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 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 ∈ [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 decentralized 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:

\[ \text{minimize } f(x) = \sum_{i=1}^{n} f_i(x), \]

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:

\[ \text{minimize } s(x) = \sum_{i=1}^{n} (f_i(x) + r_i(x)), \]

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 [68], [9], [29], [58], [55], [61]; (iii) distributed and parallel machine learning [55], [15], [21], [33], [43]. More specifically, in these applications, each function \( f_i \) 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 \( \infty \) otherwise; 3) nonconvex penalties such as ℓ_q quasi-norm \((0 \leq q < 1)\) [49], [11], smoothly clipped absolute deviation (SCAD) penalty [16] and the minimax concave penalty (MCP) [67].

When \( f_i \)'s are convex, the existing algorithms include the (sub)gradient methods [8], [10], [24], [37], [40], [59], [65], [46], and the primal-dual domain methods such as the decentralized alternating direction method of multipliers (DADMM) [51], [52], [9], DLM [31], and EXTRA [53], [54]. When \( f_i \)'s are nonconvex, some existing results include [4], [5], [18], [35], [36], [56], [57], [27], [60], [62], [68]. 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 [47], [64], [38], [23], 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 gradients, [10] shows that decreasing step sizes \( \alpha_k = \frac{1}{\sqrt{k}} \) lead to a convergence rate \( O \left( \frac{\ln k}{k} \right) \) of the running best of objective errors. [24] uses nested loops and shows an outer-loop convergence rate \( O \left( \frac{1}{k^2} \right) \) 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 \( O \left( \frac{\ln k}{k} \right) \).
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 $\alpha$ 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 $\alpha$.

(b) When a decreasing step size $\alpha_k = O(1/(k+1)^r)$ is used, where $0 < r \leq 1$ 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 $O(1/(k+1)^r)$. Moreover, we show the convergence of DGD to a stationary point of the original problem, and derive the convergence rates of DGD with different $\epsilon$ 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 $\{f(\bar{x}^k) - f(x_{\text{opt}})\}$ or $\{s(\bar{x}^k) - s(x_{\text{opt}})\}$, where $\bar{x}^k = \frac{1}{n} \sum_{i=1}^{n} x_{i}^k[i]$ is the average of the $k$th iterate and $x_{\text{opt}}$ 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 [10], [24], [40], [4], [18], [35], [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 [33], [21], [22], [19]. We will give detailed comparisons with these work latter.

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 shows some numerical experiments to verify the developed theoretical results. Section VI presents the proofs of our main results. We conclude this paper in Section VII.

**Notation:** Let $I$ denote the identity matrix of the size $n \times n$, and $1 \in \mathbb{R}^n$ 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 $\|X\|_F = \sqrt{\sum_{i,j} X_{ij}^2}$ is its Frobenius norm, which simplifies to the Euclidean norm when $X$ is a vector. Given a symmetric, positive semidefinite matrix $G \in \mathbb{R}^{n \times n}$, we let $\|X\|_G^2 \triangleq \langle X, GX \rangle$ be the induced semi-norm. Given a function $h$, $\text{dom}(f)$ 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 $x_{(i)} \in \mathbb{R}^p$ denote the local copy of $x$ at node $i$. We reformulate the consensus problem (1) into the equivalent problem:

\[
\begin{align*}
\text{minimize} \quad & \mathbf{1}^T f(x) \triangleq \sum_{i=1}^{n} f_i(x_{(i)}), \\
\text{subject to} \quad & x_{(i)} = x_{(j)}, \quad \forall (i, j) \in \mathcal{E}.
\end{align*}
\]

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

\[
\begin{align*}
x \triangleq \left( \begin{array}{c}
-x_{(1)}^T \\
-x_{(2)}^T \\
\vdots \\
-x_{(n)}^T
\end{array} \right), \quad f(x) \triangleq \left( \begin{array}{c}
f_1(x_{(1)}) \\
f_2(x_{(2)}) \\
\vdots \\
f_n(x_{(n)})
\end{array} \right) \in \mathbb{R}^{n \times p}.
\end{align*}
\]

In addition, the gradient of $f(x)$ is

\[
\nabla f(x) \triangleq \left( \begin{array}{c}
-\nabla f_1(x_{(1)})^T \\
-\nabla f_2(x_{(2)})^T \\
\vdots \\
-\nabla f_n(x_{(n)})^T
\end{array} \right) \in \mathbb{R}^{n \times p}.
\]

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

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

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

\[
x^{k+1} \leftarrow W x^k - \alpha_k \nabla f(x^k),
\]

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:

\[
\begin{align*}
\text{minimize} \quad & \sum_{i=1}^{n} (f_i(x_{(i)}) + r_i(x_{(i)})), \\
\text{subject to} \quad & x_{(i)} = x_{(j)}, \quad \forall (i, j) \in \mathcal{E}.
\end{align*}
\]

Let $r(x) \triangleq \sum_{i=1}^{n} r_i(x_{(i)})$. The algorithm Prox-DGD can be applied to the above problem (6):
TABLE I
COMPARISONS ON DIFFERENT ALGORITHMS FOR CONSENSUS SMOOTH OPTIMIZATION PROBLEM (1)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fixed step size</th>
<th>Decreasing step sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i$</td>
<td>convex only</td>
<td>(non)convex</td>
</tr>
<tr>
<td>$\nabla f_i$</td>
<td>Lipschitz</td>
<td>Lipshitz, bounded</td>
</tr>
<tr>
<td>Step size</td>
<td>$0 &lt; \alpha &lt; \frac{1 + \lambda_n(W)}{L_f}$</td>
<td>$O\left(\frac{1}{k}\right)$ with Nesterov acc.</td>
</tr>
<tr>
<td>Consensus</td>
<td>error $O(\alpha)$</td>
<td>$O\left(\frac{1}{k}\right)$</td>
</tr>
<tr>
<td>$\min_{j \leq k}</td>
<td></td>
<td>x^{j+1} - x^j</td>
</tr>
<tr>
<td>Global objective error</td>
<td>$O\left(\frac{1}{k}\right)$ until error $O\left(\frac{1}{k \frac{1-\zeta}{\zeta}}\right)$</td>
<td>Convex: $O\left(\frac{1}{k}\right)$ until error $O\left(\frac{1}{k \lambda n(W)}\right)$; Nonconvex: no rate $O\left(\frac{\ln k}{k}\right)$</td>
</tr>
</tbody>
</table>

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.

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fixed step size</th>
<th>Decreasing step sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i, r_i$</td>
<td>convex only</td>
<td>(non)convex</td>
</tr>
<tr>
<td>$\nabla f_i$</td>
<td>Lipschitz</td>
<td>Lipshitz, bounded</td>
</tr>
<tr>
<td>$\partial r_i$</td>
<td>bounded</td>
<td>bounded</td>
</tr>
<tr>
<td>Step size</td>
<td>$0 &lt; \alpha &lt; \frac{1}{L_f}$</td>
<td>$0 &lt; \alpha &lt; \frac{1 + \lambda_n(W)}{L_f}$</td>
</tr>
<tr>
<td>Consensus</td>
<td>$O\left(\gamma^k k^2\right), 0 &lt; \gamma &lt; 1$</td>
<td>$O\left(\frac{1}{k^{1+1/\gamma}}\right)$</td>
</tr>
<tr>
<td>$\min_{j \leq k}</td>
<td></td>
<td>x^{j+1} - x^j</td>
</tr>
<tr>
<td>Global objective error</td>
<td>$O\left(\frac{1}{k}\right)$</td>
<td>Form $\frac{D_1}{\alpha} + D_2 \alpha$, $D_1, D_2 &gt; 0$</td>
</tr>
</tbody>
</table>

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

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

where the proximal operator is

$$\text{prox}_{\alpha_k r}(x) \triangleq \arg\min_{u \in \mathbb{R}^{n \times p}} \left\{ \alpha_k r(u) + \frac{||u - x||^2}{2} \right\}. \quad (8)$$
III. ASSUMPTIONS AND MAIN RESULTS

This section presents all of our main results.

A. Definitions and assumptions

Definition 1 (Lipschitz differentiability). A function \( f \) is called Lipschitz differentiable if \( h \) is differentiable and its gradient \( \nabla h \) is Lipschitz continuous, i.e., \(|\nabla h(u) - \nabla h(v)| \leq L|u - v|\), \( \forall u, v \in \text{dom}(h) \), 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, Section 2.2] for examples) and can help obtain whole sequence convergence\(^1\) from subsequence convergence.

Definition 3 (Kurdyka-Łojasiewicz (KL) property \([34, 7, 2]\)). A function \( h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) has the KL property at \( x^* \in \text{dom}(h) \) 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 \( \varphi \) 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 KL inequality holds:

\[
\varphi'(h(x) - h(x^*)) \cdot \text{dist}(0, \partial h(x)) \geq 1.
\]

Proper lower semi-continuous functions that satisfy the KL inequality at each point of \( \text{dom}(h) \) are called KL 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 \( \sum_{i=1}^n f_i(x(i)) \) is \( L_f \)-Lipschitz differentiable with \( L_f \triangleq \max_i L_{f_i} \), (this can be easily verified via the definition of \( \nabla f(x) \) 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 \( (i, j) \notin E \), then \( w_{ij} = 0 \), otherwise, \( w_{ij} > 0 \).
2. (Symmetry) \( W = W^T \).
3. (Null space property) \( \text{null}(I - W) = \text{span}\{1\} \).
4. (Spectral property) \( I \preceq W \preceq I \).

By Assumption 2, a solution \( x_{\text{opt}} \) to problem (3) satisfies \( (I - W)x_{\text{opt}} = 0 \). 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 \( i \)th largest eigenvalue of \( W \). Then by Assumption 2,

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

Let \( \zeta \) be the second largest magnitude eigenvalue of \( W \). Then

\[
\zeta = \max\{\lambda_2(W), |\lambda_n(W)|\}.
\]

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]:

\[
L_\alpha(x) \triangleq 1^T f(x) + \frac{1}{2\alpha} \|x\|_W^2 - \lambda_W.
\]

It is worth reminding that convexity is not assumed.

Theorem 1 (Global convergence). Let \( \{x^k\} \) be the sequence generated by DGD (5) with the step size \( 0 < \alpha < \frac{1 + \lambda_2(W)}{L_f} \).

Let Assumptions 1 and 2 hold. Then \( \{x^k\} \) has at least one accumulation point \( x^* \), and any such point is a stationary point of \( L_\alpha(x) \).

Furthermore, the running best rates\(^2\) of

---

\(^1\)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 \( b_k = o(1/k) \).

\(^2\)Whole sequence convergence from any starting point is referred to as “global convergence” in the literature. Its limit is not necessarily a global solution.
the sequences \( \alpha^k \) \( \{\|x^{k+1} - x^k\|^2\} \), and \( \{\|\nabla \mathcal{L}_\alpha(x^k)\|^2\} \), and \( \{\|\frac{1}{n} I^T \nabla f(x^k)\|^2\} \) are \( O(\frac{1}{k}) \). The convergence rate of the sequence \( \{\frac{1}{k} \sum_{k=0}^{n} \|\frac{1}{n} I^T \nabla f(x^k)\|^2\} \) is \( O(\frac{1}{k}) \).

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

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

**Proposition 1** (Consensual bound on \( x^* \)). For each iteration \( k \), define \( \tilde{x}^k \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^k \). Then, it holds for each node \( i \) that
\[
\|x_i^k - \tilde{x}^k\| \leq \frac{\alpha D}{1 - \zeta},
\]
(13)
where \( \alpha \) is a universal bound of \( \|\nabla f(x^k)\| \) defined in Lemma 6 below, \( \zeta \) is the second largest magnitude eigenvalue of \( W \) specified in (10). As \( k \to \infty \), (13) yields the consensual bound
\[
\|x_i^* - \tilde{x}^*\| \leq \frac{\alpha D}{1 - \zeta},
\]
where \( \tilde{x}^* \equiv \frac{1}{n} \sum_{i=1}^{n} x_i^* \).

In Proposition 1, the consensual bound is proportional to the step size \( \alpha \) 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 (5) yields the following comparison:

DGD averaged:
\[
\tilde{x}^{k+1} \leftarrow \tilde{x}^k - \alpha \left( \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_i^k) \right).
\]
(14)

Centralized:
\[
\tilde{x}^{k+1} \leftarrow \tilde{x}^k - \alpha \left( \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{x}_i^k) \right).
\]
(15)

Applying the error of this approximation as
\[
\frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x_i^k) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{x}_i^k) \| \\
\leq \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x_i^k) - \nabla f_i(\tilde{x}_i^k)\| \leq \frac{\alpha DL_{f}}{1 - \zeta}.
\]

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 KL assumption on \( \mathcal{L}_\alpha \) in Theorem 1 can be satisfied if each \( f_i \) is a sub-analytic function. Since \( \|x\|^2_{1 - W} \) 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, Section 2.2] for more details and examples.

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

(a) If \( \theta = 0 \), \( x^k \) converges to \( x^* \) in finitely many iterations.

(b) If \( \theta \in (0, \frac{1}{2}) \), \( \|x^k - x^*\| \leq C_0 x^k \) for all \( k \geq k^* \) for some \( k^* > 0 \), \( C_0 > 0 \), \( \tau \in (0,1) \).

(c) If \( \theta \in (\frac{1}{2}, 1) \), \( \|x^k - x^*\| \leq C_0 k^{(1-\eta)/(2\theta-1)} \) for all \( k \geq k^* \), 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 \( x^* \) of \( \mathcal{L}_\alpha \) is not a stationary point of the original problem. We only have a consensual bound on \( x^* \). To address this issue, the next subsection 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 \( \alpha \), motivates the use of properly decreasing step sizes \( \alpha_k = O\left(\frac{1}{k^{1-\eta}}\right) \), for some \( 0 < \epsilon \leq 1 \), to diminish the consensual bound to 0. As a result, any accumulation point \( x^* \) 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 \( k \), \( \nabla f(x^k) \) is uniformly bounded by some constant \( B > 0 \), i.e., \( \|\nabla f(x^k)\| \leq B \).

Note that the bounded gradient assumption is a regular assumption in the convergence analysis of decentralized gradient methods (see, [4], [5], [18], [35], [36], [56], [57], [27], [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:
\[
\alpha_k = \frac{1}{L_f(k + 1)\tau^2}, \quad 0 < \epsilon \leq 1, \quad (16)
\]
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
\[
x^k = W^k x^0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j} \nabla f(x^j) \quad (17)
\]

**Proposition 3** (Asymptotic consensus rate). Let Assumptions 2 and 3 hold. Let DGD use (16). Let \( \tilde{x}^k \equiv \frac{1}{n} 11^T x^k \). Then, \( \|x^k - \tilde{x}^k\| \) converges to 0 at the rate of \( O(1/(k + 1)^\epsilon) \).

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 \( \epsilon \), faster decaying step sizes generally imply a faster asymptotic consensus rate. Note that \( (I - W)x^k = 0 \) and thus \( \|x^k\|^2_{1 - W} = \|x^k - \tilde{x}^k\|^2_{1 - W} \). Therefore, the above proposition implies the following result.
Corollary 1. Apply the setting of Proposition 3. \( \|x^k\|_W^2 \) converges to 0 at the rate of \( \mathcal{O}(1/(k+1)^{2\epsilon}) \).

Corollary 1 shows that the sequence \( \{x^k\} \) in the \((I-W)\) semi-norm can decay to 0 at a sublinear rate. For any global consensus solution \( x_{opt} \) to problem (3), we have \( \|x^k - x_{opt}\|_W^2 = \|x^k\|_W^2 \) so, if \( \{x^k\} \) does converge to \( x_{opt} \), 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) \( \{\mathcal{L}_{\alpha_k}(x^k)\} \) and \( \{1/Tf(x^k)\} \) converge to the same limit;
(b) \( \lim_{k \to \infty} 1/T \nabla f(x^k) = 0 \), and any limit point of \( \{x^k\} \) is a stationary point of problem (3);
(c) In addition, if there exists an isolated accumulation point, then \( \{x^k\} \) converges.

In the proof of Theorem 2, we will establish
\[
\sum_{k=0}^{\infty} \left( \alpha_k^{-1}(1 + \lambda_n(W)) - L_f \right) \|x^{k+1} - x^k\|^2 < \infty,
\]
which implies that the running best rate of the sequence \( \{\|x^{k+1} - x^k\|^2\} \) is \( o(1/k^{2\epsilon}) \). Theorem 2 shows that the objective sequence converges, and any limit point of \( \{x^k\} \) 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 \( \epsilon \).

For any \( x \in \mathbb{R}^n \), let \( f(x) \triangleq \sum_{i=1}^n f_i(x_i) \). Even if \( f_i \)'s are convex, the solution to (3) may be non-unique. Thus, let \( f^* \) be the set of solutions to (3). Given \( x^k \), we pick the solution \( x_{opt} = \text{Proj}_f(x^k) \in f^* \). Also let \( f_{opt} = f(x_{opt}) \) be the optimal value of (1). Define the ergodic objective:
\[
\tilde{f}^k = \frac{\sum_{k=0}^{K} \alpha_k f(x^{k+1})}{\sum_{k=0}^{K} \alpha_k},
\]
where \( x^{k+1} = \frac{1}{n}(1/T x^k + 1) \). Obviously, \( \tilde{f}^k \geq \min_{k=1,...,K+1} f(x^k) \).

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 \( \{\tilde{f}^k\} \) defined in (18) converges to the optimal objective value \( f_{opt} \) at the following rates:
(a) if \( 0 < \epsilon < 1/2 \), the rate is \( \mathcal{O}(1/K^{1/\epsilon}) \);
(b) if \( \epsilon = 1/2 \), the rate is \( \mathcal{O}(\ln K / \sqrt{K}) \);
(c) if \( 1/2 < \epsilon < 1 \), the rate is \( \mathcal{O}(1/K^{1/\epsilon}) \);
(d) if \( \epsilon = 1 \), the rate is \( \mathcal{O}(1/(\ln K)) \).

The convergence rates established in Proposition 4 almost as good as \( \mathcal{O}(1/K^{1/\epsilon}) \) when \( \epsilon = 1/2 \). As \( \epsilon \) 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 \( \epsilon \) implies a faster consensus rate.

Therefore, there is a tradeoff to choose an appropriate \( \epsilon \) in the practical implementation of DGD.

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 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 \( \mathcal{O}(\frac{1}{K} + \frac{1}{\sqrt{K}}) \)-ergodic convergence rate was established assuming \( K \) is sufficiently large and the step size \( \alpha \) is sufficiently small. When applied to the setting of this paper, [33, Theorem 1] implies that the sequence \( \{\frac{1}{K} \sum_{k=0}^{K-1} \|1/T \nabla f(x^k)\|^2\} \) converges to zero at the rate \( \mathcal{O}(\frac{1}{K}) \) if the step size \( 0 < \alpha < \frac{1-\zeta}{(1+\zeta)\sqrt{K}} \), where \( \zeta \) 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], [57], [28].

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
\[
x^{k+1} = \text{prox}_{\alpha_k}(x^k - \alpha_k \nabla f_{\alpha_k}(x^k)) \tag{20}
\]
based on which, we define the Lyapunov function
\[
\mathcal{L}_{\alpha_k}(x) \triangleq \mathcal{L}_{\alpha_k}(x) + r(x),
\]
where we recall \( \mathcal{L}_{\alpha_k}(x) = \sum_{i=1}^{n} f_i(x_i) + \frac{1}{\alpha_k} \|x\|_W^2 \). Clearly (20) is then the forward-backward splitting (a.k.a., prox-gradient) iteration for minimizing \( \mathcal{L}_{\alpha_k}(x) \). Specifically, (20) first performs gradient descent to the differentiable function \( \mathcal{L}_{\alpha_k}(x) \) and then computes the proximal of \( r(x) \).

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 \( L_i > 0 \).
(2) Each \( (f_i + r_i) \) is proper, lower semi-continuous, coercive.

As before, \( \sum_{i=1}^{n} f_i(x_i) \) is \( L_i \)-Lipschitz differentiable for \( L_i \triangleq \max_i L_i \).

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 \( \{x^k\} \) be the sequence generated by Prox-DGD (7) where the step size \( \alpha \) satisfies \( 0 < \alpha < \frac{1+\lambda_n(W)}{L_f} \) when \( r_i \)'s are convex:
and $0 < \alpha < \frac{\lambda_n(W)}{L_1}$, when $r_i$'s are not necessarily convex (this case requires $\lambda_n(W) > 0$). Let Assumptions 2 and 4 hold. Then $\{x^k\}$ has at least one accumulation point $x^*$, and any accumulation point is a stationary point of $L_\alpha(x)$. Furthermore, the running best rates of the sequences $\{\|x^{k+1} - x^k\|^2\}$, $\{\|g^{k+1}\|^2\}$ and $\{\|\frac{1}{n} 1^T \nabla f(x^k) + \xi_k\|^2\}$ (where $g^{k+1}$ is defined in Lemma 18, and $\xi_k$ is defined in Lemma 19) are $o(1)$. In addition, if $L_\alpha$ satisfies the KL property at an accumulation point $x^*$, then $\{x^k\}$ converges to $x^*$.

The rate of convergence of Prox-DGD can be also established by leveraging the KL property.

**Proposition 5** (Rate of convergence of Prox-DGD). Under assumptions of Theorem 3, suppose that $L_\alpha$ satisfies the KL 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 $\theta = 0$, $x^k$ converges to $x^*$ in finitely many iterations.
(b) If $\theta \in (0, \frac{1}{2}]$, $\|x^k - x^*\| \leq C_1 x^k$ for all $k \geq k^*$ for some $k^* > 0$, $C_1 > 0$, $\tau \in (0, 1]$.
(c) If $\theta \in (\frac{1}{2}, 1]$, $\|x^k - x^*\| \leq C_1(k^{-(1-\theta)/(2\theta-1)} - 1)$ 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 revisited as follows.

**Assumption 5** (Bounded composite subgradient). For each $i$, $\nabla f_i$ 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, $\|\xi_i\| \leq B_i$ for any $\xi_i \in \partial r_i(x)$ and $x \in \mathbb{R}^p$, $i = 1, \ldots, n$.

Let $\bar{B} \triangleq \sum_{i=1}^n (B_i + B_i)$. Then $\nabla f(x) + \xi$ (where $\xi \in \partial r(x)$ for any $x \in \mathbb{R}^{n \times p}$) 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 19 below, 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). Then

$$\|x^k - x^*\| \leq C (\|x^0\| + \bar{B} \sum_{j=0}^{k-1} \alpha_j (\xi^{k-1-j}),$$

and $\|x^k - x^*\|$ converges to 0 at the rate of $\mathcal{O}(1/(k + 1)^\epsilon)$. Moreover, let $x^*$ be any global solution of the problem (6). Then $\|x^k - x^*\|_{L_1} \leq C \|x^k - x^*\|_{W} \leq C \|x^k - x^*\|_{L_2} \leq \mathcal{O}(1/(k + 1)^\epsilon)$ converges to 0 at the rate of $\mathcal{O}(1/(k + 1)^\epsilon)$.

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

$$s^K = \sum_{k=0}^{K} \alpha_k s(x^{k+1}) + \sum_{k=0}^{K} \alpha_k.$$

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

(a) $\{L_\alpha(x^k)\}$ and $\{\sum_{i=1}^n f_i(x^k_i) + r_i(x^k_i)\}$ converge to the same limit;
(b) $\sum_{k=0}^{\infty} (\alpha_k^{-1} - \lambda_n(W) - \mathcal{L}_f) \|x^{k+1} - x^k\|^2 \geq \infty \text{ when } r_i$'s are convex; or, $\sum_{k=0}^{\infty} (\alpha_k^{-1} \lambda_n(W) - \mathcal{L}_f) \|x^{k+1} - x^k\|^2 \geq \infty \text{ when } r_i$'s are not necessarily convex (this case requires $\lambda_n(W) > 0$);
(c) if $\{c_k\}$ satisfies $\|\xi_k^{1} - \xi_k\| \leq L_r \|x^k - x^k\|$ for each $k \geq k_0$, some constant $L_r$, $0$, and a sufficiently large integer $k_0 > 0$, then

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

where $\xi_k$ is uniformly bounded by some constant $\lambda_{n_1}(W)$ is the one determined by the proximal operator (8), and any limit point is a stationary point of problem (6).

In addition, if there exists an isolated accumulation point, then $\{x^k\}$ converges.

**Remark 5.** A typical proximal function $r$ satisfying the assumption of Theorem 4 (c) is the $\ell_q$ 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 $\ell_q$ 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 $\|x^{k+1} - x^k\|^2 \to 0$ as $k \to \infty$. Together with the lower bound property, we can easily obtain the finite support and sign convergence of $\{x^k\}$, 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 $\tau$, where $\tau$ denotes the lower bound. Besides $\ell_q$ quasi-norm, there are some other typical cases like SCAD [16] and MCP [67] 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 pro-
posed 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, Theorem 1]). However, there are some differences between [4] and this paper. In short, Proj SGD in [4] uses convex constraints, which correspond to setting \( \alpha = 0 \) for any \( \alpha \). 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 \( \alpha_k = \mathcal{O}(k^{-1}) \) 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\(^1\) for nonconvex (1) under some regularity assumptions. They also introduced random perturbations to avoid local minima. The network considered in [57] is time-varying and directed, and specific column stochastic matrices and diminishing step sizes are used. 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, Theorem 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]. 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.

\(^1\)The original form of this algorithm, push-sum, was proposed in [25] for the average consensus problem. It was modified and analyzed in [41] for convex consensus optimization problem over time-varying directed graphs.

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. NUMERICAL EXPERIMENTS

In this section, we describe a set of numerical experiments mainly to verify our theoretical findings for DGD and Prox-DGD. The comparisons between DGD (or Prox-DGD) and the other existing algorithms can be referred to these literature like [35], [21], [22], [19].

A. Convergence of DGD

We verify the performance of DGD using both fixed and diminishing step sizes in the experimental setting identical to [57]. The following one dimensional decentralized optimization is considered in [57],

\[
\min_{x \in \mathbb{R}} f(x) = f_1(x) + f_2(x) + f_3(x),
\]

where

\[
f_1(x) = \begin{cases} 
(x^3 - 16x)(x + 2), & \text{if } |x| \leq 10, \\
4248x - 32400, & \text{if } x > 10, \\
-3112x - 25040, & \text{if } x < -10,
\end{cases}
\]

\[
f_2(x) = \begin{cases} 
(0.5x^3 + x^2)(x - 4), & \text{if } |x| \leq 10, \\
1620x - 12600, & \text{if } x > 10, \\
-2220x - 16600, & \text{if } x < -10,
\end{cases}
\]

\[
f_3(x) = \begin{cases} 
(x + 2)^2(x - 4), & \text{if } |x| \leq 10, \\
288x - 2016, & \text{if } x > 10, \\
288x - 1984, & \text{if } x < -10.
\end{cases}
\]

We plot the function \( f \) over the intervals \([-15, 15]\) and \([-6, 6]\) as shown in Fig. 1. The function achieves its global minimizer at \( x = 2.02 \) and has a local minimizer at \( x = -2.49 \), as well as a local maximizer at \( x = 1.12 \). It is easy to compute the Lipschitz constants of \( \nabla f_1, \nabla f_2 \) and \( \nabla f_3 \) as \( L_1 = 1288, \)

\( \nabla f_2 = \nabla f_3 = 0 \).
The experiment results are reported in Fig. 1. The plots of objective function $f$.

$$L_2 = 532, \text{ and } L_3 = 60, \text{ respectively. Thus, } L_f = \max_i L_i = 1288, \text{ which is used in our theoretical analysis. Moreover, } \nabla f \text{ is obviously bounded. We consider one of the following three connected networks:}$$

$$(1 \leftrightarrow 2, 2 \leftrightarrow 3), \quad (1 \leftrightarrow 3, 3 \leftrightarrow 2), \quad (2 \leftrightarrow 1, 1 \leftrightarrow 3).$$

In the experiment, the mixing matrix $W$ is taken as

$$W = \begin{pmatrix}
\frac{1}{2} & 0 & \frac{1}{2} \\
0 & \frac{1}{2} & \frac{1}{2} \\
\frac{1}{2} & \frac{1}{2} & 0
\end{pmatrix},$$

which has the eigenvalues: 1, 0.5, -0.5. The mixing matrix $W$ is not positive definite. All the assumptions used in our theory are satisfied.

We test the performance of DGD with a theoretically fixed step size and several different kinds of decreasing step sizes, starting iterations from two different initial points:

$$x_1^0 := (0, 0, 0), \quad \text{and } x_2^0 := (-1, -1.2, -1.1).$$

The second initialization is a “dangerous” point since it is close to local maximum and between two local minima (one of them is global). The experiment results are reported in Fig. 2. From these figures, DGD successfully converges to the global minimum and achieves the consensus starting from both initial points though the objective function is nonconvex.

B. Prox-DGD for decentralized $L_0$ regularization

We apply Prox-DGD to solve the following nonconvex decentralized $L_0$ regularization problem:

$$x^* \leftarrow \arg\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x),$$

where $f_i(x) = \frac{1}{2} \|B(i)x - b(i)\|_2^2 + \lambda_i \|x\|_0, B_i \in \mathbb{R}^{m_i \times p}, b_i \in \mathbb{R}^{m_i},$ for $i = 1, \ldots, n,$ and $\|x\|_0$ is called the $\ell_0$ quasi-norm, which yields the number of the nonzero components of $x$. In this experiment, we take $n = 10, p = 256,$ and $m_i = 150$ for $i = 1, \ldots, n$. In this case, the proximal operator of $\ell_0$ quasi-norm is the well-known hard thresholding mapping:

$$\text{prox}_{\alpha \| \cdot \|_0}(x) = \begin{cases}
x, & \text{if } x < \sqrt{\alpha}, \\
0, & \text{otherwise}.
\end{cases}$$

We do not consider the model selection problem, but just test the performance of Prox-DGD applied to such a given deterministic model. Take $\lambda_i = 0.5$ for each agent $i$. The connected network is shown in Fig. 3.

We use several step-size strategies, including three fixed step sizes and two decreasing step sizes, to test Prox-DGD. The initializations in all cases are 0. The experiment results are illustrated in Fig. 4. By Fig. 4(a) and (b), when a fixed step size is adopted (obviously, we should assume that such step size is sufficiently small to satisfy the theoretical restriction), a larger fixed step size generally implies a faster convergence rate (just converges to a stationary point of the
related Lyapunov function minimization problem but not the original problem) while causes a larger consensus error. These phenomena are reasonable and further verify the established results in Proposition 1. It can be also observed from Fig. 4(a) that Prox-DGD almost performs at linear rates on all the three fixed step sizes. It is expected since the $L_0$ regularization function satisfies the Kurdyka-Łojasiewicz (KL) inequality with $\psi(s) = C_1s^{1-\theta}$ for $\theta \in (0, 1/2]$ (see [2]). Thus, by Proposition 5(ii), Prox-DGD has eventual linear convergence. Once the initial guess (in this case, 0 may be a good initial guess) is good enough, Prox-DGD starts decaying linearly starting from early iterations. Similar phenomenon can be also observed by Fig. 4(c) and (d) under the decreasing step sizes. Specifically, if $\alpha_t = O(\frac{1}{t^2})$, a smaller $\epsilon$ (larger step size) generally implies a faster convergence rate but a slower consensus rate. This indeed verifies our Proposition 5, which states that a larger $\epsilon$ generally means a faster consensus rate. Moreover, from Fig. 4(b) and (d), under a fixed step size, the consensus error does not vanish but settles to a deterministic value, which can be overcome by using decreasing step sizes. This gives the main motivation of using decreasing step sizes.

VI. PROOFS

In this section, we present the 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 $L_\alpha$, following the argument in [65]; then, the properties of sufficient descent, lower boundedness, and bounded gradients are established for the sequence $\{L_\alpha(x^k)\}$, giving subsequence convergence of the DGD iterates; finally, whole sequence convergence of the DGD iterates follows from the KL property of $L_\alpha$.

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

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

$$x^{k+1} = x^k - \alpha(\nabla f(x^k) + \alpha^{-1}(I - W)x^k) = x^k - \alpha \nabla L_\alpha(x^k).$$

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

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

$$L_\alpha(x^{k+1}) \leq L_\alpha(x^k) - \frac{1}{2}\left(\alpha^{-1}(1 + \lambda_n(W)) - L_f\right)\|x^{k+1} - x^k\|^2, \quad \forall k \in \mathbb{N}.$$

**Proof:** From $x^{k+1} = x^k - \alpha \nabla L_\alpha(x^k)$, it follows that

$$\langle \nabla L_\alpha(x^k), x^{k+1} - x^k \rangle = -\frac{\|x^{k+1} - x^k\|^2}{\alpha}.$$  \hfill (25)

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

$$L_\alpha(x^{k+1}) \leq L_\alpha(x^k) + \langle \nabla L_\alpha(x^k), x^{k+1} - x^k \rangle + \frac{L^*}{2}\|x^{k+1} - x^k\|^2.$$  \hfill (26)

Combining (25) and (26) yields (24).

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

**Proof:** The lower boundedness of $L_\alpha(x^k)$ 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 $\alpha$, $L_\alpha(x^k)$ is nonincreasing and upper bounded by $L_\alpha(x^0) < +\infty$. Hence, $1^T f(x^k) \leq L_\alpha(x^0)$ implies that $x^k$ is bounded due to the coercivity of $1^T f(x)$ (Assumption 1 Part 2).

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

**Lemma 4** ($L^2$-summable and asymptotic regularity). It holds that

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 < +\infty \quad \text{and} \quad \|x^{k+1} - x^k\| \to 0 \quad \text{as} \quad k \to \infty.$$
From (23), the result below directly follows:

**Lemma 5 (Gradient bound),** \( \| \nabla L_\alpha(x^k) \| \leq \alpha^{-1} \| x^{k+1} - x^k \| \).

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

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

The running best rate of the sequence \( \{ \| x^{k+1} - x^k \|^2 \} \) follows from [13, Lemma 1.2] or [26, Theorem 3.3.1]. By Lemma 5, the running best rate of the sequence \( \{ \| \nabla L_\alpha(x^k) \|^2 \} \) is \( o(1/k) \).

By (11), \( \nabla L_\alpha(x^k) = \nabla f(x^k) + \alpha^{-1}(I - W)x^k \), which implies \( \frac{1}{n} \sum_{i=1}^{n} \nabla f(x^k) \) due to \( \frac{1}{n} \sum_{i=1}^{n} (I - W) = 0 \). Thus,

\[
\| \frac{1}{n} \sum_{i=1}^{n} \nabla f(x^k) \| \leq \frac{1}{n} \sum_{i=1}^{n} \nabla L_\alpha(x^k) \| \leq \| \nabla L_\alpha(x^k) \|,
\]

which implies the running best rate of \( \{ \frac{1}{n} \sum_{i=1}^{n} \nabla f(x^k) \} \) is also \( o(1/k) \).

By Lemmas 2 and 5, it holds

\[
\| \nabla L_\alpha(x^k) \|^2 \leq \frac{2}{\alpha(1 + \lambda_n(W) - \alpha L_f)} (L_\alpha(x^k) - L_\alpha(x^{k+1}))
\]

which implies

\[
\frac{1}{K} \sum_{k=0}^{K-1} \| \nabla L_\alpha(x^k) \|^2 \leq \frac{2(\alpha L_\alpha(x^0) - L^\ast)}{\alpha(1 + \lambda_n(W) - \alpha L_f) K}.
\]

Moreover, note that \( \frac{1}{n} \sum_{i=1}^{n} \nabla f(x^k) \| \leq \| \nabla L_\alpha(x^k) \| \). Thus, the convergence rate of \( \frac{1}{K} \sum_{k=0}^{K-1} \| \nabla f(x^k) \|^2 \) is \( O(1/K) \).

**B. Proof for Proposition 2**

**Proof:** Note that

\[
\| \nabla L_\alpha(x^{k+1}) \| \leq \| \nabla L_\alpha(x^{k+1}) - \nabla L_\alpha(x^k) \| + \| \nabla L_\alpha(x^k) \|
\]

\[
\leq L^\ast \| x^{k+1} - x^k \| + \alpha^{-1} \| x^{k+1} - x^k \|
\]

\[
= \alpha^{-1}(2 - \lambda_n(W)) + L_f \| x^{k+1} - x^k \|,
\]

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

\[
2 \| x^{k+1} - x^k \| \leq \| x^k - x^{k-1} \| + \frac{cB}{a} \times
\]

\[
(\mathcal{L}_\alpha(x^k) - \mathcal{L}_\alpha(x^\ast))^{1-\theta} - (\mathcal{L}_\alpha(x^{k+1}) - \mathcal{L}_\alpha(x^\ast))^{1-\theta},
\]

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

\[
\sum_{k=k_0}^{k} \| x^{k+1} - x^k \| \leq \| x^{k_0} - x^{k_0-1} \| + \frac{cB}{a} \times
\]

\[
(\mathcal{L}_\alpha(x^{k_0}) - \mathcal{L}_\alpha(x^\ast))^{1-\theta} - (\mathcal{L}_\alpha(x^{k+1}) - \mathcal{L}_\alpha(x^\ast))^{1-\theta}.
\]

Following a derivation similar to the proof of [1, Theorem 5], we can estimate the rate of convergence of \( \{x^k\} \) in the different cases of \( \theta \).

**C. Proof for Proposition 3**

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

**Lemma 7.** (40, Proposition 1) Let \( W_k = \sum_{i=1}^{k} W_i \) be the power of \( W \) with degree \( k \) for any \( k \in \mathbb{N} \). Under Assumption 2, it holds

\[
\| W^k - \frac{1}{n} 11^T \| \leq C\zeta^k
\]

for some constant \( C > 0 \), where \( \zeta \) 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

\[
\lim_{k \to \infty} \sum_{j=0}^{k-1} \beta^{k-1-j} \gamma = \frac{1}{1 - \beta}\gamma.
\]

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

\[
x^k - x^0 = (W^k - \frac{1}{n} 11^T)x^0
\]

\[
- \sum_{j=0}^{k-1} \alpha_j (W^{k-1-j} - \frac{1}{n} 11^T) \nabla f(x^j).
\]
Further by Lemma 7 and Assumption 3, we obtain
\[
\|x^k - \bar{x}^k\| \leq \|(W^k - \frac{1}{n}1^T)\|\|x^0\| \\
+ \sum_{j=0}^{k-1} \alpha_j\|W^{k-1-j} - \frac{1}{n}1^T\| \cdot \|\nabla f(x^j)\| \\
\leq C \left(\|x^0\|\zeta^k + B \sum_{j=0}^{k-1} \alpha_j\zeta^{k-1-j}\right).
\] (31)

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

Let \(b_k \triangleq (k + 1)^{-\epsilon}\). To show the rate of \(\|x^k - \bar{x}^k\|\), we only need to show that
\[
\lim_{k \to \infty} b_k^{-1}\|x^k - \bar{x}^k\| \leq C^*\]
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
\[
b_k^{-1}\|x^k - \bar{x}^k\| \\
\leq Cb_k^{-1} \left(\|x^0\|\zeta^k + B \sum_{j=0}^{k-1} \alpha_j\zeta^{k-1-j}\right) \\
= C\|x^0\|b_k^{-1}\zeta^k + CBb_k^{-1} \sum_{j=0}^{j'_k} \alpha_j\zeta^{k-1-j} \\
+ CBb_k^{-1} \sum_{j=j'_k+1}^{k-1} \alpha_j\zeta^{k-1-j} \\
\triangleq T_1 + T_2 + T_3,
\] (32)
where the first inequality holds because of (31).

In the following, we will estimate the above three terms in the right-hand side of (32), respectively. First, by the definition of \(j'_k\), for any \(j \leq j'_k\), we have
\[
b_k^{-1}\zeta^{k-1-j'} \leq b_k^{-1}\zeta^{k-1-j} \leq 1.
\]
Thus,
\[
T_2 \leq CB \sum_{j=0}^{j'_k} \alpha_j\zeta^{(k-1-j)/2}.
\] (33)

Second, for \(j'_k < j \leq k - 1\),
\[
b_k^{-1}\alpha_j \leq \frac{(k + 1)\epsilon}{L_f(j'_k + 1)^\epsilon} \leq \frac{(k + 1)\epsilon}{L_f(k - 1 + 2\log\zeta(k + 1))^\epsilon},
\]
and also
\[
b_k^{-1}\alpha_j \geq \frac{(k + 1)\epsilon}{L_f(k + 1)^\epsilon} = \frac{1}{L_f},
\]
Thus, for any \(j'_k < j \leq k - 1\),
\[
\lim_{k \to \infty} b_k^{-1}\alpha_j = \frac{1}{L_f}.
\] (34)

Furthermore, note that
\[
\lim_{k \to \infty} b_k^{-1}\zeta^{k/2} = 0.
\] (35)

Therefore, there exists a \(k^*\) such that for \(k \geq k^*\)
\[
b_k^{-1}\alpha_j \leq \frac{2}{L_f},
\] (36)
\[
b_k^{-1}\zeta^{k/2} \leq 1.
\] (37)

The above two inequalities imply that for sufficiently large \(k\),
\[
T_1 \leq C\|x^0\|\zeta^{k/2},
\] (38)
\[
T_3 \leq 2CB \sum_{j=j'_k+1}^{k-1} \zeta^{k-1-j}.
\] (39)

From (33), (38) and (39), we get
\[
b_k^{-1}\|x^k - \bar{x}^k\| \leq C\|x^0\|\zeta^{k/2} \\
+ CB \left(\sum_{j=0}^{j'_k} \alpha_j\zeta^{(k-1-j)/2} + \frac{2}{L_f} \sum_{j=j'_k+1}^{k-1} \zeta^{k-1-j}\right).
\] (40)

By Lemma 8 and (40), there exists a \(C^* > 0\) such that
\[
\lim_{k \to \infty} b_k^{-1}\|x^k - \bar{x}^k\| \leq C^*.
\] (41)
We have completed the proof of this proposition.

\[\blacksquare\]

D. Proof for Theorem 2

To prove Theorem 2, we first note that similar to (23), the DGD iterates under decreasing step sizes can be rewritten as
\[
x^{k+1} = x^k - \alpha_k \nabla L_\alpha_k(x^k),
\] (42)
where \(L_\alpha_k(x) = 1^T f(x) + \frac{1}{2\alpha_k}\|x\|^2_W\), and we also need the following lemmas.

**Lemma 9** ([50]). Let \(\{v_t\}\) be a nonnegative scalar sequence such that
\[
v_{t+1} \leq (1 + b_t)v_t - u_t + c_t
\]
for all \(t \in \mathbb{N}\), where \(b_t \geq 0, \ u_t \geq 0\) and \(c_t \geq 0\) with \(\sum_{t=0}^{\infty} b_t < \infty\) and \(\sum_{t=0}^{\infty} c_t < \infty\). Then the sequence \(\{v_t\}\) converges to some \(v \geq 0\) and \(\sum_{t=0}^{\infty} u_t < \infty\).

**Lemma 10.** Let \(\alpha_k\) satisfy (16). Then it holds
\[
\alpha_{k+1}^{-1} - \alpha_k^{-1} \leq 2\epsilon L_f(k + 1)^{-\epsilon - 1}.
\]

**Proof:** We first prove that
\[
(1 + x)^\epsilon - 1 \leq 2\epsilon x, \quad \forall x \in [0, 1],
\] (43)
Let \(g(x) = (1 + x)^\epsilon - 1 - 2\epsilon x\). Then its derivative 
\[
g'(x) = \epsilon(1 + x)^\epsilon - 2\epsilon < 0, \quad \forall x \in [0, 1].
\]
It implies \(g(x) \leq g(0) = 0\) for any \(x \in [0, 1]\), that is, the inequality (43) holds.

Note that
\[
\alpha_{k+1}^{-1} - \alpha_k^{-1} = L_f((k + 2)^\epsilon - (k + 1)^\epsilon) \\
= L_f(k + 1)^\epsilon\left((1 + \frac{1}{k + 1})^\epsilon - 1\right) \\
\leq 2\epsilon L_f(k + 1)^{-\epsilon - 1},
\] (44)
where the last inequality holds for (43).

\[\blacksquare\]
The following lemma shows that \( \{(\alpha_k^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{f-W}\} \) is summable.

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

**Proof:** Note that
\[
\|x^{k+1}\|^2_{f-W} = \|x^{k+1} - x^{k+1}\|^2_{f-W} \\
\leq (1 - \lambda_n(W))\|x^{k+1} - x^{k+1}\|^2.
\] (45)

By Lemma 10,
\[
(\alpha_k^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{f-W} \\
\leq 2\epsilon L_f (k + 1)\|x^{k+1}\|^2_{f-W} \\
\leq 2\epsilon L_f (k + 1)\|x^{k+1} - x^{k+1}\|^2.
\] (46)

Furthermore, by (46) and Proposition 3, the sequence \( \{(\alpha_k^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{f-W}\} \) converges to 0 at the rate of \( O(1/(k+1)^{\gamma}) \), which implies that the sequence \( \{(\alpha_k^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{f-W}\} \) is \( \ell_1 \)-summable, i.e., \( \sum_{k=0}^{\infty} (\alpha_k^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{f-W} < \infty \).

**Lemma 12** (convergence of weakly summable sequence). Let \( \{\beta_k\} \) and \( \{\gamma_k\} \) be two nonnegative scalar sequences such that
(a) \( \gamma_k = \frac{1}{\|x^{k+1}\|^2_{f-W}} \) for some \( \epsilon \in (0,1], \; k \in \mathbb{N} \); 
(b) \( \sum_{k=0}^{\infty} \gamma_k \beta_k < \infty \); 
(c) \( \beta_k - \beta_k \leq \gamma_k \),
where “\( \leq \)” means that \( |\beta_{k+1} - \beta_k| \leq M \gamma_k \) for some constant \( M > 0 \), then \( \lim_{k \to \infty} \beta_k = 0 \).

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 \( \beta_k \) converges to 0. However, if the distance of two successive steps of \( \{\beta_k\} \) with the same order of the multiplied sequence \( \gamma_k \), then we can claim the convergence of \( \beta_k \). A special case with \( \epsilon = 1/2 \) has been observed in [12].

**Proof:** By condition (b), we have
\[
\sum_{i=k}^{k+k'} \gamma_i \beta_i \to 0,
\] (47)
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., \( \beta_k \to 0 \) as \( k \to \infty \), then \( \lim \sup_{k \to \infty} \beta_k \leq C^* > 0 \). Thus, for every \( N > k_0 \), there exists a \( k > N \) such that \( \beta_k > \frac{C^*}{4} \). Let
\[
k' \triangleq \left\lfloor \frac{C^*}{4M} (k + 1) \right\rfloor,
\]
where \( \left\lfloor x \right\rfloor \) 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
\[
\beta_{k+i} \geq \frac{C^*}{4}, \quad \forall i \in \{0, 1, \ldots, k'\}.
\] (48)

Hence,
\[
\sum_{j=k}^{k+k'} \gamma_j \beta_j \geq \frac{C^*}{4} \sum_{j=k}^{k+k'} \gamma_j \geq \frac{C^*}{4} \int_k^{k+k'} (x+1)^{-\epsilon} \, dx \tag{49}
\]
\[
= \left\{ \begin{array}{ll}
\frac{C^*}{4(1-\epsilon)} ((k + k' + 1) - (k + 1))^{-\epsilon}, & \epsilon \in (0, 1), \\
\frac{C^*}{k} (\ln(k + k' + 1) - \ln(k + 1)), & \epsilon = 1.
\end{array} \right.
\]

Note that when \( \epsilon \in (0, 1) \), the term \( (k + k' + 1) - (k + 1)^{-\epsilon} \) 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
\[
\ln(k + k' + 1) - \ln(k + 1) = \ln \left( 1 + \frac{k'}{k + 1} \right) = \ln \left( 1 + \frac{C^*}{k+1} \right),
\]
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 (47). Therefore, \( \lim_{k \to \infty} \beta_k = 0 \).

**Proof of Theorem 2:** We first develop the following inequality
\[
L_{\alpha_k+1}(x^{k+1}) \leq L_{\alpha_k}(x^k) + \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{f-W} \\
- \frac{1}{2}(\alpha_k^{-1}(1 + \lambda_n(W))) - L_f \|x^{k+1} - x^k\|^2,
\] (50)
and then claim the convergence of the sequences \( \{L_{\alpha_k}(x^k)\} \) and \( \{x^k\} \) based on this inequality.

(a) **Development of (50):** From \( x^{k+1} = x^k - \alpha_k \nabla L_{\alpha_k}(x^k) \), it follows that
\[
\langle \nabla L_{\alpha_k}(x^k), x^{k+1} - x^k \rangle = -\frac{\|x^{k+1} - x^k\|^2}{\alpha_k}.
\] (51)

Since \( \sum_{i=1}^{n} \nabla f_i(x_i(k)) \) is \( L_f \)-Lipschitz, \( \nabla L_{\alpha_k} \) is Lipschitz with the constant \( L_k \triangleq L_f + \alpha_k^{-1}\lambda_{\text{max}}(I - W) = L_f + \alpha_k^{-1}(1 - \lambda_n(W)) \), implying
\[
L_{\alpha_k}(x^{k+1}) \leq L_{\alpha_k}(x^k) + \langle \nabla L_{\alpha_k}(x^k), x^{k+1} - x^k \rangle + \frac{L_k}{2}\|x^{k+1} - x^k\|^2
\] (52)
\[
= L_{\alpha_k}(x^k) - \frac{1}{2}(\alpha_k^{-1}(1 + \lambda_n(W)) - L_f)\|x^{k+1} - x^k\|^2.
\]

Moreover,
\[
L_{\alpha_{k+1}}(x^{k+1}) = L_{\alpha_k}(x^{k+1}) + \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{f-W}.
\] (53)
Combining (52) and (53) yields (50).

(b) **Convergence of objective sequence:** By Lemma 11 and Lemma 9, (50) yields the convergence of \( \{L_{\alpha_k}(x^k)\} \) and
\[
\sum_{k=0}^{\infty} (\alpha_k^{-1}(1 + \lambda_n(W))) - L_f \|x^{k+1} - x^k\|^2 < \infty
\] (54)
which implies that \( \|x^{k+1} - x^k\|^2 \) converges to 0 at the rate of \( o(k^{-\epsilon}) \) and \( \{x^k\} \) is asymptotic regular. Moreover, notice that
\[
\alpha_k^{-1}\|x^k\|^2_{f-W} = \alpha_k^{-1}\|x^k - \bar{x}^k\|^2_{f-W}
\]
\[
\leq (1 - \lambda_n(W))L_f(k + 1)^{\epsilon}\|x^k - \bar{x}^k\|^2.
\]
By Proposition 3, the term $\alpha_k^{-1}\|x^k\|^2 - W$ converges to 0 as $k \to \infty$. As a consequence,

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

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

$$
\alpha_k^{-1}(1 + \lambda_n(W)) - L_f = \alpha_k^{-1}(1 + \lambda_n(W)) - L_f \alpha_k
\geq \alpha_k^{-1} \left( 1 + \lambda_n(W) - \frac{1}{(k_0 + 1)^{\frac{1}{2}}} \right)
$$

for all $k > k_0$, where $k_0 = \left[ (1 + \lambda_n(W))^{-\frac{1}{2}} \right]$, i.e., the integer part of $(1 + \lambda_n(W))^{-\frac{1}{2}}$. Note that

$$
\|\tilde{x}^{k+1} - \tilde{x}^k\| = \frac{1}{n} I^T (x^{k+1} - x^k)
\leq \|x^{k+1} - x^k\|.
$$

Thus, (54), (55) and (56) yield

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

By the iterate (5) of DGD, we have

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

Plugging (58) into (57) yields

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

Moreover,

$$
\|\nabla f(x^{k+1})\|^2 - \|\nabla f(x^k)\|^2
\leq \|\nabla f(x^{k+1}) - \nabla f(x^k)\| \cdot (\|\nabla f(x^{k+1})\| + \|\nabla f(x^k)\|)
\leq 2B\|\nabla f(x^{k+1}) - \nabla f(x^k)\|
\leq 2BL_f\|x^{k+1} - x^k\|,
$$

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

$$
\|x^{k+1} - x^k\|
= \|x^{k+1} - x^k + \tilde{x}^{k+1} - \tilde{x}^k + \tilde{x}^k - x^k\|
\leq \|x^{k+1} - x^{k+1}\| + \|x^k - x^k\| + \alpha_k \|\nabla f(x^k)\|
\lesssim \alpha_k,
$$

where the first inequality holds for the triangle inequality and (58), and the last inequality holds for Proposition 3 and the bounded assumption of $\nabla f$. Thus, (60) and (61) imply

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

By the specific form (16) of $\alpha_k$, (59), (62) and Lemma 12, it holds

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

As a consequence,

$$
\lim_{k \to \infty} I^T \nabla f(x^k) = 0.
$$

Furthermore, by the coercivity of $f_i$ for each $i$ and the convergence of $\{I^T f(x^k)\}$, $\{x^k\}$ is bounded. Therefore, there exists a convergent subsequence of $\{x^k\}$. Let $x^*$ be any limit point of $\{x^k\}$. By (63) and the continuity of $\nabla f$, it holds

$$
I^T \nabla f(x^*) = 0.
$$

Moreover, by Proposition 3, $x^*$ is consensusal. As a consequence, $x^*$ is a stationary point of problem (3).

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

E. Proof for Proposition 4

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

Lemma 13 (Accumulated consensus of iterates). Under conditions of Proposition 3, we have

$$
\sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^k\| \leq D_1 + D_2 \sum_{k=0}^{K} \alpha_k^2,
$$

where $D_1 = C\|x_0\| \frac{\zeta}{2(1 - \zeta)}$, $D_2 = C \left( \frac{\|x_0\|}{2} + \frac{B}{1 - \zeta} \right)$, and $B$ is specified in Assumption 3.

Proof: By (31),

$$
\sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^k\| \leq C \|x_0\| \zeta \sum_{k=0}^{K} \alpha_k \zeta^k
+ CB \sum_{k=0}^{K} \sum_{j=0}^{K} \zeta^{k-j} \alpha_k \alpha_j.
$$

In the following, we estimate these two terms in the right-hand side of (66), respectively. Note that

$$
\sum_{k=0}^{K} \alpha_k \zeta^k \leq \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2 + \frac{1}{2} \sum_{k=0}^{K} \zeta^{2k}
\leq \frac{1}{2(1 - \zeta)} + \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2,
$$

and

$$
\sum_{k=0}^{K} \sum_{j=0}^{K} \zeta^{k-j} \alpha_k \alpha_j \leq \frac{1}{2} \sum_{k=0}^{K} \sum_{j=0}^{K} \zeta^{k-j} (\alpha_k^2 + \alpha_j^2)
= \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2 \sum_{j=0}^{K} \zeta^{k-j} + \frac{1}{2} \sum_{j=0}^{K} \alpha_j^2 \sum_{k=0}^{K} \zeta^{k-j}
\leq \frac{1}{2 - \zeta} \sum_{k=0}^{K} \alpha_k^2.
$$

Plugging (67) and (68) into (66) yields (65).
Besides Lemma 13, we also need the following two lemmas, which have appeared in the literature (cf. [10]).

**Lemma 14** ([10]). Let \( \gamma_k = \frac{1}{k^2} \) for some \( 0 < \epsilon \leq 1 \). Then the following hold:

(a) if \( 0 < \epsilon < 1/2 \), \[ \frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1 - \epsilon}{K^{1+\epsilon}} = O(\frac{1}{K^{1+\epsilon}}), \]
\[ \sum_{k=1}^{K} \frac{\gamma_k^2}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1 - \epsilon}{1 - 2\epsilon}, \quad K^{1-2\epsilon} = O(\frac{1}{K^{1-\epsilon}}). \]

(b) if \( \epsilon = 1/2 \), \[ \frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1 - \epsilon}{K^{1+\epsilon}} = O(\frac{1}{\sqrt{K}}), \]
\[ \sum_{k=1}^{K} \frac{\gamma_k^2}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1 + \ln K}{2(K^{1/2} - 1)} = O(\frac{\ln K}{\sqrt{K}}). \]

(c) if \( 1/2 < \epsilon < 1 \), \[ \frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1 - \epsilon}{K^{1+\epsilon}} = O(\frac{1}{K^{1+\epsilon}}), \]
\[ \sum_{k=1}^{K} \frac{\gamma_k^2}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1 - \epsilon}{2\epsilon - 1}, \quad K^{2\epsilon - 1} = O(\frac{1}{K^{1-\epsilon}}). \]

(d) if \( \epsilon = 1 \), \[ \frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1}{\ln K} = O(\frac{1}{\ln K}), \]
\[ \sum_{k=1}^{K} \frac{\gamma_k^2}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1 - 1/K}{2\ln K - 1} = O(\frac{1}{\ln K}). \]

**Lemma 15.** ([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 \( x, y, u \in \mathbb{R}^p \),
\[
h(u) \geq h(x) + \langle \nabla h(y), u - x \rangle - \frac{L_h}{2} \|x - y\|^2.
\]

**Proof of Proposition 4:** To prove this proposition, we first develop the following inequality,
\[
\mathcal{L}_{\alpha_k}(x^{k+1}) - \mathcal{L}_{\alpha_k}(u) \leq \frac{1}{2\alpha_k} \left( \|x^k - u\|^2 - \|x^{k+1} - u\|^2 \right)
\]  
(69)
for any \( u \in \mathbb{R}^{n \times p} \). By Lemma 15, we have
\[
\mathcal{L}_{\alpha_k}(u) \geq \mathcal{L}_{\alpha_k}(x^{k+1}) + \langle \nabla \mathcal{L}_{\alpha_k}(x^{k+1}), u - x^{k+1} \rangle - \frac{L^*}{2} \|x^{k+1} - x^k\|^2,
\]  
(70)
where \( L^* = L_f + \alpha_k^{-1}(1 - \lambda_n(W)) \), and by (23), we have \( \nabla \mathcal{L}_{\alpha_k}(x^k) = \alpha_k^{-1}(x^k - x^{k+1}) \). Then (70) implies
\[
\mathcal{L}_{\alpha_k}(u) \geq \mathcal{L}_{\alpha_k}(x^{k+1}) + \alpha_k^{-1}(x^k - x^{k+1}, u - x^{k+1}) - \frac{L^*}{2} \|x^{k+1} - x^k\|^2.
\]  
(71)
Note that the specific form of \( \alpha_k \) (\( \frac{1}{L_f(k+1)^{1/2}} \)), there exists an integer \( k_0 > 0 \) such that \( L^* \leq \alpha_k^{-1} \) for all \( k > k_0 \). Actually, for the simplicity of the proof, we can take \( \alpha_k < \frac{\lambda_n(W)}{L_f} \) starting from the first step so that \( L^* \leq \alpha_k^{-1} \) holds from the initial step. Thus, (71) implies
\[
\mathcal{L}_{\alpha_k}(u) \geq \mathcal{L}_{\alpha_k}(x^{k+1}) + \alpha_k^{-1}(x^k - x^{k+1}, u - x^{k+1}) - \frac{L^*}{2} \|x^{k+1} - x^k\|^2.
\]  
(72)
Recall that for any two vectors \( a \) and \( b \), it holds \( 2\langle a, b \rangle - \|a\|^2 = \|b\|^2 - \|a - b\|^2 \). Therefore,
\[
\mathcal{L}_{\alpha_k}(u) \geq \mathcal{L}_{\alpha_k}(x^{k+1}) + \frac{1}{2\alpha_k} \|u - x^{k+1}\|^2 - \|u - x^k\|^2.
\]
As a consequence, we get the basic inequality (69).

Note that the optimal solution \( x_{opt} \) is consensual and \( \|x_{opt}\|^2 = 0 \). Therefore, \( \mathcal{L}_{\alpha_k}(x_{opt}) = f(x_{opt}) = f_{opt} \). By (69), we have
\[
\alpha_k (\mathcal{L}_{\alpha_k}(x^{k+1}) - f_{opt}) \leq (\|x^k - x_{opt}\|^2 - \|x^{k+1} - x_{opt}\|^2)/2.
\]
Summing the above inequality over \( k = 0, 1, \ldots, K \) yields
\[
\sum_{k=0}^{K} \alpha_k (\mathcal{L}_{\alpha_k}(x^{k+1}) - f_{opt}) \leq \|x^0 - x_{opt}\|^2/2.
\]  
(73)
Moreover, noting that \( \mathcal{L}_{\alpha_k}(x^{k+1}) = f(x^{k+1}) \) and by the convexity of \( \mathcal{L}_{\alpha_k} \),
\[
\mathcal{L}_{\alpha_k}(x^{k+1}) \geq f(x^{k+1}) + \langle \nabla \mathcal{L}_{\alpha_k}(x^{k+1}), x^{k+1} - x^{k+1} \rangle
\]
\[
\geq f(x^{k+1}) - B\|x^{k+1} - x^{k+1}\|,
\]  
(74)
where the second inequality holds by the bounded assumption of gradient (cf. Assumption 3). Plugging (74) into (73) yields
\[
\sum_{k=0}^{K} \alpha_k (f(x^{k+1}) - f_{opt}) \leq \frac{1}{2} \|x^0 - x_{opt}\|^2 + B \sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^{k+1}\|.
\]
By the definition of \( f^K \) (18), then (75) implies
\[
(f^K - f_{opt}) \sum_{k=0}^{K} \alpha_k \leq \frac{1}{2} \|x^0 - x_{opt}\|^2 + B \sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^{k+1}\|
\]  
(76)
\leq D_3 + D_4 \sum_{k=0}^{K} \alpha_k^2,
\]  
(77)
where \( D_3 = \frac{1}{2} \|x^0 - x_{opt}\|^2 + BD_1, D_4 = BD_2, D_1 \) and \( D_2 \) are specified in Lemma 13, and the second inequality holds for Lemma 13. As a consequence,
\[
f^K - f_{opt} \leq D_3 + D_4 \sum_{k=0}^{K} \alpha_k^2.
\]  
(78)
Furthermore, by Lemma 14, we get the claims of this proposition.

**F. Proofs for Theorem 3 and Proposition 5**

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

**Lemma 16** (Sufficient descent of \( \{ \mathcal{L}_{a}(x^k) \} \)). Let Assumptions 2 and 4 hold. Results are given in two cases below:
Lemma 17 (Convexity). Let $\{\mathcal{L}_\alpha(x)\}$ be a sequence of functions, and let $r$ be a proper step size.

\[ r(x^{k+1}) - r(x^k) \leq \frac{1}{\alpha} \mathbb{E}(x^{k+1} - x^k)^2. \]

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

By Lemma 16 and under a proper step size, $\mathcal{L}_\alpha(x^k)$ is nonincreasing and upper bounded by $\mathcal{L}_\alpha(x^0)$, hence $\mathcal{L}_\alpha(x^k)$ is upper bounded by $\mathcal{L}_\alpha(x^0)$. Consequently, $\{x^k\}$ is bounded due to the coercivity of each $f_i + r_i$ (see Assumption 4 Part 2).

Lemma 18 (Bounded subgradient). Let $\partial \mathcal{L}_\alpha(x^k)$ denote the (limiting) subdifferential of $\mathcal{L}_\alpha$, which is assumed to exist for all $k \in \mathbb{N}$. Then, there exists $g^k \in \partial \mathcal{L}_\alpha(x^k)$ such that

\[ g^k \leq \mathbb{E} - \left(2 + \alpha^2(1 - \lambda_{n}/(W)) + L_f\right)\mathbb{E} - \left(1 + \alpha^2(1 - \lambda_{n}/(W)) + L_f\right)\mathbb{E} - \left(1 + \alpha^2(1 - \lambda_{n}/(W)) + L_f\right)\mathbb{E}. \]

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

\[ 0 \in \mathbb{E} - \left(2 + \alpha^2(1 - \lambda_{n}/(W)) + L_f\right)\mathbb{E} - \left(1 + \alpha^2(1 - \lambda_{n}/(W)) + L_f\right)\mathbb{E} - \left(1 + \alpha^2(1 - \lambda_{n}/(W)) + L_f\right)\mathbb{E}. \]

Thus, the claim of Lemma 18 holds.

Based on Lemmas 16–18, 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 (28), $\mathbb{E} = \mathbb{E}(\alpha^2(1 - \lambda_{n}/(W)) + L_f)$ if $r_i$'s are convex, while $\mathbb{E} = \mathbb{E}(\alpha^2(1 - \lambda_{n}/(W)) + L_f)$ if $r_i$'s are not necessarily convex and $\lambda_{n}/(W) > 0$.

**G. 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 19 (Recursion of $\{x^k\}$).** For any $k \in \mathbb{N}$,

\[ x^k = W^k x^0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j} (\nabla f(x^j) + \epsilon^{j+1}), \]

where $\epsilon^{j+1} \in \partial r(x^{j+1})$ 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

\[ x^{k+1} + \alpha_k \epsilon^{k+1} = W x^k - \alpha_k \nabla f(x^k), \]
where \( \xi^{k+1} \in \partial r(x^{k+1}) \), and thus
\[
x^{k+1} = Wx^k - \alpha_k (\nabla f(x^k) + \xi^{k+1}).
\] (85)

By (85), we can easily derive the recursion (83).

**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 \( \nabla f(x^k) + \xi^{k+1} \) is uniformly bounded by the constant \( B \) for any \( j \). Thus, we omit it here.

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

**Lemma 20.** 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 \( k \in \mathbb{N} \),
\[
\hat{L}_{\alpha_{k+1}}(x^{k+1}) \leq \hat{L}_{\alpha_k}(x^k) + \frac{1}{2}(\alpha_{k+1} - \alpha_k)\|x^{k+1}\|^2_{I-W} - \frac{1}{2}(\alpha_k^{-1}(1 + \lambda_n(W)) - L_f)\|x^{k+1} - x^k\|^2.
\] (86)

**C2:** \( r_i \)'s are not necessarily convex. For any \( k \in \mathbb{N} \),
\[
\hat{L}_{\alpha_{k+1}}(x^{k+1}) \leq \hat{L}_{\alpha_k}(x^k) + \frac{1}{2}(\alpha_{k+1} - \alpha_k)\|x^{k+1}\|^2_{I-W} - \frac{1}{2}(\alpha_k^{-1}\lambda_n(W) - L_f)\|x^{k+1} - x^k\|^2.
\] (87)

**Proof:** The proof of this lemma is similar to that of Lemma 16 via noting that
\[
\hat{L}_{\alpha_{k+1}}(x^{k+1}) = \hat{L}_{\alpha_k}(x^k) + (\hat{L}_{\alpha_{k+1}}(x^{k+1}) - \hat{L}_{\alpha_k}(x^{k+1})) + (\hat{L}_{\alpha_k}(x^{k+1}) - \hat{L}_{\alpha_k}(x^k)),
\]
and
\[
\hat{L}_{\alpha_{k+1}}(x^{k+1}) - \hat{L}_{\alpha_k}(x^k) = \frac{1}{2}(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|x^{k+1}\|^2_{I-W}.
\]

While the term \( \hat{L}_{\alpha_k}(x^{k+1}) - \hat{L}_{\alpha_k}(x^k) \) can be estimated similarly by the proof of Lemma 16.

**Lemma 21.** 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 \( u \in \mathbb{R}^{n \times p} \), we have
\[
\hat{L}_{\alpha_k}(x^{k+1}) - \hat{L}_{\alpha_k}(u) \leq \frac{1}{2\alpha_k}(\|x^k - u\|^2 - \|x^{k+1} - u\|^2).
\]

**Proof:** By Lemma 15, we have
\[
\hat{L}_{\alpha_k}(u) \geq \hat{L}_{\alpha_k}(x^{k+1}) + \langle \nabla \hat{L}_{\alpha_k}(x^k), u - x^{k+1} \rangle - \frac{L^*}{2}\|x^{k+1} - x^k\|^2,
\] (88)
where \( L^* = L_f + \alpha_k^{-1}(1 - \lambda_n(W)) \), and by the convexity of \( r \), we have
\[
r(u) \geq r(x^{k+1}) + \langle \xi^{k+1}, u - x^{k+1} \rangle,
\] (89)
where \( \xi^{k+1} \in \partial r(x^{k+1}) \) is the one determined by the proximal operator (8). By (85), it follows
\[
\hat{L}_{\alpha_k}(u) \geq \hat{L}_{\alpha_k}(x^{k+1}) + \alpha_k^{-1}(x^k - x^{k+1}) - \nabla \hat{L}_{\alpha_k}(x^k).
\] (90)

Plugging (90) into (89), and then summing up (88) and (89) yield
\[
\hat{L}_{\alpha_k}(u) \geq \hat{L}_{\alpha_k}(x^{k+1}) + \alpha_k^{-1}(x^k - x^{k+1}, u - x^{k+1}) - \frac{L^*}{2}\|x^{k+1} - x^k\|^2.
\] (91)

Similar to the rest proof of the inequality (69), we can prove this lemma based on (91).

**Proof of Theorem 4:** Based on Lemma 20 and Lemma 21, we can proof Theorem 4. The proof of Theorem 4(a)-(d) is similar to that of Theorem 2, where one minor difference is that (60) in the proof of Theorem 2 should be
\[
\|\nabla f(x^{k+1}) + \xi^{k+1}\|^2 - \|\nabla f(x^k) + \xi^k\|^2 \\
\leq \|\nabla f(x^{k+1}) + \xi^{k+1} - (\nabla f(x^k) + \xi^k)\| \times
\|\nabla f(x^{k+1}) + \xi^{k+1} + \xi^k\|
\leq 2B\|\nabla f(x^{k+1}) + \xi^{k+1} - (\nabla f(x^k) + \xi^k)\|
\leq 2B\|\nabla f(x^{k+1}) + \xi^{k+1} - (\nabla f(x^k) + \xi^k)\|
\leq 2B(L_f + L_e)\|x^{k+1} - x^k\|,
\] (92)
where \( \xi^k \triangleq \frac{1}{n} 1^n \xi^k \), 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.

**VII. 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.

**ACKNOWLEDGMENTS**

The work of J. Zeng has been supported in part by the NSF grants (61603162, 11501440) and the Doctoral start-up foundation of Jiangxi Normal University. The work of W. Yin has been supported in part by the NSF grant ECCS-1462398 and ONR grants N000141410683 and N000141210838.

**REFERENCES**
