On Nonconvex Decentralized Gradient Descent

Jinshan Zeng and Wotao Yin

***Abstract*—Consensus optimization has received considerable attention in recent years. A number of decentralized algorithms have been proposed for convex consensus optimization. However, to the behaviors or consensus *nonconvex* optimization, our understanding is more limited. When we lose convexity, we cannot hope that our algorithms always return global solutions though they sometimes still do. Somewhat surprisingly, the decentralized consensus algorithms, DGD and Prox-DGD, retain most other properties that are known in the convex setting. In particular, when diminishing (or constant) step sizes are used, we can prove convergence to a (or a neighborhood of) consensus stationary solution under some regular assumptions. It is worth noting that Prox-DGD can handle nonconvex nonsmooth functions if their proximal operators can be computed. Such functions include SCAD, MCP, and ℓ_q quasi-norms, $q \in [0, 1)$. Similarly, Prox-DGD can take the constraint to a nonconvex set with an easy projection. To establish these properties, we have to introduce a completely different line of analysis, as well as modify existing proofs that were used in the convex setting.**

***Index Terms*—Nonconvex dencentralized computing, consensus optimization, decentralized gradient descent method, proximal decentralized gradient descent.**

I. Introduction

We consider an undirected, connected network of n agents and the following consensus optimization problem defined on the network:

$$\underset{x\in\mathbb{R}^p}{\text{minimize}}\, f(x) \triangleq \sum_{i=1}^{n} f_i(x), \tag{1}$$

where f_i is a differentiable function only known to the agent i. We also consider the consensus optimization problem in the following differentiable+proximable* form:

$$\underset{x\in\mathbb{R}^p}{\text{minimize}}\ s(x) \triangleq \sum_{i=1}^{n}(f_i(x) + r_i(x)), \tag{2}$$

where f_i, r_i are differentiable and proximable functions, respectively, only known to the agent i. Each function r_i is possibly non-differentiable or nonconvex, or both.

The models (1) and (2) find applications in decentralized averaging, learning, estimation, and control. Some typical applications include: (i) the distributed compressed sensing problems [14], [30], [39], [45], [49]; (ii) distributed consensus [9], [29], [55], [58], [61], [69]; (iii) distributed and parallel machine learning [15], [21], [33], [43], [55]. More specifically, in these applications, each f_i can be: 1) the data-fidelity term (possibly nonconvex) in statistical learning and machine learning [15], [62]; 2) nonconvex utility functions used in applications such as resource allocation [6], [20]; 3) empirical risk of deep neural networks with nonlinear activation functions [3]. The proximal function r_i can be taken as: 1) convex penalties such as nonsmooth ℓ_1-norm or smooth ℓ_2-norm; 2) the indicator function for a closed convex set (or a nonconvex set with an easy projection) [4], that is, $r_i(x) = 0$ if x satisfies the constraint and ∞ otherwise; 3) nonconvex penalties such as ℓ_q quasi-norm ($0 \le q < 1$) [11], [49], smoothly clipped absolute deviation (SCAD) penalty [16] and the minimax concave penalty (MCP) [68].

When f_i's are convex, the existing algorithms include the (sub)gradient methods [8], [10], [24], [37], [40], [46], [59], [65], and the primal-dual domain methods such as the decentralized alternating direction method of multipliers (DADMM) [9], [51], [52], DLM [31], and EXTRA [53], [54]. When f_i's are nonconvex, some existing results include [4], [5], [18], [27], [35], [36], [56], [57], [60], [62], [69]. In spite of the algorithms and their analysis in these works, the convergence of the simple algorithm Decentralized Gradient Descent (DGD) [40] under nonconvex f_i's is still unknown. Furthermore, although DGD is slower than DADMM, DLM and EXTRA on convex problems, DGD is simpler and thus easier to extend to a variety of settings such as [23], [38], [47], [64], where online processing and delay tolerance are considered. Therefore, we expect our results to motivate future adoptions of nonconvex DGD.

This paper studies the convergence of two algorithms: DGD for solving problem (1) and Prox-DGD for problem (2). In each DGD iteration, every agent locally computes a gradient and then updates its variable by combining the average of its neighbors' with the negative gradient step. In each Prox-DGD iteration, every agent locally computes a gradient of f_i and a proximal map of r_i, as well as exchanges information with its neighbors. Both algorithms can use either a fixed step size or a sequence of decreasing step sizes.

When the problem is convex and a fixed step size is used, DGD does not converge to a solution of the original problem (1) but a point in its neighborhood [65]. This motivates the use of decreasing step sizes such as in [10], [24]. Assuming f_i's are convex and have Lipschitz continuous and bounded

Manuscript received September 18, 2017; revised January 26, 2018; accepted March 13, 2018. Date of publication March 21, 2018; date of current version April 19, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joao Xavier. The work of J. Zeng was supported in part by the NSFC grants (61603162, 11501440, 61772246, 61603163) and the doctoral start-up foundation of Jiangxi Normal Univerity. The work of W. Yin was supported in part by NSF grants DMS-1720237 and ECCS-1462398, and ONR Grant N00014171216. *(Corresponding author: Wotao Yin.)*

J. Zeng is with the School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330027, China (e-mail: jsh.zeng@gmail.com).

W. Yin is with the Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095 USA (e-mail: wotaoyin@ucla.edu).

Digital Object Identifier 10.1109/TSP.2018.2818081

*We call a function proximable if its *proximal operator* $\text{prox}_{\alpha f}(y) \triangleq \text{argmin}_x\{\alpha f(x) + \frac{1}{2}\|x - y\|^2\}$ is easy to compute.

TABLE I
COMPARISONS ON DIFFERENT ALGORITHMS FOR CONSENSUS SMOOTH OPTIMIZATION PROBLEM (1)

	Fixed step size		Decreasing step sizes	
algorithm	DGD [65]	DGD (this paper)	D-NG [24]	DGD (this paper)
f_i	convex only	(non)convex	convex only	(non)convex
∇f_i	Lipschitz		Lipschitz, bounded	
step size	$0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$		$\mathcal{O}(\frac{1}{k})$ with Nesterov acc.	$\mathcal{O}(\frac{1}{k^\epsilon})$ $\epsilon \in (0,1]$
consensus	error $\mathcal{O}(\alpha)$		$\mathcal{O}(\frac{1}{k})$	$\mathcal{O}(\frac{1}{k^\epsilon})$
$\min_{j\le k}\|\mathbf{x}^{j+1}-\mathbf{x}^j\|^2$	$o(\frac{1}{k})$		no rate	$o(\frac{1}{k^{1+\epsilon}})$
global objective error	$\mathcal{O}(\frac{1}{k})$ until error $\mathcal{O}(\frac{\alpha}{1-\zeta})$	Convex: $\mathcal{O}(\frac{1}{k})$ until error $\mathcal{O}(\frac{\alpha}{1-\zeta})$; Nonconvex: no rate	$\mathcal{O}(\frac{\ln k}{k})$	Convex[b]: $\mathcal{O}(\frac{\ln k}{\sqrt{k}})(\epsilon=1/2)$, $\mathcal{O}(\frac{1}{\ln k})(\epsilon=1)$, $\mathcal{O}(\frac{1}{k^{\min\{\epsilon,1-\epsilon\}}})$(other ϵ); Nonconvex: no rate

[b]The objective error rates of DGD and Prox-DGD obtained in this paper and those in convex DProx-Grad [10] are ergodic or running best rates.

gradients, [10] shows that decreasing step sizes $\alpha_k = \frac{1}{\sqrt{k}}$ lead to a convergence rate $\mathcal{O}(\frac{\ln k}{k})$ of the running best of objective errors. [24] uses nested loops and shows an outer-loop convergence rate $\mathcal{O}(\frac{1}{k^2})$ of objective errors, utilizing Nesterov's acceleration, provided that the inner loop performs substantial consensus computation. Without a substantial inner loop, their single-loop algorithm using the decreasing step sizes $\alpha_k = \frac{1}{k^{1/3}}$ has a reduced rate $\mathcal{O}(\frac{\ln k}{k})$.

The objective of this paper is two-fold: (a) we aim to show, other than losing global optimality, most existing convergence results of DGD and Prox-DGD that are known in the convex setting remain valid in the nonconvex setting, and (b) to achieve (a), we illustrate how to tailor nonconvex analysis tools for decentralized optimization. In particular, our asymptotic exact and inexact consensus results require new treatments because they are special to decentralized algorithms.

The analytic results of this paper can be summarized as follows.

a) When a fixed step size α is used and properly bounded, the DGD iterates converge to a stationary point of a Lyapunov function. The difference between each local estimate of x and the global average of all local estimates is bounded, and the bound is proportional to α.
b) When a decreasing step size $\alpha_k = \mathcal{O}(1/(k+1)^\epsilon)$ is used, where $0 < \epsilon \le 1$ and k is the iteration number, the objective sequence converges, and the iterates of DGD are asymptotically consensual (i.e., become equal one another), and they achieve this at the rate of $\mathcal{O}(1/(k+1)^\epsilon)$. Moreover, we show the convergence of DGD to a stationary point of the original problem, and derive the convergence rates of DGD with different ϵ for objective functions that are convex.
c) The convergence analysis of DGD can be extended to the algorithm Prox-DGD for solving problem (2). However, when the proximable functions r_i's are nonconvex, the mixing matrix is required to be positive definite and a smaller step size is also required. (Otherwise, the mixing matrix can be non-definite.)

The detailed comparisons between our results and the existing results on DGD and Prox-DGD are presented in Tables I and II. The global objective error rate in these two tables refers to the rate of $\{f(\bar{x}^k) - f(x_{\rm opt})\}$ or $\{s(\bar{x}^k) - s(x_{\rm opt})\}$, where $\bar{x}^k = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{(i)}^k$ is the average of the kth iterate and $x_{\rm opt}$ is a global solution. The comparisons beyond DGD and Prox-DGD are presented in Section IV and Table III.

New proof techniques are introduced in this paper, particularly, in the analysis of convergence of DGD and Prox-DGD with decreasing step sizes. Specifically, the convergence of objective sequence and convergence to a stationary point of the original problem with decreasing step sizes are justified via taking a Lyapunov function and several new lemmas (cf. Lemmas 9, 12, and the proof of Theorem 2). Moreover, we estimate the consensus rate by introducing an auxiliary sequence and then showing both sequences have the same rates (cf. the proof of Proposition 3). All these proof techniques are new and distinguish our paper from the existing works such as [4], [10], [18], [24], [35], [40], [57], [62]. It should be mentioned that during the revision of this paper, we found some recent, related but independent work on the convergence of nonconvex decentralized algorithms including [19], [21], [22], [33]. We will give detailed comparisons with these work latter. Some numerical results can be found in [67] due to page limit.

The rest of this paper is organized as follows. Section II describes the problem setup and reviews the algorithms. Section III presents our assumptions and main results. Section IV discusses related works. Section V presents the proofs of our main results. We conclude this paper in Section VI.

Notation: Let I denote the identity matrix of the size $n \times n$, and $\mathbf{1} \in \mathbb{R}^n$ denote the vector of all 1's. For the matrix X, X^T denotes its transpose, X_{ij} denotes its (i,j)th component, and $\|X\| \triangleq \sqrt{\langle X, X\rangle} = \sqrt{\sum_{i,j} X_{ij}^2}$ is its Frobenius norm, which simplifies to the Euclidean norm when X is a vector. Given a symmetric, positive semidefinite matrix $G \in \mathbf{R}^{n\times n}$, we let $\|X\|_G^2 \triangleq \langle X, GX\rangle$ be the induced semi-norm. Given a function h, $\mathrm{dom}(f)$ denotes its domain.

II. PROBLEM SETUP AND ALGORITHM REVIEW

Consider a connected *undirected* network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ is a set of n nodes and $\mathcal{E}$ is the edge set. Any edge $(i,j) \in \mathcal{E}$ represents a communication link between nodes i and j. Let

TABLE II
COMPARISONS ON DIFFERENT ALGORITHMS FOR CONSENSUS COMPOSITE OPTIMIZATION PROBLEM (2)

	Fixed step size			Decreasing step sizes	
algorithm	AccDProx-Grad [8]	DProx-Grad [10]	Prox-DGD (this paper)	DProx-Grad [10]	Prox-DGD (this paper)
f_i, r_i	convex only		(non)convex	convex only	(non)convex
∇f_i	Lipschitz, bounded		Lipschitz	Lipschitz, bounded	
∂r_i	bounded		–	bounded	
step size	$0 < \alpha < \frac{1}{L_f}$		$0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$ (convex r_i); $0 < \alpha < \frac{\lambda_n(W)}{L_f}$ (nonconvex r_i, $\lambda_n(W) > 0$)	$\mathcal{O}(\frac{1}{(k+1)^{1/2}})$	$\mathcal{O}(\frac{1}{(k+1)^{\epsilon}})$ $\epsilon \in (0, 1]$
consensus	$\mathcal{O}(\gamma^k k^2), 0 < \gamma < 1$	error $\mathcal{O}(\alpha)$		$\mathcal{O}(\frac{1}{k^{1/2}})$	$\mathcal{O}(\frac{1}{k^{\epsilon}})$
$\min_{j\le k} \|\mathbf{x}^{j+1} - \mathbf{x}^j\|^2$	no rate	no rate	$o(\frac{1}{k})$	no rate	$o(\frac{1}{k^{1+\epsilon}})$
global objective error	$\mathcal{O}(\frac{1}{k})$	Form $\frac{D_1}{\alpha} + D_2\alpha$, $D_1, D_2 > 0$	Convex: form $\frac{D_3}{\alpha} + D_4\alpha$, $D_3, D_4 > 0$; Nonconvex: no rate	$\mathcal{O}(\frac{\ln k}{k})^{\dagger}$	Convex†: $\mathcal{O}(\frac{\ln k}{\sqrt{k}})(\epsilon = 1/2)$, $\mathcal{O}(\frac{1}{\ln k})(\epsilon = 1)$, $\mathcal{O}(\frac{1}{k^{\min\{\epsilon, 1-\epsilon\}}})$(other ϵ), Nonconvex: no rate

TABLE III
COMPARISONS ON SCENARIOS APPLIED FOR DIFFERENT NONCONVEX DECENTRALIZED ALGORITHMS$^{\natural}$

	f_i	nonsmooth r_i		step size		network (W)		algorithm type		fusion scheme	
algorithm	smooth	cvx	ncvx	fixed	diminish	static	dynamic	determin	stochastic	ATC	CTA
DGD (this paper)	$\surd$			$\surd$	$\surd$	$\surd$ (doubly)	$--$	$\surd$	$--$	$--$	$\surd$
Perturbed Push-sum [57]	$\surd$			$--$	$\surd$	$--$	$\surd$ (column)	$\surd$	$\surd$	$--$	$\surd$
ZENITH [18]	$\surd$			$\surd$	$--$	$\surd$ (doubly)	$--$	$\surd$	$--$	$--$	$\surd$
Prox-DGD (this paper)	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$ (doubly)	$--$	$\surd$	$--$	$--$	$\surd$
NEXT [35]	$\surd$	$\surd$	$--$	$--$	$\surd$	$--$	$\surd$ (doubly)	$\surd$	$--$	$\surd$	$--$
DeFW [62]	$\surd$	$\surd$	$--$	$--$	$\surd$	$\surd$ (doubly)	$--$	$\surd$	$--$	$\surd$	$--$
Proj SGD [4]	$\surd$	$\surd$	$--$	$--$	$\surd$	$--$	$\surd$ (row)	$--$	$\surd$	$\surd$	$--$

$^{\natural}$ In this table, the full names of these abbreviations are list as follows: cvx (convex), ncvx (nonconvex), diminish (diminishing), determin (deterministic), ATC (adaptive-then-combine), CTA (combine-then-adaptive), doubly (doubly stochastic), column (column stochastic), row (row stochastic), where vocabularies in the brackets are the full names. A row, or column, or double stochastic W means that: $W\mathbf{1} = \mathbf{1}$, or $W^T\mathbf{1} = \mathbf{1}$, or both hold.

$x_{(i)} \in \mathbb{R}^p$ denote the *local copy* of x at node i. We reformulate the consensus problem (1) into the **equivalent problem**:

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \;\; \mathbf{1}^T\mathbf{f}(\mathbf{x}) \triangleq \sum_{i=1}^{n} f_i(\mathbf{x}_{(i)}), \\ &\text{subject to } \mathbf{x}_{(i)} = \mathbf{x}_{(j)}, \forall (i,j) \in \mathcal{E}, \end{aligned} \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^{n\times p}$, $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^n$ with

$$\mathbf{x} \triangleq \begin{pmatrix} - & \mathbf{x}_{(1)}^T & - \\ - & \mathbf{x}_{(2)}^T & - \\ & \vdots & \\ - & \mathbf{x}_{(n)}^T & - \end{pmatrix}, \;\; \mathbf{f}(\mathbf{x}) \triangleq \begin{pmatrix} f_1(\mathbf{x}_{(1)}) \\ f_2(\mathbf{x}_{(2)}) \\ \vdots \\ f_n(\mathbf{x}_{(n)}) \end{pmatrix}.$$

In addition, the gradient of $\mathbf{f}(\mathbf{x})$ is

$$\nabla\mathbf{f}(\mathbf{x}) \triangleq \begin{pmatrix} - & \nabla f_1(\mathbf{x}_{(1)})^T & - \\ - & \nabla f_2(\mathbf{x}_{(2)})^T & - \\ & \vdots & \\ - & \nabla f_n(\mathbf{x}_{(n)})^T & - \end{pmatrix} \in \mathbb{R}^{n\times p}. \tag{4}$$

The ith rows of the matrices $\mathbf{x}$ and $\nabla\mathbf{f}(\mathbf{x})$, and vector $\mathbf{f}(\mathbf{x})$, correspond to agent i. The analysis in this paper applies to any integer $p \ge 1$. **For simplicity, one can let $p = 1$ and treat $\mathbf{x}$ and $\nabla\mathbf{f}(\mathbf{x})$ as vectors (rather than matrices).**

The **algorithm DGD** [40] for (3) is described as follows:

Pick an arbitrary $\mathbf{x}^0$. For $k = 0, 1, \ldots$, compute

$$\mathbf{x}^{k+1} \leftarrow W\mathbf{x}^k - \alpha_k\nabla\mathbf{f}(\mathbf{x}^k), \tag{5}$$

where W is a mixing matrix and $\alpha_k > 0$ is a step-size parameter.

Similarly, we can reformulate the composite problem (2) as the following equivalent form:

$$\begin{aligned} &\underset{\mathbf{x}}{\text{minimize}} \;\; \sum_{i=1}^{n}(f_i(\mathbf{x}_{(i)}) + r_i(\mathbf{x}_{(i)})), \\ &\text{subject to } \mathbf{x}_{(i)} = \mathbf{x}_{(j)}, \; \forall (i,j) \in \mathcal{E}. \end{aligned} \tag{6}$$

Let $r(\mathbf{x}) \triangleq \sum_{i=1}^n r_i(\mathbf{x}_{(i)})$. The algorithm Prox-DGD can be applied to the above problem (6):

Prox-DGD: Take an arbitrary $\mathbf{x}^0$. For $k = 0, 1, \ldots$, perform

$$\mathbf{x}^{k+1} \leftarrow \text{prox}_{\alpha_k r}(W\mathbf{x}^k - \alpha_k\nabla\mathbf{f}(\mathbf{x}^k)), \tag{7}$$

where the proximal operator is

$$\text{prox}_{\alpha_k r}(\mathbf{x}) \triangleq \underset{\mathbf{u}\in\mathbb{R}^{n\times p}}{\text{argmin}} \left\{ \alpha_k r(\mathbf{u}) + \frac{\|\mathbf{u} - \mathbf{x}\|^2}{2} \right\}. \tag{8}$$

III. Assumptions and Main Results

This section presents all of our main results.

A. Definitions and Assumptions

Definition 1 (Lipschitz differentiability): A function h is called Lipschitz differentiable if h is differentiable and its gradient ∇h is Lipschitz continuous, i.e., $\|\nabla h(u) - \nabla h(v)\| \leq L\|u - v\|, \forall u, v \in \mathrm{dom}(h)$, where $L > 0$ is its Lipschitz constant.

Definition 2 (Coercivity): A function h is called coercive if $\|u\| \to +\infty$ implies $h(u) \to +\infty$.

The next definition is a property that many functions have (see [63, Sec. 2.2] for examples) and can help obtain whole sequence convergence[†] from subsequence convergence.

Definition 3 (Kurdyka-Łojasiewicz (KŁ) property [2], [7], [34]): A function $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ has the KŁ property at $x^* \in \mathrm{dom}(\partial h)$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of x^*, and a continuous concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:

i) $\varphi(0) = 0$ and φ is differentiable on $(0, \eta)$;
ii) for all $s \in (0, \eta)$, $\varphi'(s) > 0$;
iii) for all x in $U \cap \{x : h(x^*) < h(x) < h(x^*) + \eta\}$, the KŁ inequality holds

$$\varphi'\big(h(x) - h(x^*)\big) \cdot \mathrm{dist}\big(0, \partial h(x)\big) \geq 1. \tag{9}$$

Proper lower semi-continuous functions that satisfy the KŁ inequality at each point of $\mathrm{dom}(\partial h)$ are called KŁ functions.

Assumption 1 (Objective): The objective functions $f_i : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, $i = 1, \ldots, n$, satisfy the following:

1) f_i is Lipschitz differentiable with constant $L_{f_i} > 0$.
2) f_i is proper (i.e., not everywhere infinite) and coercive.

The sum $\sum_{i=1}^n f_i(\mathbf{x}_{(i)})$ is L_f-Lipschitz differentiable with $L_f \triangleq \max_i L_{f_i}$ (this can be easily verified via the definition of $\nabla \mathbf{f}(\mathbf{x})$ as shown in (4)). In addition, each f_i is lower bounded following Part (2) of the above assumption.

Assumption 2 (Mixing matrix): The mixing matrix $W = [w_{ij}] \in \mathbb{R}^{n \times n}$ has the following properties:

1) (Graph) If $i \neq j$ and $(i, j) \notin \mathcal{E}$, then $w_{ij} = 0$, otherwise, $w_{ij} > 0$.
2) (Symmetry) $W = W^T$.
3) (Null space property) $\mathrm{null}\{I - W\} = \mathrm{span}\{\mathbf{1}\}$.
4) (Spectral property) $I \succeq W \succ -I$.

By Assumption 2, a solution $\mathbf{x}_{\mathrm{opt}}$ to problem (3) satisfies $(I - W)\mathbf{x}_{\mathrm{opt}} = 0$. Due to the symmetric assumption of W, its eigenvalues are real and can be sorted in the nonincreasing order. Let $\lambda_i(W)$ denote the ith largest eigenvalue of W. Then by Assumption 2,

$$\lambda_1(W) = 1 > \lambda_2(W) \geq \cdots \geq \lambda_n(W) > -1.$$

Let ζ be the second largest magnitude eigenvalue of W. Then

$$\zeta = \max\{|\lambda_2(W)|, |\lambda_n(W)|\}. \tag{10}$$

[†]Whole sequence convergence from any starting point is referred to as "global convergence" in the literature. Its limit is not necessarily a global solution.

B. Convergence Results of DGD

We consider the convergence of DGD with both a fixed step size and a sequence of decreasing step sizes.

1) Convergence Results of DGD With a Fixed Step Size: The convergence result of DGD with a fixed step size (i.e., $\alpha_k \equiv \alpha$) is established based on the Lyapunov function [65]:

$$\mathcal{L}_\alpha(\mathbf{x}) \triangleq \mathbf{1}^T \mathbf{f}(\mathbf{x}) + \frac{1}{2\alpha}\|\mathbf{x}\|^2_{I-W}. \tag{11}$$

It is worth reminding that convexity is *not* assumed.

Theorem 1 (Global convergence): Let $\{\mathbf{x}^k\}$ be the sequence generated by DGD (5) with the step size $0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$. Let Assumptions 1 and 2 hold. Then $\{\mathbf{x}^k\}$ has at least one accumulation point $\mathbf{x}^*$, and any such point is a stationary point of $\mathcal{L}_\alpha(\mathbf{x})$. Furthermore, the running best rates[‡] of the sequences[§] $\{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\}$, and $\{\|\nabla \mathcal{L}_\alpha(\mathbf{x}^k)\|^2\}$, and $\{\|\frac{1}{n}\mathbf{1}^T \nabla \mathbf{f}(\mathbf{x}^k)\|^2\}$ are $o(\frac{1}{k})$. The convergence rate of the sequence $\{\frac{1}{K}\sum_{k=0}^{K-1} \|\frac{1}{n}\mathbf{1}^T \nabla \mathbf{f}(\mathbf{x}^k)\|^2\}$ is $\mathcal{O}(\frac{1}{K})$.

In addition, if $\mathcal{L}_\alpha$ satisfies the KŁ property at an accumulation point $\mathbf{x}^*$, then $\{\mathbf{x}^k\}$ globally converges to $\mathbf{x}^*$.

Remark 1: Let $\mathbf{x}^*$ be a stationary point of $\mathcal{L}_\alpha(\mathbf{x})$, and thus

$$0 = \nabla \mathbf{f}(\mathbf{x}^*) + \alpha^{-1}(I - W)\mathbf{x}^*. \tag{12}$$

Since $\mathbf{1}^T(I - W) = 0$, (12) yields $0 = \mathbf{1}^T \nabla \mathbf{f}(\mathbf{x}^*)$, indicating that $\mathbf{x}^*$ is also a stationary point to the separable function $\sum_{i=1}^n f_i(\mathbf{x}_{(i)})$. Since the rows of $\mathbf{x}^*$ are not necessarily identical, we *cannot* say $\mathbf{x}^*$ is a stationary point to Problem (3). However, the differences between the rows of $\mathbf{x}^*$ are bounded, following our next result below adapted from [65]:

Proposition 1 (Consensual bound on $\mathbf{x}^*$): For each iteration k, define $\bar{x}^k \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{x}^k_{(i)}$. Then, it holds for each node i that

$$\|\mathbf{x}^k_{(i)} - \bar{x}^k\| \leq \frac{\alpha D}{1 - \zeta}, \tag{13}$$

where D is a universal bound of $\|\nabla \mathbf{f}(\mathbf{x}^k)\|$ defined in Lemma 6 (Section V.A), ζ is the second largest magnitude eigenvalue of W specified in (10). As $k \to \infty$, (13) yields the consensual bound

$$\|\mathbf{x}^*_{(i)} - \bar{x}^*\| \leq \frac{\alpha D}{1 - \zeta},$$

where $\bar{x}^* \triangleq \frac{1}{n}\sum_{i=1}^n \mathbf{x}^*_{(i)}$.

Take $\mathbf{x}^0 = 0$ for proof simplicity. This proposition can be proved by applying Lemma 7 (Section V.C) to

$$\mathbf{x}^k - \bar{\mathbf{x}}^k = -\alpha \sum_{j=0}^{k-1} \left(W^{k-1-j} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) \nabla \mathbf{f}(\mathbf{x}^j).$$

In Proposition 1, the consensual bound is proportional to the step size α and inversely proportional to the gap between the largest and the second largest magnitude eigenvalues of W.

Let us compare the DGD iteration with the iteration of *centralized gradient descent* (15) for $f(x)$. Averaging the rows of

[‡]Given a nonnegative sequence a_k, its running best sequence is $b_k = \min\{a_i : i \leq k\}$. We say a_k has a running best rate of $o(1/k)$ if $b_k = o(1/k)$.

[§]These quantities naturally appear in the analysis, so we keep the squares.

(5) yields the following comparison:

$$\text{DGD averaged:} \quad \bar{x}^{k+1} \leftarrow \bar{x}^k - \alpha\left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_{(i)}^k)\right). \quad (14)$$

$$\text{Centralized:} \quad \bar{x}^{k+1} \leftarrow \bar{x}^k - \alpha\left(\frac{1}{n}\sum_{i=1}^n \nabla f_i(\bar{x}^k)\right). \quad (15)$$

Apparently, DGD approximates centralized gradient descent by evaluating $\nabla f_{(i)}$ at local variables $\mathbf{x}_{(i)}^k$ instead of the global average. We can estimate the error of this approximation as

$$\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_{(i)}^k) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\bar{x}^k)\right\| \le \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_{(i)}^k) - \nabla f_i(\bar{x}^k)\| \le \frac{\alpha D L_f}{1-\zeta}.$$

Unlike the convex analysis in [65], it is impossible to bound the difference between the sequences of (14) and (15) without convexity because the two sequences may converge to different stationary points of $\mathcal{L}_\alpha$.

Remark 2: The KŁ assumption on $\mathcal{L}_\alpha$ in Theorem 1 can be satisfied if each f_i is a sub-analytic function. Since $\|\mathbf{x}\|_{I-W}^2$ is obviously sub-analytic and the sum of two sub-analytic functions remains sub-analytic, $\mathcal{L}_\alpha$ is sub-analytic if each f_i is so. See [63, Sec. 2.2] for more details and examples.

Proposition 2 (KŁ convergence rates): Let the assumptions of Theorem 1 hold. Suppose that $\mathcal{L}_\alpha$ satisfies the KŁ inequality at an accumulation point $\mathbf{x}^*$ with $\psi(s) = cs^{1-\theta}$ for some constant $c > 0$. Then, the following convergence rates hold:

a) If $\theta = 0$, $\mathbf{x}^k$ converges to $\mathbf{x}^*$ in finitely many iterations.
b) If $\theta \in (0, \frac{1}{2}]$, $\|\mathbf{x}^k - \mathbf{x}^*\| \le C_0\tau^k$ for all $k \ge k^*$ for some $k^* > 0, C_0 > 0, \tau \in [0, 1)$.
c) If $\theta \in (\frac{1}{2}, 1)$, $\|\mathbf{x}^k - \mathbf{x}^*\| \le C_0 k^{-(1-\theta)/(2\theta-1)}$ for all $k \ge k^*$, for certain $k^* > 0, C_0 > 0$.

Note that the rates in parts (b) and (c) of Proposition 2 are of the *eventual* type.

Using fixed step sizes, our results are limited because the stationary point $\mathbf{x}^*$ of $\mathcal{L}_\alpha$ is not a stationary point of the original problem. We only have a consensual bound on $\mathbf{x}^*$. To address this issue, the next section uses decreasing step sizes and presents better convergence results.

2) Convergence of DGD With Decreasing Step Sizes: The positive consensual error bound in Proposition 1, which is proportional to the constant step size α, motivates the use of properly decreasing step sizes $\alpha_k = \mathcal{O}(\frac{1}{(k+1)^\epsilon})$, for some $0 < \epsilon \le 1$, to diminish the consensual bound to 0. As a result, any accumulation point $\mathbf{x}^*$ becomes a stationary point of the original problem (3). To analyze DGD with decreasing step sizes, we add the following assumption.

Assumption 3 (Bounded gradient): For any k, $\nabla\mathbf{f}(\mathbf{x}^k)$ is uniformly bounded by some constant $B > 0$, i.e., $\|\nabla\mathbf{f}(\mathbf{x}^k)\| \le B$.

Note that the bounded gradient assumption is a regular assumption in the convergence analysis of decentralized gradient methods (see, [4], [5], [18], [27], [35], [36], [56], [57], [62] for example), even in the convex setting [24] and also [10], though it is not required for centralized gradient descent.

We take the step size sequence:

$$\alpha_k = \frac{1}{L_f(k+1)^\epsilon}, \quad 0 < \epsilon \le 1, \quad (16)$$

throughout the rest part of this section. (The numerator 1 can be replaced by any positive constant.) By iteratively applying iteration (5), we obtain the following expression

$$\mathbf{x}^k = W^k\mathbf{x}^0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j}\nabla\mathbf{f}(\mathbf{x}^j). \quad (17)$$

Proposition 3 (Asymptotic consensus rate): Let Assumptions 2 and 3 hold. Let DGD use (16). Let $\bar{\mathbf{x}}^k \triangleq \frac{1}{n}\mathbf{1}\mathbf{1}^T\mathbf{x}^k$. Then, $\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|$ converges to 0 at the rate of $\mathcal{O}(1/(k+1)^\epsilon)$.

According to Proposition 3, the iterates of DGD with decreasing step sizes can reach consensus asymptotically (compared to a nonzero bound in the fixed step size case in Proposition 1). Moreover, with a larger ϵ, faster decaying step sizes generally imply a faster asymptotic consensus rate. Note that $(I-W)\bar{\mathbf{x}}^k = 0$ and thus $\|\mathbf{x}^k\|_{I-W}^2 = \|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_{I-W}^2$. Therefore, the above proposition implies the following result.

Corollary 1: Apply the setting of Proposition 3. $\|\mathbf{x}^k\|_{I-W}^2$ converges to 0 at the rate of $\mathcal{O}(1/(k+1)^{2\epsilon})$.

Corollary 1 shows that the sequence $\{\mathbf{x}^k\}$ in the $(I-W)$ semi-norm can decay to 0 at a sublinear rate. For any *global* consensual solution $\mathbf{x}_{\rm opt}$ to problem (3), we have $\|\mathbf{x}^k - \mathbf{x}_{\rm opt}\|_{I-W}^2 = \|\mathbf{x}^k\|_{I-W}^2$ so, if $\{\mathbf{x}^k\}$ does converge to $\mathbf{x}_{\rm opt}$, then their distance in the same semi-norm decays at $\mathcal{O}(1/k^{2\epsilon})$.

Theorem 2 (Convergence): Let Assumptions 1, 2 and 3 hold. Let DGD use step sizes (16). Then

a) $\{\mathcal{L}_{\alpha_k}(\mathbf{x}^k)\}$ and $\{\mathbf{1}^T\mathbf{f}(\mathbf{x}^k)\}$ converge to the same limit;
b) $\lim_{k\to\infty}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k) = 0$, and any limit point of $\{\mathbf{x}^k\}$ is a stationary point of problem (3);
c) In addition, if there exists an isolated accumulation point, then $\{\mathbf{x}^k\}$ converges.

In the proof of Theorem 2, we will establish

$$\sum_{k=0}^{\infty}\left(\alpha_k^{-1}(1+\lambda_n(W)) - L_f\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 < \infty,$$

which implies that the running best rate of the sequence $\{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\}$ is $o(1/k^{1+\epsilon})$. Theorem 2 shows that the objective sequence converges, and any limit point of $\{\mathbf{x}^k\}$ is a stationary point of the original problem. However, there is no result on the convergence rate of the objective sequence to an optimal value, and it is generally difficult to get such a rate without convexity.

Although our primary focus is nonconvexity, next we assume convexity and present the objective convergence rate, which has an interesting relation with ϵ.

For any $\mathbf{x} \in \mathbb{R}^{n\times p}$, let $\bar{f}(\mathbf{x}) \triangleq \sum_{i=1}^n f_i(\mathbf{x}_{(i)})$. Even if f_i's are convex, the solution to (3) may be non-unique. Thus, let $\mathcal{X}^*$ be the set of solutions to (3). Given $\mathbf{x}^k$, we pick the solution $\mathbf{x}_{\rm opt} = {\rm Proj}_{\mathcal{X}^*}(\mathbf{x}^k) \in \mathcal{X}^*$. Also let $f_{\rm opt} = \bar{f}(\mathbf{x}_{\rm opt})$ be the optimal value of (1). Define the ergodic objective:

$$\bar{f}^K = \frac{\sum_{k=0}^K \alpha_k \bar{f}(\bar{\mathbf{x}}^{k+1})}{\sum_{k=0}^K \alpha_k}, \quad (18)$$

where $\bar{\mathbf{x}}^{k+1} = \frac{1}{n}(\mathbf{1}^T\mathbf{x}^{k+1})\mathbf{1}$. Obviously,

$$\bar{f}^K \geq \min_{k=1,\ldots,K+1} \bar{f}(\bar{\mathbf{x}}^k). \tag{19}$$

Proposition 4 (Convergence rates under convexity): Let Assumptions 1, 2 and 3 hold. Let DGD use step sizes (16). If $\lambda_n(W) > 0$ and each f_i is convex, then $\{\bar{f}^K\}$ defined in (18) converges to the optimal objective value $f_{\rm opt}$ at the following rates:

a) if $0 < \epsilon < 1/2$, the rate is $\mathcal{O}(\frac{1}{K^\epsilon})$;
b) if $\epsilon = 1/2$, the rate is $\mathcal{O}(\frac{\ln K}{\sqrt{K}})$;
c) if $1/2 < \epsilon < 1$, the rate is $\mathcal{O}(\frac{1}{K^{1-\epsilon}})$;
d) if $\epsilon = 1$, the rate is $\mathcal{O}(\frac{1}{\ln K})$.

The convergence rates established in Proposition 4 are almost as good as $\mathcal{O}(\frac{1}{\sqrt{K}})$ when $\epsilon = \frac{1}{2}$. As ϵ goes to either 0 or 1, the rates become slower, and $\epsilon = 1/2$ may be the optimal choice in terms of the convergence rate. However, by Proposition 3, a larger ϵ implies a faster consensus rate. Therefore, there is a tradeoff to choose an appropriate ϵ in the practical implementation of DGD. The proof of this proposition can be found in [67] due to page limit.

Remark 3: A related algorithm is the perturbed push-sum algorithm, also called subgradient-push, which was proposed in [25] for average consensus problem over time-varying network. Its convergence in the convex setting was developed in [41]. Recently, its convergence (to a critical point) in the nonconvex setting was established in [57] under some regularity assumptions. Moreover, by utilizing perturbations on the update process and the assumption of no saddle-point existence, almost sure convergence to a local minimum of its perturbed variant was also shown in [57].

Remark 4: Another recent algorithm is decentralized stochastic gradient descent (D-PSGD) in [33] with support to nonconvex large-sum objectives. An $\mathcal{O}(\frac{1}{K} + \frac{1}{\sqrt{nK}})$-ergodic convergence rate was established assuming K is sufficiently large and the step size α is sufficiently small. When applied to the setting of this paper, [33, Th. 1] implies that the sequence $\{\frac{1}{K}\sum_{k=0}^{K-1} \|\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k)\|^2\}$ converges to zero at the rate $\mathcal{O}(\frac{1}{K})$ if the step size $0 < \alpha < \frac{1-\zeta}{6L_f\sqrt{n}}$, where ζ is defined in (10). From Theorem 1, we can also establish such an $\mathcal{O}(\frac{1}{K})$-ergodic convergence rate of DGD as long as $0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$. Similar rates of convergence to a stationary point have also been shown for different nonconvex algorithms in [18], [28], [57].

C. Convergence Results of Prox-DGD

Similarly, we consider the convergence of Prox-DGD with both a fixed step size and decreasing step sizes. The iteration (7) can be reformulated as

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\alpha_k r}(\mathbf{x}^k - \alpha_k\nabla\mathcal{L}_{\alpha_k}(\mathbf{x}^k)) \tag{20}$$

based on which, we define the Lyapunov function

$$\hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}) \triangleq \mathcal{L}_{\alpha_k}(\mathbf{x}) + r(\mathbf{x}),$$

where we recall $\mathcal{L}_{\alpha_k}(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x}_{(i)}) + \frac{1}{2\alpha_k}\|\mathbf{x}\|^2_{I-W}$. Then (20) is clearly the forward-backward splitting (a.k.a., prox-gradient) iteration for $\mathrm{minimize}_{\mathbf{x}}\ \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x})$. Specifically, (20) first performs gradient descent to the differentiable function $\mathcal{L}_{\alpha_k}(\mathbf{x})$ and then computes the proximal of $r(\mathbf{x})$.

To analyze Prox-DGD, we should revise Assumption 1 as follows.

Assumption 4 (Composite objective): The objective function of (6) satisfies the following:

1) Each f_i is Lipschitz differentiable with constant $L_{f_i} > 0$.
2) Each $(f_i + r_i)$ is proper, lower semi-continuous, coercive.

As before, $\sum_{i=1}^n f_i(\mathbf{x}_{(i)})$ is L_f-Lipschitz differentiable for $L_f \triangleq \max_i L_{f_i}$.

1) Convergence Results of Prox-DGD With a Fixed Step Size: Based on the above assumptions, we can get the global convergence of Prox-DGD as follows.

Theorem 3 (Global convergence of Prox-DGD): Let $\{\mathbf{x}^k\}$ be the sequence generated by Prox-DGD (7) where the step size α satisfies $0 < \alpha < \frac{1+\lambda_n(W)}{L_f}$ when r_i's are convex; and $0 < \alpha < \frac{\lambda_n(W)}{L_f}$, when r_i's are not necessarily convex (this case requires $\lambda_n(W) > 0$). Let Assumptions 2 and 4 hold. Then $\{\mathbf{x}^k\}$ has at least one accumulation point $\mathbf{x}^*$, and any accumulation point is a stationary point of $\hat{\mathcal{L}}_\alpha(\mathbf{x})$. Furthermore, the running best rates of the sequences $\{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\}$, $\{\|\mathbf{g}^{k+1}\|^2\}$ and $\{\|\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k) + \frac{1}{n}\mathbf{1}^T\xi^k\|^2\}$ (where $\mathbf{g}^{k+1}$ is defined in Lemma 16, and ξ^k is defined in Lemma 17) are $o(\frac{1}{k})$. The convergence rate of the sequence $\{\frac{1}{K}\sum_{k=0}^{K-1}\|\frac{1}{n}\mathbf{1}^T(\nabla\mathbf{f}(\mathbf{x}^k) + \xi^k)\|^2\}$ is $\mathcal{O}(\frac{1}{K})$.

In addition, if $\hat{\mathcal{L}}_\alpha$ satisfies the KŁ property at an accumulation point $\mathbf{x}^*$, then $\{\mathbf{x}^k\}$ converges to $\mathbf{x}^*$.

Theorem 3 assumes $\lambda_n(W) > 0$ when r_i's are nonconvex. If this fails to hold, we can establish with W being replaced by $\frac{I+W}{2}$. As this changes the spectral property of W, it may slow down the convergence rate. The rate of convergence of Prox-DGD can be also established by leveraging the KŁ property.

Proposition 5 (Rate of convergence of Prox-DGD): Under assumptions of Theorem 3, suppose that $\hat{\mathcal{L}}_\alpha$ satisfies the KŁ inequality at an accumulation point x^* with $\psi(s) = c_1 s^{1-\theta}$ for some constant $c_1 > 0$. Then the following hold:

a) If $\theta = 0$, $\mathbf{x}^k$ converges to $\mathbf{x}^*$ in finitely many iterations.
b) If $\theta \in (0, \frac{1}{2}]$, $\|\mathbf{x}^k - \mathbf{x}^*\| \leq C_1\tau^k$ for all $k \geq k^*$ for some $k^* > 0, C_1 > 0, \tau \in [0, 1)$.
c) If $\theta \in (\frac{1}{2}, 1)$, $\|\mathbf{x}^k - \mathbf{x}^*\| \leq C_1 k^{-(1-\theta)/(2\theta-1)}$ for all $k \geq k^*$, for certain $k^* > 0, C_1 > 0$.

2) Convergence of Prox-DGD With Decreasing Step Sizes: In Prox-DGD, we also use the decreasing step size (16). To investigate its convergence, the bounded gradient Assumption 3 should be revised as follows.

Assumption 5 (Bounded composite subgradient): For each i, ∇f_i is uniformly bounded by some constant $B_i > 0$, i.e., $\|\nabla f_i(x)\| \leq B_i$ for any $x \in \mathbb{R}^p$. Moreover, $\|\xi_i\| \leq B_{r_i}$ for any $\xi_i \in \partial r_i(x)$ and $x \in \mathbb{R}^p$, $i = 1\ldots, n$.

Let $\bar{B} \triangleq \sum_{i=1}^n (B_i + B_{r_i})$. Then $\nabla\mathbf{f}(\mathbf{x}) + \xi$ (where $\xi \in \partial r(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{R}^{n\times p}$) is uniformly bounded by $\bar{B}$. Note that the same assumption is used to analyze the convergence of distributed proximal-gradient method in the convex setting [8], [10], and also is widely used to analyze the convergence of nonconvex decentralized algorithms like in [35], [36]. In light of Lemma 17 (Section V.F), the claims in Proposition 3 and Corollary 1 also hold for Prox-DGD.

Proposition 6 (Asymptotic consensus and rate): Let Assumptions 2 and 5 hold. In Prox-DGD, use the step sizes (16). There hold

$$\|\mathbf{x}^k - \bar{\mathbf{x}}^k\| \leq C\left(\|\mathbf{x}^0\|\zeta^k + \bar{B}\sum_{j=0}^{k-1}\alpha_j\zeta^{k-1-j}\right),$$

and $\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|$ converges to 0 at the rate of $\mathcal{O}(1/(k+1)^\epsilon)$. Moreover, let $\mathbf{x}^*$ be any *global* solution of the problem (6). Then $\|\mathbf{x}^k - \mathbf{x}^*\|_{I-W}^2 = \|\mathbf{x}^k\|_{I-W}^2 = \|\mathbf{x}^k - \bar{\mathbf{x}}^*\|_{I-W}^2$ converges to 0 at the rate of $\mathcal{O}(1/(k+1)^{2\epsilon})$.

For any $\mathbf{x} \in \mathbb{R}^{n\times p}$, define $\bar{s}(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x}_{(i)}) + r_i(\mathbf{x}_{(i)})$. Let $\mathcal{X}^\dagger$ be a set of solutions of (6), $\mathbf{x}_{\rm opt} = \mathrm{Proj}_{\mathcal{X}^\dagger}(\mathbf{x}^k) \in \mathcal{X}^\dagger$, and $s_{\rm opt} = \bar{s}(\mathbf{x}_{\rm opt})$ be the optimal value of (6). Define

$$\bar{s}^K = \frac{\sum_{k=0}^K \alpha_k \bar{s}(\bar{\mathbf{x}}^{k+1})}{\sum_{k=0}^K \alpha_k}. \tag{21}$$

Theorem 4 (Convergence and rate): Let Assumptions 2, 4 and 5 hold. In Prox-DGD, use the step sizes (16). Then

a) $\{\hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^k)\}$ and $\{\sum_{i=1}^n f_i(\mathbf{x}_{(i)}^k) + r_i(\mathbf{x}_{(i)}^k)\}$ converge to the same limit;
b) $\sum_{k=0}^\infty \left(\alpha_k^{-1}(1+\lambda_n(W)) - L_f\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 < \infty$ when r_i's are convex; or, $\sum_{k=0}^\infty \left(\alpha_k^{-1}\lambda_n(W) - L_f\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 < \infty$ when r_i's are not necessarily convex (this case requires $\lambda_n(W) > 0$);
c) if $\{\xi^k\}$ satisfies $\|\xi^{k+1} - \xi^k\| \leq L_r\|\mathbf{x}^{k+1} - \mathbf{x}^k\|$ for each $k > k_0$, some constant $L_r > 0$, and a sufficiently large integer $k_0 > 0$, then

$$\lim_{k\to\infty} \mathbf{1}^T(\nabla\mathbf{f}(\mathbf{x}^k) + \xi^{k+1}) = 0,$$

where $\xi^{k+1} \in \partial r(\mathbf{x}^{k+1})$ is the one determined by the proximal operator (8), and any limit point is a stationary point of problem (6).
d) in addition, if there exists an isolated accumulation point, then $\{\mathbf{x}^k\}$ converges.
e) furthermore, if f_i and r_i are convex and $\lambda_n(W) > 0$, then the claims on the rates of $\{\bar{f}^K\}$ in Proposition 4 hold for the sequence $\{\bar{s}^K\}$ defined in (21).

Theorem 4(b) implies that the running best rate of $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2$ is $o(\frac{1}{k^{1+\epsilon}})$. The additional condition imposed on $\{\xi^k\}$ in Theorem 4(c) is some type of restricted continuous regularity of the subgradient ∂r with respect to the generated sequence. This condition is only used to establish the desired inequality (68). If ∂r is locally Lipschitz continuous in a neighborhood of a limit point, then such Lipschitz condition on $\{\xi^k\}$ can generally be satisfied, since $\{\mathbf{x}^k\}$ is asymptotic regular, and thus $\mathbf{x}^k$ will lies in such neighborhood of this limit point when k is sufficiently large. There are many kinds of proximal functions satisfying such assumption as studied in [66] (see, Remark 5 for detailed information). Theorem 4(e) gives the convergence rates of Prox-DGD in the convex setting.

Remark 5: A typical proximal function r satisfying the assumption of Theorem 4 (c) is the ℓ_q quasi-norm ($0 \leq q < 1$) widely studied in sparse optimization, which takes the form $r(x) = \sum_{i=1}^p |x_i|^q$.[¶] From [11] and [66], there is a positive lower bound for the absolute values of non-zero components of the solutions of ℓ_q regularized optimization problem. Furthermore, as shown by [66, Property 1(b)], the sequence generated by Prox-DGD also has the similar lower bound property. Moreover, by Theorem 4(b), we have $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \to 0$ as $k \to \infty$. Together with the lower bound property, we can easily obtain the finite support and sign convergence of $\{\mathbf{x}^k\}$, that is, the supports and signs of the non-zero components will freeze for sufficiently large k. When restricted to such nonzero subspace, the gradient of $r_i(u) = |u|^q$ is Lipschitz continuous for any $|u| \geq \tau$ and some positive constant τ, where τ denotes the lower bound. Besides ℓ_q quasi-norm, there are some other typical cases like SCAD [16] and MCP [68] widely used in statistical learning, satisfying the condition (c) of this theorem.

Remark 6: One tightly related algorithm of Prox-DGD is the projected stochastic gradient descent (Proj SGD) method proposed by [4] for solving the constrained multi-agent optimization problem with a convex constraint set. When restricted to the deterministic case as studied in this paper, the convergence results of Proj SGD are very similar to that of Prox-DGD (see, Theorem 4 (c)–(d) in this paper and [4, Th. 1]). However, there are some differences between [4] and this paper. In short, Proj SGD in [4] uses **convex constraints**, which correspond to setting $r(x)$ in our paper as indicator functions of those convex sets. Our paper also considers **nonconvex functions** like ℓ_q quasi-norm ($0 \leq q < 1$), SCAD, and MCP, which are widely used in statistical learning. Another difference is that Proj SGD of [4] uses **adaptive-then-combine (ATC)** and Prox-DGD of this paper does **combine-then-adaptive (CTA)**. By [4, Assumption 2], Proj SGD uses decreasing step sizes like $\mathcal{O}(k^{-\epsilon})$ for some $\epsilon > 1/2$. We study the step size $\alpha_k = \mathcal{O}(k^{-\epsilon})$ for any $0 < \epsilon \leq 1$ for Prox-DGD, as well as a fixed step size.

IV. Related Works and Discussions

We summarize some recent nonconvex decentralized algorithms in Table III. Most of them apply to either the smooth optimization problem (1) or the composite optimization problem (2) and use diminishing step sizes. Although (1) is a special case of (2) via letting $r_i(x) = 0$, there are still differences in both algorithm design and theoretical analysis. Therefore, we divide their comparisons.

We first discuss the algorithms for (1). In [57], the authors proved the convergence of perturbed push-sum for nonconvex (1) under some regularity assumptions. The convergence results for the deterministic perturbed push-sum algorithm obtained in [57] are similar to those of DGD developed in this paper under similar assumptions (see, Theorem 2 above and [57, Th. 3]). The detailed comparisons between two algorithms are illustrated in Remark 3. In [33], the sublinear convergence to a stationary point of D-PSGD algorithm was developed under the nonconvex setting. DGD studied in this paper can be viewed a special D-PSGD with a zero variance. In [18], a primal-dual approximate gradient algorithm called ZENITH was developed for (1). The convergence of ZENITH was given in the expectation of constraint violation under the Lipschitz differentiable assumption and other assumptions. The last one is the proximal primal-dual algorithm (Prox-PDA) recently proposed in [21]. The $\mathcal{O}(\frac{1}{k})$-rate of convergence to a stationary point was established in [21].

[¶]When $q = 0$, we denote $0^0 = 0$.

Latter, a perturbed variant of Prox-PDA was proposed in [22] for constrained composite (smooth+nonsmooth) optimization problem with a linear equality constraint.

Table III includes three algorithms for solving the composite problem (2), which are related to ours. All of them only deal with convex r_i (whereas r_i in this paper can also be nonconvex). In [36], the authors proposed NEXT based on the previous successive convex approximation (SCA) technique. The iterates of NEXT include two stages, a local SCA stage to update local variables and a consensus update stage to fuse the information between agents. While NEXT has results similar to Prox-DGD using diminishing step sizes. Another interesting algorithm is decentralized Frank-Wolfe (DeFW) proposed in [62] for nonconvex, smooth, constrained decentralized optimization, where a bounded convex constraint set is imposed. There are three steps at each iteration of DeFW: average gradient computation, local variable evaluation by Frank-Wolfe, and information fusion between agents. In [62], the authors established convergence results similar to Prox-DGD under diminishing step sizes. The stochastic version of DeFW has also been developed in [27] for high-dimensional convex sparse optimization. The next one is projected stochastic gradient algorithm (Proj SGD) [4] for constrained, nonconvex, smooth consensus optimization with a convex constrained set. The detailed comparison between Proj SGD and Prox-DGD are shown in Remark 6.

Based on the above analysis, the convergence results of DGD and Prox-DGD with decreasing step sizes of this paper are comparable with most of the existing ones. However, we allow nonconvex nonsmooth r_i and are able to obtain the estimates of asymptotic consensus rates. We also establish global convergence using a fixed step size while it is only found in ZENITH.

V. Proofs

In this section, we present the main proofs of our main theorems and propositions.

A. Proof for Theorem 1

The sketch of the proof is as follows: DGD is interpreted as the gradient descent algorithm applied to the Lyapunov function $\mathcal{L}_\alpha$, following the argument in [65]; then, the properties of sufficient descent, lower boundedness, and bounded gradients are established for the sequence $\{\mathcal{L}_\alpha(\mathbf{x}^k)\}$, giving subsequence convergence of the DGD iterates; finally, whole sequence convergence of the DGD iterates follows from the KŁ property of $\mathcal{L}_\alpha$.

Lemma 1 (Gradient descent interpretation): The sequence $\{\mathbf{x}^k\}$ generated by the DGD iteration (5) is the same sequence generated by applying gradient descent with the fixed step size α to the objective function $\mathcal{L}_\alpha(\mathbf{x})$.

A proof of this lemma is given in [65], and it is based on reformulating (5) as the iteration:

$$\begin{aligned}\mathbf{x}^{k+1} &= \mathbf{x}^k - \alpha(\nabla \mathbf{f}(\mathbf{x}^k) + \alpha^{-1}(I-W)\mathbf{x}^k)\\ &= \mathbf{x}^k - \alpha\nabla\mathcal{L}_\alpha(\mathbf{x}^k).\end{aligned} \tag{22}$$

Although the sequence $\{\mathbf{x}^k\}$ generated by the DGD iteration (5) can be interpreted as a centralized gradient descent sequence of function $\mathcal{L}_\alpha(\mathbf{x})$, it is different to the gradient descent of the original problem (3).

Lemma 2 (Sufficient descent of $\{\mathcal{L}_\alpha(\mathbf{x}^k)\}$): Let Assumptions 1 and 2 hold. Set the step size $0<\alpha<\frac{1+\lambda_n(W)}{L_f}$. It holds that

$$\begin{aligned}&\mathcal{L}_\alpha(\mathbf{x}^{k+1}) \le \mathcal{L}_\alpha(\mathbf{x}^k)\\ &-\frac{1}{2}\big(\alpha^{-1}(1+\lambda_n(W)) - L_f\big)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2, \quad \forall k\in\mathbb{N}.\end{aligned} \tag{23}$$

Proof: From $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha\nabla\mathcal{L}_\alpha(\mathbf{x}^k)$, it follows that

$$\langle \nabla\mathcal{L}_\alpha(\mathbf{x}^k), \mathbf{x}^{k+1}-\mathbf{x}^k\rangle = -\frac{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2}{\alpha}. \tag{24}$$

Since $\sum_{i=1}^n \nabla f_i(\mathbf{x}_{(i)})$ is L_f-Lipschitz, $\nabla\mathcal{L}_\alpha$ is Lipschitz with the constant $L^* \triangleq L_f + \alpha^{-1}\lambda_{\max}(I-W) = L_f + \alpha^{-1}(1-\lambda_n(W))$, implying

$$\begin{aligned}\mathcal{L}_\alpha(\mathbf{x}^{k+1}) \le \mathcal{L}_\alpha(\mathbf{x}^k) + \langle\nabla\mathcal{L}_\alpha(\mathbf{x}^k), \mathbf{x}^{k+1}-\mathbf{x}^k\rangle\\ + \frac{L^*}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2.\end{aligned} \tag{25}$$

Combining (24) and (25) yields (23). ■

Lemma 3 (Boundedness): Under Assumptions 1 and 2, if $0<\alpha<\frac{1+\lambda_n(W)}{L_f}$, then the sequence $\{\mathcal{L}_\alpha(\mathbf{x}^k)\}$ is lower bounded, and the sequence $\{\mathbf{x}^k\}$ is bounded, i.e., there exists a constant $\mathcal{B}>0$ such that $\|\mathbf{x}^k\|<\mathcal{B}$ for all k.

Proof: The lower boundedness of $\mathcal{L}_\alpha(\mathbf{x}^k)$ is due to the lower boundedness of each f_i as it is proper and coercive (Assumption 1 Part (2)).

By Lemma 2 and the choice of α, $\mathcal{L}_\alpha(\mathbf{x}^k)$ is nonincreasing and upper bounded by $\mathcal{L}_\alpha(\mathbf{x}^0)<+\infty$. Hence, $\mathbf{1}^T\mathbf{f}(\mathbf{x}^k)\le \mathcal{L}_\alpha(\mathbf{x}^0)$ implies that $\mathbf{x}^k$ is bounded due to the coercivity of $\mathbf{1}^T\mathbf{f}(\mathbf{x})$ (Assumption 1 Part (2)). ■

From Lemmas 2 and 3, we immediately obtain the following lemma.

Lemma 4 (ℓ_2^2-summable and asymptotic regularity[‖]): It holds that $\sum_{k=0}^\infty \|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2<+\infty$ and that $\|\mathbf{x}^{k+1}-\mathbf{x}^k\|\to 0$ as $k\to\infty$.

From (22), the result below directly follows:

Lemma 5 (Gradient bound): $\|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\| \le \alpha^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|$.

Based on the above lemmas, we get the global convergence of DGD.

Proof of Theorem 1: By Lemma 3, the sequence $\{\mathbf{x}^k\}$ is bounded, so there exist a convergent subsequence and a limit point, denoted by $\{\mathbf{x}^{k_s}\}_{s\in\mathbb{N}}\to\mathbf{x}^*$ as $s\to+\infty$. By Lemmas 2 and 3, $\mathcal{L}_\alpha(\mathbf{x}^k)$ is monotonically nonincreasing and lower bounded, and therefore $\mathcal{L}_\alpha(\mathbf{x}^k)\to\mathcal{L}^*$ for some $\mathcal{L}^*$ and $\|\mathbf{x}^{k+1}-\mathbf{x}^k\|\to 0$ as $k\to\infty$. Based on Lemma 5, $\|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\|\to 0$ as $k\to\infty$. In particular, $\|\nabla\mathcal{L}_\alpha(\mathbf{x}^{k_s})\|\to 0$ as $s\to\infty$. Hence, we have $\nabla\mathcal{L}_\alpha(\mathbf{x}^*)=0$.

The running best rate of the sequence $\{\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\}$ follows from [13, Lemma 1.2] or [26, Th. 3.3.1]. By Lemma 5, the running best rate of the sequence $\{\|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\|^2\}$ is $o(\frac{1}{k})$.

[‖] A sequence $\{a_k\}$ is said to be asymptotic regular if $\|a_{k+1}-a_k\|\to 0$ as $k\to\infty$.

By (11), $\nabla\mathcal{L}_\alpha(\mathbf{x}^k) = \nabla\mathbf{f}(\mathbf{x}^k) + \alpha^{-1}(I-W)\mathbf{x}^k$, which implies $\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k) = \frac{1}{n}\mathbf{1}^T\nabla\mathcal{L}_\alpha(\mathbf{x}^k)$ due to $\frac{1}{n}\mathbf{1}^T(I-W)=0$. Thus,

$$\left\|\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k)\right\|^2 = \left\|\frac{1}{n}\mathbf{1}^T\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\right\|^2 \le \|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\|^2,$$

which implies the running best rate of $\{\|\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k)\|^2\}$ is also $o(\frac{1}{k})$.

By Lemmas 2 and 5, it holds

$$\|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\|^2 \le \frac{2}{\alpha(1+\lambda_n(W)-\alpha L_f)}(\mathcal{L}_\alpha(\mathbf{x}^k)-\mathcal{L}_\alpha(\mathbf{x}^{k+1})),$$

which implies

$$\frac{1}{K}\sum_{k=0}^{K-1}\|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\|^2 \le \frac{2(\mathcal{L}_\alpha(\mathbf{x}^0)-\mathcal{L}^*)}{\alpha(1+\lambda_n(W)-\alpha L_f)K}.$$

Moreover, note that $\|\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k)\|^2 \le \|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\|^2$. Thus, the convergence rate of $\{\frac{1}{K}\sum_{k=0}^{K-1}\|\frac{1}{n}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k)\|^2\}$ is $\mathcal{O}(\frac{1}{K})$.

Similar to [2, Th. 2.9], we can claim the global convergence of the considered sequence $\{\mathbf{x}^k\}_{k\in\mathbb{N}}$ under the KŁ assumption of $\mathcal{L}_\alpha$. ■

Next, we derive a bound on the gradient sequence $\{\nabla\mathbf{f}(\mathbf{x}^k)\}$, which is used in Proposition 1.

Lemma 6: Under Assumption 1, there exists a point $\mathbf{y}^*$ satisfying $\nabla\mathbf{f}(\mathbf{y}^*)=0$, and the following bound holds

$$\|\nabla\mathbf{f}(\mathbf{x}^k)\| \le D \triangleq L_f(\mathcal{B}+\|\mathbf{y}^*\|), \quad \forall k\in\mathbb{N}, \tag{26}$$

where $\mathcal{B}$ is the bound of $\|\mathbf{x}^k\|$ given in Lemma 3.

Proof: By the lower boundedness assumption (Assumption 1 Part (2)), the minimizer of $\mathbf{1}^T\mathbf{f}(\mathbf{y})$ exists. Let $\mathbf{y}^*$ be a minimizer. Then by Lipschitz differentiability of each f_i (Assumption 1 Part (1)), we have that $\nabla\mathbf{f}(\mathbf{y}^*)=0$.

Then, for any k, we have

$$\begin{aligned}\|\nabla\mathbf{f}(\mathbf{x}^k)\| &= \|\nabla\mathbf{f}(\mathbf{x}^k)-\nabla\mathbf{f}(\mathbf{y}^*)\| \le L_f\|\mathbf{x}^k-\mathbf{y}^*\| \\ \text{(Lemma 3)} \quad &\le L_f(\mathcal{B}+\|\mathbf{y}^*\|).\end{aligned}$$

Therefore, we have proven this lemma. ■

B. Proof for Proposition 2

Proof: Note that

$$\begin{aligned}\|\nabla\mathcal{L}_\alpha(\mathbf{x}^{k+1})\| &\le \|\nabla\mathcal{L}_\alpha(\mathbf{x}^{k+1})-\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\| + \|\nabla\mathcal{L}_\alpha(\mathbf{x}^k)\| \\ &\le L^*\|\mathbf{x}^{k+1}-\mathbf{x}^k\| + \alpha^{-1}\|\mathbf{x}^{k+1}-\mathbf{x}^k\| \\ &= (\alpha^{-1}(2-\lambda_n(W))+L_f)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|,\end{aligned}$$

where the second inequality holds for Lemma 5 and the Lipschitz continuity of $\nabla\mathcal{L}_\alpha$ with constant $L^* = L_f + \alpha^{-1}(1-\lambda_n(W))$. Thus, it shows that $\{\mathbf{x}^k\}$ satisfies the so-called relative error condition as list in [2]. Moreover, by Lemmas 2 and 3, $\{\mathbf{x}^k\}$ also satisfies the so-called sufficient decrease and continuity conditions as listed in [2]. Under such three conditions and the KŁ property of $\mathcal{L}_\alpha$ at $\mathbf{x}^*$ with $\psi(s)=cs^{1-\theta}$, following the proof of [2, Lemma 2.6], there exists $k_0>0$ such that for all $k\ge k_0$, we have

$$\begin{aligned}2\|\mathbf{x}^{k+1}-\mathbf{x}^k\| &\le \|\mathbf{x}^k-\mathbf{x}^{k-1}\| + \frac{cb}{a} \\ &\times\left((\mathcal{L}_\alpha(\mathbf{x}^k)-\mathcal{L}_\alpha(\mathbf{x}^*))^{1-\theta} - (\mathcal{L}_\alpha(\mathbf{x}^{k+1})-\mathcal{L}_\alpha(\mathbf{x}^*))^{1-\theta}\right),\end{aligned} \tag{27}$$

where $a \triangleq \frac{1}{2}(\alpha^{-1}(1+\lambda_n(W))-L_f)$ and $b\triangleq\alpha^{-1}(2-\lambda_n(W)) + L_f$. Then, an easy induction yields

$$\begin{aligned}\sum_{t=k_0}^{k}\|\mathbf{x}^{t+1}-\mathbf{x}^t\| &\le \|\mathbf{x}^{k_0}-\mathbf{x}^{k_0-1}\| + \frac{cb}{a} \\ &\times\left((\mathcal{L}_\alpha(\mathbf{x}^{k_0})-\mathcal{L}_\alpha(\mathbf{x}^*))^{1-\theta} - (\mathcal{L}_\alpha(\mathbf{x}^{k+1})-\mathcal{L}_\alpha(\mathbf{x}^*))^{1-\theta}\right).\end{aligned}$$

Following a derivation similar to the proof of [1, Th. 5], we can estimate the rate of convergence of $\{\mathbf{x}^k\}$ in the different cases of θ. ■

C. Proof for Proposition 3

In order to prove Proposition 3, we also need the following lemmas.

Lemma 7: ([40, Proposition 1]) Let $W^k \triangleq \overbrace{W\cdots W}^{k}$ be the power of W with degree k for any $k\in\mathbb{N}$. Under Assumption 2, it holds

$$\left\|W^k - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right\| \le C\zeta^k \tag{28}$$

for some constant $C>0$, where ζ is the second largest magnitude eigenvalue of W as specified in (10).

Lemma 8: ([48, Lemma 3.1]) Let $\{\gamma_k\}$ be a scalar sequence. If $\lim_{k\to\infty}\gamma_k=\gamma$ and $0<\beta<1$, then $\lim_{k\to\infty}\sum_{l=0}^{k}\beta^{k-l}\gamma_l = \frac{\gamma}{1-\beta}$.

Proof of Proposition 3: By the recursion (17), note that

$$\begin{aligned}\mathbf{x}^k-\bar{\mathbf{x}}^k = &\left(W^k-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbf{x}^0 \\ &-\sum_{j=0}^{k-1}\alpha_j\left(W^{k-1-j}-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\nabla\mathbf{f}(\mathbf{x}^j).\end{aligned} \tag{29}$$

Further by Lemma 7 and Assumption 3, we obtain

$$\begin{aligned}\|\mathbf{x}^k-\bar{\mathbf{x}}^k\| &\le \left\|\left(W^k-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right\|\|\mathbf{x}^0\| \\ &\quad+\sum_{j=0}^{k-1}\alpha_j\left\|W^{k-1-j}-\frac{1}{n}\mathbf{1}\mathbf{1}^T\right\|\cdot\|\nabla\mathbf{f}(\mathbf{x}^j)\| \\ &\le C\left(\|\mathbf{x}^0\|\zeta^k + B\sum_{j=0}^{k-1}\alpha_j\zeta^{k-1-j}\right).\end{aligned} \tag{30}$$

Furthermore, by Lemma 8 and step sizes (16), we get $\lim_{k\to\infty}\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|=0$.

Let $b_k \triangleq (k+1)^{-\epsilon}$. To show the rate of $\|\mathbf{x}^k-\bar{\mathbf{x}}^k\|$, we only need to show that

$$\lim_{k\to\infty} b_k^{-1}\|\mathbf{x}^k-\bar{\mathbf{x}}^k\| \le C^*$$

for some $0 < C^* < \infty$. Let $j'_k \triangleq [k-1+2\log_\zeta(b_k^{-1})]$ (where $[x]$ denotes the integer part of x for any $x \in \mathbb{R}$). Note that

$$\begin{aligned} & b_k^{-1}\|\mathbf{x}^k - \bar{\mathbf{x}}^k\| \\ & \le Cb_k^{-1}\left(\|\mathbf{x}^0\|\zeta^k + B\sum_{j=0}^{k-1}\alpha_j\zeta^{k-1-j}\right) \\ & = C\|\mathbf{x}^0\|b_k^{-1}\zeta^k + CBb_k^{-1}\sum_{j=0}^{j'_k}\alpha_j\zeta^{k-1-j} \\ & \quad + CBb_k^{-1}\sum_{j=j'_k+1}^{k-1}\alpha_j\zeta^{k-1-j} \\ & \triangleq T_1 + T_2 + T_3, \end{aligned} \tag{31}$$

where the first inequality holds because of (30).

In the following, we will estimate the above three terms in the right-hand side of (31), respectively. First, by the definition of j'_k, for any $j \le j'_k$, we have $b_k^{-1}\zeta^{\frac{k-1-j}{2}} \le b_k^{-1}\zeta^{\frac{k-1-j'_k}{2}} \le 1$. Thus,

$$T_2 \le CB\sum_{j=0}^{j'_k}\alpha_j\zeta^{(k-1-j)/2}. \tag{32}$$

Second, for $j'_k < j \le k-1$,

$$b_k^{-1}\alpha_j \le \frac{(k+1)^\epsilon}{L_f(j'_k+1)^\epsilon} \le \frac{(k+1)^\epsilon}{L_f(k-1+2\epsilon\log_\zeta(k+1))^\epsilon},$$

and also $b_k^{-1}\alpha_j \ge \frac{(k+1)^\epsilon}{L_f(k+1)^\epsilon} = \frac{1}{L_f}$. Thus, for any $j'_k < j \le k-1$, $\lim_{k\to\infty} b_k^{-1}\alpha_j = \frac{1}{L_f}$. Furthermore, note that $\lim_{k\to\infty} b_k^{-1}\zeta^{k/2} = 0$. Therefore, there exists a k^* such that for $k \ge k^*$, $b_k^{-1}\alpha_j \le \frac{2}{L_f}$ and $b_k^{-1}\zeta^{k/2} \le 1$. The above two inequalities imply that for sufficiently large k,

$$T_1 \le C\|\mathbf{x}^0\|\zeta^{k/2}, \quad T_3 \le \frac{2CB}{L_f}\sum_{j=j'_k+1}^{k-1}\zeta^{k-1-j}. \tag{33}$$

From (32) and (33), we get

$$\begin{aligned} b_k^{-1}\|\mathbf{x}^k - \bar{\mathbf{x}}^k\| &\le C\|\mathbf{x}^0\|\zeta^{k/2} \\ &\quad + CB\left(\sum_{j=0}^{j'_k}\alpha_j\zeta^{(k-1-j)/2} + \frac{2}{L_f}\sum_{j=j'_k+1}^{k-1}\zeta^{k-1-j}\right). \end{aligned} \tag{34}$$

By Lemma 8 and (34), there exists a $C^* > 0$ such that

$$\lim_{k\to\infty} b_k^{-1}\|\mathbf{x}^k - \bar{\mathbf{x}}^k\| \le C^*. \tag{35}$$

We have completed the proof of this proposition. ■

D. Proof for Theorem 2

To prove Theorem 2, we first note that similar to (22), the DGD iterates under decreasing step sizes can be rewritten as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k\nabla\mathcal{L}_{\alpha_k}(\mathbf{x}^k), \tag{36}$$

where $\mathcal{L}_{\alpha_k}(\mathbf{x}) = \mathbf{1}^T\mathbf{f}(\mathbf{x}) + \frac{1}{2\alpha_k}\|\mathbf{x}\|^2_{I-W}$, and we also need the following lemmas.

Lemma 9 ([50]): Let $\{v_t\}$ be a nonnegative scalar sequence such that

$$v_{t+1} \le (1+b_t)v_t - u_t + c_t$$

for all $t \in \mathbb{N}$, where $b_t \ge 0$, $u_t \ge 0$ and $c_t \ge 0$ with $\sum_{t=0}^\infty b_t < \infty$ and $\sum_{t=0}^\infty c_t < \infty$. Then the sequence $\{v_t\}$ converges to some $v \ge 0$ and $\sum_{t=0}^\infty u_t < \infty$.

Lemma 10: Let α_k satisfy (16). Then it holds

$$\alpha_{k+1}^{-1} - \alpha_k^{-1} \le 2\epsilon L_f(k+1)^{\epsilon-1}.$$

Proof: We first prove that

$$(1+x)^\epsilon - 1 \le 2\epsilon x, \quad \forall x \in [0,1]. \tag{37}$$

Let $g(x) = (1+x)^\epsilon - 1 - 2\epsilon x$. Then its derivative

$$g'(x) = \epsilon(1+x)^{\epsilon-1} - 2\epsilon < 0, \quad \forall x \in [0,1].$$

It implies $g(x) \le g(0) = 0$ for any $x \in [0,1]$, that is, the inequality (37) holds.

Note that

$$\begin{aligned} \alpha_{k+1}^{-1} - \alpha_k^{-1} &= L_f\left((k+2)^\epsilon - (k+1)^\epsilon\right) \\ &= L_f(k+1)^\epsilon\left(\left(1+\frac{1}{k+1}\right)^\epsilon - 1\right) \\ &\le 2\epsilon L_f(k+1)^{\epsilon-1}, \end{aligned} \tag{38}$$

where the last inequality holds for (37). ■

The following shows that $\{(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W}\}$ is summable.

Lemma 11: Let Assumptions 1, 2, and 3 hold. In DGD, use step sizes α_k in (16). Then $\{(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W}\}$ is summable, i.e., $\sum_{k=0}^\infty(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W} < \infty$.

Proof: Note that

$$\begin{aligned} \|\mathbf{x}^{k+1}\|^2_{I-W} &= \|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\|^2_{I-W} \\ &\le (1-\lambda_n(W))\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\|^2. \end{aligned} \tag{39}$$

By Lemma 10,

$$\begin{aligned} &(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W} \le 2\epsilon L_f(k+1)^{\epsilon-1}\|\mathbf{x}^{k+1}\|^2_{I-W} \\ &\le 2\epsilon L_f(k+1)^{\epsilon-1}(1-\lambda_n(W))\|\mathbf{x}^{k+1} - \bar{\mathbf{x}}^{k+1}\|^2. \end{aligned} \tag{40}$$

Furthermore, by (40) and Proposition 3, the sequence $\{(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W}\}$ converges to 0 at the rate of $\mathcal{O}(1/(k+1)^{1+\epsilon})$, which implies that the sequence $\{(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W}\}$ is ℓ_1-summable, i.e., $\sum_{k=0}^\infty(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W} < \infty$. ■

Lemma 12 (convergence of weakly summable sequence): Let $\{\beta_k\}$ and $\{\gamma_k\}$ be two nonnegative scalar sequences such that

a) $\gamma_k = \frac{1}{(k+1)^\epsilon}$, for some $\epsilon \in (0,1]$, $k \in \mathbb{N}$;
b) $\sum_{k=0}^\infty \gamma_k\beta_k < \infty$;
c) $|\beta_{k+1} - \beta_k| \lesssim \gamma_k$,

where "$\lesssim$"means that $|\beta_{k+1} - \beta_k| \le M\gamma_k$ for some constant $M > 0$, then $\lim_{k\to\infty}\beta_k \to 0$.

We call a sequence $\{\beta_k\}$ satisfying Lemma 12 (a) and (b) a *weakly summable* sequence since itself is not necessarily

summable but becomes summable via multiplying another non-summable, diminishing sequence $\{\gamma_k\}$. It is generally impossible to claim that β_k converges to 0. However, if the distance of two successive steps of $\{\beta_k\}$ with the same order of the multiplied sequence γ_k, then we can claim the convergence of β_k. A special case with $\epsilon = 1/2$ has been observed in [12].

Proof: By condition (b), we have

$$\sum_{i=k}^{k+k'} \gamma_i \beta_i \to 0, \tag{41}$$

as $k \to \infty$ and for any $k' \in \mathbb{N}$.

In the following, we will show $\lim_{k\to\infty} \beta_k = 0$ by contradiction. Assume this is not the case, i.e., $\beta_k \nrightarrow 0$ as $k \to \infty$, then $\limsup_{k\to\infty} \beta_k \triangleq C^* > 0$. Thus, for every $N > k_0$, there exists a $k > N$ such that $\beta_k > \frac{C^*}{2}$. Let

$$k' \triangleq \left[\frac{C^*}{4M}(k+1)^{\epsilon}\right],$$

where $[x]$ denotes the integer part of x for any $x \in \mathbb{R}$. By condition (c), i.e., $|\beta_{j+1} - \beta_j| \leq M\gamma_j$ for any $j \in \mathbb{N}$, then

$$\beta_{k+i} \geq \frac{C^*}{4}, \quad \forall i \in \{0, 1, \ldots, k'\}. \tag{42}$$

Hence,

$$\sum_{j=k}^{k+k'} \gamma_j \beta_j \geq \frac{C^*}{4} \sum_{j=k}^{k+k'} \gamma_j \geq \frac{C^*}{4} \int_k^{k+k'} (x+1)^{-\epsilon} dx \tag{43}$$

$$= \begin{cases} \frac{C^*}{4(1-\epsilon)} \left((k+k'+1)^{1-\epsilon} - (k+1)^{1-\epsilon}\right), & \epsilon \in (0,1), \\ \frac{C^*}{4} \left(\ln(k+k'+1) - \ln(k+1)\right), & \epsilon = 1. \end{cases}$$

Note that when $\epsilon \in (0,1)$, the term $(k+k'+1)^{1-\epsilon} - (k+1)^{1-\epsilon}$ is monotonically increasing with respect to k, which implies that $\sum_{j=k}^{k+k'} \gamma_j \beta_j$ is lower bounded by a positive constant when $\epsilon \in (0,1)$. While when $\epsilon = 1$, noting that the specific form of k', we have

$$\ln(k+k'+1) - \ln(k+1) = \ln\left(1 + \frac{k'}{k+1}\right) = \ln\left(1 + \frac{C^*}{4M}\right),$$

which is a positive constant. As a consequence, $\sum_{j=k}^{k+k'} \gamma_j \beta_j$ will not go to 0 as $k \to 0$, which contradicts with (41). Therefore, $\lim_{k\to\infty} \beta_k = 0$. ∎

Proof of Theorem 2: We first develop the following inequality

$$\mathcal{L}_{\alpha_{k+1}}(\mathbf{x}^{k+1}) \leq \mathcal{L}_{\alpha_k}(\mathbf{x}^k) + \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|_{I-W}^2$$
$$- \frac{1}{2}\big(\alpha_k^{-1}(1 + \lambda_n(W)) - L_f\big)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2, \tag{44}$$

and then claim the convergence of the sequences $\{\mathcal{L}_{\alpha_k}(\mathbf{x}^k)\}$, $\{\mathbf{1}^T\mathbf{f}(\mathbf{x}^k)\}$ and $\{\mathbf{x}^k\}$ based on this inequality.

a) Development of (44): From $\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla\mathcal{L}_{\alpha_k}(\mathbf{x}^k)$, it follows that

$$\langle \nabla\mathcal{L}_{\alpha_k}(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle = -\frac{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2}{\alpha_k}. \tag{45}$$

Since $\sum_{i=1}^n \nabla f_i(\mathbf{x}_{(i)})$ is L_f-Lipschitz, $\nabla\mathcal{L}_{\alpha_k}$ is Lipschitz with the constant $L_k \triangleq L_f + \alpha_k^{-1}\lambda_{\max}(I - W) = L_f + \alpha_k^{-1}(1 - \lambda_n(W))$, implying

$$\mathcal{L}_{\alpha_k}(\mathbf{x}^{k+1})$$
$$\leq \mathcal{L}_{\alpha_k}(\mathbf{x}^k) + \langle \nabla\mathcal{L}_{\alpha_k}(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_k}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2$$
$$= \mathcal{L}_{\alpha_k}(\mathbf{x}^k) - \frac{1}{2}\big(\alpha_k^{-1}(1 + \lambda_n(W)) - L_f\big)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2. \tag{46}$$

Moreover,

$$\mathcal{L}_{\alpha_{k+1}}(\mathbf{x}^{k+1})$$
$$= \mathcal{L}_{\alpha_k}(\mathbf{x}^{k+1}) + \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|_{I-W}^2. \tag{47}$$

Combining (46) and (47) yields (44).

b) Convergence of objective sequence: By Lemma 11 and Lemma 9, (44) yields the convergence of $\{\mathcal{L}_{\alpha_k}(\mathbf{x}^k)\}$ and

$$\sum_{k=0}^{\infty} \big(\alpha_k^{-1}(1 + \lambda_n(W)) - L_f\big)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 < \infty \tag{48}$$

which implies that $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2$ converges to 0 at the rate of $o(k^{-\epsilon})$ and $\{\mathbf{x}^k\}$ is asymptotic regular Moreover, notice that

$$\alpha_k^{-1}\|\mathbf{x}^k\|_{I-W}^2 = \alpha_k^{-1}\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|_{I-W}^2$$
$$\leq (1 - \lambda_n(W))L_f(k+1)^{\epsilon}\|\mathbf{x}^k - \bar{\mathbf{x}}^k\|^2.$$

By Proposition 3, the term $\alpha_k^{-1}\|\mathbf{x}^k\|_{I-W}^2$ converges to 0 as $k \to \infty$. As a consequence,

$$\lim_{k\to\infty} \mathbf{1}^T\mathbf{f}(\mathbf{x}^k) = \lim_{k\to\infty} \left(\mathcal{L}_{\alpha_k}(\mathbf{x}^k) - \frac{\|\mathbf{x}^k\|_{I-W}^2}{2\alpha_k}\right)$$
$$= \lim_{k\to\infty} \mathcal{L}_{\alpha_k}(\mathbf{x}^k).$$

c) Convergence to a stationary point: Let $\bar{\nabla}\mathbf{f}(\mathbf{x}^k) \triangleq \frac{1}{n}\mathbf{1}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k)$. By the specific form (16) of α_k, we have

$$\alpha_k^{-1}(1 + \lambda_n(W)) - L_f$$
$$= \alpha_k^{-1}(1 + \lambda_n(W) - L_f\alpha_k)$$
$$\geq \alpha_k^{-1}\left(1 + \lambda_n(W) - \frac{1}{(k_0+1)^{\epsilon}}\right) \tag{49}$$

for all $k > k_0$, where $k_0 = [(1 + \lambda_n(W))^{-\frac{1}{\epsilon}}]$, i.e., the integer part of $(1 + \lambda_n(W))^{-\frac{1}{\epsilon}}$. Note that

$$\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\| = \left\|\frac{1}{n}\mathbf{1}\mathbf{1}^T(\mathbf{x}^{k+1} - \mathbf{x}^k)\right\|$$
$$\leq \|\mathbf{x}^{k+1} - \mathbf{x}^k\|. \tag{50}$$

Thus, (48), (49) and (50) yield

$$\sum_{k=0}^{\infty} \alpha_k^{-1}\|\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k\|^2 < \infty. \tag{51}$$

By the iterate (5) of DGD, we have

$$\bar{\mathbf{x}}^{k+1} - \bar{\mathbf{x}}^k = -\alpha_k\bar{\nabla}\mathbf{f}(\mathbf{x}^k). \tag{52}$$

Plugging (52) into (51) yields

$$\sum_{k=0}^{\infty}\alpha_k\|\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|^2<\infty. \tag{53}$$

Moreover,

$$\begin{aligned}&|\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})\|^2-\|\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|^2|\\&\le\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})-\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|\cdot(\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})\|+\|\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|)\\&\le 2B\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})-\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|\\&\le 2B\|\nabla\mathbf{f}(\mathbf{x}^{k+1})-\nabla\mathbf{f}(\mathbf{x}^k)\|\\&\le 2BL_f\|\mathbf{x}^{k+1}-\mathbf{x}^k\|,\end{aligned} \tag{54}$$

where the second inequality holds by the bounded gradient assumption (Assumption 3), the third inequality holds by the specific form of $\bar{\nabla}\mathbf{f}(\mathbf{x}^k)$, and the last inequality holds by the Lipschitz continuity of $\nabla\mathbf{f}$. Note that

$$\begin{aligned}&\|\mathbf{x}^{k+1}-\mathbf{x}^k\|\\&=\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}+\bar{\mathbf{x}}^{k+1}-\bar{\mathbf{x}}^k+\bar{\mathbf{x}}^k-\mathbf{x}^k\|\\&\le\|\mathbf{x}^{k+1}-\bar{\mathbf{x}}^{k+1}\|+\|\bar{\mathbf{x}}^k-\mathbf{x}^k\|+\alpha_k\|\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|\\&\lesssim\alpha_k,\end{aligned} \tag{55}$$

where the first inequality holds for the triangle inequality and (52), and the last inequality holds for Proposition 3 and the bounded assumption of $\nabla\mathbf{f}$. Thus, (54) and (55) imply

$$|\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})\|^2-\|\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|^2|\lesssim\alpha_k. \tag{56}$$

By the specific form (16) of α_k, (53), (56) and Lemma 12, it holds

$$\lim_{k\to\infty}\|\bar{\nabla}\mathbf{f}(\mathbf{x}^k)\|^2=0. \tag{57}$$

As a consequence,

$$\lim_{k\to\infty}\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^k)=0. \tag{58}$$

Furthermore, by the coercivity of f_i for each i and the convergence of $\{\mathbf{1}^T\mathbf{f}(\mathbf{x}^k)\}$, $\{\mathbf{x}^k\}$ is bounded. Therefore, there exists a convergent subsequence of $\{\mathbf{x}^k\}$. Let $\mathbf{x}^*$ be any limit point of $\{\mathbf{x}^k\}$. By (57) and the continuity of $\nabla\mathbf{f}$, it holds

$$\mathbf{1}^T\nabla\mathbf{f}(\mathbf{x}^*)=0.$$

Moreover, by Proposition 3, $\mathbf{x}^*$ is consensual. As a consequence, $\mathbf{x}^*$ is a stationary point of problem (3).

In addition, if $\mathbf{x}^*$ is isolated, then by the asymptotic regularity of $\{\mathbf{x}^k\}$ (Lemma 4), $\{\mathbf{x}^k\}$ converges to $\mathbf{x}^*$ [44]. ■

E. Proofs for Theorem 3 and Proposition 5

In order to prove Theorem 3, we need the following lemmas.

Lemma 13: ([10, Proposition 3]) Let $h:\mathbb{R}^d\to\mathbb{R}$ be a continuously differentiable function whose gradient is Lipschitz continuous with constant L_h. Then for any $x,y,u\in\mathbb{R}^p$,

$$h(u)\ge h(x)+\langle\nabla h(y),u-x\rangle-\frac{L_h}{2}\|x-y\|^2.$$

Lemma 14 (Sufficient descent of $\{\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)\}$): Let Assumptions 2 and 4 hold. Results are given in two cases below:

C1: r_i's are convex. Set $0<\alpha<\frac{1+\lambda_n(W)}{L_f}$.

$$\begin{aligned}&\hat{\mathcal{L}}_\alpha(\mathbf{x}^{k+1})\le\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)\\&-\frac{1}{2}\big(\alpha^{-1}(1+\lambda_n(W))-L_f\big)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2,\forall k\in\mathbb{N}.\end{aligned} \tag{59}$$

C2: r_i's are not necessarily convex (in this case, we assume $\lambda_n(W)>0$). Set $0<\alpha<\frac{\lambda_n(W)}{L_f}$.

$$\begin{aligned}&\hat{\mathcal{L}}_\alpha(\mathbf{x}^{k+1})\le\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)\\&-\frac{1}{2}\big(\alpha^{-1}\lambda_n(W)-L_f\big)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2,\forall k\in\mathbb{N}.\end{aligned} \tag{60}$$

Proof: Recall from Lemma 2 that $\nabla\mathcal{L}_\alpha(\mathbf{x})$ is L^*-Lipschitz continuous for $L^*=L_f+\alpha^{-1}(1-\lambda_n(W))$, and thus

$$\begin{aligned}&\hat{\mathcal{L}}_\alpha(\mathbf{x}^{k+1})-\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)\\&=\mathcal{L}_\alpha(\mathbf{x}^{k+1})-\mathcal{L}_\alpha(\mathbf{x}^k)+r(\mathbf{x}^{k+1})-r(\mathbf{x}^k)\\&\le\langle\nabla\mathcal{L}_\alpha(\mathbf{x}^{\mathbf{k}}),\mathbf{x}^{k+1}-\mathbf{x}^k\rangle+\frac{L^*}{2}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\\&+r(\mathbf{x}^{k+1})-r(\mathbf{x}^k).\end{aligned} \tag{61}$$

C1: From the convexity of r, (8), and (20), it follows that

$$0=\xi^{k+1}+\frac{1}{\alpha}\big(\mathbf{x}^{k+1}-\mathbf{x}^k+\alpha\nabla\mathcal{L}_\alpha(\mathbf{x}^{\mathbf{k}})\big),\ \xi^{k+1}\in\partial r(\mathbf{x}^{k+1}).$$

This and the convexity of r further give us

$$\begin{aligned}&r(\mathbf{x}^{k+1})-r(\mathbf{x}^k)\le\langle\xi^{k+1},\mathbf{x}^{k+1}-\mathbf{x}^k\rangle\\&=-\frac{1}{\alpha}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2-\langle\nabla\mathcal{L}_\alpha(\mathbf{x}^{\mathbf{k}}),\mathbf{x}^{k+1}-\mathbf{x}^k\rangle.\end{aligned}$$

Substituting this inequality into the inequality (61) and then expanding $L^*=L_f+\alpha^{-1}(1-\lambda_n(W))$ yield

$$\begin{aligned}&\hat{\mathcal{L}}_\alpha(\mathbf{x}^{k+1})-\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)\le-\left(\frac{1}{\alpha}-\frac{L^*}{2}\right)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2\\&=-\frac{1}{2}\big(\alpha^{-1}(1+\lambda_n(W))-L_f\big)\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2.\end{aligned}$$

Sufficient descent requires the last term to be negative, thus $0<\alpha<\frac{1+\lambda_n(W)}{L_f}$.

C2: From (8) and (20), it follows that the function $r(\mathbf{u})+\frac{\|\mathbf{u}-(\mathbf{x}^k-\alpha\nabla\mathcal{L}_\alpha(\mathbf{x}^{\mathbf{k}}))\|^2}{2\alpha}$ reaches its minimum at $\mathbf{u}=\mathbf{x}^{k+1}$. Comparing the values of this function at $\mathbf{x}^{k+1}$ and $\mathbf{x}^k$ yields

$$\begin{aligned}&r(\mathbf{x}^{k+1})-r(\mathbf{x}^k)\le\frac{1}{2\alpha}\|\mathbf{x}^k-(\mathbf{x}^k-\alpha\nabla\mathcal{L}_\alpha(\mathbf{x}^{\mathbf{k}}))\|^2\\&-\frac{1}{2\alpha}\|\mathbf{x}^{k+1}-(\mathbf{x}^k-\alpha\nabla\mathcal{L}_\alpha(\mathbf{x}^{\mathbf{k}}))\|^2\\&=-\frac{1}{2\alpha}\|\mathbf{x}^{k+1}-\mathbf{x}^k\|^2-\langle\nabla\mathcal{L}_\alpha(\mathbf{x}^{\mathbf{k}}),\mathbf{x}^{k+1}-\mathbf{x}^k\rangle.\end{aligned}$$

Substituting this inequality into (61) and expanding L^* yield

$$\hat{\mathcal{L}}_\alpha(\mathbf{x}^{k+1}) - \hat{\mathcal{L}}_\alpha(\mathbf{x}^k) \le -\left(\frac{1}{2\alpha} - \frac{L^*}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 = -\frac{1}{2}\big(\alpha^{-1}\lambda_n(W) - L_f\big)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2.$$

Hence, sufficient descent requires $0 < \alpha < \frac{\lambda_n(W)}{L_f}$. ■

Lemma 15 (Boundedness): Under the conditions of Lemma 14, the sequence $\{\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)\}$ is lower bounded, and the sequence $\{\mathbf{x}^k\}$ is bounded.

Proof: The lower boundedness of $\{\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)\}$ is due to Assumption 4 Part (2).

By Lemma 14 and under a proper step size, $\hat{\mathcal{L}}_\alpha(\mathbf{x}^k)$ is nonincreasing and upper bounded by $\hat{\mathcal{L}}_\alpha(\mathbf{x}^0)$. Hence, $\sum_{i=1}^n (f_i(\mathbf{x}^k_{(i)}) + r_i(\mathbf{x}^k_{(i)}))$ is upper bounded by $\hat{\mathcal{L}}_\alpha(\mathbf{x}^0)$. Consequently, $\{\mathbf{x}^k\}$ is bounded due to the coercivity of each $f_i + r_i$ (see Assumption 4 Part (2)). ■

Lemma 16 (Bounded subgradient): Let $\partial\hat{\mathcal{L}}_\alpha(\mathbf{x}^{k+1})$ denote the (limiting) subdifferential of $\hat{\mathcal{L}}_\alpha$, which is assumed to exist for all $k \in \mathbb{N}$. Then, there exists $\mathbf{g}^{k+1} \in \partial\hat{\mathcal{L}}_\alpha(\mathbf{x}^{k+1})$ such that

$$\|\mathbf{g}^{k+1}\| \le (\alpha^{-1}(2 - \lambda_n(W)) + L_f)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|.$$

Proof: By the iterate (20), the following optimality condition holds

$$0 \in \alpha^{-1}(\mathbf{x}^{k+1} - \mathbf{x}^k + \alpha\nabla\mathcal{L}_\alpha(\mathbf{x}^k)) + \partial r(\mathbf{x}^{k+1}), \tag{62}$$

where $\partial r(\mathbf{x}^{k+1})$ denotes the (limiting) subdifferential of r at $\mathbf{x}^{k+1}$. For any $\xi^{k+1} \in \partial r(\mathbf{x}^{k+1})$, it follows from (62) that

$$\nabla\mathcal{L}_\alpha(\mathbf{x}^{k+1}) + \xi^{k+1} = \alpha^{-1}(\mathbf{x}^k - \mathbf{x}^{k+1}) + (\nabla\mathcal{L}_\alpha(\mathbf{x}^{k+1}) - \nabla\mathcal{L}_\alpha(\mathbf{x}^k)),$$

which immediate yields

$$\begin{aligned}&\|\nabla\mathcal{L}_\alpha(\mathbf{x}^{k+1}) + \xi^{k+1}\|\\ &\le \alpha^{-1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\| + \|\nabla\mathcal{L}_\alpha(\mathbf{x}^{k+1}) - \nabla\mathcal{L}_\alpha(\mathbf{x}^k)\|\\ &\le (\alpha^{-1} + L^*)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|\\ &\le (\alpha^{-1}(2 - \lambda_n(W)) + L_f)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|.\end{aligned}$$

Thus, then the claim of Lemma 16 holds. ■

Based on Lemmas 14–16, we can easily prove Theorem 3 and Proposition 5.

Proof of Theorem 3: The proof of this theorem is similar to that of Theorem 1 and thus is omitted. ■

Proof of Proposition 5: The proof is similar to that of Proposition 2. We shall however note that in (27), $a = \frac{1}{2}(\alpha^{-1}(1 + \lambda_n(W)) - L_f)$ if r_i's are convex, while $a = \frac{1}{2}(\alpha^{-1}\lambda_n(W) - L_f)$ if r_i's are not necessarily convex and $\lambda_n(W) > 0$. ■

F. Proofs for Theorem 4 and Proposition 6

Based on the iterate (7) of Prox-DGD, we derive the following recursion of the iterates of Prox-DGD, which is similar to (17).

Lemma 17 (Recursion of $\{\mathbf{x}^k\}$): For any $k \in \mathbb{N}$,

$$\mathbf{x}^k = W^k\mathbf{x}^0 - \sum_{j=0}^{k-1}\alpha_j W^{k-1-j}(\nabla\mathbf{f}(\mathbf{x}^j) + \xi^{j+1}), \tag{63}$$

where $\xi^{j+1} \in \partial r(\mathbf{x}^{j+1})$ is the one determined by the proximal operator (8), for any $j = 0, \ldots, k-1$.

Proof: By the definition of the proximal operator (8), the iterate (7) implies

$$\mathbf{x}^{k+1} + \alpha_k\xi^{k+1} = W\mathbf{x}^k - \alpha_k\nabla\mathbf{f}(\mathbf{x}^k), \tag{64}$$

where $\xi^{k+1} \in \partial r(\mathbf{x}^{k+1})$, and thus

$$\mathbf{x}^{k+1} = W\mathbf{x}^k - \alpha_k(\nabla\mathbf{f}(\mathbf{x}^k) + \xi^{k+1}). \tag{65}$$

By (65), we can easily derive the recursion (63). ■

Proof of Proposition 6: The proof of this proposition is similar to that of Proposition 3. It only needs to note that the subgradient term $\nabla\mathbf{f}(\mathbf{x}^j) + \xi^{j+1}$ is uniformly bounded by the constant $\bar{B}$ for any j. Thus, we omit it here. ■

To prove Theorem 4, we still need the following lemmas.

Lemma 18: Let Assumptions 2 and 4 hold. In Prox-DGD, use the step sizes (16). Results are given in two cases below:

C1: r_i's are convex. For any $k \in \mathbb{N}$,

$$\hat{\mathcal{L}}_{\alpha_{k+1}}(\mathbf{x}^{k+1}) \le \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^k) + \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W} - \frac{1}{2}\big(\alpha_k^{-1}(1 + \lambda_n(W)) - L_f\big)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2. \tag{66}$$

C2: r_i's are not necessarily convex. For any $k \in \mathbb{N}$,

$$\hat{\mathcal{L}}_{\alpha_{k+1}}(\mathbf{x}^{k+1}) \le \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^k) + \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W} - \frac{1}{2}\big(\alpha_k^{-1}\lambda_n(W) - L_f\big)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2. \tag{67}$$

Proof: The proof of this lemma is similar to that of Lemma 14 via noting that

$$\hat{\mathcal{L}}_{\alpha_{k+1}}(\mathbf{x}^{k+1}) = \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^k) + (\hat{\mathcal{L}}_{\alpha_{k+1}}(\mathbf{x}^{k+1}) - \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^{k+1})) + (\hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^{k+1}) - \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^k)),$$

and

$$\hat{\mathcal{L}}_{\alpha_{k+1}}(\mathbf{x}^{k+1}) - \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^{k+1}) = \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|\mathbf{x}^{k+1}\|^2_{I-W}.$$

While the term $\hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^{k+1}) - \hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^k)$ can be estimated similarly by the proof of Lemma 14. ■

Lemma 19: Let Assumptions 2, 4 and 5 hold. In Prox-DGD, use the step sizes (16). If further each f_i and r_i are convex, then for any $\mathbf{u} \in \mathbb{R}^{n\times p}$, we have

$$\hat{\mathcal{L}}_{\alpha_k}(\mathbf{x}^{k+1}) - \hat{\mathcal{L}}_{\alpha_k}(\mathbf{u}) \le \frac{1}{2\alpha_k}(\|\mathbf{x}^k - \mathbf{u}\|^2 - \|\mathbf{x}^{k+1} - \mathbf{u}\|^2).$$

The proof of this lemma can be found in [67] due to page limit.

Proof of Theorem 4: Based on Lemmas 18 and 19, we can prove Theorem 4. The proof of Theorem 4(a)–(d) is similar to that of Theorem 2, where one minor difference is that (54) in

the proof of Theorem 2 is updated as

$$\begin{aligned}
&|\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})+\bar{\xi}^{k+1}\|^2-\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k})+\bar{\xi}^{k}\|^2|\\
&\le \|(\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})+\bar{\xi}^{k+1})-(\bar{\nabla}\mathbf{f}(\mathbf{x}^{k})+\bar{\xi}^{k})\|\\
&\quad\times(\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1}\|+\|\bar{\nabla}\mathbf{f}(\mathbf{x}^{k})+\bar{\xi}^{k}\|)\\
&\le 2\bar{B}\|(\bar{\nabla}\mathbf{f}(\mathbf{x}^{k+1})+\bar{\xi}^{k+1})-(\bar{\nabla}\mathbf{f}(\mathbf{x}^{k})+\bar{\xi}^{k})\|\\
&\le 2\bar{B}\|(\nabla\mathbf{f}(\mathbf{x}^{k+1})+\xi^{k+1})-(\nabla\mathbf{f}(\mathbf{x}^{k})+\xi^{k})\|\\
&\le 2\bar{B}(L_f+L_r)\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|, \qquad (68)
\end{aligned}$$

where $\bar{\xi}^k \triangleq \frac{1}{n}\mathbf{1}\mathbf{1}^T\xi^k$, and the final inequality holds for the Lipschitz assumption on $\{\xi^k\}$ for large k in Theorem 4(c).

The proof of Theorem 4(e) is very similar to that of Proposition 4. ■

VI. Conclusion

In this paper, we study the convergence behavior of the algorithm DGD for smooth, possibly nonconvex consensus optimization. We consider both fixed and decreasing step sizes. When using a fixed step size, we show that the iterates of DGD converge to a stationary point of a Lyapunov function, which approximates to one of the original problem. Moreover, we estimate the bound between each local point and its global average, which is proportional to the step size and inversely proportional to the gap between the largest and the second largest magnitude eigenvalues of the mixing matrix. This motivate us to study the algorithm DGD with decreasing step sizes. When using decreasing step sizes, we show that the iterates of DGD reach consensus asymptotically at a sublinear rate and converge to a stationary point of the original problem. We also estimate the convergence rates of objective sequence in the convex setting using different diminishing step size strategies. Furthermore, we extend these convergence results to Prox-DGD designed for minimizing the sum of a differentiable function and a proximal function. Both functions can be nonconvex. If the proximal function is convex, a larger fixed step size is allowed. These results are obtained by applying both existing and new proof techniques.

References

[1] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Math. Program.*, vol. 116, pp. 5–16, 2009.

[2] H. Attouch, J. Bolte, and B. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods," *Math. Program. A*, vol. 137, pp. 91–129, 2013.

[3] Y. Bengio, Y. LeCun, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[4] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for nonconvex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, Feb. 2013.

[5] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7405–7418, Nov. 2013.

[6] E. Bjornson and E. Jorswieck, "Optimal resource allocation in coordinated multi-cell systems," *Found. Trends Commun. Inf. Theory*, vol. 9, no. 2/3, pp. 113–381, 2012.

[7] J. Bolte, A. Daniilidis, and A. Lewis, "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems," *SIAM J. Optim.*, vol. 17, no. 4, pp. 1205–1223, 2007.

[8] A. Chen and A. Ozdaglar, "A fast distributed proximal gradient method," in *Proc. 50th Allerton Conf. Commun., Control Comput.*, Moticello, IL, Oct. 2012, pp. 601–608.

[9] T. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 482–497, Jan. 2015.

[10] A. Chen, "Fast distributed first-order methods," Master's thesis, Dept. Elect. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, MA, 2012.

[11] X. Chen, F. Xu, and Y. Ye, "Lower bound theory of nonzero entries in solutions of $\ell_2-\ell_p$ minimization," *SIAM J. Sci. Comput.*, vol. 32, no. 5, pp. 2832–2852, 2010.

[12] Y. T. Chow, T. Wu, and W. Yin, "Cyclic coordinate update algorithms for fixed-point problems: Analysis and applications." *SIAM J. Sci. Comput.*, vol. 39, no. 4, pp. A1280–A1300, 2017.

[13] W. Deng, M. Lai, Z. Peng, and W. Yin, "Parallel multi-block ADMM with o(1/k) convergence," *J. Sci. Comput.*, vol. 71, no. 2, pp. 712–736, 2017.

[14] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. Conf. Rec. 39th Asilomar Conf. Signals, Syst. Comput.*, 2005, pp. 1058–6393.

[15] F. Facchinei, G. Scutari, and S. Sagratella, "Parallel selective algorithms for nonconvex big data optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1874–1889, Apr. 2015.

[16] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc., Theory Method*, vol. 96, no. 456, pp. 1348–1360, 2001.

[17] E. Hazan, K. Y. Levy, and S. Shalev-Shwarz, "On graduated optimization for stochastic nonconvex problems," in *Proc. 33rd Int. Conf. Mach. Learning*, New York, NY, USA, 2016, pp. 1833–1841.

[18] D. Hajinezhad, M. Hong, and A. Garcia, "ZENTH: A zeroth-order distributed algorithm for multi-agent nonconvex optimization," arXiv Preprint arXiv:1710.09997, 2017.

[19] D. Hajinezhad, M. Hong, and A. Garcia, *Zeroth Order Nonconvex Multi-Agent Optimization Over Networks*, arXiv:1710.09997, 2017.

[20] M. Hong and Z.-Q. Luo, "Signal processing and optimal resource allocation for the interference channel," in *Academic Press Library in Signal Processing: Communications and Radar Signal Processing*. New York, NY, USA: Academic, 2013, vol. 2, ch. 8, pp. 409–462.

[21] 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 *Proc. 34th Int. Conf. Mach. Learning*, 2017, vol. 70, pp. 1529–1538.

[22] D. Hajinezhad and M. Hong, "Perturbed proximal primal dual algorithm for nonconvex nonsmooth optimization," Tech. Rep, 2017. [Online]. Available: http://people.ece.umn.edu/ mhong/PProx_PDA.pdf

[23] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed optimization on dynamic networks, *IEEE Trans. Autom. Control*, vol. 61, no. 11, pp. 3545–3550, Nov. 2016.

[24] D. Jakovetic, J. Xavier, and J. Moura, "Fast distributed gradient methods," *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1131–1146, May 2014.

[25] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *Proc. 44th Annu. IEEE Symp. Found. Comput. Sci.*, 2003, pp. 482–491.

[26] K. Knopp, *Infinite Sequences and Series*. North Chelmsford, MA, USA: Courier Corporation, 1956.

[27] J. Lafond, H. T. Wai, and E. Moulines, "D-FW: Communication efficient distributed algorithms for high-dimensional sparse optimization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, 2016, pp. 4144–4148, doi: 10.1109/ICASSP.2016.7472457.

[28] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," *Math. Programming*, vol. 156, no. 1/2, pp. 59–99, 2016.

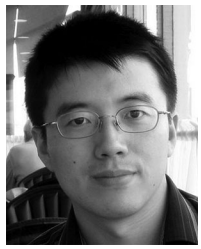
[29] S. Lee and A. Nedic, "Distributed random projection algorithm for convex optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 221–229, Apr. 2013.

[30] Q. Ling and Z. Tian, "Decentralized sparse signal recovery for compressive sleeping wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3816–3827, Jul. 2010.

[31] Q. Ling, W. Shi, G. Wu, and A. Ribeiro, "DLM: Decentralized linearized alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 4051–4064, Aug. 2015.

[32] X. Lian, Y. Huang, Y. Li, and J. Liu, "Asynchronous parallel stochastic gradient for nonconvex optimization," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, vol. 2, pp. 2737–2745.
[33] 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 *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017.
[34] S. Łojasiewicz, "Sur la géométrie semi-et sous-analytique," *Ann. Inst. Fourier (Grenoble)* vol. 43, no. 5, pp. 1575–1595, 1993.
[35] P. D. Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 2, no. 2, pp. 120–136, Jun. 2016.
[36] P. D. Lorenzo and G. Scutari, "Distributed nonconvex optimization over time-varying networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016.
[37] I. Matei and J. Baras, "Performance evaluation of the consensus-based distributed subgradient method under random communication topologies," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 754–771, Aug. 2011.
[38] H. McMahan and M. Streeter, "Delay-Tolerant algorithms for asynchronous distributed online learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, Canada, 2014.
[39] G. Mateos, J. Bazerque, and G. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.
[40] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.
[41] A. Nedic and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 60, no. 3, pp. 601–615, Mar. 2015.
[42] M. Nevelson and R. Z. Khasminskii, *Stochastic Approximation and Recursive Estimation* [translated from the Russian by Israel Program for Scientific Translations; translation edited by B. Silver]. Providence, NY, USA: Ame. Math. Soc., 1973.
[43] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proc. 34th Int. Conf. Machine Learning*, 2017, vol. 70, pp. 2681–2690.
[44] A. M. Ostrowski, *Solution of Equations in Euclidean and Banach Spaces*. New York, NY, USA: Academic, 1973.
[45] Stacy Patterson, Yonina C. Eldar, and I. Keidar, "Distributed compressed sensing for static and time-varying networks," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 4931–4946, Oct. 2014.
[46] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, to be published.
[47] M. Raginsky, N. Kiarashi, and R. Willett, "Decentralized online convex programming with local information," in *Proc. Amer. Control Conf.*, San Francisco, CA, USA, 2011, pp. 5363–5369, doi: 10.1109/ACC.2011.5991212.
[48] S. Ram, A. Nedic, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, pp. 516–545, 2010.
[49] C. Ravazzi, S. M. Fosson, and E. Magli, "Distributed iterative thresholding for ℓ_0/ℓ_1-regularized linear inverse problems," *IEEE Trans. Signal Process.*, vol. 61, no. 4, pp. 2081–2100, Apr. 2015.
[50] H. Robbins and D. Siegmund, "A convergence theorem for nonnegative almost supermartingales and some applications," in *Proc. Optim, Methods Statist.*, 1971, pp. 233–257.
[51] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in ad hoc WSNs with noisy links—Part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, Jan. 2008.
[52] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, Apr. 2014.
[53] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015.
[54] W. Shi, Q. Ling, G. Wu, and W. Yin, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, Nov. 2015.
[55] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, Feb. 2014.
[56] T. Tatarenko and B. Touri, "On local analysis of distributed optimization," in *Proc. Amer. Control Conf.*, Boston, MA, 2016, pp. 5626–5631, doi: 10.1109/ACC.2016.7526552.
[57] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 3744–3757, Aug. 2017.
[58] G. Tychogiorgos, A. Gkelias, and K. K. Leung, "A non-convex distributed optimization framework and its application to wireless ad-hoc networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4286–4296, Sep. 2013.
[59] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. AC-31, no. 9, pp. 803–812, Sep. 1986.
[60] H. T. Wai, T. H. Chang, and A. Scaglione, "A consensus-based decentralized algorithm for nonconvex optimization with application to dictionary learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, South Brisbane, QLD, 2015, pp. 3546–3550, doi: 10.1109/ICASSP.2015.7178631.
[61] H. Wai and A. Scaglione, "Consensus on state and time: Decentralized regression with asynchronous sampling," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2972–2985, Jun. 2015.
[62] H. Wai, A. Scaglione, J. Lafond, and E. Moulines, "Decentralized Frank-Wolfe algorithm for convex and nonconvex problems," *IEEE Trans. Automat. Control*, vol. 62, no. 11, pp. 5522–5537, Nov. 2017.
[63] Y. Xu and W. Yin, "A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion," *SIAM J. Imag. Sci.*, vol. 6, pp. 1758–1789, 2013.
[64] F. Yan, S. Sundaram, S. Vishwanathan, and Y. Qi, "Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties," *IEEE T Knowl. Data Eng.*, vol. 25, no. 11, pp. 2483–2493, Nov. 2013.
[65] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.
[66] J. Zeng, S. Lin, and Z. Xu, "Sparse regularization: Convergence of iterative jumping thresholding algorithm," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5160–5118, Oct. 2016.
[67] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," arXiv:1608.05766.
[68] C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.
[69] M. Zhu and S. Martinez, "An approximate dual subgradient algorithm for multi-agent non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 6, pp. 1534–1539, Jun. 2013.

Jinshan Zeng received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2015. He is currently an Assistant Professor with the School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China.



Wotao Yin received the Ph.D. degree in operations research from Columbia University, New York, NY, USA, in 2006, respectively. He is currently a Professor with the Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, USA. His research interests include computational optimization and its applications in signal processing, machine learning, and other data science problems. During 2006–2013, he was with Rice University. He was the NSF CAREER award in 2008, the Alfred P. Sloan Research Fellowship in 2009, and the Morningside Gold Medal in 2016, and has coauthored five papers receiving best-paper-type awards. He invented fast algorithms for sparse optimization and has been leading the research of optimization algorithms for large-scale problems.