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 sometimes. 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 and have guaranteed rates of convergence. It is worth noting that Prox-DGD can handle nonconvex nonsmooth functions if their proximal operators can be computed. Such functions include SCAD and ℓq 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 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:

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

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:

$$\text{minimize } s(x) = \sum_{i=1}^{n} (f_i(x) + r_i(x)), \quad (2)$$

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 specific examples include: (i) the distributed compressed sensing and machine learning problems, where $f_i$ is the data-fidelity term, which is often differentiable, and $r_i$ is a sparsity-promoting regularizer such as the ℓq (quasi)-norm with $0 \leq q \leq 1$ [21, 27]; (ii) optimization problems with per-agent constraints, where $f_i$ is a differentiable objective function of agent $i$ and $r_i$ is the indicator function of the constraint set of agent $i$, that is, $r_i(x) = 0$ if $x$ satisfies the constraint and $\infty$ otherwise [7, 20]. When $f_i$’s are convex, the existing algorithms include the (sub)gradient methods [6, 8, 16, 23, 28, 41, 46, 51], and the primal-dual domain methods such as the decentralized alternating direction method of multipliers (DADMM) [35, 36, 7], and EXTRA [37, 38]. However, when $f_i$’s are nonconvex, few algorithms have convergence guarantees. Some existing results include [3, 4, 13, 23, 24, 39, 40, 19, 42, 43, 48]. In spite of the algorithms and their analysis in these works, the convergence of the simple algorithm Decentralized Gradient Descent (DGD) [28] under nonconvex $f_i$’s is still unknown. Furthermore, although DGD is slower than D-ADMM and EXTRA on convex problems, DGD is simpler and thus easier to extend to a variety of settings such as [32, 45, 26, 15], 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 [46]. This motivates the use of decreasing step sizes such as in [8, 16]. Assuming $f_i$’s are convex and have Lipschitz continuous and bounded gradients, [8] shows that decreasing step sizes $\alpha_k = \frac{1}{\sqrt{k}}$ lead to a convergence rate $\mathcal{O}(\ln k)$ of the running best of objective errors. [16] uses nested loops and shows an outer-loop convergence rate $\mathcal{O}(\frac{1}{\sqrt{k}})$ 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 $\mathcal{O}(\ln k)$.

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 \( \alpha \) 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 \( \alpha_k = O(1/(k+1)^\epsilon) \) is used, where \( 0 < \epsilon \leq 1 \) 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 \( O(1/(k+1)^\epsilon) \). Moreover, we show the convergence of DGD to a stationary point of the original problem, and derive the convergence rates of DGD with different \( \epsilon \) 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 \( \bar{x}^k = \frac{1}{n} \sum_{i=1}^{n} x_i^k \) 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 V 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 2, 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 8, 16, 23, 3, 13, 23, 40, 45.

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 \( 1 \in \mathbb{R}^n \) 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 \( G \in \mathbb{R}^{n \times n} \), we let \( \|X\|^2_2 \triangleq \langle X, GX \rangle \) 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 \( x_{ij} \in \mathbb{R}^p \) denote the local copy of \( x \) at node \( i \). We reformulate the consensus problem (1) into the equivalent problem:

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

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

\[
\begin{pmatrix}
- & \nabla f_1(x_{11}) & - \\
- & \nabla f_2(x_{22}) & - \\
& \vdots & \ddots & \vdots \\
- & \nabla f_n(x_{nn}) & - \\
\end{pmatrix} \in \mathbb{R}^{n \times p}.
\]

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 [23] for (3) is described as follows:

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

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

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:

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

Let \( r(x) \triangleq \sum_{i=1}^{n} r_i(x_{ij}) \). The algorithm Prox-DGD can be applied to the above problem (5).

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

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

where the proximal operator is

\[
\text{prox}_{\alpha_k r}(x) \triangleq \arg\min_{u \in \mathbb{R}^{n \times p}} \left\{ \alpha_k r(u) + \frac{1}{2} \| u - x \|^2 \right\}.
\]

III. Assumptions and main results

This section presents all of our main results.
Decomposing the step size

**Definition 1 (Lipschitz differentiability).** A function \( h \) is called \( L \)-Lipschitz differentiable if \( h \) is differentiable and its gradient \( \nabla h \) is \( L \)-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 \( h \) is bounded from below and unbounded from above.

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

\[
\text{Definition 2 (Coercivity).} \quad \text{A function } h \text{ is called coercive if } \text{h} \text{ is bounded from below and unbounded from above.}
\]

**A. Definitions and assumptions**

\[
v \parallel, \forall u, v \in \text{dom}(h), \text{ where } L > 0 \text{ is its Lipschitz constant.}
\]
TABLE III

COMPARISONS ON SCENARIOS APPLIED FOR DIFFERENT NONCONVEX DECENTRALIZED ALGORITHMS.

<table>
<thead>
<tr>
<th>algorithm</th>
<th>smooth</th>
<th>nonsmooth</th>
<th>r_i</th>
<th>step size</th>
<th>network (W)</th>
<th>algorithm type</th>
<th>fusion scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>DGD (this paper)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>diminish</td>
<td>static</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓ (doubly)</td>
<td>dynamic</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>determin</td>
<td>deterministic</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>stochastic</td>
<td>ATC</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CTA</td>
<td>✓</td>
</tr>
<tr>
<td>Perturbed Push-sum [40]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓ (doubly)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ZENITH [13]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓ (doubly)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Proj SGD [3]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓ (doubly)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

† In this table, the full names of these abbreviations are listed 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: W1 = 1, or Wᵀ1 = 1, or both hold.

The next definition is a property that many functions have (see [44] Section 2.2) for examples) and can help obtain whole sequence convergence[†] from subsequence convergence.

**Definition 3** (Kurdyka–Łojasiewicz (KL) property [22, 5, 2]). A function \( h : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\} \) has the KL property at \( x^* \in \text{dom}(\partial h) \) if there exist \( \eta \in (0, +\infty) \), a neighborhood \( U \) of \( x^* \), and a continuous concave function \( \varphi : [0, \eta) \rightarrow \mathbb{R}_+ \) 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 \( U \cap \{ h(x^*) < h(x) < h(x^*) + \eta \} \), the KL inequality holds

\[
\varphi'(h(x) - h(x^*)) \cdot \text{dist}(0, \partial h(x)) \geq 1. \tag{8}
\]

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

**Assumption 1** (Objective). The objective functions \( f_i : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\}, i = 1, \ldots, n \), 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 \( L_f \triangleq \max_i L_{f_i} \). 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) \( \text{null}\{ I - W \} = \text{span}\{1\} \).
4. (Spectral property) \( I \succeq W \succ -I \).

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

\[
\lambda_1(W) = 1 > \lambda_2(W) \geq \cdots \geq \lambda_n(W) > -1. \tag{9}
\]

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

\[
\zeta = \max\{|\lambda_2(W)|, |\lambda_n(W)|\}. \tag{9}
\]

**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 [46]:

\[
L_{\alpha}(x) \triangleq 1^T f(x) + \frac{1}{2\alpha} \|x\|^2 - W. \tag{10}
\]

It is worth reminding that convexity is not assumed.

**Theorem 1** (Global convergence). Let \( \{x^k\} \) be the sequence generated by DGD [4] with the step size \( 0 < \alpha < \frac{1}{\lambda_2(W)} \).

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 rate of the sequence\[4\] of \( \|x^k + 1 - x^k\|^2 \) and \( \|\nabla L_{\alpha}(x^k)\|^2 \) is \( o(1) \).

In addition, if \( L_{\alpha} \) 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 \( L_{\alpha}(x) \), and thus

\[
0 = \nabla f(x^*) + \alpha^{-1}(I - W)x^*. \tag{11}
\]

Since \( 1^T(I - W) = 0 \), \( \{1\} \) yields \( 0 = 1^T \nabla f(x^*) \), indicating that \( x^* \) is also a stationary point to the separable function

\[
\sum_{i=1}^n f_i(x_i). \tag{11}
\]

Since the rows of \( x^* \) are not necessarily

\[4\] 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 \( b_k = o(1/k) \).

\[5\] These quantities naturally appear in the analysis, so we keep the squares.
identical, we cannot say $x^*$ is a stationary point to Problem (5). However, the differences between the rows of $x^*$ are bounded, following our next result below adapted from [36]:

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

$$||x_i^k - \bar{x}|| \leq \frac{\alpha D}{1 - \zeta},$$

(12)

where $D$ is a universal bound of $||\nabla f_i(x^k)||$ defined in Lemma 6 below, $\zeta$ is the second largest magnitude eigenvalue of $W$ specified in [9]. As $k \to \infty$, (12) yields the consensual bound

$$||x_i^k - \bar{x}^*|| \leq \frac{\alpha D}{1 - \zeta},$$

(13)

where $\bar{x}^* \triangleq \frac{1}{n} \sum_{i=1}^n x_i^*$. In Proposition 4, the consensal 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 (14) for $f(x)$. Averaging the rows of (4) yields the following comparison:

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

(13)

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

(14)

Apparenty, DGD approximates centralized gradient descent by evaluating $\nabla f_i(x^k)$ at local variables $x_i^k$ instead of the global average. We can estimate the error of this approximation as

$$\frac{1}{n} \sum_{i=1}^n \nabla f_i(x_i^k) - \frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k)$$

$$\leq \frac{1}{n} \sum_{i=1}^n ||\nabla f_i(x_i^k) - \nabla f_i(x^k)|| \leq \frac{\alpha D L_f}{1 - \zeta}.$$

Unlike the convex analysis in [40], it is impossible to bound the difference between the sequences of (15) and (14) without convexity because the two sequences may converge to different stationary points of $L_\alpha$.

**Remark 2.** The KL assumption on $L_\alpha$ in Theorem 7 can be satisfied if each $f_i$ is a sub-analytic function. Since $||x||_W^{2-W}$ is obviously sub-analytic and the sum of two sub-analytic functions remains sub-analytic, $L_\alpha$ is sub-analytic if each $f_i$ is so. See [44 Section 2.2] for more details and examples.

**Proposition 2** (KL convergence rates). Let the assumptions of Theorem 7 hold. Suppose that $L_\alpha$ satisfies the KL inequality at an accumulation point $x^*$ with $\psi(s) = cs^{1-\theta}$ 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 $\theta \in (0, \frac{1}{2}]$, $||x^k - x^*|| \leq C_0\tau^k$ for all $k \geq k^*$ for some $k^* > 0$, $C_0 > 0$, $\tau \in (0, 1)$. 
(c) If $\theta \in (\frac{1}{2}, 1)$, $||x^k - x^*|| \leq C_0\tau^{(1-\theta)/(2\theta-1)}$ 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 $L_\alpha$ 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 $\alpha_k = O\left(\frac{1}{(k+1)^\epsilon}\right)$, 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 (5). 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, [31], [4], [13], [23], [24], [39], [40], [19], [43] for example), even in the convex setting [16] and also [28], though it is not required for centralized gradient descent.

We take the step size sequence:

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

(15)

throughout the rest part of this section. (The numerator 1 can be replaced by any positive constant.) By iteratively applying iteration (4), we obtain the following expression

$$x^k = W^k x_0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j} \nabla f(x^j).$$

(16)

**Proposition 3** (Asymptotic consensus). Let Assumptions 2 and 3 hold. Let DGD use (15). Let $\tilde{x}^k \triangleq \frac{1}{n} 1^T x^k$. Then, $\|x^k - x^*\|_W$ converges to 0 at the rate of $O(1/(k+1)^\epsilon)$.

According to Proposition 3, the itterates 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 $\|x^k\|_W = \|x^k - \tilde{x}^k\|_W$. Therefore, the above proposition implies the following result.

**Corollary 1.** Apply the setting of Proposition 3. $\|x^k\|_W$ converges to 0 at the rate of $O(1/(k+1)^{2\epsilon})$.

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

**Theorem 2** (Convergence). Let Assumptions 1, 2 and 3 hold. Let DGD use step sizes (15). Then

(a) $\{L_\alpha(x^k)\}$ and $\{1^T f(x^k)\}$ 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 \( o(1/k^1+) \). 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 \( x \in \mathbb{R}^{n \times n} \), let \( \bar{f}(x) \triangleq \sum_{i=1}^{n} f_i(x_{(i)}) \). Even if \( f_i \)'s are convex, the solution to (5) may be non-unique. Thus, let \( \lambda^* \) be the set of solutions to (5). Given \( x^k \), we pick the solution \( x_{opt} = \text{Proj}_{\lambda^*}(x^k) \in \lambda^* \). Also let \( f_{opt} = \bar{f}(x_{opt}) \) be the optimal value of (1). Define the ergodic objective:
\[
\bar{f}^K = \sum_{k=0}^{K} \frac{\alpha_k}{\sum_{k=0}^{K} \alpha_k} \bar{f}(x^{k+1}),
\]
where \( x^{k+1} = \frac{1}{n}(I^T x^{k+1}) I \). Obviously,
\[
\bar{f}^K \geq \min_{k=1,\ldots,K+1} \bar{f}(x^k).
\]

**Proposition 4** (Convergence rates under convexity). Let Assumptions 1, 2 and 3 hold. Let DGD use step sizes (15). If \( \lambda_n(W) > 0 \) and each \( f_i \) is convex, then \( \{\bar{f}^K\} \) defined in (17) converges to the optimal objective value \( f_{opt} \) at the following rates:

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

The convergence rates established in Proposition 4 almost as good as \( O\left(\frac{1}{\sqrt{K}}\right) \) when \( \epsilon = \frac{1}{2} \). 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.

**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 (5) can be reformulated as
\[
x^{k+1} = \text{prox}_{\alpha_k r}(x^k - \alpha_k \nabla L_{\alpha_k}(x^k))
\]

based on which, we define the Lyapunov function
\[
\hat{L}_{\alpha_k}(x) \triangleq L_{\alpha_k}(x) + r(x),
\]
where we recall \( L_{\alpha_k}(x) = \sum_{i=1}^{n} f_i(x_{(i)}) + \frac{\alpha_k}{2}\|x\|^2_{-W} \).

Further, (19) is clearly the forward-backward splitting (a.k.a., prox-gradient) iteration for minimizing \( L_{\alpha_k}(x) \). Specifically, (19) first performs gradient descent to the differentiable function \( L_{\alpha_k}(x) \) 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 (5) 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, \( \sum_{i=1}^{n} f_i(x_{(i)}) \) 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 \( \{x^k\} \) be the sequence generated by Prox-DGD (6) where the step size \( \alpha \) satisfies \( 0 < \alpha < \frac{1 + \lambda_n(W)}{L_f} \) when \( r_i \)'s are convex; and \( 0 < \alpha < \frac{\lambda_n(W)}{L_f} \), when \( r_i \)'s are not necessarily convex (this case requires \( \lambda_n(W) > 0 \)). Let Assumptions 2, 4 and 3 hold. Then \( \{x^k\} \) has at least one accumulation point \( x^\ast \), and any accumulation point is a stationary point of \( \hat{L}_{\alpha_k}(x) \). Furthermore, the running best rate \( \bar{f}^K \) of the sequences \( \{\|x^{k+1} - x^k\|^2\} \) and \( \|g^{k+1}\|^2 \) is defined in Lemma 8 are both \( o\left(\frac{1}{K}\right) \).

In addition, if \( \hat{L}_{\alpha_k} \) satisfies the KL property at an accumulation point \( x^\ast \), then \( \{x^k\} \) converges to \( x^\ast \).

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 \( \hat{L}_{\alpha_k} \) satisfies the KL inequality at an accumulation point \( x^\ast \) with \( v(x) = c_1 s^{1-\theta} \) for some constant \( c_1 > 0 \). Then the following hold:

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

(b) If \( \theta \in (0, \frac{1}{2}) \), \( \|x^k - x^\ast\| \leq C_1 r^k \) for all \( k \geq k^* \) for some \( k^* > 0 \), \( C_1 > 0 \), \( r \in (0, 1) \).

(c) If \( \theta \in (\frac{1}{2}, 1) \), \( \|x^k - x^\ast\| \leq C_1 k^{-(1-\theta)/(2\theta - 1)} \) 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 (15). To investigate its convergence, the bounded gradient Assumption 3 should be revised 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, \( \|\xi_i\| \leq B_{r_i} \) for any \( \xi_i \in \partial r_i(x) \) and \( x \in \mathbb{R}^p \). Let \( B \triangleq \sum_{i=1}^{n} (B_i + B_{r_i}) \). Then \( \nabla f(x) + \xi \) (where \( \xi \in \partial r(x) \) for any \( x \in \mathbb{R}^{n \times n} \)) is uniformly bounded by \( B \). Note that the same assumption is used to analyze the convergence

\(^5\)A nonnegative sequence \( a_k \) induces its running best sequence \( b_k = \min\{a_i : i \leq k\} \); therefore, \( a_k \) has running best rate of \( o(1/k) \) if \( b_k = o(1/k) \).
Let $x^k - \bar{x}$ hold for any $x \in \mathbb{R}^{n \times p}$, define $\tilde{s}(x) = \sum_{i=1}^{n} f_i(x(i)) + r_i(x(i))$. Let $\lambda^l$ be a set of solutions of (5), $x_{\text{opt}} = \text{Proj}_{\lambda^l}(x^k) \in \lambda^l$, and $s_{\text{opt}} = \tilde{s}(x_{\text{opt}})$. Define

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

(20)

Theorem 4 (Convergence and rate). Let Assumptions 2, 4 and 5 hold. In Prox-DGD, use the step sizes (15). Then

(a) $\{\hat{L}_{\alpha_k}(x^k)\}$ and $\{\sum_{i=1}^{n} f_i(x(i)) + r_i(x(i))\}$ converge to the same limit;

(b) $\sum_{k=0}^{\infty} (\alpha_k^{-1}(1 + \lambda_k(W)) - L_f) \| x^{k+1} - x^k \|^2 < \infty$ when $r_i$'s are convex; or, $\sum_{k=0}^{\infty} (\alpha_k^{-1} \lambda_n(W) - L_f) \| x^{k+1} - x^k \|^2 < \infty$ when $r_i$'s are not necessarily convex (this case requires $\lambda_n(W) > 0$);

(c) if $\{x^k\}$ satisfies $\| x^{k+1} - x^k \| \leq L_r \| x^{k+1} - x^k \|$ for each $k > k_0$, some constant $L_r > 0$, and a sufficiently large integer $k_0 > 0$, then

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

where $\xi^{k+1} \in \partial r_i(x^{k+1})$ is the one determined by the proximal operator (7), and any limit point is a stationary point of problem (5).

(d) in addition, if there exists an isolated accumulation point, then $\{x^k\}$ converges,

(e) furthermore, if $f_i$ and $r_i$ are convex and $\lambda_n(W) > 0$, then the claims on the rates of $\{\hat{f}_i\}$ in Proposition 2 hold for the sequence $\{\tilde{s}^K\}$ defined in (20).

Theorem 3(b) implies that the best running rate of $\| x^{k+1} - x^k \|^2$ is of $O(1/k)$. The additional condition imposed on $\{\xi^k\}$ in Theorem 4(c) is some type of restricted continuous regularity of the subgradient $\partial r$ with respect to the generated sequence, which may be held for a class of proximal functions as studied in (7). If $\partial r$ is locally Lipschitz continuous in a neighborhood of a limit point, then such condition can generally be satisfied, since $\{x^k\}$ is asymptotic regular, and thus $x^k$ will lies in such neighborhood of this limit point when $k$ is sufficiently large. Theorem 4(e) gives the convergence rates of Prox-DGD in the convex setting.

IV. Related works and discussions

We summarize some recent nonconvex decentralized algorithms in Table [11]. 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 [40], the authors proved the convergence of perturbed push-sum for nonconvex (1) under some regularity assumptions. They also introduced random perturbations to avoid local minima. The network considered in [40] is time-varying and directed, and specific column stochastic matrices and diminishing step sizes are used. Their algorithm is an extension of DGD with diminishing step sizes of this paper. The convergence results for the deterministic perturbed push-sum algorithm obtained in [40] are similar to those of DGD developed in this paper under similar assumptions (see, Theorem 2 above and 40 Theorem 3). (However, in this paper, we obtain the asymptotic consensus and convergence to a stationary point of DGD via a Lyapunov function and developing several new results such as Lemma 12 for the convergence of the so-called weakly-summable sequence.) The proofs in [40] are mainly based on [30] Theorem 2.7.3. In [13], 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.

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 [24], 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, Prox-DGD is simpler than NEXT. Another interesting algorithm is decentralized Frank-Wolfe (DeFW) proposed in [43] 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 [43], the authors established convergence results similar to Prox-DGD under diminishing step sizes. The stochastic version of DeFW has also been developed in [19] for high-dimensional convex sparse optimization. The last one is projected stochastic gradient algorithm (Proj SGD) [3] for constrained, nonconvex, smooth consensus optimization. It has two steps at each iteration: a projected stochastic gradient step to update local variables and a consensus step to exchange the information between local agents. The mixing matrix used in this algorithm is

1 The original form of this algorithm, push-sum, was proposed in [47] for the average consensus problem. It was modified and analyzed in [29] for convex consensus optimization problem over time-varying directed graphs.
random and row stochastic, but its expectation is column stochastic. Asymptotic consensus and convergence to the set of Karush-Kuhn-Tucker points were proved under diminishing step sizes, smooth objective function, some mean and variance restrictions to the stochastic direction, and other assumptions on the mixing matrices and the constraint set.

Based on the above analysis, the convergence results of DGD and Prox-DGD with diminishing step sizes of this paper are comparable with most of the existing ones, which involve more complicated methods. However, we allow non-convex 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 proofs of our main theorems and propositions.

A. Proof for Theorem 7

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 [46]; 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 (4) 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 [46], and it is based on reformulating (4) as the iteration:

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

(21)

Although the sequence $\{x^k\}$ generated by the DGD iteration (4) 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 7 and 2 hold. Set the step size $0 < \alpha < \frac{1 + \lambda_n(W)}{L_f}$. It holds that

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

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

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

(23)

Since $\sum_{k=1}^{\infty} \nabla f_i(x(t))$ 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

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

(24)

Combining (23) and (24) yields (22).

Lemma 3 (Boundedness). Under Assumptions 7 and 2 if $0 < \alpha < \frac{1 + \lambda_n(W)}{L_f}$, then the sequence $\{L_\alpha(x^k)\}$ 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 (\(f^2\)-summable and asymptotic regularity). It holds that $\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2 < +\infty$ and that $\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$.

From (21), 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 $\{x^k\} \to x^* \text{ as } s \to +\infty$. By Lemmas 4 and 3, $L_\alpha(x^k)$ is monotonically nonincreasing and lower bounded, and therefore $\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$. Based on Lemma 5, $\|\nabla L_\alpha(x^*)\| \to 0$ as $s \to +\infty$. Hence, we have $\nabla L_\alpha(x^*) = 0$.

The running best rate of the sequence $\{\|x^{k+1} - x^k\|^2\}$ follows from (10) Lemma 1.2 or (18) Theorem 3.3.1. By Lemma 5 the running best rate of the sequence $\{\|\nabla L_\alpha(x^k)\|^2\}$ is $o(\frac{1}{k})$.

Similar to [2] Theorem 2.9, we can claim the global convergence of the considered sequence $\{x^k\}_{k \in \mathbb{N}}$ under the KL assumption of $L_\alpha$.

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

Lemma 6. Under Assumption 7 there exists a point $y^*$ satisfying $\nabla f(y^*) = 0$, and the following bound holds

$$\|\nabla f(x^k)\| \leq D \triangleq L_f(B + \|y^*\|), \; \forall k \in \mathbb{N},$$

(25)

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

**A sequence $\{a_k\}$ is said to be asymptotic regular if $\|a_{k+1} - a_k\| \to 0$ as $k \to \infty$.**
Proof. By the lower boundedness assumption (Assumption 1 Part 2)), the minimizer of \(1^T f(y)\) exists. Let \(y^*\) be a minimizer. Then by Lipschitz differentiability of each \(f_i\) (Assumption 1 Part 1)), we have that \(\nabla f(y^*) = 0\).

Then, for any \(k\), we have
\[
\|\nabla f(x^k)\| = \|\nabla f(x^k) - \nabla f(y^*)\| \leq L_f \|x^k - y^*\|
\]

(Lemma 3) \(\leq L_f (B + \|y^*\|)\).

Therefore, we have proven this lemma. \(\square\)

B. Proof for Proposition 2

Proof. Note that
\[
\|\nabla \mathcal{L}_\alpha(x^{k+1})\| \leq \|\nabla \mathcal{L}_\alpha(x^{k+1}) - \nabla \mathcal{L}_\alpha(x^k)\| + \|\nabla \mathcal{L}_\alpha(x^k)\| \\
\leq L^* \|x^{k+1} - x^k\| + \alpha \|x^{k+1} - x^k\| \\
= (\alpha^{-1} + (2 - \lambda_n(W)) + L_f) \|x^{k+1} - x^k\|
\]

where the second inequality holds for Lemma 5 and the Lipschitz continuity of \(\nabla \mathcal{L}_\alpha\) with constant \(L^* = L_f + \alpha^{-1}(1 - \lambda_n(W))\). 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 \(\mathcal{L}_\alpha\) at \(x^*\) with \(\psi(s) = cs^{1-\theta}\), following the proof of [2] Lemma 2.6], there exists \(k_0 > 0\) such that for all \(k \geq k_0\), we have
\[
2 \|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\| + \frac{cb}{a}\times
\]
\[
((\mathcal{L}_\alpha(x^k)) \mathcal{L}_\alpha(x^k) - (\mathcal{L}_\alpha(x^k))^{1-\theta} - (\mathcal{L}_\alpha(x^{k+1}) - \mathcal{L}_\alpha(x^k))^{1-\theta}),
\]

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
\[
\sum_{t=k_0}^{k} \|x^{t+1} - x^t\| \leq \|x^{k_0} - x^{k_0-1}\| + \frac{cb}{a}\times
\]
\[
((\mathcal{L}_\alpha(x^{k_0})) \mathcal{L}_\alpha(x^{k_0}) - (\mathcal{L}_\alpha(x^{k_0+1}) - \mathcal{L}_\alpha(x^{k_0}))^{1-\theta}).
\]

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 \(\theta\). \(\square\)

C. Proof for Proposition 3

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

Lemma 7. (28 Proposition 11) Let \(W_k \triangleq W \cdot \cdots \cdot W\) 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 (9).

Lemma 8. (23 Lemma 3.11) Let \(\{\gamma_k\}\) be a scalar sequence. If \(\lim_{k \to \infty} \gamma_k = \gamma\) and \(0 < \beta < 1\), then \(\lim_{k \to \infty} \sum_{l=0}^{k} \beta^{k-l} \gamma_l = \frac{1}{1 - \beta}\).

Proof of Proposition 3 By the recursion (16), note that
\[
x^k - x^k = (W^k - \frac{1}{n}11^T) x^0 \\
- \sum_{j=0}^{k-1} \alpha_j (W^{k-1-j} - \frac{1}{n}11^T) \nabla f(x^j).
\]

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

Furthermore, by Lemma 8 and step sizes (15), we get \(\lim_{k \to \infty} \|x^k - x^k\| = 0\).

Let \(b_k \triangleq (k + 1)^{-\epsilon}\). To show the rate of \(\|x^k - x^k\|\), we only need to show that
\[
\lim_{k \to \infty} b_k^{-1} \|x^k - x^k\| \leq C^\ast
\]
for some \(0 < C^\ast < \infty\). Let \(j_k \triangleq \lfloor k - 2 \log_\zeta(b_k^{-1}) \rfloor\) (where \(\lfloor x \rfloor\) denotes the integer part of \(x\) for any \(x \in \mathbb{R}\)). Note that
\[
b_k^{-1} \|x^k - x^k\| \\
\leq Cb_k^{-1} \left( \|x^0\| \zeta^k + B \sum_{j=0}^{k-1} \alpha_j \zeta^{k-j} \right) \\
= C \|x^0\| b_k^{-1} \zeta^k + CBb_k^{-1} \sum_{j=0}^{j_k} \alpha_j \zeta^{j_k-j} \\
+ Cb_k^{-1} \sum_{j=j_k+1}^{k-1} \alpha_j \zeta^{j-k-j} \\
\triangleq T_1 + T_2 + T_3,
\]
where the first inequality holds because of (29).

In the following, we will estimate the above three terms in the right-hand side of (30), respectively. First, by the definition of \(j_k\), for any \(j \leq j_k\), we have
\[
b_k^{-1} \zeta^{j-k-j} \leq b_k^{-1} \zeta^{j-k-j} \leq 1.
\]

Thus,
\[
T_2 \leq CB \sum_{j=0}^{j_k} \alpha_j \zeta^{(k-j)}/2.
\]

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
\[
b_k^{-1} \alpha_j \geq \frac{(k + 1)^\epsilon}{L_f(k + 1)^\epsilon} = \frac{1}{L_f},
\]

where \(\epsilon > 0\).
Thus, for any \( j_k' < j \leq k - 1 \)
\[
\lim_{k \to \infty} b_k^{-1} \alpha_j = \frac{1}{L_f}.
\]

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

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

The above two inequalities imply that for sufficiently large \( k \),
\[
\begin{align*}
T_1 &\leq C||x_0||\zeta^{k/2}, \\
T_3 &\leq 2CB\frac{L_f}{2} \sum_{j=j_k'+1}^{k-1} \zeta^{k-1-j}.
\end{align*}
\]

From (31), (36) and (37), we get
\[
\begin{align*}
b_k^{-1}||x_k - \hat{x}|| &\leq C||x_0||\zeta^{k/2} \\
&+ CB \left( \sum_{j=0}^{j_k'} \alpha_j \zeta^{(k-1-j)/2} + \frac{2}{L_f} \sum_{j=j_k'+1}^{k-1} \zeta^{k-1-j} \right).
\end{align*}
\]

By Lemma 8 and (38), there exists a \( C^* > 0 \) such that
\[
\lim_{k \to \infty} b_k^{-1}||x_k - \hat{x}|| \leq C^*.
\]

We have completed the proof of this proposition. \( \square \)

D. Proof for Theorem 2

To prove Theorem 2, we first note that similar to (21), the DGD iterates under decreasing step sizes can be rewritten as
\[
x_{k+1} = x_k - \alpha_k \nabla f(x_k),
\]
where \( \nabla f(x) = 1^T f(x) + \frac{1}{2\delta} ||x||^2_{F_W} \), and we also need the following lemmas.

Lemma 9 ([34]). 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 (15). Then it holds
\[
\alpha_{k+1} - \alpha_k \leq 2e L f(k+1)^{e-1}.
\]

Proof. We first prove that
\[
(1 + x)^e - 1 \leq 2e x, \quad \forall x \in [0, 1].
\]

Let \( g(x) = (1 + x)^e - 1 - 2e x. \) Then its derivative
\[
g'(x) = e(1 + x)^{e-1} - 2e < 0, \quad \forall x \in [0, 1].
\]

It implies \( g(x) \leq g(0) = 0 \) for any \( x \in [0, 1] \), that is, the inequality (41) holds.

Note that
\[
\begin{align*}
\alpha_{k+1} - \alpha_k &= L f((k+2)^e - (k+1)^e) \\
&= L f(k+1)^e(1 + \frac{1}{k+1})^e - 1) \\
&\leq 2e L f(k+1)^{e-1} - 1,
\end{align*}
\]

where the last inequality holds for (41). \( \square \)

Note that the term \( \{(\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} \} \) exists in the right hand side of the latter inequality (48). In order to apply Lemma 9 and then show the convergence of \( \{L_{\alpha_k}(x_k)\} \), we need the following lemma to guarantee that \( \{(\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} \} \) is summable.

Lemma 11. Let Assumptions 7, 2 and 8 hold. In DGD, use step sizes \( \alpha_k \) in (15). Then \( \{(\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} \} \) is summable, i.e., \( \sum_k^\infty (\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} < \infty \).

Proof. Note that
\[
||x_k||^2_{F_W} = ||x_k - x_k||^2_{F_W} \leq (1 - \lambda_n(W))||x_k - \hat{x}||^2_{F_W}.
\]

By Lemma 10,
\[
\begin{align*}
(\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} &\leq 2e L f(k+1)^{e-1}||x_k||^2_{F_W} - \lambda_n(W)||x_k - \hat{x}||^2_{F_W} \\
&\leq 2e L f(k+1)^{e-1}||x_k||^2_{F_W} - \lambda_n(W)||x_k - \hat{x}||^2_{F_W}.
\end{align*}
\]

Furthermore, by (44) and Proposition 3, the sequence \( \{(\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} \} \) converges to 0 at the rate of \( O(1/(k+1)^{1+\epsilon}) \), which implies that the sequence \( \{(\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} \} \) is \( \ell_1 \)-summable, i.e., \( \sum_{k=0}^\infty (\alpha_{k+1} - \alpha_k^-)||x_k||^2_{F_W} < \infty \). \( \square \)

Lemma 12 (convergence of weakly summable sequence). Let \( \{\beta_k\} \) be a nonnegative scalar sequence such that
\[
\begin{align*}
&\text{(a)} \quad \gamma_k = \frac{1}{(k+1)^{1+\epsilon}}, \quad \text{for some } \epsilon \in (0, 1], \quad k \in \mathbb{N}; \\
&\text{(b)} \quad \sum_{k=0}^\infty \gamma_k < \infty; \\
&\text{(c)} \quad |\beta_{k+1} - \beta_k| \leq \gamma_k,
\end{align*}
\]
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 [9].

Proof. By condition (b), we have
\[
\sum_{k=0}^\infty \gamma_i \beta_i = 0,
\]

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 \( \lim \sup_{k \to \infty} \beta_k \triangleq C^* > 0 \). 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

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

(46)

Hence,

\[
k' + k' \sum_{j=0}^{k'} \gamma_j \beta_j \geq \frac{C^*}{4} \left( \sum_{j=0}^{k'} \gamma_j \right) \geq \frac{C^*}{4} k' \left( (k + 1)^{1-\epsilon} - (k + 1) \right)
\]

\[
= \frac{C^*}{4} \epsilon \left( (k + 1)^{1-\epsilon} - (k + 1) \right), \quad \epsilon \in (0, 1),
\]

\[
\epsilon = 1.
\]

Note that when \( \epsilon \in (0, 1) \), the term \((k + 1)^{1-\epsilon} - (k + 1)\) is monotonically increasing with respect to \( k \), which implies that \( \sum_{j=0}^{k'} \gamma_j \beta_j \) 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

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

which is a positive constant. As a consequence, \( \sum_{j=0}^{k'} \gamma_j \beta_j \) will not go to zero as \( k \to 0 \), which contradicts with (45). Therefore, \( \lim_{k \to \infty} \beta_k = 0. \)

**Proof of Theorem 2** We first develop the following inequality.

\[
\mathcal{L}_{\alpha_{k+1}}(x^{k+1}) \leq \mathcal{L}_{\alpha_k}(x^k) + \left( \frac{1}{2} \alpha_k^{-1} - \omega_k^{-1} \right) \|x^{k+1} - x^k\|^2 + \frac{1}{2} \left( \alpha_k^{-1} \right) \|x^{k+1} - x^k\|^2 - \frac{1}{2} \left( \alpha_k^{-1} \right) (1 + \lambda_n(W)) - L_f \|x^{k+1} - x^k\|^2.
\]

(48)

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

(a) **Development of (48)**: From \( x^{k+1} = x^k - \alpha_k \nabla \mathcal{L}_{\alpha_k}(x^k) \), it follows that

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

(49)

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

\[
\mathcal{L}_{\alpha_k}(x^{k+1}) \leq \mathcal{L}_{\alpha_k}(x^k) + \langle \nabla \mathcal{L}_{\alpha_k}(x^k), x^{k+1} - x^k \rangle + \frac{L_k}{2} \|x^{k+1} - x^k\|^2
\]

\[
= \mathcal{L}_{\alpha_k}(x^k) - \frac{1}{2} \left( \alpha_k^{-1} \right) (1 + \lambda_n(W)) - L_f \|x^{k+1} - x^k\|^2.
\]

Moreover,

\[
\mathcal{L}_{\alpha_{k+1}}(x^{k+1}) = \mathcal{L}_{\alpha_k}(x^{k+1}) + \frac{1}{2} \left( \alpha_k^{-1} - \omega_k^{-1} \right) \|x^{k+1}\|^2_{W^{-1}}.
\]

(51)

Combining (50) and (51) yields (48).

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

\[
\sum_{k=0}^\infty (\alpha_k^{-1} (1 + \lambda_n(W)) - L_f) \|x^{k+1} - x^k\|^2 < \infty
\]

(52)

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.\[\dagger\]

Moreover, notice that

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

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

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

\[
= \lim_{k \to \infty} \mathcal{L}_{\alpha_k}(x^k).
\]

(c) **Convergence to a stationary point**: Let \( \bar{\nabla} f(x^k) \triangleq 1^1 T \bar{\nabla} f(x^k) \). By the specific form [13] of \( \alpha_k \), we have

\[
\alpha_k^{-1} (1 + \lambda_n(W)) - L_f \alpha_k^{-1} (1 + \lambda_n(W)) - L_f \alpha_k \leq \alpha_k^{-1} (1 + \lambda_n(W)) - \frac{1}{(k_0 + 1)^\epsilon}
\]

(53)

Thus, (52), (53) and (54) yield

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

(55)

By the iterate (41) of DGD, we have

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

(56)

Plugging (56) into (55) yields

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

(57)

Moreover,

\[
\|\bar{\nabla} f(x^{k+1})\|^2 - \|\bar{\nabla} f(x^k)\|^2 \leq \|\bar{\nabla} f(x^{k+1}) - \bar{\nabla} f(x^k)\| (\|\bar{\nabla} f(x^{k+1})\| + \|\bar{\nabla} f(x^k)\|)
\]

\[
\leq 2B \|\bar{\nabla} f(x^{k+1}) - \bar{\nabla} f(x^k)\| \leq 2B \|\bar{\nabla} f(x^{k+1}) - \bar{\nabla} f(x^k)\| \leq 2BL_f \|x^{k+1} - x^k\|.
\]

(58)

where the second inequality holds by the bounded gradient assumption (Assumption 3), the third inequality holds by the

\[\dagger\] A sequence \( \{\alpha_k\} \) is said to be asymptotic regular if \( \|\alpha_{k+1} - \alpha_k\| \to 0 \) as \( k \to \infty. \)
specific form of $\nabla f(x^k)$, and the last inequality holds by the Lipschitz continuity of $\nabla f$. Note that
\[
\|x^{k+1} - x^k\| = \|x^{k+1} - \bar{x}^{k+1} + \bar{x}^{k+1} - \bar{x}^k + \bar{x}^k - x^k\|
\leq \|x^{k+1} - \bar{x}^{k+1}\| + \|\bar{x}^k - x^k\| + \alpha_k \|\nabla f(x^k)\| \lesssim \alpha_k, \tag{59}
\]
where the first inequality holds for the triangle inequality and (56), and the last inequality holds for Proposition 3 and the bounded assumption of $\nabla f$. Thus, (58) and (59) imply
\[
\|\nabla f(x^{k+1})\|^2 - \|\nabla f(x^k)\|^2 \lesssim \alpha_k. \tag{60}
\]
By the specific form of $\alpha_k$, (57), (60) and Lemma 12, it holds
\[
\lim_{k \to \infty} \|\nabla f(x^k)\|^2 = 0. \tag{61}
\]
As a consequence,
\[
\lim_{k \to \infty} 1^T \nabla f(x^k) = 0. \tag{62}
\]
Furthermore, by the coercivity of $f_i$ for each $i$ and the convergence of $\{1^T f(x^k)\}$, $\{x^k\}$ is bounded. Therefore, there exists a convergent subsequence of $\{x^k\}$. Let $x^*$ be any limit point of $\{x^k\}$. By (61) and the continuity of $\nabla f$, it holds
\[
1^T \nabla f(x^*) = 0.
\]
Moreover, by Proposition 3 $x^*$ is consensual. 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 $x^*$.

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
\[
\sum_{k=0}^{K} \alpha_k \|x^{k+1} - \bar{x}^{k+1}\| \leq D_1 + D_2 \sum_{k=0}^{K} \alpha_k^2, \tag{63}
\]
where $D_1 = C \frac{\|x^0\|}{2(1-\zeta)}$, $D_2 = C \left( \frac{\|x^0\|}{2} + \frac{K}{1-\zeta} \right)$, and $B$ is specified in Assumption 3.

Proof. By (29),
\[
\sum_{k=0}^{K} \alpha_k \|x^{k+1} - \bar{x}^{k+1}\| \leq C\|x^0\| \zeta \sum_{k=0}^{K} \alpha_k \zeta^k
\]
\[+ CB \sum_{k=0}^{K} \sum_{j=0}^{k} \zeta^{k-j} \alpha_k \alpha_j. \tag{64}\]

In the following, we estimate these two terms in the right-hand side of (64), respectively. Note that
\[
\sum_{k=0}^{K} \alpha_k \zeta^k \leq \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2 + \frac{1}{2} \sum_{k=0}^{K} \zeta^{2k} \leq \frac{1}{2(1-\zeta)} + \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2, \tag{65}\]
\[
\sum_{k=0}^{K} \sum_{j=0}^{k} \zeta^{k-j} \alpha_k \alpha_j \leq \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2 \sum_{j=0}^{K} \zeta^{k-j} \leq \frac{1}{2} \sum_{k=0}^{K} \alpha_k \alpha_k \sum_{j=0}^{K} \zeta^{k-j} \leq \alpha_k, \tag{66}\]
Plugging (65) and (66) into (64) yields (63).

Besides Lemma 13, we also need the following two lemmas, which have appeared in the literature (cf. [8]).

Lemma 14 (8). Let $\gamma_k = \frac{1}{k^r}$ for some $0 < \epsilon \leq 1$. Then the following hold
(a) if $0 < \epsilon < 1/2$, \[
\sum_{k=0}^{K} \gamma_k^2 \leq \frac{1}{K^{1-\epsilon}} = O\left(\frac{1}{K^{1-\epsilon}}\right), \tag{67}
\]
(b) if $\epsilon = 1/2$, \[
\sum_{k=0}^{K} \gamma_k^2 \leq \frac{1}{2^{1/2}} = O\left(\frac{1}{\sqrt{K}}\right), \tag{68}
\]
(c) if $1/2 < \epsilon < 1$, \[
\sum_{k=1}^{K} \gamma_k^2 \leq \frac{1-\epsilon}{2 - \epsilon} \frac{2\epsilon - 1/K^{2\epsilon-1}}{K^{2\epsilon-1} - 1} = O\left(\frac{1}{K^{1-\epsilon}}\right), \tag{69}
\]
(d) if $\epsilon = 1$, \[
\sum_{k=1}^{K} \gamma_k^2 \leq \frac{1}{K} \frac{1}{K+1} = O\left(\frac{1}{K}\right), \tag{70}
\]
\[
\sum_{k=1}^{K} \gamma_k \leq \frac{1}{K} = O\left(\frac{1}{K}\right), \tag{71}
\]
\[
\sum_{k=1}^{K} \gamma_k^2 \leq \frac{1}{K+1} - \frac{1}{K} = O\left(\frac{1}{K}\right). \tag{72}
\]

Proof. of Proposition 4. To prove this proposition, we first develop the following inequality,
\[
L_{\alpha_k}(x^{k+1}) - L_{\alpha_k}(u) \leq \frac{1}{2\alpha_k} (\|x^k - u\|^2 - \|x^{k+1} - u\|^2) \tag{67}
\]
for any $u \in \mathbb{R}^{n \times p}$. By Lemma 13 we have
\[
L_{\alpha_k}(u) \geq L_{\alpha_k}(x^{k+1}) + \langle \nabla L_{\alpha_k}(x^{k+1}), u - x^{k+1} \rangle - \frac{L^*}{2} \|x^{k+1} - x^k\|^2, \tag{68}
\]
where $L^* = L_f + \alpha_k^{-1}(1 - \lambda_n(W))$, and by (21), we have $\nabla L_{\alpha_k}(x^{k+1}) = \alpha_k^{-1}(x^k - x^{k+1})$. Then (68) implies
\[
L_{\alpha_k}(u) \geq L_{\alpha_k}(x^{k+1}) \tag{69}
\]
\[+ \alpha_k^{-1}(x^k - x^{k+1}, u - x^{k+1}) - \frac{L^*}{2} \|x^{k+1} - x^k\|^2.
\]
Note that the specific form of $\alpha_k = \frac{1}{L(r_{k+1})}$, 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, \eqref{eq:77} implies
\begin{equation}
\mathcal{L}_{\alpha_k}(u) \geq \mathcal{L}_{\alpha_k}(x^{k+1}) + \alpha_k^{-1} \langle x^k - x^{k+1}, u - x^{k+1} \rangle + \frac{1}{2\alpha_k} \|x^{k+1} - x^k\|^2.
\end{equation}

Recall that for any two vectors $a$ and $b$, it holds $2\langle a, b \rangle - \|a\|^2 = \|b\|^2 - \|a - b\|^2$. Therefore,
\begin{equation}
\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).
\end{equation}

As a consequence, we get the basic inequality \eqref{eq:77}.

Note that the optimal solution $x_{\text{opt}}$ is consensual and thus, $\|x_{\text{opt}}\|_W^2 = 0$. Therefore, $\mathcal{L}_{\alpha_k}(x_{\text{opt}}) = f(x_{\text{opt}}) = f_{\text{opt}}$. By \eqref{eq:77}, we have
\begin{equation}
\alpha_k (\mathcal{L}_{\alpha_k}(x^{k+1}) - f_{\text{opt}}) \leq (\|x^k - x_{\text{opt}}\|^2 - \|x^{k+1} - x_{\text{opt}}\|^2)/2.
\end{equation}

Summing the above inequality over $k = 0, 1, \ldots, K$ yields
\begin{equation}
\sum_{k=0}^{K} \alpha_k (\mathcal{L}_{\alpha_k}(x^{k+1}) - f_{\text{opt}}) \leq \|x^0 - x_{\text{opt}}\|^2/2.
\end{equation}

Moreover, noting that $\mathcal{L}_{\alpha_k}(x^{k+1}) = \bar{f} (x^{k+1})$ and by the convexity of $\mathcal{L}_{\alpha_k}$,
\begin{equation}
\mathcal{L}_{\alpha_k}(x^{k+1}) \geq \bar{f} (x^{k+1}) + \langle \nabla \mathcal{L}_{\alpha_k}(x^{k+1}), x^{k+1} - x^{k+1} \rangle
\end{equation}

where the second inequality holds by the bounded assumption of gradient (cf. Assumption 3). Plugging \eqref{eq:72} into \eqref{eq:71} yields
\begin{equation}
\sum_{k=0}^{K} \alpha_k (\bar{f} (x^{k+1}) - f_{\text{opt}}) \leq \frac{1}{2} \|x^0 - x_{\text{opt}}\|^2 + B \sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^{k+1}\|.
\end{equation}

By the definition of $\bar{f} K$ \eqref{eq:17}, then \eqref{eq:73} implies
\begin{equation}
(\bar{f} K - f_{\text{opt}}) \sum_{k=0}^{K} \alpha_k \leq \frac{1}{2} \|x^0 - x_{\text{opt}}\|^2 + B \sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^{k+1}\|
\end{equation}

where $D_3 = \frac{1}{2} \|x^0 - x_{\text{opt}}\|^2 + BD_1$, $D_4 = BD_2$, $D_1$ and $D_2$ are specified in Lemma 13 and the second inequality holds for Lemma 13. As a consequence,
\begin{equation}
\bar{f} K - f_{\text{opt}} \leq \frac{D_3 + D_4 \sum_{k=0}^{K} \alpha_k^2}{\sum_{k=0}^{K} \alpha_k^2}.
\end{equation}

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

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

\begin{equation}
\hat{\mathcal{L}}_\alpha(x^{k+1}) \leq \hat{\mathcal{L}}_\alpha(x^k)
\end{equation}

\begin{equation}
-\frac{1}{2} (\alpha - 1 + \lambda_n(W) - L_f) \|x^{k+1} - x^k\|^2, \forall k \in N.
\end{equation}

**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{equation}
\hat{\mathcal{L}}_\alpha(x^{k+1}) \leq \hat{\mathcal{L}}_\alpha(x^k)
\end{equation}

\begin{equation}
-\frac{1}{2} (\alpha - \lambda_n(W) - L_f) \|x^{k+1} - x^k\|^2, \forall k \in N.
\end{equation}

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

\begin{equation}
= \mathcal{L}_\alpha(x^{k+1}) - \mathcal{L}_\alpha(x^k) + r(x^{k+1}) - r(x^k)
\end{equation}

\begin{equation}
\leq \langle \nabla \mathcal{L}_\alpha(x^k), x^{k+1} - x^k \rangle + \frac{L_f}{2} \|x^{k+1} - x^k\|^2 + r(x^{k+1}) - r(x^k).
\end{equation}

**C1:** From the convexity of $r$, \eqref{eq:77}, and \eqref{eq:19}, it follows that
\begin{equation}
0 = \xi^{k+1} + \frac{1}{\alpha} (x^{k+1} - x^k + \alpha \nabla \mathcal{L}_\alpha(x^k)), \xi^{k+1} \in \partial r(x^{k+1}).
\end{equation}

This and the convexity of $r$ further give us
\begin{equation}
r(x^{k+1}) - r(x^k) \leq \langle \xi^{k+1} - x^{k+1}, x^{k+1} - x^k \rangle
\end{equation}

\begin{equation}
= -\frac{1}{\alpha} \|x^{k+1} - x^k\|^2 - \langle \nabla \mathcal{L}_\alpha(x^k), x^{k+1} - x^k \rangle.
\end{equation}

Substituting this inequality into the inequality \eqref{eq:79} and then expanding $L^* = L_f + \alpha^{-1} (1 - \lambda_n(W))$ yield
\begin{equation}
\hat{\mathcal{L}}_\alpha(x^{k+1}) - \hat{\mathcal{L}}_\alpha(x^k) \leq -\frac{1}{\alpha} \left(-\frac{L_f}{2}\right) \|x^{k+1} - x^k\|^2
\end{equation}

\begin{equation}
= -\frac{1}{2} \left(\alpha - 1 + \lambda_n(W) - L_f\right) \|x^{k+1} - x^k\|^2.
\end{equation}

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

**C2:** From \eqref{eq:77} and \eqref{eq:19}, it follows that the function $r(u) + \frac{\|u - (x^k - \alpha \nabla \mathcal{L}_\alpha(x^k))\|_W^2}{2\alpha}$ reaches its minimum at $u = x^{k+1}$. Comparing the values of this function at $x^{k+1}$ and $x^k$ yields
\begin{equation}
r(x^{k+1}) - r(x^k) \leq \frac{1}{2\alpha} \|x^k - (x^k - \alpha \nabla \mathcal{L}_\alpha(x^k))\|_W^2
\end{equation}

\begin{equation}
= -\frac{1}{2\alpha} \|x^k - \alpha \nabla \mathcal{L}_\alpha(x^k)\|_W^2
\end{equation}

\begin{equation}
= -\frac{1}{2\alpha} \|x^{k+1} - x^k\|^2 - \langle \nabla \mathcal{L}_\alpha(x^k), x^{k+1} - x^k \rangle.
\end{equation}

Substituting this inequality into \eqref{eq:79} and expanding $L^*$ yield
\begin{equation}
\hat{\mathcal{L}}_\alpha(x^{k+1}) - \hat{\mathcal{L}}_\alpha(x^k) \leq -\frac{1}{2\alpha} \left(-\frac{L_f}{2}\right) \|x^{k+1} - x^k\|^2
\end{equation}

\begin{equation}
= -\frac{1}{2} (\alpha - \lambda_n(W) - L_f) \|x^{k+1} - x^k\|^2.
\end{equation}

Hence, sufficient descent requires $0 < \alpha < \frac{\lambda_n(W)}{L_f}$. \qed
Lemma 17 (Boundedness). Under the conditions of Lemma 16 the sequence \( \{ \hat{L}_\alpha(x^k) \} \) is lower bounded, and the sequence \( \{ x^k \} \) is bounded.

Proof. The lower boundedness of \( \{ \hat{L}_\alpha(x^k) \} \) is due to Assumption 3 Part (2).

By Lemma 16 and under a proper step size, \( \hat{L}_\alpha(x^k) \) is nonincreasing and upper bounded by \( \hat{L}_\alpha(x^0) \). Hence, \( \sum_{i=1}^n r_i(x_{i-1}^k) + r_i(x_{i-1}^k) \) is upper bounded by \( \hat{L}_\alpha(x^0) \). 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 \( \partial \hat{L}_\alpha(x^{k+1}) \) denote the (limiting) subdifferential of \( \hat{L}_\alpha \), which is assumed to exist for all \( k \in \mathbb{N} \). Then, there exists \( g^{k+1} \in \partial \hat{L}_\alpha(x^{k+1}) \) such that
\[
\| g^{k+1} \| \leq (\alpha^{-1} (2 - \lambda_n(W)) + L_f) \| x^{k+1} - x^k \|.
\]

Proof. By the iterate (19), the following optimality condition holds
\[
0 \in \alpha^{-1} (x^{k+1} - x^k + \alpha \nabla \hat{L}_\alpha(x^k)) + \partial r(x^{k+1}), \tag{80}
\]
where \( \partial r(x^{k+1}) \) denotes the (limiting) subdifferential of \( r \) at \( x^{k+1} \). For any \( \xi^{k+1} \in \partial r(x^{k+1}) \), it follows from (80) that
\[
\nabla \hat{L}_\alpha(x^{k+1}) + \xi^{k+1} = \alpha^{-1} (x^{k+1} - x^k) + \nabla \hat{L}_\alpha(x^k) - \nabla \hat{L}_\alpha(x^{k+1}),
\]
which immediately yields
\[
\| \nabla \hat{L}_\alpha(x^{k+1}) + \xi^{k+1} \| \\
\leq \alpha^{-1} \| x^{k+1} - x^k \| + \| \nabla \hat{L}_\alpha(x^k) - \nabla \hat{L}_\alpha(x^{k+1}) \| \\
\leq (\alpha^{-1} + L_f) \| x^{k+1} - x^k \| \\
\leq (\alpha^{-1} (2 - \lambda_n(W)) + L_f) \| x^{k+1} - x^k \|.
\]
Thus, then the claim of Lemma 18 holds.

Based on Lemmas 16 and 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 11 and thus is omitted.

Proof of Proposition 5. The proof is similar to that of Proposition 2. We shall however note that in (26), \( a = \frac{1}{2} (\alpha^{-1} (1 + \lambda_n(W)) - L_f) \) 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 \( \lambda_n(W) > 0 \).

G. Proofs for Theorem 4 and Proposition 6

Based on the iterate (6) of Prox-DGD, we derive the following recursion of the iterates of Prox-DGD, which is similar to (16).

Lemma 19 (Recursion of \( \{ x^k \} \)). For any \( k \in \mathbb{N} \),
\[
x^k = W x^0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j} \nabla f(x^j + \xi^{j+1}), \tag{81}
\]
where \( \xi^{j+1} \in \partial r(x^{j+1}) \) is the one determined by the proximal operator (7), for any \( j = 0, \ldots, k - 1 \).

Proof. By the definition of the proximal operator (7), the iterate (6) implies
\[
x^{k+1} + \alpha_k \xi^{k+1} = W x^k - \alpha_k \nabla f(x^k), \tag{82}
\]
where \( \xi^{k+1} \in \partial r(x^{k+1}) \), and thus
\[
x^{k+1} = W x^k - \alpha_k (\nabla f(x^k) + \xi^{k+1}). \tag{83}
\]
By (83), we can easily derive the recursion (81).

Proof of Proposition 6. The proof of this proposition is similar to that of Proposition 5. It only needs to note that the subgradient term \( \nabla f(x^j) + \xi^{j+1} \) 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 (15). Results are given in two cases below:

C1: \( r_i \)'s are convex. For any \( k \in \mathbb{N} \),
\[
\hat{L}_{\alpha^{k+1}}(x^{k+1}) \leq \hat{L}_{\alpha_k}(x^k) + \frac{1}{2} (\alpha^{-1} (1 + \lambda_n(W)) - L_f) \| x^{k+1} - x^k \|^2. \tag{84}
\]

C2: \( r_i \)'s are not necessarily convex. For any \( k \in \mathbb{N} \),
\[
\hat{L}_{\alpha^{k+1}}(x^{k+1}) \leq \hat{L}_{\alpha_k}(x^k) + \frac{1}{2} (\alpha^{-1} \lambda_n(W) - L_f) \| x^{k+1} - x^k \|^2. \tag{85}
\]

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
\[
\hat{L}_{\alpha^{k+1}}(x^{k+1}) - \hat{L}_{\alpha_k}(x^k) \leq \frac{1}{2} (\alpha^{-1} \| x^k - u \|^2 - \| x^{k+1} - u \|^2). \tag{86}
\]

Proof. By Lemma 15 we have
\[
L_{\alpha_k}(u) \geq L_{\alpha_k}(x^{k+1}) + (\nabla L_{\alpha_k}(x^{k+1}), u - x^{k+1}) - \frac{L_f}{2} \| x^{k+1} - x^k \|^2, \tag{87}
\]
where \( L_f = L_f + \alpha^{-1} (1 - \lambda_n(W)) \), and by the convexity of \( r \), we have
\[
r(u) \geq r(x^{k+1}) + \langle \xi^{k+1}, u - x^{k+1} \rangle, \tag{88}
\]
where \( \xi^{k+1} \in \partial r(x^{k+1}) \) is the one determined by the proximal operator (7). By (83), it follows
\[
\xi^{k+1} = \alpha^{-1} (x^{k+1} - x^k) - \nabla L_{\alpha_k}(x^{k+1}). \tag{89}
\]
Plugging (88) into (87), and then summing up (86) and (87) yield
\[
\tilde{L}_{\omega_k}(u) \geq \tilde{L}_{\omega_k}(x^{k+1}) + \alpha_k^{-1}(x^k - x^{k+1}, u - x^{k+1}) - \frac{L^*}{2} \|x^{k+1} - x^k\|^2.
\] (89)

Similar to the rest proof of the inequality (67), we can prove this lemma based on (89). □

**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 while the proof of Theorem 4(d) is very similar to that of Proposition 4 and thus the proof of this theorem is omitted. □

**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 motivates 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 N000141712162.

**REFERENCES**


