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, on consensus optimization with nonconvex objective functions, our understanding to the behavior of these algorithms is limited. When we lose convexity, we cannot hope for obtaining globally optimal solutions (though we still do sometimes). Somewhat surprisingly, we retain most other properties from the convex setting for the decentralized consensus algorithms DGD and Prox-DGD, such as convergence to a (neighborhood of) consensus stationary solution and the rates of convergence when diminishing (constant) step sizes are used. It is worth noting that the Prox-DGD algorithm can handle nonconvex nonsmooth functions assume that their proximal operators can be computed.

To establish some of these properties, the existing proofs from the convex setting need changes, and some results require a completely different line of analysis.

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:

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

(1)

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

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

(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. When $f_i$'s are convex, the existing algorithms include the (sub)gradient methods [5], [6], [10], [13], [15], [23], [27], and the primal-dual domain methods such as the decentralized alternating direction method of multipliers (DADMM) [19], [20] and EXTRA [21], [22]. However, when $f_i$'s are nonconvex, very few algorithms have convergence guarantees. Some existing results include [3], [24], [29]. In spite of the algorithms and their analysis in these works, the convergence of the simple algorithm Decentralized Gradient Descent (DGD) [15] 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 [16], [26], [14], [19], where online processing and delay tolerance are considered. Therefore, we expect our results to motivate future adaptations of nonconvex DGD.

This paper studies the convergence of DGD for solving problem (1) and a related algorithm, Prox-DGD, for problem (2). At each iteration of DGD, each agent locally computes a gradient and then updates its variable by combining the average of its neighbors’ with the negative gradient step. While at each iteration of Prox-DGD, each agent locally computes a gradient of the smooth part of its objective and a proximal map of the nonsmooth part, as well as exchanges information with its neighbors. They can use both a fixed step size and a sequence of decreasing step sizes. In the convex setting, when using a fixed step size, DGD does not converge to a solution of the original problem (1) but a point in its neighborhood regardless of whether the $f_i$'s are differentiable or not [27]. This motivates the use of certain decreasing step sizes such as in [6], [10]. For the general convex case and under the Lipschitz continuous and bounded gradient assumption, [6] shows that decreasing step sizes $\alpha_k = \frac{1}{\sqrt{k}}$ lead to a convergence rate of $O\left(\frac{\ln k}{k}\right)$ in terms of the running best of objective error. For the general convex case, under assumptions of fixed step size and Lipschitz continuous, bounded gradient, [10] shows an outer-loop convergence rate of $O\left(\frac{\ln k}{k}\right)$ in terms of objective error, utilizing Nesterov’s acceleration, provided that the inner loop performs substantial consensus computation. Without a substantial inner loop, the decreasing step sizes $\alpha_k = \frac{1}{k}$ lead to a reduced rate of $O\left(\frac{\ln k}{k}\right)$.

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 hold in the convex setting remain valid in the nonconvex setting, and (b) to achieve the goal in part (a), we illustrate how to tailor existing nonconvex analysis tools to decentralized optimization. In particular, the property of asymptotic consensus, or consensus up to a bounded error, requires special 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 is used and properly bounded, the DGD iterates converge to a stationary point of a certain Lyapunov function. The difference between each local point and the global average of all points is bounded, and the bound is proportional to the step size.

(b) When a decreasing step size $\alpha_k = O\left(1/(k+1)^{\epsilon}\right)$ is
used, where $0 < \epsilon \leq 1$ and $k$ is the iteration number, the objective sequence converges, and the iterates of DGD are asymptotically consensus (i.e., become equal one another), and they achieve this at the rate of $O(1/(k+1)^\gamma)$. Moreover, for any sequence generated by DGD, there exists a convergent subsequence and its limit point is a stationary point of the original problem. For a convex objective, we can derive the convergence rates of DGD with different $\epsilon$.

(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, a smaller step size is required.

To the best of our knowledge, this paper presents the first analysis of the DGD and prox-DGD algorithms in the non-convex setting. The detailed comparisons between our results with the existing related works are presented in Tables I and II. The global objective error rate in these two tables is 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.

To develop these theoretical results, new proof techniques are introduced in this paper, particularly, in the analysis of the convergence of DGD and Prox-DGD with the 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 leveraging a simple 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 [6], [15].

The rest of this paper is organized as follows. Section 2 specifies the problem setup and reviews DGD. Section 3 presents our assumptions and main results. Section 4 presents the proofs of the main results. Section 5 extends the results to Prox-DGD. We conclude this paper in Section 6.

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 any 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.

II. PROBLEM SETUP AND DGD REVIEW

Consider an undirected network $G = \{V, E\}$, where $V$ is a set of $n$ nodes and $E$ is the edge set. Any edge $(i,j) \in E$ 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 & \quad x^T f(x) \triangleq \sum_{i=1}^{n} f_i(x(i)), \\
\text{subject to} \quad & \quad x(i) = x(j), \quad \forall (i,j) \in E,
\end{align*}

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

\[
\begin{pmatrix}
- x_{11}^T \\
- x_{21}^T \\
\vdots \\
- x_{n1}^T
\end{pmatrix}
\quad \text{and}
\quad
\begin{pmatrix}
f_1(x(1)) \\
f_2(x(2)) \\
\vdots \\
f_n(x(n))
\end{pmatrix}.
\]

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

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

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 [13] 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 stepsize parameter.

III. ASSUMPTIONS AND MAIN RESULTS

This section presents all of our main results.

A. Definitions and assumptions

Definition 1 (Lipschitz differentiability). A function $h$ is called Lipschitz differentiable if $h$ is differentiable and its gradient $\nabla h$ 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 $\|h\| \to +\infty$ implies $h(u) \to +\infty$.

The next definition is a property that is commonly used to improve subsequence convergence to whole sequence convergence, which is often referred to global convergence (not necessarily to the global solution).

Definition 3 (Kurdyka-Łojasiewicz (KL) property [12, 4]). A function $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ 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 $\varphi : [0, \eta) \to \mathbb{R}_+$ such that:

(i) $\varphi(0) = 0$ and $\varphi$ is $C^1$ 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

\[
\varphi'(h(x) - h(x^*)) \cdot \text{dist}(0, \partial h(x)) \geq 1.
\]
TABLE I
COMPARISONS ON DIFFERENT ALGORITHMS FOR CONSENSUS SMOOTH OPTIMIZATION PROBLEM (1).

| Algorithm          | Fixed step size | Decreasing step sizes |
|--------------------|----------------|-----------------------|
|                    | DGD [27]       | DGD (this paper)      |
|                    | D-NG [10]      | DGD (this paper)      |
| $f_i$              | convex only    | (non)convex           |
| $\nabla f_i$       | Lipschitz      | Lipschitz, bounded    |
| step size          | $0 < \alpha < \frac{1 + \lambda_n(W)}{L_f}$ | $\mathcal{O}(\frac{1}{\sqrt{k}})$ with Nesterov acc. |
| consensus          | error $\mathcal{O}(\alpha)$ | $\mathcal{O}(\frac{1}{k})$ |
| error $\mathcal{O}(\alpha)$ | $\mathcal{O}(\frac{1}{k})$ |
| $\min_{j \leq k} \|x^{j+1} - x^j\|^2$ | $O(\frac{1}{k})$ | no rate | $O(\frac{1}{k^{1+\gamma}})$ |
| global objective error | $\mathcal{O}(\frac{1}{k})$ until error $\mathcal{O}(\frac{1}{(1-\frac{1}{k})^{1/2}})$ | Convex: $\mathcal{O}(\frac{1}{k})$ until error $\mathcal{O}(\frac{1}{(1-\frac{1}{k})^{1/2}})$; Nonconvex: no rate | $\mathcal{O}(\ln k \frac{1}{k^{1 \gamma}})$† |

†The objective rates of DGD and Prox-DGD obtained in this paper and those in convex DProx-Grad [6] are ergodic or running best rates.

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

| Algorithm          | Fixed step size | Decreasing step sizes |
|--------------------|----------------|-----------------------|
|                    | AccDProx-Grad [3] | DProx-Grad [6]        |
|                    | DProx-Grad (this paper) | DProx-Grad [9]        |
|                    | DProx-Grad (this paper) | DProx-Grad (this paper) |
| $f_i, r_i$         | convex only    | (non)convex           |
| $\nabla f_i$       | Lipschitz      | Lipschitz             |
| $\partial r_i$    | bounded        | bounded               |
| step size          | $0 < \alpha < \frac{1}{L_f}$ | $0 < \alpha < \frac{1 + \lambda_n(W)}{L_f}$ (convex $r_i$); $0 < \alpha < \frac{\lambda_n(W)}{L_f}$ (nonconvex $r_i$) |
| consensus          | $\mathcal{O}(\gamma^k k^2), 0 < \gamma < 1$ | error $\mathcal{O}(\alpha)$ |
| error $\mathcal{O}(\alpha)$ | $\mathcal{O}(\frac{1}{k^{1/2}})$ |
| $\min_{j \leq k} \|x^{j+1} - x^j\|^2$ | no rate | no rate | $O(\frac{1}{k})$ |
| global objective error | $\mathcal{O}(\frac{1}{k})$ | $\mathcal{O}(\frac{D_1 + D_2 \alpha}{D_1, D_2 > 0})$ | Convex: $\frac{D_1}{\alpha} + D_2 \alpha$, $D_1, D_2 > 0$; Nonconvex: no rate | $\mathcal{O}(\ln k \frac{1}{k^{1 \gamma}})$† |

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 $L_f > 0$.
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 $(i, j) \notin E$, then $w_{ij} = 0$. 

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.
(2) (Symmetry) $W = W^T$.
(3) (Null space property) $\text{null}\{I - W\} = \text{span}\{1\}$.
(4) (Spectral property) $I \succeq W > 0$.

Assumption 4 can be always met by increasing the diagonals of $W$. By Assumption 3 a solution $x_{\text{opt}}$ to problem (3) satisfies $(I - W)x_{\text{opt}} = 0$. Let $\lambda_i(W)$ denote the $i$th largest eigenvalue of $W$. Then by Assumption 2 it holds that $\lambda_1(W) = 1 > \lambda_2(W) \geq \cdots \geq \lambda_n(W) > 0$.

B. 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:

$$L_\alpha(x) \triangleq 1^T f(x) + \frac{1}{2\alpha} \|x\|^2_{I - W}.$$ (6)

It is worth reminding that convexity is not assumed.

Theorem 1 (Global convergence). Let $\{x^k\}$ be the sequence generated by DGD with the step size $0 < \alpha < \frac{1 + \lambda_n(W)}{L_f}$. Let Assumptions 7 and 2 hold. Then $\{x^k\}$ has at least one accumulation point $x^\ast$, and any such point is a stationary point of $L_\alpha(x)$. Furthermore, the running best rate of the sequence $\{\|x^{k+1} - x^k\|\}$ and $\{\|L_\alpha(x^k)\|^2\}$ are $o(1/k)$.

In addition, if $L_\alpha$ satisfies the KL property at an accumulation point $x^\ast$, then $\{x^k\}$ globally converges to $x^\ast$.

Remark 1. Let $x^\ast$ be a stationary point of $L_\alpha(x)$, and thus

$$0 = \nabla f(x^\ast) + \alpha^{-1}(I - W)x^\ast.$$ (7)

Since $1^T (I - W) = 0$, (7) yields $0 = 1^T \nabla f(x^\ast)$, indicating that $x^\ast$ is also a stationary point to the separable function $\sum_{i=1}^n f_i(x^\ast(i))$. Since the rows of $x^\ast$ are not necessarily identical, we cannot say $x^\ast$ is a stationary point to Problem (3). However, the differences between the rows of $x^\ast$ are bounded, following our next result below, adapted from [27]:

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

$$\|x^k(i) - \bar{x}^k\| \leq \frac{aD}{1 - \lambda_2(W)},$$ (8)

where $D$ is a universal bound of $\|\nabla f(x^\ast)\|$ defined in Lemma 6. As $k \to \infty$, (8) yields the consensual bound

$$\|x^\ast(i) - \bar{x}^\ast\| \leq \frac{aD}{1 - \lambda_2(W)},$$

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

In Proposition 1 the consensual bound is proportional to the step size $\alpha$ and is inversely proportional to the gap between the largest and the second largest eigenvalues of $W$.

Let us compare the DGD iteration with the iteration of centralized gradient descent for $f(x)$. Averaging the rows of (4) yields the following comparison:

DGD averaged: $\bar{x}^{k+1} \leftarrow \bar{x}^k - \alpha \left( \frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k(i)) \right)$. (9)

Centralized: $x^{k+1} \leftarrow x^k - \alpha \left( \frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k) \right)$. (10)

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

$$\left\| \frac{1}{n} \sum_{i=1}^n \nabla f_i(x^k(i)) - \frac{1}{n} \sum_{i=1}^n \nabla f_i(\bar{x}) \right\|$$

$$\leq \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^k(i)) - \nabla f_i(\bar{x})\| \leq \frac{\alpha DL_f}{1 - \lambda_2(W)}.$$ (9)

Unlike the convex analysis in [27], it is impossible to bound the difference between the sequences of (9) and (10) 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 $\|x\|_{I-W}$ 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 [25, 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 $x^\ast$ 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 $x^\ast$ in finitely many iterations.
(b) If $\theta \in (0, \frac{1}{2}]$, $\|x^k - x^\ast\| \leq C_0k^\tau$ for all $k \geq k^\ast$ for some $C_0 > 0$, $\tau \in (0, 1)$.
(c) If $\theta \in (\frac{1}{2}, 1)$, $\|x^k - x^\ast\| \leq C_0k^{(1-\theta)/(2\theta-1)}$ for all $k \geq k^\ast$, for certain $C_0 > 0$.

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 $x^\ast$ of $L_\alpha$ is not a stationary point of the original problem. We only have a consensual bound on $x^\ast$. To address this issue, the next subsection uses decreasing step sizes and presents better convergence results.

C. 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 $\alpha_i = O(\frac{1}{\sqrt{i+1}})$, for some $0 < \epsilon \leq 1$, to diminish the consensual bound to $0$. As a result, any accumulation point $x^\ast$ 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., $\|\nabla f(x^k)\| \leq B$. 

---

1 A nonnegative sequence $a_k$ induces its running best sequence $b_k = \min\{a_i : i \leq k\}$; therefore, $a_k$ has a running best rate of $o(1/k)$ if $b_k = o(1/k)$.

1 These quantities naturally appear in the analysis, so we keep the squares.
Note that the bounded gradient assumption is a regular assumption in the convergence analysis of decentralized gradient methods even in the convex setting [10] and also [6], though it is not required for centralized gradient descent.

We take the step size sequence:
\[
\alpha_k = \frac{1}{L_f(k+1)^\epsilon}, \quad 0 < \epsilon \leq 1,
\]
throughout the rest part of this section though we can also use another constant rather than 1. By iteratively applying iteration [4], we obtain the following expression
\[
x^k = W^k x^0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j} \nabla f(x^j).
\]
(12)

**Proposition 3** (Asymptotic consensus rate). Let Assumptions [2] and [3] hold. Let DGD use [11]. Let \( \hat{x}^k \triangleq \frac{1}{n} \sum_{i=1}^{n} x^k \). Then, \( \|x^k - \hat{x}^k\| \) 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 [11]). Moreover, with a larger \( \epsilon \), faster decaying step sizes generally imply a faster asymptotic consensus rate. Note that \( (I - W^T)\hat{x}^k = 0 \) and thus \( \|x^k - \hat{x}^k\|^2_{I - W} = \|x^k - \hat{x}^k\|^2_{\nabla^T} \). Therefore, the above proposition implies the following result.

**Corollary 1.** Apply the setting of Proposition [3], \( \|x^k\|^2_{\nabla^T} \) converges to 0 at the rate of \( O(1/(k+1)^{2\epsilon}) \).

Corollary [11] shows that the sequence \( \{x^k\} \) in \( "(I - W)" \)-norm” can decay to 0 at a sublinear rate. For any global consensual solution \( x_{opt} \) to problem [3], we have \( \|x^k - x_{opt}\|^2_{\nabla^T} = \|x^k - x_{opt}\|^2_{\nabla^T} \) so, if \( \{x^k\} \) does converge to \( x_{opt} \), then their distance in \( "(I - W)" \)-norm” decays at \( O(1/k^{2\epsilon}) \).

**Theorem 2** (Convergence). Let Assumptions [1][2] and [3] hold. Let DGD use step sizes [11]. Then:
(a) \( \{L_{opt}(x^k)\} \) 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 \( o(1/k^{1+\epsilon}) \). 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 impossible 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 interesting relation with \( \epsilon \), also the proof techniques appears to be new and potentially useful in other settings.

For any \( x \in \mathbb{R}^{n \times p} \), let \( \bar{f}(x) \triangleq \sum_{i=1}^{n} f_i(x(i)) \). Even if \( f_i \)'s are convex, the solution to [3] may be non-unique. Thus, let \( \mathcal{X} \) be the set of solutions to [3]. Given \( x^k \), we pick the solution \( x_{opt} = \text{Proj}_{\mathcal{X}}(x^k) \in \mathcal{X} \). Also let \( f_{opt} = \bar{f}(x_{opt}) \) be the optimal value of [11]. Define the ergodic objective:
\[
\bar{f}^K = \frac{\sum_{k=0}^{K} \alpha_k \bar{f}(x^{k+1})}{\sum_{k=0}^{K} \alpha_k},
\]
where \( x^{k+1} = \frac{1}{n}(1^T x^k) \). Obviously,
\[
\bar{f}^K \geq \min_{k=1,...,K+1} \bar{f}(x^k).
\]
(14)

**Theorem 3** (Convergence rates under convexity). Let Assumptions [7][2] and [3] hold. Let DGD use step sizes [11]. If each \( f_i \) is convex, then \( \{f^K\} \) defined in (13) converges to the optimal objective value \( f_{opt} \) at the following rates:
(a) if \( 0 < \epsilon < 1/2 \), the rate is \( O(\frac{1}{k^{1+\epsilon}}) \);
(b) if \( \epsilon = 1/2 \), the rate is \( O(\frac{\ln K}{K}) \);
(c) if \( 1/2 < \epsilon < 1 \), the rate is \( O(\frac{1}{K^{1-\epsilon}}) \);
(d) if \( \epsilon = 1 \), the rate is \( O(\frac{1}{K}) \).

The convergence rates established in Theorem [3] almost as good as \( O(\frac{1}{K}) \) 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], larger \( \epsilon \) implies a faster consensus rate. Therefore, there is a tradeoff to choose an appropriate \( \epsilon \) in the practical implementation of DGD.

The results for Prox-DGD are given in Section [5] after we first present the proofs for DGD in the next section.

IV. PROOFS

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 \( L_{opt} \), following the argument in [27]; then, the properties of sufficient descent, lower boundedness, and bounded gradients are established for the sequence \( \{L_{opt}(x^k)\} \), giving subsequent convergence of the DGD iterates; finally, whole sequence convergence of the DGD iterates follows from the KL property of \( L_{opt} \).

Lemma 1 (Gradient descent interpretation). The sequence \( \{x^k\} \) generated by the DGD iteration [3] is the same sequence generated by applying gradient descent with the fixed step size \( \alpha \) to the objective function \( L_{opt}(x) \).

A proof of this lemma is given in [27], and it is based on reformulating [3] as the iteration:
\[
x^{k+1} = x^k - \alpha (\nabla f(x^k) + 1/(I - W) x^k) = x^k - \alpha \nabla L_{opt}(x^k).
\]
(15)

Although the sequence \( \{x^k\} \) generated by the DGD iteration [4] can be interpreted as a centralized gradient descent sequence of function \( L_{opt}(x) \), it is different with the gradient descent of the original problem [3].

Lemma 2 (Sufficient descent of \( \{L_{opt}(x^k)\} \)). Let Assumptions [1] and [2] hold. Set the step size \( 0 < \alpha < \frac{1 + \lambda_n(W)}{L_f} \). It holds
that
\[ L_\alpha(x^{k+1}) \leq L_\alpha(x^k) \] (16)
\[- \frac{1}{2}(\alpha^{-1}(1 + \lambda_n(W)) - L_f)\|x^{k+1} - x^k\|^2, \quad \forall k \in \mathbb{N}. \]

Proof: From \(x^{k+1} = x^k - \alpha \nabla L_\alpha(x^k)\), it follows that
\[
\langle \nabla L_\alpha(x^k), x^{k+1} - x^k \rangle = -\frac{\|x^{k+1} - x^k\|^2}{\alpha}. \tag{17}
\]
Since \(\sum_{i=1}^{n} \nabla f_i(x^{(i)}) = L_f \cdot \text{Lipschitz}, \) \(\nabla L_\alpha\) is Lipschitz with the constant \(L^* = \frac{L_f + \alpha^{-1}\lambda_{\text{max}}(I - W)}{L_f + \alpha^{-1}(1 - \lambda_n(W))}\), implying
\[
L_\alpha(x^{k+1}) \leq L_\alpha(x^k) + \langle \nabla L_\alpha(x^k), x^{k+1} - x^k \rangle + \frac{L^*}{2}\|x^{k+1} - x^k\|^2. \tag{18}
\]
Combining (17) and (18) yields (16).

Lemma 3 (Boundedness). Under Assumptions 2 and 1, if \(0 < \alpha < \frac{1}{L_f \hspace{1mm} (18)}\), 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, \(1^T f(x^k) \leq L_\alpha(x^k)\) 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 (\(\ell_2\)-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 (15), 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 \(\{x^k_s\}_{s \in \mathbb{N}} \to x^*\) as \(s \to +\infty\). By Lemmas 2 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, \(\|\nabla L_\alpha(x^k)\| \to 0\) as \(k \to \infty\). In particular, \(\|\nabla L_\alpha(x^*)\| \to 0\) 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 Lemma 1.2 or Theorem 3.3.1. By Lemma 5, the running best rate of the sequence \(\|\nabla L_\alpha(x^k)\|^2\) is \(o(1/\beta)\).

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

Finally, we derive a bound of 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
\[
\|\nabla L(x^k)\| \leq D \triangleq L_f(B + \|y^*\|), \quad \forall k \in \mathbb{N}, \tag{19}
\]
where \(B\) is the bound of \(\|x^k\|\) given in Lemma 3.

Proof: By the lower boundedness assumption (Assumption 1 Part (2)), the minimizer of \(1^TF(y)\) 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
\[
\|\nabla f(x^k)\| = \|\nabla f(x^k) - \nabla f(y^*)\| \leq L_f\|x^k - y^*\| \tag{Lemma 3} \leq L_f(B + \|y^*\|).
\]

Therefore, we have proven this lemma.

B. Proof for Proposition 2

Proof: Note that
\[
\|\nabla L_\alpha(x^{k+1})\| \leq \|\nabla L_\alpha(x^{k+1}) - \nabla L_\alpha(x^k)\| + \|\nabla L_\alpha(x^k)\|
\leq L^*\|x^{k+1} - x^k\| + \alpha^{-1}\|x^{k+1} - x^k\|
= (\alpha^{-1}(2 - \lambda_n(W)) + L_f)\|x^{k+1} - x^k\|,
\]
where the second inequality holds for Lemma 5 and the Lipschitz continuity of \(\nabla L_\alpha\) with constant \(L^* = L_f + \alpha^{-1}(1 - \lambda_n(W))\). Thus, it shows that \(\{x^k\}\) satisfies the so-called relative error condition as in 2. Moreover, by Lemmas 2 and 5, \(\{x^k\}\) 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 \(x^*\) with \(\psi(s) = cs^{\theta-1}\), following the proof of 2, there exists \(k_0 > 0\) such that for all \(k \geq k_0\), we have
\[
2\|x^{k+1} - x^k\| \leq \|x^k - x^{k-1}\| + \frac{cb}{a} \times ((L_\alpha(x^k) - L_\alpha(x^*))^{1-\theta} - (L_\alpha(x^{k+1}) - L_\alpha(x^*))^{1-\theta}),
\]
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
\[
\sum_{i=k_0}^{k} \|x^{i+1} - x^i\| \leq \|x^{k_0} - x^{k_0-1}\| + \frac{cb}{a} \times ((L_\alpha(x^{k_0}) - L_\alpha(x^*))^{1-\theta} - (L_\alpha(x^{k+1}) - L_\alpha(x^*))^{1-\theta}).
\]

Following a derivation similar to the proof of 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.

Moreover, we list some lemmas which will be frequently used in the proofs of this section.

Lemma 7. (11, Proposition 1) Let \(W^k \triangleq W \cdots W = W^k\) be the power of \(W\) with degree \(k\) for any \(k \in \mathbb{N}\). Under Assumption 2, it holds
\[
\|W^k - \frac{1}{n} 11^T\| \leq C(\lambda_2(W))^k \tag{21}
\]
for some constant $C > 0$.

**Lemma 8.** ([17] Lemma 3.1) Let $\{c_k\}$ be a scalar sequence. If $\lim_{k \to \infty} c_k = \gamma$ and $0 < \beta < 1$, then $\lim_{k \to \infty} \sum_{i=0}^{k} \beta^{k-i} \gamma_i = \gamma / (1 - \beta).$

**Proof of Proposition 3:** By the recursion (12), note that

$$x^k - \tilde{x}^k = (W^k - \frac{1}{n} 11^T)x^0 - \sum_{j=0}^{k-1} \alpha_j (W^{k-1-j} - \frac{1}{n} 11^T) \nabla f(x^j).$$

Further by Lemma 7 and Assumption 3, we obtain

$$\|x^k - \tilde{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\| \gamma_2(W) + B \sum_{j=0}^{k-1} \alpha_j \gamma_2^{k-1-j}(W) \right).$$

(23)

Furthermore, by Lemma 8 and step sizes (11), we get $\lim_{k \to \infty} \|x^k - \tilde{x}^k\| = 0.$

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

$$\lim_{k \to \infty} b_k^{-1} \|x^k - \tilde{x}^k\| \leq C^*$$

for some $0 < C^* < \infty$. Let $j'_k \triangleq [k - 1 + 2 \log \gamma_2(W)(b_k^{-1})]$ (where $[x]$ denotes the integer part of $x$ for any $x \in \mathbb{R}$). Note that

$$b_k^{-1} \|x^k - \tilde{x}^k\| \leq Cb_k^{-1} \left( \|x^0\| \gamma_2(W) + B \sum_{j=0}^{k-1} \alpha_j \gamma_2^{k-1-j}(W) \right)$$

$$= C \|x^0\|b_k^{-1} \gamma_2(W)$$

$$+ CBb_k^{-1} \sum_{j=0}^{j'_k} \alpha_j \gamma_2^{k-1-j}(W)$$

$$+ CBb_k^{-1} \sum_{j=j'_k+1}^{k-1} \alpha_j \gamma_2^{k-1-j}(W)$$

$$\triangleq T_1 + T_2 + T_3,$$

(24)

where the first inequality holds because of (23).

In the following, we will estimate the above three terms in the right-hand side of (24), respectively. First, by the definition of $j'_k$, for any $j \leq j'_k$, we have

$$b_k^{-1} \gamma_2^{k-1-j}(W) \leq b_k^{-1} \gamma_2^{j'_k}(W) \leq 1.$$

Thus,

$$T_2 \leq CB \sum_{j=0}^{j'_k} \alpha_j \gamma_2^{(k-1-j)/2}(W),$$

(25)

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

$$b_k^{-1} \alpha_j \leq \frac{(k+1)^\epsilon}{L_f(j'_k+1)\epsilon} \leq \frac{(k+1)^\epsilon}{L_f(k-1+2\epsilon \log \gamma_2(W)(k+1)^\epsilon)},$$

and also

$$b_k^{-1} \alpha_j \geq \frac{(k+1)^\epsilon}{L_f(k+1)\epsilon} = \frac{1}{L_f}.$$

Thus, for any $j'_k < j \leq k - 1$

$$\lim_{k \to \infty} b_k^{-1} \alpha_j = \frac{1}{L_f}.$$

(26)

Furthermore, note that

$$\lim_{k \to \infty} b_k^{-1} \lambda_2^{k/2}(W) = 0.$$

(27)

Therefore, there exists a $k^*$ such that for $k \geq k^*$

$$b_k^{-1} \alpha_j \leq \frac{2}{L_f},$$

(28)

$$b_k^{-1} \lambda_2^{k/2}(W) \leq 1.$$

(29)

The above two inequalities imply that for sufficiently large $k$,

$$T_1 \leq C \|x^0\| \lambda_2^{k/2}(W),$$

(30)

$$T_3 \leq \frac{2CB}{L_f} \sum_{j=j'_k+1}^{k-1} \lambda_2^{k-1-j}(W).$$

(31)

From (25), (30) and (31), we get

$$b_k^{-1} \|x^k - \tilde{x}^k\| \leq C \|x^0\| \lambda_2^{k/2}(W)$$

$$+ CB \left( \sum_{j=0}^{j'_k} \alpha_j \lambda_2^{(k-1-j)/2}(W) + \frac{2}{L_f} \sum_{j=j'_k+1}^{k-1} \lambda_2^{k-1-j}(W) \right).$$

By Lemma 8 and (22), there exists a $C^* > 0$ such that

$$\lim_{k \to \infty} b_k^{-1} \|x^k - \tilde{x}^k\| \leq C^*.$$

(33)

We have completed the proof of this proposition.

**D. Proof for Theorem 2**

To prove Theorem 2 we first note that similar to (15), the DGD iterates under decreasing step sizes can be rewritten as

$$x^{k+1} = x^k - \alpha_k \nabla L_{\alpha_k}(x^k),$$

(34)

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.** ([18]) Let $\{v_t\}$ be a nonnegative scalar sequence such that

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

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 (11). Then it holds

$$\alpha_k^{-1} - \alpha_k^{-1} < 2\epsilon L_f(k+1)^\epsilon.$$
\begin{proof}
We first prove that
\[(1 + x)^c - 1 \leq 2cx, \quad \forall x \in [0, 1]. \quad (35)\]
Let \(g(x) = (1 + x)^c - 1 - 2cx\). Then its derivative
\[g'(x) = c(1 + x)^{c-1} - 2c < 0, \quad \forall x \in [0, 1].\]
It implies \(g(x) \leq g(0) = 0\) for any \(x \in [0, 1]\), that is, the inequality \((35)\) holds.

Note that
\[
\alpha_{k+1}^{-1} - \alpha_k^{-1} = L_f((k + 2)^c - (k + 1)^c) \\
= L_f(k + 1)^c ((1 + 1/k + k + 1)^c - 1) \\
\leq 2\epsilon L_f(k + 1)^{c-1}, \quad (36)
\]
where the last inequality holds for \((35)\).
\end{proof}

\begin{lemma}
Let \(\alpha_k\) and \(\beta_k\) be two nonnegative scalar sequences such that

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

where \(\eta_k\) is an arbitrary sequence, we need the following lemma to guarantee that \(\{\alpha_k - \alpha_{k-1}\}|x^{k+1}||i-W| \leq \infty\).
\end{lemma}

\begin{proof}
Note that
\[
\|x^{k+1}||i-W| = \|x^{k+1} - x^{k+1}\|_{i-W}^2 \leq (1 - \lambda_n(W))\|x^{k+1} - x^{k+1}\|^2. \quad (37)
\]
By Lemma 10
\[
(\alpha_{k+1}^{-1} - \alpha_k^{-1})\|x^{k+1}\|_{i-W}^2 \leq 2\epsilon L_f(k + 1)^{c-1}\|x^{k+1}\|_{i-W}^2 \leq 2\epsilon L_f(k + 1)^{c-1}(1 - \lambda_n(W))\|x^{k+1} - x^{k+1}\|^2. \quad (38)
\]
Furthermore, by \((38)\) and Proposition 3 the sequence \(\{\alpha_k - \alpha_{k-1}\}|x^{k+1}||i-W| \leq 1 - \lambda_n(W)\|x^{k+1} - x^{k+1}\| < \infty\).
\end{proof}

\begin{lemma}
(Convergence of weakly summable sequence). Let \(\beta_k\) and \(\gamma_k\) be two nonnegative scalar sequences such that

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

where \(\eta_k\) is an arbitrary sequence, we need the following lemma to guarantee that \(\{\alpha_k - \alpha_{k-1}\}|x^{k+1}||i-W| \leq \infty\).
\end{lemma}

\begin{proof}
We call a sequence \(\{\beta_k\}\) satisfying Lemma \(12\) (a) and (b) a \textit{weakly summable} sequence since itself is not necessarily summable but becomes summable by 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 \(7\).
\end{proof}

\begin{proof}
By condition \((b)\), we have
\[
\sum_{i=k}^{k+k'} \gamma_i \beta_i \to 0, \quad (39)
\]
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 \(\lim_{k \to \infty} \beta_k = C > 0\). Thus, for every \(N > k_0\), there exists a \(k > N\) such that \(\beta_k > \frac{C}{4M}\). Let
\[
k' \leq \frac{C^*}{4M}(k + 1)^c, \quad (40)
\]
where \([x]\) denotes the integer part of \(x\) for any \(x \in \mathbb{R}\). By condition \((c)\), \(|\beta_{j+1} - \beta_j| \leq M \gamma_j\) for any \(j \in \mathbb{N}\), then
\[
\beta_{k+1} \geq \frac{C^*}{4}, \quad \forall i \in \{0, 1, \ldots, k'\}. \quad (40)
\]
Hence,
\[
\sum_{j=k}^{k+k'} \sum_{j=k}^{k+k'} \gamma_j \beta_j \geq \frac{C^*}{4} \sum_{j=k}^{k+k'} \gamma_j \geq \frac{C^*}{4} \int_k^{k+k'} (x + 1)^{c-1} dx \quad (41)
\]
\[
= \left\{ \frac{C^*}{4(c-1)} ((k + k' + 1)^{c-1} - (k + 1)^{c-1}) , \quad \epsilon \in (0, 1), \right. \left. \frac{C^*}{4} (\ln(k + k' + 1) - \ln(k + 1)) , \quad \epsilon = 1. \right. \quad (41)
\]
Note that when \(\epsilon \in (0, 1)\), the term \((k + k' + 1)^{c-1} - (k + 1)^{c-1}\) 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
\[
\ln(k + k' + 1) - \ln(k + 1) = \ln(1 + \frac{k'}{k + 1}) = \ln(1 + \frac{C^*}{M}),
\]
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 \((39)\).
\end{proof}
Moreover,
\[ L_{\alpha_{k+1}}(x^{k+1}) = L_{\alpha_k}(x^{k+1}) + \frac{1}{2}(\alpha_{k+1} - \alpha_k)\|x^{k+1}\|_W^2. \] (45)

Combining (44) and (45) yields (42).

(b) Convergence of objective sequence: By Lemma 11 and Lemma 9, (46) yields the convergence of \( \{L_{\alpha_k}(x^k)\} \)
\[ \sum_{k=0}^{\infty} (\alpha_k^{-1}(1 + \lambda_n(W)) - L_f)\|x^{k+1} - x_k\|^2 < \infty \] (46)
which implies that \( \|x^{k+1} - x^k\| \) converges to 0 at the rate of \( o(k^{-\epsilon/2}) \) and \( \{x^k\} \) is asymptotic regular. Moreover, notice that
\[ \alpha_k^{-1}\|x^{k+1} - x_k\|_W^2 = \alpha_k^{-1}\|x^k - x_k\|_W^2 - L_f = (1 - \lambda_n(W))L_f(k+1)^r\|x^k - x_k\|^2. \]

By Proposition 3, the term \( \alpha_k^{-1}\|x^{k+1} - x_k\|_W^2 \) converges to 0 as \( k \to \infty \). As a consequence,
\[ \lim_{k \to \infty} I^F(x^k) = \lim_{k \to \infty} \left( L_{\alpha_k}(x^k) - \frac{\|x^{k+1}\|_W^2}{2\alpha_k} \right) = \lim_{k \to \infty} \mathcal{L}_{\alpha_k}(x^k). \]

(c) Convergence to a stationary point: Let \( \nabla f(x^k) = \frac{1}{n} I^T \nabla f(x^k) \). By the specific form (11) of \( \alpha_k \), we have
\[ \alpha_k^{-1}(1 + \lambda_n(W)) - L_f = \alpha_k^{-1}(1 + \lambda_n(W)) - L_f \alpha_k \geq \lambda_n(W)\alpha_k^{-1}. \] (47)

Note that
\[ \|x^{k+1} - x^k\| = \frac{1}{n} \|I^T(x^{k+1} - x^k)\| \leq \|x^{k+1} - x^k\|. \] (48)

Thus, (46), (47) and (48) yield
\[ \sum_{k=0}^{\infty} \alpha_k^{-1}\|x^{k+1} - x_k\|^2 < \infty. \] (49)

By the iterate (4) of DGD, we have
\[ x^{k+1} - x^k = -\alpha_k \nabla f(x^k). \] (50)

Plugging (50) into (49) yields
\[ \sum_{k=0}^{\infty} \alpha_k \|\nabla f(x^k)\|^2 < \infty. \] (51)

Moreover,
\[ \|\nabla f(x^{k+1})\|^2 - \|\nabla f(x^k)\|^2 \leq 2B\|\nabla f(x^{k+1}) - \nabla f(x^k)\|\|\nabla f(x^k)\| = \|\nabla f(x^k)\|^2 \]
\[ \leq 2B\|\nabla f(x^{k+1}) - \nabla f(x^k)\| \leq 2B\|\nabla f(x^{k+1}) - \nabla f(x^k)\| \leq 2BL_f\|x^{k+1} - x^k\| , \] (52)

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
\[ \|x^{k+1} - x^k\| = \|x^{k+1} - \bar{x}^{k+1} + \bar{x}^{k+1} - \bar{x}^k + \bar{x}^k - x^k\| \leq \|x^{k+1} - \bar{x}^{k+1}\| + \|x^k - x^k\| + \alpha_k \|\nabla f(x^k)\| \leq \alpha_k, \] (53)
where the first inequality holds for the triangle inequality and (50), and the last inequality holds for Proposition 3 and the bounded assumption of \( \nabla f \). Thus, (52) and (53) imply
\[ \|\nabla f(x^{k+1})\|^2 - \|\nabla f(x^k)\|^2 \leq \alpha_k. \] (54)

By the specific form (11) of \( \alpha_k \), (51) and Lemma 12, it holds
\[ \lim_{k \to \infty} \|\nabla f(x^k)\|^2 = 0. \] (55)

As a consequence,
\[ \lim_{k \to \infty} I^f(x^k) = 0. \] (56)

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 (55) and the continuity of \( \nabla f \), it holds
\[ I^f(x^*) = 0. \]

Moreover, by Proposition 3 \( x^* \) is consensual. As a consequence, \( x^* \) is a stationary point of problem (6).

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

E. Proof for Theorem 7

To prove Theorem 3 we still need the following lemmas.

Lemma 13 (Accumulated consensus of iterates). Under conditions of Proposition 5 we have
\[ \sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^{k+1}\| \leq D_1 + D_2 \sum_{k=0}^{K} \alpha_k^2, \] (57)
where \( D_1 = C\|x^0\|\lambda_2(W)\|, \) \( D_2 = C \left( \frac{\|x^0\|\lambda_2(W)}{2(1 - \lambda_2(W))} + \frac{B}{1 - \lambda_2(W)} \right), \) and \( B \) is specified in Assumption 3.

Proof: By Proposition 5,
\[ \sum_{k=0}^{K} \alpha_k \|x^{k+1} - x^{k+1}\| \leq C\|x^0\|\lambda_2(W) \sum_{k=0}^{K} \alpha_k \lambda_k(W) \]
\[ + CB \sum_{k=0}^{K} \sum_{j=0}^{k} \lambda_k \lambda_{k-j}(W) \alpha_k \alpha_j. \] (58)
In the following, we estimate these two terms in the right-hand side of (58), respectively. Note that
\[
\sum_{k=0}^{K} \alpha_k \lambda_2^2(W) \leq \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2 + \frac{1}{2} \sum_{k=0}^{K} \lambda_2^2(W) \\
\leq \frac{1}{2(1 - \lambda_2(W))} + \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2,
\]
and
\[
\sum_{k=0}^{K} \sum_{j=0}^{k} \lambda_2^{k-j}(W) \alpha_k \alpha_j \leq \frac{1}{2} \sum_{k=0}^{K} \sum_{j=0}^{k} \lambda_2^{k-j}(W) (\alpha_k^2 + \alpha_j^2) \\
= \frac{1}{2} \sum_{k=0}^{K} \alpha_k^2 \sum_{j=0}^{k} \lambda_2^{k-j}(W) + \frac{1}{2} \sum_{j=0}^{K} \sum_{k=j}^{K} \lambda_2^{k-j}(W) \\
\leq \frac{1}{1 - \lambda_2(W)} \sum_{k=0}^{K} \alpha_k^2.
\]
Plugging (59) and (60) into (58) yields (57).

Besides Lemma 13, we also need the following two lemmas, which have appeared in the literature (cf. [6]).

Lemma 14 ([6]). Let \( \gamma_k = \frac{1}{k} \) for some \( 0 < \epsilon \leq 1 \). Then the following hold

(a) if \( 0 < \epsilon < 1/2 \),
\[
\frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1}{1 - \epsilon} \leq O\left(\frac{1}{K^2}\right),
\]
(b) if \( \epsilon = 1/2 \),
\[
\frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1}{2^2(K/2 - 1)} = O\left(\frac{1}{K^2}\right),
\]
(c) if \( 1/2 < \epsilon < 1 \),
\[
\frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1}{2 - \epsilon} \leq O\left(\frac{1}{K^{2-\epsilon}}\right),
\]
(d) if \( \epsilon = 1 \),
\[
\frac{1}{\sum_{k=1}^{K} \gamma_k} \leq \frac{1}{2^{K-1}} = O\left(\frac{1}{K}\right),
\]
where \( L^* = L_f + \alpha_k^{-1}(1 - \lambda_n(W)) \), and by (15), we have

\[
\nabla \mathcal{L}_{\alpha_k}(x^k) = \alpha_k^{-1}(x^k - x^{k+1}).
\]

Then (62) implies
\[
\mathcal{L}_{\alpha_k}(u) \geq \mathcal{L}_{\alpha_k}(x^{k+1}) + \alpha_k^{-1}(x^k - x^{k+1}, u - x^{k+1}) - \frac{L^*}{2} \|x^{k+1} - x^k\|^2.
\]

Note that the specific form of \( \alpha_k = \frac{1}{L_f(x^{k+1})} \), 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 \( \alpha_k < \frac{1}{L_f(W)} \) starting from the first step so that \( L^* \leq \alpha_k^{-1} \) holds from the initial step. Thus, (63) implies
\[
\mathcal{L}_{\alpha_k}(u) \geq \mathcal{L}_{\alpha_k}(x^{k+1}) + \alpha_k^{-1}(x^k - x^{k+1}, u - x^{k+1}) - \frac{1}{2 \alpha_k} \|x^{k+1} - x^k\|^2.
\]

Recall that for any two vectors \( a, b \), it holds \( 2(a, b) - ||a||^2 = ||b||^2 - ||a - b||^2 \). 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 (61).

Note that the optimal solution \( x_{\text{opt}} \) is consensual and thus, \( ||x_{\text{opt}}||_W^2 = 0 \). Therefore, \( \mathcal{L}_{\alpha_k}(x_{\text{opt}}) = \bar{f}(x_{\text{opt}}) = f_{\text{opt}} \). By (61), we have
\[
\alpha_k (\mathcal{L}_{\alpha_k}(x^{k+1}) - f_{\text{opt}}) \leq (||x^k - x_{\text{opt}}||^2 - ||x^{k+1} - x_{\text{opt}}||^2)/2.
\]

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

Moreover, noting that \( \mathcal{L}_{\alpha_k}(x^{k+1}) = \bar{f}(x^{k+1}) \) and by the convexity of \( \mathcal{L}_{\alpha_k} \),
\[
\mathcal{L}_{\alpha_k}(x^{k+1}) \geq \bar{f}(x^{k+1}) + \langle \nabla \mathcal{L}_{\alpha_k}(x^{k+1}), x^{k+1} - x^{k+1} \rangle \\
\geq \bar{f}(x^{k+1}) - B ||x^{k+1} - x^{k+1}||,
\]
where the second inequality holds by the bounded assumption of gradient (cf. Assumption 3). Plugging (60) into (65) yields
\[
\sum_{k=0}^{K} \alpha_k (\bar{f}(x^{k+1}) - f_{\text{opt}}) \leq \frac{1}{2} ||x^0 - x_{\text{opt}}||^2 + B \sum_{k=0}^{K} \alpha_k ||x^{k+1} - x^{k+1}||.
\]

By the definition of \( \bar{f}^K(15) \), then (67) implies
\[
(f^K - f_{\text{opt}}) \sum_{k=0}^{K} \alpha_k \leq \frac{1}{2} ||x^0 - x_{\text{opt}}||^2 + B \sum_{k=0}^{K} \alpha_k ||x^{k+1} - x^{k+1}|| \\
\leq D_3 + D_4 \sum_{k=0}^{K} \alpha_k^2,
\]

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 14. As a consequence,

$$
\tilde{f}^K - f_{\text{opt}} \leq \frac{D_3 + D_4 \sum_{k=0}^{K} \alpha_k^2}{\sum_{k=0}^{K} \alpha_k}.
$$

(70)

Furthermore, by Lemma 14 we get the claims of this theorem.

V. EXTENSION TO PROX-DGD

Consider the following differentiable+proximable multi-agent consensus optimization problem:

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

(71)

where $f_i$ is $L_{f_i}$-Lipschitz differentiable (as is assumed before) and $r_i$ is possibly nonsmooth, nonconvex, or both. Let

$$
r(x) \triangleq \sum_{i=1}^{n} r_i(x(i)).
$$

The algorithm Prox-DGD can be applied to the above problem (71):

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

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

(72)

where the proximal operator is

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

(73)

As is done before, the iteration (72) can be reformulated as

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

(74)

based on which, we define the Lyapunov function

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

where we recall $\mathcal{L}_{\alpha_k}(x) = \sum_{i=1}^{n} f_i(x(i)) + \frac{1}{2\alpha_k} \|x\|_{W^{-1}}^2$. Then (74) is clearly the forward-backward splitting (a.k.a., prox-gradient) iteration for minimizing $\mathcal{L}_{\alpha_k}(x)$. Specifically, (74) 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 (71) 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 $L_f \triangleq \max_i L_{f_i}$.

A. Prox-DGD with a fixed step size: $\alpha_k \equiv \alpha$

**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 $0 < \alpha < \frac{1}{4\lambda_n(W)}$.

$$
\begin{align*}
\hat{\mathcal{L}}_{\alpha}(x^{k+1}) & \leq \hat{\mathcal{L}}_{\alpha}(x^k) \\
& - \frac{1}{2} \left( \alpha^{-1} (1 + \lambda_n(W)) - L_f \right) \|x^{k+1} - x^k\|^2, \forall k \in \mathbb{N}.
\end{align*}
$$

(75)

**C2:** $r_i$’s are not necessarily convex. Set $0 < \alpha < \frac{\lambda_n(W)}{L_f}$.

$$
\begin{align*}
\hat{\mathcal{L}}_{\alpha}(x^{k+1}) & \leq \hat{\mathcal{L}}_{\alpha}(x^k) \\
& - \frac{1}{2} \left( \alpha^{-1} \lambda_n(W) - L_f \right) \|x^{k+1} - x^k\|^2, \forall k \in \mathbb{N}.
\end{align*}
$$

(76)

**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

$$
\begin{align*}
\hat{\mathcal{L}}_{\alpha}(x^{k+1}) - \hat{\mathcal{L}}_{\alpha}(x^k) & = \mathcal{L}_\alpha(x^{k+1}) - \mathcal{L}_\alpha(x^k) + r(x^{k+1}) - r(x^k) \\
& \leq \langle \nabla \mathcal{L}_\alpha(x^k), x^{k+1} - x^k \rangle + \frac{L^*}{2} \|x^{k+1} - x^k\|^2 \\
& \quad + r(x^{k+1}) - r(x^k).
\end{align*}
$$

(77)

**C1:** From the convexity of $r_i$ (23), and (74), it follows that

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

This and the convexity of $r$ further give us

$$
r(x^{k+1}) - r(x^k) \leq \langle \xi^{k+1}, x^{k+1} - x^k \rangle
$$

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

Substituting this inequality into the inequality (77) and then expanding $L^* = L_f + \alpha^{-1}(1 - \lambda_n(W))$ yield

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

(78)

Sufficient descent requires the last term to be negative, thus

$$
0 < \alpha < \frac{1}{4\lambda_n(W)}.
$$

**C2:** From (73) and (74), it follows that the function $r(u) + \|u - (x^k - \alpha \nabla \mathcal{L}_{\alpha}(x^k))\|^2$ reaches its minimum at $u = x^{k+1}$. Comparing the values of this function at $x^{k+1}$ and $x^k$ yields

$$
r(x^{k+1}) - r(x^k) \leq \frac{1}{2\alpha} \|x^k - (x^k - \alpha \nabla \mathcal{L}_{\alpha}(x^k))\|^2 - \frac{1}{2\alpha} \|x^{k+1} - (x^k - \alpha \nabla \mathcal{L}_{\alpha}(x^k))\|^2
$$

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

Substituting this inequality into (77) and expanding $L^*$ yield

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

(79)

Hence, sufficient descent requires $0 < \alpha < \frac{\lambda_n(W)}{L_f}$. ■
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 4 (Part 2).

By Lemma 16 and under a proper step size, \( \hat{L}_\alpha(x^k) \) is nonincreasing and upper bounded by \( \hat{L}_\alpha(0) \). Hence, \( \sum_{i=1}^{n} (f_i(x^k_i) + r_i(x^k_i)) \) is upper bounded by \( \hat{L}_\alpha(0) \). Consequently, \( \{ x^k \} \) is bounded due to the coercivity of each \( f_i + r_i \) (see Assumption 3 (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 (74), the following optimality condition holds
\[
0 \in \alpha^{-1}(x^{k+1} - x^k + \alpha \nabla \hat{L}_\alpha(x^k)) + \partial \hat{r}(x^{k+1}),
\]
which immediate yields
\[
\| \nabla \hat{L}_\alpha(x^{k+1}) + \xi^{k+1} \|
\leq (\alpha^{-1}\| x^{k+1} - x^k \| + \| \nabla \hat{L}_\alpha(x^{k+1}) - \nabla \hat{L}_\alpha(x^k) \|)
\leq (\alpha^{-1} + L) \| x^{k+1} - x^k \|
\leq (\alpha^{-1}(2 - \lambda_n(W)) + L_f) \| x^{k+1} - x^k \|.
\]
Thus, then the claim of Lemma 18 holds.

Based on Lemmas 16-18, we establish the global convergence of Prox-DGD.

Theorem 4 (Global convergence of Prox-DGD). Let \( \{ x^k \} \) be the sequence generated by Prox-DGD (72) where the step size \( \alpha \) satisfies \( 0 < \alpha < \frac{1}{\lambda_n(W) + L_f} \). 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 \( \hat{L}_\alpha(x) \). Furthermore, the running best rate of the sequences \( \{ \| x^{k+1} - x^k \| \} \) and \( \{ g^{k+1} \| \} \) (where \( g^{k+1} \) is defined in Lemma 18) are both \( o(1) \).

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

The proof of this theorem is similar to that of Theorem 1 and thus is omitted.

Proposition 4 (Rate of convergence of Prox-DGD). Under assumptions of Theorem 2, \( \alpha \) satisfies the KL inequality at an accumulation point \( x^* \) with \( \nu(s) = c_1 s^{1-\theta} \) 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 \( \theta \in (0, \frac{1}{2}] \), \( \| x^k - x^* \| \leq C \tau \) for all \( k \geq k^* \) for some \( k^* > 0 \), \( C_1 > 0 \), \( \tau \in [0, 1) \).

(c) If \( \theta \in (\frac{1}{2}, 1) \), \( \| x^k - x^* \| \leq C_1 k^{-\theta/(2\theta - 1)} \) for all \( k \geq k^* \), for certain \( k^* > 0 \), \( C_1 > 0 \).

Proof: The proof is similar to that of Proposition 2. We shall however note that in (20), \( a = \frac{1}{2}(\alpha^{-1}(1+\lambda_n(W)) - L_f) \) if \( r_i \)’s are convex, while \( a = \frac{1}{2}(\alpha^{-1}(W) - L_f) \) if \( r_i \)’s are not necessarily convex.

B. Prox-DGD with decreasing step sizes

In Prox-DGD, we also use the decreasing step size (11). To investigate the convergence of Prox-DGD with decreasing step sizes, the bounded gradient Assumption 3 should be revised as follows.

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

Let \( \tilde{B} = \sum_{i=1}^{n} (B_i + B_r) \). Then \( \nabla f(x) + \xi \) (where \( \xi \in \partial r(x) \) for any \( x \in \mathbb{R}^{n \times p} \)) is uniformly bounded by \( \tilde{B} \). Note that the same assumption is used to analyze the convergence of distributed proximal-gradient method in the convex setting [5, 6].

Based on the iterate (72) of Prox-DGD, we derive the following recursion of the iterates of Prox-DGD, which is similar to (12).

Lemma 19 (Recursion of \( \{ x^k \} \)). For any \( k \in \mathbb{N} \),
\[
x^k = W^k x^0 - \sum_{j=0}^{k-1} \alpha_j W^{k-1-j}(\nabla f(x^j) + \xi^{j+1}),
\]
where \( \xi^{j+1} \in \partial r(x^{j+1}) \) is the one determined by the proximal operator (73), for any \( j = 0, \ldots, k-1 \).

Proof: By the definition of the proximal operator (73), the iterate (72) implies
\[
x^{k+1} + \alpha_k \xi^{k+1} = W x^k - \alpha_k \nabla f(x^k),
\]
where \( \xi^{k+1} \in \partial r(x^{k+1}) \), and thus
\[
x^{k+1} = W x^k - \alpha_k (\nabla f(x^k) + \xi^{k+1}).
\]
By (81), we can easily derive the recursion (79).

In light of Lemma 19 the claims in Proposition 3 and Corollary 1 also hold for Prox-DGD.

Proposition 5 (Asymptotic consensus and rate). Let Assumptions 2 and 5 hold. In Prox-DGD, use the step sizes (11). There hold
\[
\| x^k - x^* \| \leq C \| x^0 \| \lambda_n(W) + B \sum_{j=0}^{k-1} \alpha_j \lambda_n^{k-1-j}(W),
\]
and \( \| x^k - \bar{x}^* \| \) converges to 0 at the rate of \( O(1/(k+1)) \). Moreover, let \( x^* \) be any global solution of the problem (71). Then \( \| x^k - x^* \| \) converges to 0 at the rate of \( O(1/(k+1)^2) \).
The proof of this proposition is similar to that of Proposition 3. It only needs to note that the subgradient term $\nabla f(x^i) + \xi_{j+1}$ is uniformly bounded by the constant $B$ for any $j$. Thus, we omit the proof.

**Lemma 20.** Let Assumptions 2 and 4 hold. In Prox-DGD, use the step sizes $\left\{\frac{1}{2}\alpha_k - \alpha_k\right\}$. By (81), it follows that $\lambda^i$ is a solution of (11). Results are given in two cases below:

**C1:** $\bar{r}_i$'s are convex. For any $k \in \mathbb{N}$,

\[
\hat{\lambda}_{\alpha_k+1}(x^{k+1}) \leq \hat{\lambda}_{\alpha_k}(x^k) + \left(\frac{1}{2}(\alpha_k - 1)\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(x^k) \leq \left(\frac{1}{2}(\alpha_k - 1)\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(x^k).
\]

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

\[
\hat{\lambda}_{\alpha_k+1}(x^{k+1}) \leq \hat{\lambda}_{\alpha_k}(x^k) + \left(\frac{1}{2}(\alpha_k - 1)\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(x^k) \leq \left(\frac{1}{2}(\alpha_k - 1)\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(x^k).
\]

The proof of this lemma is similar to that of Lemma 16 via noting that

\[
\hat{\lambda}_{\alpha_k+1}(x^{k+1}) = \hat{\lambda}_{\alpha_k}(x^k) + (\hat{\lambda}_{\alpha_k+1}(x^{k+1}) - \hat{\lambda}_{\alpha_k}(x^{k+1})) + (\hat{\lambda}_{\alpha_k}(x^{k+1}) - \hat{\lambda}_{\alpha_k}(x^k)),
\]

and

\[
\hat{\lambda}_{\alpha_k+1}(x^{k+1}) - \hat{\lambda}_{\alpha_k}(x^{k+1}) = \left(\frac{1}{2}(\alpha_k - 1)\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(x^k).
\]

While the term $\hat{\lambda}_{\alpha_k}(x^{k+1}) - \hat{\lambda}_{\alpha_k}(x^k)$ 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 $\left\{\frac{1}{2}\alpha_k - \alpha_k\right\}$. If further each $\bar{f}_i$ and $\bar{r}_i$ are convex, then for any $u \in \mathbb{R}^{n \times p}$, we have

\[
\hat{\lambda}_{\alpha_k}(x^{k+1}) - \hat{\lambda}_{\alpha_k}(u) \leq \left(\frac{1}{2}\alpha_k - \alpha_k\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(u) \leq \left(\frac{1}{2}\alpha_k - \alpha_k\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(u).
\]

**Proof:** By Lemma 18 we have

\[
\hat{\lambda}_{\alpha_k}(u) = \hat{\lambda}_{\alpha_k}(x^{k+1}) + \left(\frac{1}{2}(\alpha_k - 1)\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(u) \leq \left(\frac{1}{2}\alpha_k - \alpha_k\right)x^{k+1} - \hat{\lambda}_{\alpha_k}(u).
\]

We now have

\[
\hat{\lambda}_{\alpha_k}(u) \geq \hat{\lambda}_{\alpha_k}(x^{k+1}) + \langle \nabla \hat{\lambda}_{\alpha_k}(x^{k+1}), u - x^{k+1} \rangle - \frac{L^*}{2}x^{k+1} - x^{k+1} ||^2,
\]

where $L^* = L_f + \alpha_k(1 - \lambda_n(W))$, and by the convexity of $\bar{r}$, we have

\[
\bar{r}(u) \geq \bar{r}(x^{k+1}) + \langle \xi^{k+1}, u - x^{k+1} \rangle,
\]

where $\xi^{k+1} \in \partial \bar{r}(x^{k+1})$ is the one determined by the proximal operator (73). By (81), it follows

\[
\hat{\lambda}_{\alpha_k}(u) \geq \hat{\lambda}_{\alpha_k}(x^{k+1}) + \alpha_k(x^{k+1} - x^{k+1}) - \nabla \hat{\lambda}_{\alpha_k}(x^{k+1}).
\]

Plugging (86) into (85), and then summing up (84) and (85) yield

\[
\hat{\lambda}_{\alpha_k}(u) \geq \hat{\lambda}_{\alpha_k}(x^{k+1}) + \alpha_k(x^{k+1} - x^{k+1}) - \frac{L^*}{2}x^{k+1} - x^{k+1} ||^2.
\]

Similar to the rest proof of the inequality (61), we can prove this lemma based on (87).

For any $x \in \mathbb{R}^{n \times p}$, define $s(x) = \sum_{i=1}^{n} f_i(x_i) + r_i(x_i)$. Let $\lambda^i$ be a set of solutions of (71), $x_{opt} = Proj_{x_i} - \lambda^i(x_i) \in \lambda^i$, and $s_{opt} = s(x_{opt})$ be the optimal value of (71). Define

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

Based on Lemma 20 and Lemma 21, we can extend Theorems 2 and 3 to the algorithm Prox-DGD as shown in the following.

**Theorem 5.** Convergence and rate. Let Assumptions 2, 4 and 5 hold. In Prox-DGD, use the step sizes $\left\{\frac{1}{2}\alpha_k - \alpha_k\right\}$. Then

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

(b) $\sum_{k=0}^{\infty} \alpha_k(1 + \lambda_n(W)) - L_f \|x^{k+1} - x^k\|^2 < \infty$ when $\bar{r}_i$'s are convex or, $\sum_{k=0}^{\infty} \alpha_k(1 + \lambda_n(W)) - L_f \|x^{k+1} - x^k\|^2 < \infty$ when $\bar{r}_i$'s are not necessarily convex;

(c) if $\{\xi^k\}$ satisfies $\|\xi^{k+1} - \xi^k\| \leq L_r \|x^{k+1} - x^k\|^2$ for all $k > k_0$, some constant $L_r > 0$ and a sufficiently large integer $k_0 > 0$, then

\[
\lim_{k \to \infty} \frac{1}{T} \|\xi^{k+1} - \xi^k\| = 0,
\]

where $\xi^{k+1} \in \bar{r}(x^{k+1})$ is the one determined by the proximal operator (73), and any limit point is a stationary point of problem (71).

(d) In addition, if there exists an isolated accumulation point, then $\{x^{k+1}\}$ converges.

(e) if further each $f_i$ and $r_i$ are convex, then the claims on the rates of $\{f^{K}\}$ in Theorem 5 also hold for the sequence $\{s^{K}\}$ defined in (87).

The proof of Theorem 5(a)-(d) is similar to that of Theorem 2 while the proof of Theorem 5(e) is very similar to that of Theorem 3 and thus the proof of this theorem is omitted. Theorem 5(b) implies that the best running rate of $\|x^{k+1} - x^k\|^2$ is $o\left(\frac{1}{k^2}\right)$. The additional condition imposed on $\{\xi^k\}$ in Theorem 5(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 28. 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 lie in such neighborhood of this limit point when $k$ is sufficiently large. Theorem 5(e) gives the convergence rates of Prox-DGD in the convex setting.

**VI. CONCLUSION**

In this paper, we study the convergence behaviour of DGD algorithm for solving the smooth, possibly nonconvex consensus optimization problem. We consider both fixed and decreasing step sizes. When using the fixed step size, we show that under certain assumptions, the iterations 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 is inversely proportional to the gap between the largest and the second largest eigenvalue 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 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 the algorithm Prox-DGD designed for minimizing the sum of a differentiable function and a proximal function. Both functions can be nonconvex. While if the proximal function is convex, a larger fixed step size is allowed. These results are obtained by combining both existing and new proof techniques.

ACKNOWLEDGMENTS

The work of J. Zeng has been supported in part by the NSF grants 61603162, 61603163, 11501440. The work of W. Yin has been supported in part by the NSF grant ECCS-1462398 and ONR grants N000141410683 and N000141210838.

REFERENCES

[1] H. Attouch, and J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, Math. Program., 116: 5-16, 2009.
[2] H. Attouch, J. Bolte and B. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods, Math. Program., Ser. A, 137: 91-129, 2013.
[3] P. Bianchi and J. Jakubowicz, Convergence of a multi-agent projected stochastic gradient algorithm for nonconvex optimization, IEEE Trans. Automatic Control, 58(2):391-405, 2013.
[4] J. Bolte, A. Daniilidis and A. Lewis, The Lojasiewicz inequality for non-smooth subanalytic functions with applications to subgradient dynamical systems, SIAM Journal on Optimization, 17(4):1205-1223, 2007.
[5] A. Chen and A. Ozdaglar, A fast distributed proximal gradient method, in Proc. 50th Allerton Conf. Commun., Control Comput., Moticello, IL, Oct. 2012, pp. 601-608.
[6] A. Chen, Fast Distributed First-Order Methods, Master’s thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2012.
[7] Y.T. Chow, T. Wu and W. Yin, Cyclic Coordinate Update Algorithms for Fixed-Point Problems: Analysis and Applications. UCLA CAM 16-78, 2016.
[8] W. Deng, M. Lai, Z. Peng and W. Yin, Parallel multi-block admm with o(1/k) convergence, Journal of Scientific Computing, DOI 10.1007/s10915-016-0318-2, 2016.
[9] S. Hosseini, A. Chapman and M. Mesbahi, Online distributed optimization on dynamic networks, IEEE T. Auto. Control, 61(11): 3545–3550, 2016.
[10] D. Jakovetic, J. Xavier and J. Moura, Fast distributed gradient methods, IEEE Trans. Automatic Control, 59:1131-1146, 2014.
[11] K. Knopp, Infinite sequences and series, Courier Corporation, 1956.
[12] S. Lojasiewicz, Sur la géométrie semi-et sous-analytique, Ann. Inst. Fourier (Grenoble) 43(5): 1575–1595, 1993.
[13] I. Matei and J. Baras, Performance evaluation of the consensus-based distributed subgradient method under random communication topologies, IEEE J. Sel. Top. Signal Process., 5:754-771, 2011.
[14] H. McMahan and M. Streeter, Delay-Tolerant algorithms for asynchronous distributed online learning, In: Advances in Neural Information Processing Systems (NIPS), 2014.
[15] A. Nedic and A. Ozdaglar, Distributed subgradient methods for multi-agent optimization, IEEE Trans. Automatic Control, 54:48-61, 2009.
[16] M. Raginsky, N. Kiarashi and R. Willett, Decentralized online convex programming with local information, In: 2011 American Control Conference, San Francisco, CA, USA, 2011.
[17] S. Ram, A. Nedic and V. Veeravalli, Distributed stochastic subgradient projection algorithms for convex optimization, J. Optim. Theory Appl., 147: 516-545, 2010.
[18] H. Robbins and D. Siegmund, A convergence theorem for nonnegative almost supermartingales and some applications, in Proc. Optim. Methods Stat., 233-257, 1971.