^k\right\|_F^2+\tfrac{L}{2}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2\\
&\quad+\tfrac{L}{2}\left(\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2+\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2\right)\\
&\quad-\tfrac{1}{\eta}\left\langle\mathbf{V}_0^k,\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\rangle\\
&\overset{(35)}{\leq}\tfrac{2}{L}\epsilon_k+\tfrac{L}{2}\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2+L\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2\\
&\quad-\tfrac{1}{2\eta}\left(\left\|\mathbf{V}_0^k\right\|_F^2+\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2-\left\|\mathbf{X}_0^k-\mathbf{X}_0^{k-1}\right\|_F^2\right)
\end{aligned}
$$

where the last inequality comes from $\epsilon_{k+1}\leq\epsilon_k$ for all $k\geq 1$. Combining like terms results in the right hand side of (41); further using the equality established in (B.10) completes the proof. □

*Proof* [Of Lemma 6] By Lemmas 3 and 4 and also using $\epsilon_{k+1}\leq\epsilon_k$, we have

$$
\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^{k+1};\mathbf{Y}^{k+1},\mathbf{Z}^{k+1})-\mathcal{L}_\eta(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)\\
&\leq\tfrac{(8L^2(1-\rho)+16L^2)\eta^2+2L(1-\rho)\eta-(1-\rho)}{2(1-\rho)\eta}\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2\\
&\quad-\tfrac{1}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2+\tfrac{10+4(1-\rho)}{(1-\rho)\eta}\left\|\mathbf{V}_0^k\right\|_F^2+\tfrac{32L\eta+16L(1-\rho)\eta+(1-\rho)}{2L(1-\rho)}\epsilon_k.
\end{aligned}
$$

Now multiplying $C>0$ to both sides of (41) and adding to the above inequality, we have

$$
\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^{k+1};\mathbf{Y}^{k+1},\mathbf{Z}^{k+1})+\tfrac{C}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_{\mathbf{W}}^2+\tfrac{C}{2\eta}\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2\\
&\quad+\tfrac{C}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2+\tfrac{C}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{X}_0^{k+1}-\mathbf{X}_0^k)\right\|_F^2\\
&\leq\mathcal{L}_\eta(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)+\tfrac{C}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^k\right\|_F^2\\
&\quad+\tfrac{C}{2\eta}\left\|\mathbf{X}_0^k-\mathbf{X}_0^{k-1}\right\|_{\mathbf{W}}^2+\tfrac{C}{2\eta}\left\|\mathbf{X}_0^k-\mathbf{X}_0^{k-1}\right\|_F^2\\
&\quad+\left(\tfrac{(8L^2(1-\rho)+16L^2)\eta^2+(C+2)L(1-\rho)\eta-(1-\rho)}{2(1-\rho)\eta}\right)\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2\\
&\quad+\tfrac{2CL\eta-C-1}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2+\tfrac{20+8(1-\rho)-C(1-\rho)}{2(1-\rho)\eta}\left\|\mathbf{V}_0^k\right\|_F^2\\
&\quad+\tfrac{(1-\rho)+(32L+16L(1-\rho))\eta+4C(1-\rho)}{2L(1-\rho)}\epsilon_k.
\end{aligned}
\tag{B.11}
$$

Since the minimum eigenvalue of $\mathbf{I}+\mathbf{W}$ is $\rho_N\triangleq 1+\lambda_N(\mathbf{W})>0$, it holds $\frac{20+8(1-\rho)}{(1-\rho)}\mathbf{I}\preccurlyeq C(\mathbf{I}+\mathbf{W})$ when $C\geq\frac{20+8(1-\rho)}{(1-\rho)\rho_N}$. Hence, we have $0\leq\frac{C(1-\rho)-20-8(1-\rho)}{2(1-\rho)\eta}\left\|\mathbf{V}_0^k\right\|_F^2+\frac{C}{2\eta}\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2$ by noticing

$$
\begin{aligned}
&\tfrac{C(1-\rho)-20-8(1-\rho)}{2(1-\rho)\eta}\left\|\mathbf{V}_0^k\right\|_F^2+\tfrac{C}{2\eta}\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2\\
&=\left\|\mathbf{V}_0^k\right\|^2_{\frac{C(1-\rho)-20-8(1-\rho)}{2(1-\rho)\eta}\mathbf{I}+\frac{C}{2\eta}\mathbf{W}}\geq 0.
\end{aligned}
\tag{B.12}
$$

Noticing that $\frac{1}{(1-\rho)}\geq\frac{1}{\rho_N}$, we have $C\geq\frac{20+8(1-\rho)}{(1-\rho)^2}$ also satisfies the above requirement. In addition, it holds $\frac{C}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{X}_0^{k+1}-\mathbf{X}_0^k)\right\|_F^2+\frac{C}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_{\mathbf{W}}^2=\frac{C}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_{\mathbf{I}-\mathbf{W}}^2+\frac{C}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_{\mathbf{W}}^2=\frac{C}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2$. Furthermore, noting $\frac{C}{2\eta}\left\|\mathbf{X}_0^k-\mathbf{X}_0^{k-1}\right\|_{\mathbf{W}}^2\leq\frac{C}{2\eta}\left\|\mathbf{X}_0^k-\mathbf{X}_0^{k-1}\right\|_F^2$, we obtain the desired result from (B.11). □

## Appendix C
## Supporting Lemmas and proofs for ADAPD-OG

The proof of the following Lemma can be found in the longer version of this work [3] (see Lemma 10).

*Lemma C.1:* Provided that $\eta<\frac{1}{L}$, we have

$$
\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)-\mathcal{L}_\eta(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)\\
&\leq\tfrac{L\eta-1}{2\eta}\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2
\end{aligned}
\tag{C.1}
$$

for all $k\geq 0$.

*Lemma C.2:* Let $\{(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)\}$ be obtained from Alg. 2 or equivalently by updates (31) and (24)-(26). If $\eta<\frac{1}{L}$, then it holds for all $k\geq 0$,

$$
\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^{k+1};\mathbf{Y}^{k+1},\mathbf{Z}^{k+1})-\mathcal{L}_\eta(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)\\
&\leq\tfrac{L\eta-1}{2\eta}\left\|\mathbf{X}^{k+1}-\mathbf{X}^k\right\|_F^2-\tfrac{1}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2\\
&\quad+\eta\left\|\mathbf{Y}^{k+1}-\mathbf{Y}^k\right\|_F^2+\eta\left\|\mathbf{Z}^{k+1}-\mathbf{Z}^k\right\|_F^2.
\end{aligned}
\tag{C.2}
$$

*Proof* The proof follows from using the same technique as in Lemma 3 but with (C.1), (B.2), (B.3), and (B.4). □

*Lemma C.3:* Under the assumptions of Lemma C.2, it holds that for all $k\geq 0$,

$$
\eta\left\|\mathbf{Y}^{k+1}-\mathbf{Y}^k\right\|_F^2\leq 2L^2\eta\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F^2+\tfrac{2}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2,
\tag{C.3}
$$

$$
\eta\left\|\mathbf{Z}^{k+1}-\mathbf{Z}^k\right\|_F^2\leq\tfrac{4L^2\eta}{(1-\rho)}\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F^2+\tfrac{6}{(1-\rho)\eta}\left\|\mathbf{V}_0^k\right\|_F^2,
\tag{C.4}
$$

where $\mathbf{V}_0^k$ is defined in (33).

*Proof* To prove (C.3), we have from (49) that $\eta\left\|\mathbf{Y}^{k+1}-\mathbf{Y}^k\right\|_F^2=\eta\left\|-\nabla F(\mathbf{X}^k)+\nabla F(\mathbf{X}^{k-1})-\frac{1}{\eta}\mathbf{V}_0^k\right\|_F^2\overset{(35),(10)}{\leq}2L^2\eta\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F^2+\frac{2}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2$. To prove (C.4), we start from (B.8) to have

$$
\begin{aligned}
&\eta\left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{Z}^{k+1}-\mathbf{Z}^k)\right\|_F^2\\
&\leq 2\eta\left\|\mathbf{Y}^{k+1}-\mathbf{Y}^k\right\|_F^2+\tfrac{2}{\eta}\left\|\mathbf{W}\mathbf{V}_0^k\right\|_F^2\\
&\overset{(C.3)}{\leq}4L^2\eta\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F^2+\tfrac{4}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2+\tfrac{2}{\eta}\left\|\mathbf{W}\mathbf{V}_0^k\right\|_F^2\\
&\leq 4L^2\eta\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F^2+\tfrac{6}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2,
\end{aligned}
\tag{C.5}
$$

where the last inequality uses Assumption 1(iv). Now we notice that (36) still holds for ADAPD-OG. Hence by choosing $\mathbf{Z}^0\in\operatorname{range}\big(\sqrt{\mathbf{I}-\mathbf{W}}\big)$, we have $\mathbf{Z}^k\in\operatorname{range}\big(\sqrt{\mathbf{I}-\mathbf{W}}\big)$ for all $k\geq 0$ from (19). Thus (B.5) still holds, and it together with (C.5) implies (C.4). □

*Lemma C.4:* For all $k \geq 0$, the following relation holds

$$
\begin{aligned}
&\tfrac{1}{2\eta}\left(\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2 + \left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k})\right\|_F^2 - \left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k}\right\|_F^2\right)\\
&+ \tfrac{1}{2\eta}\left(\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2 + \left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_{\mathbf{W}}^2 - \left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_{\mathbf{W}}^2\right)\\
&\leq \tfrac{L\eta-1}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2 + \tfrac{L}{2}\left\|\mathbf{X}^{k}-\mathbf{X}^{k-1}\right\|_F^2 + \tfrac{1}{2\eta}\left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_F^2\\
&- \tfrac{1}{2\eta}\left\|\mathbf{V}_0^k\right\|_F^2
\end{aligned}
\tag{C.6}
$$

where $\mathbf{V}_0^k$ is defined in (33).

*Proof* The proof of this Lemma uses the same techniques used in the proof of Lemma 5 where (B.10) is replaced by

$$
\begin{aligned}
&\left\langle \mathbf{Y}^{k+1}-\mathbf{Y}^{k}, \mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\rangle\\
&= \left\langle -\nabla F(\mathbf{X}^k) + \nabla F(\mathbf{X}^{k-1}) - \tfrac{1}{\eta}\mathbf{V}_0^k, \mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\rangle .
\end{aligned}
\tag{C.7}
$$

For explicit algebraic manipulations, see Lemma 12 in [3].□

*Lemma C.5:* Let $\{(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)\}$ be obtained from Alg. 2 or equivalently by updates (31), (24)-(26). If $\eta < \frac{1}{L}$, then

$$
\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^{k+1};\mathbf{Y}^{k+1},\mathbf{Z}^{k+1}) + \tfrac{\hat{C}}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2\\
&+ \tfrac{\hat{C}}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2\\
&\leq \mathcal{L}_\eta(\mathbf{X}^{k},\mathbf{X}_0^{k};\mathbf{Y}^{k},\mathbf{Z}^{k}) + \tfrac{\hat{C}}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k}\right\|_F^2 + \tfrac{\hat{C}}{\eta}\left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_F^2\\
&+ \left(\tfrac{L\eta-1}{2\eta}\right)\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2 + \left(\tfrac{\hat{C}L\eta-\hat{C}-1}{2\eta}\right)\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2\\
&+ \tfrac{4L^2(1-\rho)\eta+8L^2\eta+\hat{C}L(1-\rho)}{2(1-\rho)}\left\|\mathbf{X}^{k}-\mathbf{X}^{k-1}\right\|_F^2
\end{aligned}
\tag{C.8}
$$

for all $k \geq 0$, where $\hat{C}$ satisfies $\hat{C} \geq \frac{12+4(1-\rho)}{(1-\rho)^2}$.

*Proof* Using Lemma C.3 and (C.2) in conjunction with multiplying $\hat{C} > 0$ to both sides of (C.6) gives

$$
\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^{k+1};\mathbf{Y}^{k+1},\mathbf{Z}^{k+1}) + \tfrac{\hat{C}}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2\\
&+ \tfrac{\hat{C}}{2\eta}\left(\left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k})\right\|_F^2 - \left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k}\right\|_F^2\right)\\
&+ \tfrac{\hat{C}}{2\eta}\left(\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2 + \left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_{\mathbf{W}}^2 - \left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_{\mathbf{W}}^2\right)\\
&\leq \mathcal{L}_\eta(\mathbf{X}^{k},\mathbf{X}_0^{k};\mathbf{Y}^{k},\mathbf{Z}^{k}) + \tfrac{L\eta-1}{2\eta}\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2 - \tfrac{1}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2\\
&+ \tfrac{4L^2(1-\rho)\eta+8L^2\eta+\hat{C}L(1-\rho)}{2(1-\rho)}\left\|\mathbf{X}^{k}-\mathbf{X}^{k-1}\right\|_F^2 + \tfrac{\hat{C}}{2\eta}\left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_F^2\\
&+ \tfrac{\hat{C}L\eta-\hat{C}}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2 + \tfrac{12+4(1-\rho)-\hat{C}(1-\rho)}{2(1-\rho)\eta}\left\|\mathbf{V}_0^k\right\|_F^2 .
\end{aligned}
$$

Since the minimum eigenvalue of $\mathbf{I}+\mathbf{W}$ is $\rho_N > 0$ in (6), it holds $\frac{12+4(1-\rho)}{(1-\rho)}\mathbf{I} \preccurlyeq \hat{C}(\mathbf{I}+\mathbf{W})$ when $\hat{C} \geq \frac{12+4(1-\rho)}{(1-\rho)\rho_N}$. Furthermore, since $\frac{1}{1-\rho} \geq \frac{1}{\rho_N}$ for $\hat{C} \geq \frac{12+4(1-\rho)}{(1-\rho)^2}$, we have $0 \leq \frac{\hat{C}(1-\rho)-12-4(1-\rho)}{2(1-\rho)\eta}\left\|\mathbf{V}_0^k\right\|_F^2 + \frac{\hat{C}}{2\eta}\left\|\mathbf{V}_0^k\right\|_{\mathbf{W}}^2$ by similar logic as that applied to (B.12). Thus,

$$
\begin{aligned}
&\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^{k+1};\mathbf{Y}^{k+1},\mathbf{Z}^{k+1}) + \tfrac{\hat{C}}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k+1}\right\|_F^2\\
&+ \tfrac{\hat{C}}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}(\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k})\right\|_F^2 + \tfrac{\hat{C}}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_{\mathbf{W}}^2\\
&\leq \mathcal{L}_\eta(\mathbf{X}^{k},\mathbf{X}_0^{k};\mathbf{Y}^{k},\mathbf{Z}^{k}) + \tfrac{\hat{C}}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k}\right\|_F^2 + \tfrac{L\eta-1}{2\eta}\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2\\
&- \tfrac{1}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2 + \tfrac{\hat{C}}{2\eta}\left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_{\mathbf{W}}^2\\
&+ \tfrac{4L^2(1-\rho)\eta+8L^2\eta+\hat{C}L(1-\rho)}{2(1-\rho)}\left\|\mathbf{X}^{k}-\mathbf{X}^{k-1}\right\|_F^2\\
&+ \tfrac{\hat{C}L\eta-\hat{C}}{2\eta}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2 + \tfrac{\hat{C}}{2\eta}\left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_F^2 .
\end{aligned}
$$

The rest of the proof follows from the proof of Lemma 6.□

We now define a new Lyapunov function based on the results from Lemma C.5. Fix $\hat{C} \triangleq \frac{16}{(1-\rho)^2}$ used in (C.8) and define

$$
\begin{aligned}
\hat{\Phi}^k \triangleq{}& \mathcal{L}_\eta(\mathbf{X}^{k},\mathbf{X}_0^{k};\mathbf{Y}^{k},\mathbf{Z}^{k}) + \tfrac{\hat{C}}{2\eta}\left\|\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{X}_0^{k}\right\|_F^2 + \tfrac{\hat{C}}{\eta}\left\|\mathbf{X}_0^{k}-\mathbf{X}_0^{k-1}\right\|_F^2\\
&+ \tfrac{4L^2(1-\rho)\eta+8L^2\eta+\hat{C}L(1-\rho)}{2(1-\rho)}\left\|\mathbf{X}^{k}-\mathbf{X}^{k-1}\right\|_F^2 .
\end{aligned}
\tag{C.9}
$$

Using Lemma C.5, for all $k \geq 0$, we have

$$
\begin{aligned}
&\hat{\Phi}^{k+1} + \left(\tfrac{(1-\rho)-(\hat{C}+1)L(1-\rho)\eta-((1-\rho)+2)4L^2\eta^2}{2(1-\rho)\eta}\right)\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2\\
&+ \left(\tfrac{1-\hat{C}L\eta}{2\eta}\right)\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2 \leq \hat{\Phi}^k
\end{aligned}
\tag{C.10}
$$

which comes directly from adding and subtracting $\frac{4L^2(1-\rho)\eta+8L^2\eta+\hat{C}L(1-\rho)}{2(1-\rho)}\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2$ to the left hand side of (C.8), combining like terms, and using (C.9).

*Proposition C.1:* Under Assumptions 1 and 2, let $\{(\mathbf{X}^k,\mathbf{X}_0^k;\mathbf{Y}^k,\mathbf{Z}^k)\}$ be obtained from Alg. 2 or equivalently by (31) and (24)-(26). Choose $\hat{C}$ and $\eta$ such that

$$
\hat{C} \triangleq \frac{16}{(1-\rho)^2} \text{ and } \eta < \frac{1}{2\hat{C}L}. \tag{C.11}
$$

Then the Lyapunov function (C.9) is uniformly lower bounded. More specifically, for all $k \geq 0$,

$$
\hat{\Phi}^k \geq \underline{f} - 1 > -\infty, \tag{C.12}
$$

where $\underline{f}$ is defined in Assumption 2.

*Proof* First, we have $\mathcal{L}_\eta(\mathbf{X}^{k},\mathbf{X}_0^{k};\mathbf{Y}^{k},\mathbf{Z}^{k}) \leq \hat{\Phi}^k$ for all $k$, by the definition of $\hat{\Phi}^k$ in (C.9). Second, by the definition of $\underline{f}$ in (11), we have for any integer number $K \geq 1$,

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\left(\hat{\Phi}^{k+1} - \underline{f}\right) \geq \sum_{k=0}^{K-1}\left(\mathcal{L}_\eta(\mathbf{X}^{k+1},\mathbf{X}_0^{k+1};\mathbf{Y}^{k+1},\mathbf{Z}^{k+1}) - \underline{f}\right)\\
&\overset{(56)}{\geq} -\tfrac{\eta}{2}\left\|\mathbf{Y}^0\right\|_F^2 - \tfrac{\eta}{2}\left\|\mathbf{Z}^0\right\|_F^2 .
\end{aligned}
\tag{C.13}
$$

Thirdly, by (C.10) and the choice of $\hat{C}$ and $\eta$, it holds that

$$
\hat{\Phi}^{k+1} \leq \hat{\Phi}^k. \tag{C.14}
$$

Now assume that there exists a $k_0 \geq 0$ such that $\hat{\Phi}^{k_0} - \underline{f} < -1$. Then (C.14) gives $\hat{\Phi}^{k} - \underline{f} \leq \hat{\Phi}^{k_0} - \underline{f} < -1$ for all $k \geq k_0$. Hence, $\sum_{k=k_0+1}^{\infty}\left(\hat{\Phi}^k - \underline{f}\right) = -\infty$ which contradicts (C.13). Therefore, we conclude that $\hat{\Phi}^k - \underline{f} \geq -1$ for all $k \geq 0$ and complete the proof. □

*Theorem C.1:* Under the same conditions assumed in Proposition C.1, it holds that

$$
\frac{\hat{C}_1}{K}\sum_{k=0}^{K-1}\left(\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2 + \left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^{k}\right\|_F^2\right) \leq \frac{\Delta_{\hat{\Phi}}}{K} \tag{C.15}
$$

where $\Delta_{\hat{\Phi}} \triangleq \hat{\Phi}^0 - \underline{f} + 1$ and $\hat{C}_1 \triangleq \frac{L}{(1-\rho)^2} \leq \frac{(1-\rho)-(\hat{C}+1)L(1-\rho)\eta-((1-\rho)+2)4L^2\eta^2}{2(1-\rho)\eta}$.

*Proof* Summing up (C.10) from $k = 0$ to $K-1$ and dividing by $K$ and utilizing $\hat{C}_1 > 0$ yields the desired result. □

*Proof* [of Theorem 3] By (26), we have

$$\frac{1}{\eta}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}\right\|_F^2 \overset{(C.4)}{\leq} 4L^2\eta\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F^2+\frac{6}{\eta}\left\|\mathbf{V}_0^k\right\|_F^2 \tag{C.16}$$

and by (25),

$$\left\|\mathbf{X}^{k+1}-\mathbf{X}_0^{k+1}\right\|_F^2 \overset{(C.3)}{\leq} 2L^2\eta^2\left\|\mathbf{X}^k-\mathbf{X}^{k-1}\right\|_F^2+2\left\|\mathbf{V}_0^k\right\|_F^2. \tag{C.17}$$

Thus, by (60) we have $\frac{1}{K}\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2 \leq \frac{1}{(1-\rho)^2K}\sum_{k=0}^{K-1}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}^{k+1}\right\|_F^2$, hence

$$\begin{aligned}
&\frac{1}{(1-\rho)^2K}\sum_{k=0}^{K-1}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}^{k+1}\right\|_F^2\\
&\leq \frac{2}{(1-\rho)^2K}\sum_{k=0}^{K-1}\left(\left\|(\mathbf{I}-\mathbf{W})(\mathbf{X}^{k+1}-\mathbf{X}_0^{k+1})\right\|_F^2+\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}\right\|_F^2\right)\\
&\leq \frac{2}{(1-\rho)^2K}\left(4\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\mathbf{X}_0^{k+1}\right\|_F^2+\sum_{k=0}^{K-1}\left\|(\mathbf{I}-\mathbf{W})\mathbf{X}_0^{k+1}\right\|_F^2\right)\\
&\leq \frac{24L^2\eta^2}{(1-\rho)^2K}\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2+\frac{28}{(1-\rho)^2K}\sum_{k=0}^{K-1}\left\|\mathbf{V}_0^k\right\|_F^2\\
&\overset{(35)}{\leq} \frac{48L^2\eta^2}{(1-\rho)^2K}\sum_{k=0}^{K-1}\left\|\mathbf{X}^{k+1}-\mathbf{X}^{k}\right\|_F^2+\frac{112}{(1-\rho)^2K}\sum_{k=0}^{K-1}\left\|\mathbf{X}_0^{k+1}-\mathbf{X}_0^k\right\|_F^2\\
&\overset{(C.15)}{\leq} \frac{\hat{C}_2\Delta_{\hat{\Phi}}}{\hat{C}_1K}
\end{aligned} \tag{C.18}$$

where we have used the fact that $\|\mathbf{I}-\mathbf{W}\|_2 \leq 2$ and defined $\hat{C}_2 \triangleq \frac{112}{(1-\rho)^2}$. Furthermore, we use (36) and (49) to have

$$\nabla F(\mathbf{X}^k)+\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+1} = -\frac{1}{\eta}(\mathbf{I}+\mathbf{W})[\mathbf{X}_0^{k+1}-\mathbf{X}_0^k]. \tag{C.19}$$

Now, using Assumption 1(ii), we have $\mathbf{e}^\top\sqrt{\mathbf{I}-\mathbf{W}} = \mathbf{0}$. Hence,

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_F^2\\
&= \frac{1}{K}\sum_{k=0}^{K-1}\left\|\frac{1}{N}\mathbf{e}\mathbf{e}^\top\left(\nabla F(\bar{\mathbf{X}}^{k+1})+\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+2}\right)\right\|_F^2\\
&\leq \left\|\frac{1}{N}\mathbf{e}\mathbf{e}^\top\right\|_2^2\frac{1}{K}\sum_{k=0}^{K-1}\left\|F(\bar{\mathbf{X}}^{k+1})+\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+2}\right\|_F^2\\
&\leq \frac{2}{K}\sum_{k=0}^{K-1}\left(\left\|F(\mathbf{X}^{k+1})+\sqrt{\mathbf{I}-\mathbf{W}}\mathbf{Z}^{k+2}\right\|_F^2+\left\|\nabla F(\bar{\mathbf{X}}^{k+1})-\nabla F(\mathbf{X}^{k+1})\right\|_F^2\right)\\
&\overset{(C.19),(10)}{\leq} \frac{2}{K}\sum_{k=0}^{K-1}\left(\left\|-\frac{1}{\eta}(\mathbf{I}+\mathbf{W})[\mathbf{X}_0^{k+2}-\mathbf{X}_0^{k+1}]\right\|_F^2+L^2\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2\right)\\
&\overset{(C.15),(C.18)}{\leq} \frac{\left(2\hat{C}_2L^2+\hat{C}_3\right)\Delta_{\hat{\Phi}}}{\hat{C}_1K},
\end{aligned} \tag{C.20}$$

where we have used $\|\mathbf{I}+\mathbf{W}\|_2 \leq 2$ in last inequality and defined $\hat{C}_3 \triangleq \frac{8}{\eta^2}$. We complete proof by using (C.18) and (C.20), to have $\frac{1}{K}\sum_{k=0}^{k-1}\left(\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_2^2+\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2\right) \leq \frac{\left((2L^2+1)\hat{C}_2+\hat{C}_3\right)\Delta_{\hat{\Phi}}}{\hat{C}_1K}$.
□

## Appendix D
## Supporting proofs for the complexity analysis

*Proof* [of Corollary 1] By Remark 4, since one communication round is performed during each iteration of Alg. 1, the communication complexity in (52) follows from setting (51) less than or equal to $\varepsilon$ and solving for $K$. Remark 3 demonstrates the additional logarithmic dependence on the number of gradient computations. □

*Proof* [of Corollary 2] The proof follows the same logic as the proof for Corollary 1, but there is no logarithmic dependence on the number of gradient computations. □

*Proof* [of Theorem 4] By Lemma 1, the dependence on the spectrum of the graph after $R$ iterations of Alg. 3 becomes $2\left(1-\sqrt{1-\rho}\right)^R$; define this quantity to be $\rho_R \triangleq 2\left(1-\sqrt{1-\rho}\right)^R$ such that (51) becomes

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(\left\|\nabla f(\bar{\mathbf{x}}^{k+1})\right\|_2^2+\left\|\mathbf{X}^{k+1}-\bar{\mathbf{X}}^{k+1}\right\|_F^2\right) = \mathcal{O}\left(\frac{L}{(1-\rho_R)^2K}\right). \tag{D.1}$$

With $R = \lceil\frac{2}{\sqrt{1-\rho}}\rceil$, we find a $u > 0$ such that,

$$\frac{1}{\left(1-2(1-\sqrt{1-\rho})^{\lceil\frac{2}{\sqrt{1-\rho}}\rceil}\right)^2} \leq u.$$

First, we rearranging to have

$$\left(1-\sqrt{1-\rho}\right)^{\lceil\frac{2}{\sqrt{1-\rho}}\rceil} \leq \frac{\sqrt{u}-1}{2\sqrt{u}}.$$

Now, let $x = \sqrt{1-\rho} \in (0,1]$, then $(1-x) \in [0,1)$ and $\frac{2}{x} \leq \lceil\frac{2}{x}\rceil$ so that $(1-x)^{\lceil\frac{2}{x}\rceil} \leq (1-x)^{\frac{2}{x}}$. Next, we maximize this quantity with respect to $x \in (0,1]$. Define $g(x) \triangleq (1-x)^{\frac{2}{x}}$ and compute $\frac{d}{dx}g(x)$ to have

$$\frac{d}{dx}g(x) = -(1-x)^{\frac{2}{x}}\left(\frac{2}{x(1-x)}+\frac{2\ln(1-x)}{x^2}\right) < 0, \forall\, x \in (0,1).$$

Hence, $g(x)$ is decreasing on $(0,1)$. Since $g(0+) = \frac{1}{e^2}$, we have $g(x) < \frac{1}{e^2}$ for $x \in (0,1]$. Now we compute,

$$\frac{1}{e^2} \leq \frac{\sqrt{u}-1}{2\sqrt{u}},$$

which holds for all $u \geq 2$. Thus, it holds that $(1-\rho_R)^{-2} \leq 2$. Hence we have the number of gradient computations is independent of $\rho_R$ and the number of neighbor communications must be multiplied by $R = \mathcal{O}\left(\frac{1}{\sqrt{1-\rho}}\right)$. □

*Proof* [of Theorem 5] The proof follows the same logic as the proof of Theorem 4. □

## References

[1] M. Hong, D. Hajinezhad, and M.-M. Zhao, "Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70 of *Proceedings of Machine Learning Research*, (International Convention Centre, Sydney, Australia), pp. 1529–1538, PMLR, 06–11 Aug 2017. 1

[2] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014. 1

[3] G. Mancino-Ball, Y. Xu, and J. Chen, "A decentralized primal-dual framework for non-convex smooth consensus optimization," *arXiv preprint arXiv:2107.11321*, 2021. 1, 2, 3

[4] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "Fedpd: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021. 1