Distributed Zero-Order Algorithms for Nonconvex Multi-Agent Optimization

Yujie Tang\(^1\) and Na Li\(^1\)

\(^1\)School of Engineering and Applied Sciences, Harvard University

Abstract

Distributed multi-agent optimization is the core of many applications in distributed learning, control, estimation, etc. Most existing algorithms assume knowledge of first-order information of the objective and have been analyzed for convex problems. However, there are situations where the objective is nonconvex, and one can only evaluate the function values at finitely many points. In this paper we consider derivative-free distributed algorithms for nonconvex multi-agent optimization, based on recent progress in zero-order optimization. We develop two algorithms for different settings, provide detailed analysis of their convergence behavior, and compare them with existing centralized zero-order algorithms and gradient-based distribution algorithms.

1 Introduction

Consider a set of \(n\) agents connected over a network, each of which is associated with a smooth local objective function \(f_i\) that can be nonconvex. The goal is to solve the optimization problem

\[
\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)
\]

with the restriction that \(f_i\) is only known to agent \(i\) and each agent can exchange information only with its neighbors in the network during the optimization procedure. We focus on the situation where only zero-order information of \(f_i\) is available to agent \(i\).

Distributed multi-agent optimization lies at the core of a wide range of applications, and a large body of literature has been contributed to distributed multi-agent optimization algorithms. One line of research combines (sub)gradient-based methods with a consensus/averaging scheme. It has been shown that, for convex functions, the convergence rates of distributed gradient-based algorithms can match or nearly match those of centralized gradient-based algorithms. Specifically, [1][2][3] proposed and analyzed distributed algorithms with \(O(\log t/\sqrt{t})\) convergence for nonsmooth convex functions; [4][5][6] proposed distributed algorithms with \(O(1/t)\) convergence for smooth convex functions and linear convergence for strongly convex functions; [7] employed Nesterov’s gradient descent method and showed \(O(1/t^{1.4-\epsilon})\) convergence for smooth convex functions and improved linear convergence for strongly convex functions. Besides convergence rates, some works have additional focuses such
as time-varying/undirected graphs \[1, 2, 8, 5, 10\], uncoordinated step sizes \[11, 12\], stochastic (sub)gradient \[13, 14\], etc.

While distributed convex optimization has a broad applicability, nonconvex problems also appear in important applications such as distributed learning \[15\], compressed sensing \[16\], robotic networks \[17\], operation of wind farms \[18\], etc, and several works have considered nonconvex multi-agent optimization. \[19\] studied the behavior of distributed projected stochastic gradient algorithm via tools from continuous-time dynamical systems. \[20\] developed distributed algorithms based on the convexification-decomposition technique. \[21\] established convergence of the distributed push-sum algorithm for nonconvex problems and also proposed perturbations to avoid local maxima. \[22\] studied the decentralized parallel stochastic gradient descent, and showed its $O(1/\sqrt{T})$ convergence rate to stationary points. \[23\] proposed a decentralized Frank–Wolfe algorithm, and showed its $O(1/\sqrt{T})$ convergence rate of the quantity $\langle \nabla f(\bar{x}(t)), \bar{x}(t) - x^* \rangle$. \[24\] proposed the proximal primal-dual algorithm for distributed nonconvex optimization, and showed its $O(1/t)$ convergence to a stationary point. \[25\] studied decentralized gradient descent-type algorithms for nonconvex problems, established their convergence to stationary points and also provided consensus rates.

Recently there has been increasing interest in zero-order optimization, where one does not have access to the gradient of the objective. Such situations can occur, for example, when only black-box procedures are available for computing the values of the functional characteristics of the problem, or when resource limitations restrict the use of fast or automatic differentiation techniques. Many existing works on zero-order optimization are based on constructing gradient estimators using finitely many function evaluations. \[26\] proposed and analyzed a single-point gradient estimator, and \[27\] further studied the convergence rate of single-point zero-order algorithms for highly smooth objectives. \[28\] proposed two-point gradient estimators and showed that the convergence of the resulting algorithms are comparable with their first-order counterparts. \[29\] studied two-point gradient estimators in stochastic nonconvex zero-order optimization. \[30\] and \[31\] showed that for stochastic zero-order convex optimization with two-point gradient estimators, the optimal rate $O(\sqrt{d/N})$ is achievable where $N$ denotes the number of function value queries. \[32\] proposed and analyzed a zero-order stochastic Frank-Wolfe algorithm.

Some recent works have also started to combine zero-order and distributed methods. \[33\] proposed a distributed zero-order algorithm for stochastic nonconvex problems based on the method of multipliers. \[34\] proposed a zero-order ADMM algorithm for distributed online convex optimization. \[35\] proposed a distributed zero-order algorithm over random networks and established its convergence for strongly convex objectives. \[36\] considered distributed zero-order methods for constrained convex optimization. On the other hand, there are still many questions remain to be studied in distributed zero-order optimization, e.g., how zero-order and distributed methods affect the performance of each other and whether their fundamental structural properties could be kept by tuning the way of their combination. This paper aims at providing messages along this line: We propose and analyze two zero-order distributed algorithms for deterministic nonconvex optimization, and compare their convergence rates with their distributed first-order and centralized zero-order counterparts. The first algorithm employs a simple two-point gradient estimator and only does consensus on the local decision variables, while the second algorithms uses a 2d-point gradient estimator and incorporates gradient tracking. The convergence rates of the two algorithms are summarized in Table 1 and are compared with their distributed first-order and centralized counterparts. We show that for deterministic nonconvex optimization, the proposed distributed zero-order algorithms have comparable convergence behavior with their first-order and centralized counterparts. These results shed light on how zero-order evaluations affect distributed optimization...
Table 1: Comparison of different algorithms for distributed optimization and zero-order optimization.

|                          | smooth                     | gradient dominated          |
|--------------------------|----------------------------|-----------------------------|
| this paper (nonconvex)   | $O\left(\frac{d}{N} \log N\right)$ | $O\left(\frac{d}{N}\right)$ |
| Alg. 1                   | $O\left(\sqrt{\frac{d}{N}} N \log N\right)$ | $O\left(1 - c(1 - \rho)^2 \left(\frac{\mu}{L}\right)^{\frac{3}{2}}\right)^{N/d}$ |
| Alg. 2                   | $O\left(\frac{d}{N}\right)$ | $O\left(\left[1 - c(1 - \rho)^2 \left(\frac{\mu}{L}\right)^{\frac{3}{2}}\right]^{N/d}\right)$ |
| distributed              | $O\left(\log t \sqrt{\frac{d}{t}}\right)$ [2] (convex) | $O\left(\frac{\log t}{t}\right)$ [37] (convex) |
| gradient-based methods   | $O\left(\frac{1}{\sqrt{T}}\right)$ [22] (nonconvex) | $O\left(\frac{\log t}{t}\right)$ [37] (convex) |
| methods                  | $O\left(\frac{1}{t}\right)$ [6] | $O\left(\left[1 - c(1 - \rho)^2 \left(\frac{\mu}{L}\right)^{\frac{3}{2}}\right]^t\right)$ [15] |
| centralized              | $O\left(\frac{d}{N}\right)$ (nonconvex) | $O\left(\left[1 - c \frac{\mu}{\bar{d}}\right]^N\right)$ (strongly convex) |

Note: $t$ denotes the number of iterations, $N$ denotes the number of function value queries, $d$ denotes the dimension of the decision variable, and $c$’s represent numerical constants that can be different for different algorithms.

$T$ denotes the total number of iterations provided before the optimization procedure. The rate in [22] assumes knowledge of $T$ and uses $T$ to set a constant step size.

The listed convergence rates are the ergodic rates of $\|\nabla f\|^2$ for the smooth case, and the objective error rates for the gradient dominated case, respectively.

We do not include algorithms with Nesterov-type acceleration in this comparison.

and how the network structure affects zero-order algorithms.

**Notation:** We denote the $\ell_2$-norm by $\|\cdot\|$, and the standard inner product by $\langle x, y \rangle := x^T y$. The standard basis of $\mathbb{R}^d$ will be denoted by $\{e_k\}_{k=1}^d$. The closed unit ball $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$ will be denoted by $\mathbb{B}_d$, and the unit sphere $\{x \in \mathbb{R}^d : \|x\| = 1\}$ will be denoted by $\mathbb{S}_{d-1}$. The uniform distributions over $\mathbb{B}_d$ and $\mathbb{S}_{d-1}$ will be denoted by $\mathcal{U}(\mathbb{B}_d)$ and $\mathcal{U}(\mathbb{S}_{d-1})$.

# 2 Formulation and Algorithms

## 2.1 Problem Formulation

Let $\mathcal{N} = \{1, 2, \ldots, n\}$ denote the set of agents. Suppose the agents are connected by a communication network, whose topology is represented by a connected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ where $\mathcal{E}$ denotes the set of edges that represent the communication links. The graph $\mathcal{G}$ is assumed to be undirected,
meaning that the communication links of the network are bidirectional.

Each agent \(i\) is associated with a local objective function \(f_i : \mathbb{R}^d \to \mathbb{R}\). The goal of the agents is to collaboratively solve the optimization problem

\[
\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x).
\]

We assume that at each time step, agent \(i\) can only query the function values of \(f_i\) at finitely many points, and can only communicate with its neighbors in the communication network. We also assume that the queries of the function values are noise-free and error-free.

We use the following definitions of function classes throughout the paper:

\textbf{Definition 1.} 1. A function \(f : \mathbb{R}^d \to \mathbb{R}\) is said to be \(L\)-smooth if \(f\) is continuously differentiable and satisfies \(\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|\) for all \(x, y \in \mathbb{R}^d\).

2. A function \(f : \mathbb{R}^d \to \mathbb{R}\) is said to be \(G\)-Lipschitz if \(|f(x) - f(y)| \leq G\|x - y\|\) for all \(x, y \in \mathbb{R}^d\).

3. A function \(f : \mathbb{R}^d \to \mathbb{R}\) is said to be \(\mu\)-gradient dominated\(^1\) if \(f\) is differentiable, has a global minimizer \(x^*\), and

\[
2\mu(f(x) - f(x^*)) \leq \|\nabla f(x)\|^2
\]

for all \(x \in \mathbb{R}^d\).

Lipschitz continuity and smoothness are standard assumptions on the objective function in distributed optimization. The concept of gradient domination can be viewed as a nonconvex analog of strict convexity, and is important in the non-convex optimization literature [39, 38]. It has been observed that nonconvex but gradient-dominated objective functions appear in various applications [40, 41].

\subsection{2.2 Algorithms}

We propose consensus-based distributed algorithms for solving Problem (1), where each agent maintains a local copy of the global variables, and weighs its neighbors’ information to updates the local variable. Specifically, we introduce a consensus matrix \(W = [W_{ij}] \in \mathbb{R}^{n \times n}\) that satisfies the following assumption:

\textbf{Assumption 1.} 1. \(W_{ij} \geq 0\) for all \(i, j \in \mathcal{N}\) and \(W1_d = W^T1_d = 1_d\), i.e., \(W\) is a doubly stochastic matrix.

2. \(W_{ii} > 0\) for all \(i \in \mathcal{N}\), and for two distinct agents \(i\) and \(j\), \(W_{ij} > 0\) if and only if \((i, j) \in \mathcal{E}\).

It is a standard result of consensus optimization that, when Assumption 1 is satisfied, we have

\[
\rho := \sup_{\|x\|=1} \frac{\|W(x - n^{-1}1_n1_n^Tx)\|}{\|x - n^{-1}1_n1_n^Tx\|} < 1.
\]

Because each agent \(i\) can only query function values of \(f_i\) at finitely many points, we employ techniques from zero-order optimization and introduce the following maps:

\[
G_f^{(2)}(x; u, z) := \frac{d}{2u} \cdot \frac{f(x + uz) - f(x - uz)}{z},
\]

\[
G_f^{(2d)}(x; u) := \sum_{k=1}^{d} \frac{f(x + u e_k) - f(x - u e_k)}{2u} e_k.
\]

\(^1\) This definition is adopted from [38].
Algorithm 1: 2-point gradient estimator without global gradient tracking

for $t = 1, 2, 3, \ldots$ do
    foreach $i \in \mathcal{N}$ do
        1. Generate $z^i(t) \sim U(S_d - 1)$ independently from $(z^i(\tau))_{\tau=1}^{t-1}$ and $(z^j(\tau))_{\tau=1}^{t}$ for $j \neq i$.
        2. Update $x^i(t)$ by
           \[ g^i(t) = G_{f_i}^{(2)}(x^i(t-1); u_t, z^i(t)), \]
           \[ x^i(t) = \sum_{j=1}^{n} W_{ij} (x^j(t-1) - \eta_t g^j(t)). \]
    end
end

The map $G_{f_i}^{(2d)}(x; u)$ approximates $\nabla f(x)$ by difference quotients along $d$ orthogonal directions, and can be viewed as a noise-free version of the Kiefer-Wolfowitz type method [42]. The following proposition establishes the rationale of employing $G_{f_i}^{(2d)}(x; u, z)$ as an estimator of $\nabla f(x)$ when $z$ is properly randomly generated:

**Proposition 1** ([26]). Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is continuous. Then for any $u > 0$ and $x \in \mathbb{R}^d$,
\[
\mathbb{E}_{z \sim U(S_d - 1)} \left[ G_{f_i}^{(2)}(x; u, z) \right] = \nabla f^u(x),
\]
where $f^u(x) := \mathbb{E}_{y \sim U(B_d)} [f(x + uy)]$.

Basically, Proposition 1 indicates that when $z$ is randomly generated from the sphere $S_d - 1$, the expectation of $G_{f_i}(x; u, z)$ is the gradient of a “smoothed version” of $f$.

We propose two distributed algorithms for Problem (1) based on the gradient estimators (2) and (3):

1. Algorithm 1 employs the 2-point gradient estimator (2) in which $z$ is independently sampled from the uniform distribution $U(S_d - 1)$, and only involves consensus on the local decision variables which is similar to the decentralized (sub)gradient descent (DGD) method [1, 2].
2. Algorithm 2 employs the 2d-point gradient estimator (3) and also introduces auxiliary local variables $s^i(t)$ for gradient tracking. We shall see in Theorem 3 that $s^i(t)$ converges to the gradient of the global objective function as $t \to \infty$ under mild conditions.

Note that here we employ the the adapt-then-combine (ATC) strategy [43] for both algorithms, which is a commonly used variant for consensus optimization.
Algorithm 2: 2d-point gradient estimator with global gradient tracking

Set \( s^i(0) = g^i(0) = 0 \) for each \( i \in \mathcal{N} \).

for \( t = 1, 2, 3, \ldots \) do

foreach \( i \in \mathcal{N} \) do

1. Update \( s^i(t) \) by

\[
 g^i(t) = G_f^{(2d)}(x^i(t-1); u_t),
\]

\[
 s^i(t) = \sum_{j=1}^{n} W_{ij} \left( s^j(t-1) + g^i(t) - g^i(t-1) \right). \tag{7}
\]

2. Update \( x^i(t) \) by

\[
 x^i(t) = \sum_{j=1}^{n} W_{ij} (x^j(t-1) - \eta s^j(t)). \tag{8}
\]

end

end

3 Main Results

3.1 Convergence of Algorithm [1]

Let \( x^i(t) \) denote the sequence generated by Algorithm [1] where the sequence of step sizes \( \eta_t \) is positive and non-increasing. Denote

\[
 \bar{x}(t) := \frac{1}{n} \sum_{i=1}^{n} x^i(t), \quad R_0 := \sum_{i=1}^{n} \|x^i(0) - \bar{x}(0)\|^2.
\]

We first analyze the case with general nonconvex but smooth objectives.

**Theorem 1.** Assume that each local objective function \( f_i \) is uniformly \( G \)-Lipschitz and \( L \)-smooth for some positive constants \( G \) and \( L \), and that \( f^* := \inf_{x \in \mathbb{R}^d} f(x) > -\infty \).

1. Suppose \( \eta_1 L \leq 1/4 \), \( \sum_{t=1}^{\infty} \eta_t = +\infty \), \( \sum_{t=1}^{\infty} \eta_t^2 < +\infty \), and \( \sum_{t=1}^{\infty} \eta_t u_t^2 < +\infty \). Then almost surely, \( \|x^i(t) - \bar{x}(t)\| \) converges to zero for all \( i \in \mathcal{N} \), \( \nabla f(\bar{x}(t)) \) converges to zero, and the function value \( f(\bar{x}(t)) \) converges as \( t \to \infty \).

2. Suppose now that

\[
 \eta_t = \frac{\alpha_\eta}{4L\sqrt{d}} \cdot \frac{1}{t^3}, \quad u_t \leq \frac{\alpha_u G}{L\sqrt{d}} \cdot \frac{1}{t^{(\gamma-\beta)/2}}
\]

with \( \alpha_\eta \in (0, 1) \), \( \alpha_u \geq 0 \), \( \beta \in (1/2, 1) \) and \( \gamma > 1 \). Then

\[
 \frac{\sum_{t=0}^{\tau-1} \eta_{t+1} \mathbb{E}[\|\nabla f(\bar{x}(\tau))\|^2]}{\sum_{t=0}^{\tau-1} \eta_{t+1}} \leq \frac{(1-\beta)\sqrt{d}}{t^{1-\beta}} \left[ \frac{16(f(\bar{x}(0))-f^*)}{\alpha_\eta G^2/L} + \frac{12R_0 L^2/G^2}{n(1-\rho^2)\sqrt{d}} \right.
\]

\[
 + \frac{4\alpha_\eta}{(6\beta-3)n^2} + \frac{18\alpha_u^2 \kappa^2 \rho^2}{\sqrt{d}(1-\rho^2)^2} + \frac{9\alpha_u^2 \gamma}{2(\gamma-1)} + o \left( \frac{1}{t^{1-\beta}} \right), \tag{9}
\]

\[
 \sum_{t=0}^{\tau-1} \eta_{t+1} \mathbb{E}[\|\nabla f(\bar{x}(\tau))\|^2] \leq \frac{16(f(\bar{x}(0))-f^*)}{\alpha_\eta G^2/L} + \frac{12R_0 L^2/G^2}{n(1-\rho^2)\sqrt{d}} \right].
\]
and
\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \|x'(t) - \bar{x}(t)\|^2 \right] \leq \frac{\alpha_n^2 \rho^2}{(1 - \rho^2)^2} \frac{G^2}{L^2} t^{2\beta} + o(t^{-2\beta}),
\]
(10)

3. Suppose now that
\[
\eta_t = \frac{\alpha_n}{4L\sqrt{d}} \cdot \frac{1}{\sqrt{t}}, \quad u_t \leq \frac{\alpha_n G}{L\sqrt{d}} \cdot \frac{1}{t^{\gamma/2 - 1/4}}
\]
with \(\alpha_n \in (0,1]\), \(\alpha_n \geq 0\) and \(\gamma > 1\), and that every agent starts from the same initial point. Then almost surely, \(\|x'(t) - \bar{x}(t)\|\) converges to zero for all \(i\), and \(\lim_{t \to \infty} \|\nabla f(\bar{x}(t))\| = 0\). Furthermore, we have
\[
\frac{\sum_{t=0}^{\tau-1} \eta_{t+1} \mathbb{E} [\| \frac{1}{n} \nabla f(\bar{x}(t)) \|^2]}{\sum_{t=0}^{\tau-1} \eta_{t+1}} \leq \sqrt{\frac{d}{t}} \left[ \frac{\alpha_n}{3n^2} \log(2t+1) + \frac{8(f(\bar{x}(0)) - f^*)}{\alpha_n G^2 / L} + \frac{6R_0 L^2 / G^2}{n(1 - \rho^2)^{2\beta} \sqrt{d}} \right. \\
\left. + \frac{9\alpha_n^2 \rho^2}{(1 - \rho^2)^2 \sqrt{d}} + \frac{9\alpha_n^2 \gamma}{4(4\gamma - 1)} \right] + o\left(\frac{1}{\sqrt{t}}\right),
\]
(11)
and
\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \|x'(t) - \bar{x}(t)\|^2 \right] \leq \frac{\alpha_n^2 \rho^2}{(1 - \rho^2)^2} \frac{G^2}{L^2} t^{2\beta} + o(t^{-1}).
\]
(12)

Remark 1. Theorem \(\text{[\text{I}]}\) uses the squared norm of the gradient to assess the sub-optimality of the iterates, which is common for unconstrained nonconvex problems where we do not aim for global optimal solutions \([24,28]\). While Theorem \(\text{[\text{I}]}\) does not exclude the possibility of converging to saddle points, recent works \([45,46]\) show that saddle points can be avoided almost surely with a proper random initialization for many first-order methods, and we conjecture that the proposed algorithms may also share this property. Rigorous analysis is left for future work.

Remark 2. Each iteration of Algorithm \(\text{[\text{II}]}\) requires 2 queries of function values. Thus the convergence rates \(\text{(9)}\) and \(\text{(11)}\) can also be interpreted as \(O(N^{1-\beta})\) and \(O(N^{\gamma} / \sqrt{d})\) respectively where \(N\) denotes the number of function value queries. Characterizing convergence rate in terms of the number of function value queries \(N\) and the dimension \(d\) is conventional for zero-order optimization. In scenarios where zero-order methods are applied, the computation of the function values is usually one of the most time-consuming procedures. In addition, how the convergence of the algorithm scale with the problem dimension \(d\) is also of interest.

The following result shows that for a gradient dominated global objective, a faster convergence rate can be achieved by Algorithm \(\text{[\text{III}]}\).

Theorem 2. Assume that each local objective function \(f_i\) is uniformly \(G\)-Lipschitz and \(L\)-smooth for some positive constants \(G\) and \(L\). Furthermore, assume that the global objective function \(f\) is \(\mu\)-gradient dominated and has a minimum value denoted by \(f^*\). Suppose
\[
\eta_t = \frac{2\alpha_n}{\mu(t + t_0)}, \quad t_0 \geq \frac{8\alpha_n L}{\mu} - 1, \quad u_t \leq \frac{\alpha_n G}{L} \cdot \frac{1}{\sqrt{t + 1}}
\]
for some \(\alpha_n > 1\) and \(\alpha_0 \geq 0\). Then
\[
\mathbb{E} [f(\bar{x}(t)) - f^*] \leq \frac{\alpha_n G^2 d}{\mu(\alpha_n - 1)} \left( \frac{8\alpha_n L}{\mu} + 2\alpha_n^2 \right) \cdot \frac{1}{t + 1} + o(t^{-1}),
\]
(13)
and
\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \| x'(t) - x(t) \|^2 \right] \leq \frac{64\alpha_n^2 \kappa^2 \rho^2 G^2 d}{\mu^2 (1 - \rho^2)^2} \cdot \frac{1}{t^2} + o(t^{-2}). \tag{14}
\]

Remark 3. The convergence rate (14) can also be described as \( \mathbb{E} [f(x(t)) - f^*] = O(d/N) \), where \( N \) is the number of function value queries.

Table 1 shows that, while Algorithm 1 employs a randomized 2-point zero-order estimator of \( \nabla f_i \), its convergence rates are comparable with its gradient-based counterpart, the decentralized gradient descent (DGD) algorithm [22, 37]. However, its convergence rates are inferior to its centralized zero-order counterpart in [28].

3.2 Convergence of Algorithm 2

Let \( (x^i(t), s^i(t)) \) denote the sequence generated by Algorithm 2 with a constant step size \( \eta \). Denote
\[
\bar{x}(t) := \frac{1}{n} \sum_{i=1}^{n} x^i(t), \quad R_0 := \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\eta \rho^2}{2L} \| \nabla f_i(x^i(0)) \|^2 + \| x^i(0) - \bar{x}(0) \|^2 \right) + \frac{\eta \rho^2 u_i^2 L d}{4}.
\]

We first analyze the case where the local objectives are nonconvex and smooth.

Theorem 3. Assume that each local objective function \( f_i \) is uniformly \( L \)-smooth for some positive constant \( L \), and that \( f^* := \inf_{x \in \mathbb{R}^d} f(x) > -\infty \). Suppose
\[
\eta L \leq \min \left\{ \frac{1}{6}, \frac{(1 - \rho^2)^2}{4 \rho^2 (3 + 4 \rho^2)} \right\} \quad \text{and} \quad R_0 := d \sum_{i=1}^{\infty} u_i^2 < +\infty.
\]

Then \( f(\bar{x}(t)) \) converges,
\[
\frac{1}{t} \sum_{t=0}^{t-1} \| \nabla f(\bar{x}(\tau)) \|^2 \leq \frac{1}{t} \left[ \frac{3.2(\bar{f}(0)) - f^*)}{\eta} + \frac{12.8L^2 R_0}{1 - \rho^2} + 2.4R_u L^2 \right], \tag{15}
\]

and
\[
\frac{1}{t} \sum_{t=0}^{t-1} \frac{1}{n} \sum_{i=1}^{n} \| x^i(\tau) - \bar{x}(\tau) \|^2 \leq \frac{1}{t} \left[ 1.6\eta(\bar{f}(0)) - f^* \right] + 6.4(\eta L)^2 R_0 + 0.35R_u, \tag{16}
\]
\[
\frac{1}{t} \sum_{t=0}^{t-1} \frac{1}{n} \sum_{i=1}^{n} \| s^i(\tau) - \nabla f(\bar{x}(-1)) \|^2 \leq \frac{1}{t} \left[ 9.6L(\bar{f}(0)) - f^* \right] + 38.4\eta L^3 R_0 + \frac{3.35}{\eta} L R_u. \tag{17}
\]

Remark 4. Theorem 3 shows that Algorithm 2 achieves a convergence rate of \( O(1/t) \) in terms of the averaged squared norm of \( \nabla f(\bar{x}(t)) \), and has a consensus rate of \( O(1/t) \) of the averaged squared consensus error \( \| x^i(t) - \bar{x}(t) \|^2 \) and the squared gradient tracking error \( \| s^i(t) - \nabla f(\bar{x}(t-1)) \|^2 \). They match the rates for distributed convex optimization with gradient tracking [6]. On the other hand, since each iteration requires \( 2d \) queries of function values, we get a \( O(d/N) \) rate in terms of the number of function value queries \( N \). This matches the convergence rate of centralized zero-order algorithms without Nesterov-type acceleration [28].

\[\footnote{Existing convergence rates of gradient tracking algorithms are mainly objective error rates for convex problems. On the other hand, notice that \( L \)-smoothness of \( f \) implies \( \| \nabla f(x) \|^2 \leq 2L(f(x) - f^*) \). Therefore this statement should be interpreted as follows: Both Algorithm 2 and the gradient tracking algorithm in [6] achieve the \( O(1/t) \) ergodic rate of \( \| \nabla f \|^2 \) for smooth non-strongly convex problems.} \]
Now we proceed to the situation with a gradient dominated global objective.

**Theorem 4.** Assume that each local objective function $f_i$ is uniformly $L$-smooth for some positive constant $L$, and that the global objective function $f$ is $\mu$-gradient dominated and achieves it global minimum at $x^*$. Suppose the step size $\eta$ satisfies

$$\eta L = \alpha \cdot \left( \frac{\mu}{L} \right)^{\frac{1}{2}} \frac{(1 - \rho^2)^2}{14}$$

for some $\alpha \in (0, 1]$, and let $\lambda := 1 - \alpha \left( \frac{1}{\mu} \right)^{\frac{1}{2}}$. Then

$$f(\bar{x}(t)) - f(x^*) \leq O(\lambda^t) + 5(1 - \rho^2)\frac{L d}{t}\sum_{\tau=0}^{t-1} \lambda^\tau u_{t-\tau}^2,$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \|x_i(t) - \bar{x}(t)\|^2 \leq O(\lambda^t) + \frac{3\alpha(1 - \rho^2)}{10\sqrt{2}} \left( \frac{\mu}{L} \right)^{\frac{1}{2}} d \sum_{\tau=0}^{t-1} \lambda^\tau u_{t-\tau}^2,$$

$$\frac{1}{n}\sum_{i=1}^{n} \|s_i(t) - \nabla f(\bar{x}(t-1))\|^2 \leq O(\lambda^t) + \frac{7\sqrt{2}}{5(1 - \rho^2)} L^2 d \sum_{\tau=0}^{t-1} \lambda^\tau u_{t-\tau}^2.$$

**Remark 5.** If we use an exponentially decreasing sequence $u_t \propto \tilde{\lambda}^{t/2}$ with $\tilde{\lambda} < \lambda$, then both the objective error $f(\bar{x}(t)) - f(x^*)$ and the consensus errors $\|x_i(t) - \bar{x}(t)\|^2$ and $\|s_i(t) - \nabla f(\bar{x}(t-1))\|^2$ achieve exponential convergence rate $O(\lambda^t)$, or $O(\lambda^N/d)$ in terms of the number of function value queries. In addition, we notice that the decaying factor $\lambda$ given by Theorem 4 has a better dependence on $\mu/L$ than in [5] for convex problems. We point out that this is not a result of using zero-order techniques, but rather a more refined analysis of the gradient tracking procedure.

### 3.3 Comparison of the Two Algorithms

We see from the above results that Algorithm 2 converges faster than Algorithm 1 asymptotically as $N \to \infty$ in theory. However, each iteration of Algorithm 2 makes progress only after $2d$ queries of function values, which could be an issue if $d$ is very large. On the contrary, each iteration of Algorithm 1 only requires $2$ function value queries, meaning that progress can be made relatively immediately without exploring all the $d$ dimensions. This observation suggests that, neglecting communication delays, Algorithm 1 is more favorable for high-dimensional problems, while Algorithm 2 could handle problems of relatively low dimensions better with faster convergence. On the other hand, the number of local information exchanges per function value query for Algorithm 1 is $d/2$ times as large as that for Algorithm 2. This suggests that the rate of communication between agents can have a larger impact on the performance of Algorithm 1 than that of Algorithm 2 in practice.

### 3.4 Comparison with Existing Algorithms

In this subsection, we compare our algorithms and results with existing literature on distributed zero-order optimization, specifically [33, 34, 35, 36].
1. References [34, 35, 36] discuss convex problems, while [33] and our work focus on nonconvex problems.

2. In terms of the assumptions on the noisy function queries, [36] and our work consider a noise-free case. [33] considers stochastic queries but two function evaluations are available for each random sample. [34] considers an online setting but two online function evaluations are available for each time step. [35] assumes that independent noises are added on each query of function values. We expect that our proposed Algorithm 1 can be generalized to the setting used in [33] with heavier mathematical exposition. Extensions to general stochastic cases remain our ongoing work.

3. In terms of the approach to reach consensus among agents, our algorithms are similar to [35, 36], where some form of weighted average of the neighbors’ local variables is utilized, while [33] uses the method of multipliers and [34] uses ADMM to design their algorithms. We also mention that, our Algorithm 2 also employs the gradient tracking technique, which, to our best knowledge, has not been discussed in existing literature on distributed zero-order optimization yet.

4. Regarding the convergence rates for nonconvex optimization, [33] proves that its proposed algorithm achieves $O(1/T)$ rate if each iteration also employs $O(T)$ function value queries where $T$ is the number of iterations. Therefore in terms of the number of function value queries, its convergence rate is in fact $O(1/\sqrt{N})$, which is roughly comparable with Algorithm 1 and slower than Algorithm 2 in our paper. Further, [33] also does not discuss dependency on the problem dimension $d$. Moreover, our algorithms only require constant numbers ($2$ or $2d$) of function value queries which is more appealing for practical implementation when $T$ is set to be very large for achieving sufficiently accurate solutions.

4 Numerical Example

We consider a phase retrieval problem formulated as
\[
\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(x), \quad f_i(x) := \frac{1}{m} \sum_{k=1}^{m} (y_{ik}^2 - |a_{ik}^T x|^2)^2. \tag{21}
\]

We generate the complex vectors $a_{ik} = a_{ik}^R + i a_{ik}^I$ such that $(a_{ik}^R, a_{ik}^I) \sim \mathcal{N}(0, \frac{1}{2} I_{2d})$, and they are independent of each other. The scalars $y_{ik}$ are generated by
\[
y_{ik} = |a_{ik}^T x^{\ast}| + \varepsilon_{ik},
\]
where $x^{\ast} = (1, 0, 0, \ldots, 0)$, and $\varepsilon_{ik} \sim \mathcal{N}(0, 0.01^2)$ are independent Gaussian noise.

We set the dimension to be $d = 64$, the number of agents to be $n = 50$, and set $m = 30$. The graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is generated by uniformly randomly sampling $n$ points on $S_2$, and then connecting pairs of points with spherical distances less than $\pi/4$. The Metropolis-Hastings weights [47] are employed for constructing $W$:
\[
W_{ij} = \begin{cases} 
\frac{1}{1 + \max\{\deg(i), \deg(j)\}}, & (i, j) \in \mathcal{E}, \\
1 - \sum_{k: (i,k) \in \mathcal{E}} W_{ik}, & i = j, \\
0, & \text{otherwise},
\end{cases}
\]
where \( \text{deg}(i) \) denotes the degree of vertex \( i \). We randomly sample a number different initial points of \( x(0) \) from the distribution \( N(0, \frac{1}{I_{\text{nd}}}I_{\text{nd}}) \), and test Algorithm 1 and Algorithm 2 starting from these initial points.

Figure 1 illustrates the convergence of \( \| \nabla f(\bar{x}(t)) \|_2^2 \) for Algorithms 1 and 2 with the same initial point that has been generated randomly. The light blue curves represent the results of 10 random instances for Algorithm 1 and the dark blue curve represents their average. The horizontal axis has been normalized as the number of function value queries \( N \). It can be seen that, Algorithm 1 converges faster during the initial stage, but then slows down and converges at a relatively stable sublinear rate; Algorithm 2 converges relatively slowly initially, but its convergence rate does not change very much after \( N \gtrsim 5 \times 10^3 \), and finally achieves smaller squared gradient norm as \( N \gtrsim 2.6 \times 10^4 \). Therefore, if the total number of function value queries is limited by \( N \lesssim 2.6 \times 10^4 \), then Algorithm 1 gives better performance despite slower asymptotic convergence rate, while if more function value queries are allowed, then Algorithm 2 could be favored. We observe that this is related with the discussion in Section 3.3.

We refer to the appendix for more numerical results.

5 Conclusion

We proposed two distributed zero-order algorithms for nonconvex multi-agent optimization, established theoretical results on their convergence rates, and showed that they achieve comparable performance with their distributed gradient-based or centralized zero-order counterparts. We also provided a brief discussion on how the dimension of the problem and rate of communication will affect their performance in practice.

We point out some future directions that are worth exploring:

1. In this work, we assume that \( f_i(x) \) can be evaluated without noise or error, which can limit the applicability of the results here considering that in many practical scenarios the function values are
obtained through some noisy measurement procedure. We are interested in investigating distributed zero-order algorithms in this situation.

2. Some recent works \cite{45, 46, 48} show that modified centralized first-order methods with proper random initialization can escape saddle-point efficiently. As the algorithms proposed here are based on first-order methods, we are interested in whether these results can be extended to distributed zero-order methods.

3. It is interesting to see whether techniques for distributed optimization over time-varying directed graphs can be applied and give similar performance guarantees.

4. As discussed in Section 3.3, there is a trade-off between convergence rate and the ability to handle high-dimensional problems for the two proposed algorithms. On the contrary, the centralized zero-order algorithm in \cite{28} is able to handle high-dimensional problems without sacrificing convergence rate. We are interested in whether this gap between distributed and centralized zero-order algorithms can be mitigated.

References

[1] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48, 2009.

[2] I-An Chen. Fast distributed first-order methods. Master’s thesis, Massachusetts Institute of Technology, 2012.

[3] John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2012.

[4] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[5] Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[6] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

[7] Guannan Qu and Na Li. Accelerated distributed nesterov gradient descent, 2017. arXiv preprint arXiv:1705.07176.

[8] Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2014.

[9] Chenguang Xi, Ran Xin, and Usman A Khan. Add-opt: Accelerated distributed directed optimization. *IEEE Transactions on Automatic Control*, 63(5):1329–1339, 2018.

[10] Ran Xin and Usman A Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.
[11] Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In Proceedings of the 54th IEEE Conference on Decision and Control (CDC), pages 2055–2060, 2015.

[12] Angelia Nedić, Alex Olshevsky, Wei Shi, and César A Uribe. Geometrically convergent distributed optimization with uncoordinated step-sizes. In 2017 American Control Conference (ACC), pages 3950–3955. IEEE, 2017.

[13] S Sundhar Ram, Angelia Nedić, and Venugopal V Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. Journal of optimization theory and applications, 147(3):516–545, 2010.

[14] Shi Pu and Angelia Nedić. A distributed stochastic gradient tracking method. In Proceedings of the 57th IEEE Conference on Decision and Control (CDC), pages 963–968, 2018.

[15] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P. How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 2681–2690. PMLR, 2017.

[16] Stacy Patterson, Yonina C Eldar, and Idit Keidar. Distributed compressed sensing for static and time-varying networks. IEEE Transactions on Signal Processing, 62(19):4931–4946, 2014.

[17] Benjamin Charrow, Nathan Michael, and Vijay Kumar. Cooperative multi-robot estimation and control for radio source localization. The International Journal of Robotics Research, 33(4):569–580, 2014.

[18] Jason R. Marden, Shalom D. Ruben, and Lucy Y. Pao. A model-free approach to wind farm control using game theoretic methods. IEEE Transactions on Control Systems Technology, 21(4):1207–1214, 2013.

[19] Pascal Bianchi and Jérémie Jakubowicz. Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. IEEE transactions on automatic control, 58(2):391–405, 2012.

[20] Paolo Di Lorenzo and Gesualdo Scutari. NEXT: In-network nonconvex optimization. IEEE Transactions on Signal and Information Processing over Networks, 2(2):120–136, 2016.

[21] Tatiana Tatarenko and Behrouz Touri. Non-convex distributed optimization. IEEE Transactions on Automatic Control, 62(8):3744–3757, 2017.

[22] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, pages 5336–5346, 2017.

[23] Hoi-To Wai, Jean Lafond, Anna Scaglione, and Eric Moulines. Decentralized frank–wolfe algorithm for convex and nonconvex problems. IEEE Transactions on Automatic Control, 62(11):5522–5537, 2017.
[24] Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1529–1538. PMLR, 2017.

[25] Jinshan Zeng and Wotao Yin. On nonconvex decentralized gradient descent. IEEE Transactions on signal processing, 66(11):2834–2848, 2018.

[26] Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 385–394, 2005.

[27] Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In 29th Annual Conference on Learning Theory, volume 49 of Proceedings of Machine Learning Research, pages 257–283. PMLR, 2016.

[28] Yuri Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. Foundations of Computational Mathematics, 17(2):527–566, 2017.

[29] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341–2368, 2013.

[30] John C. Duchi, Michael I. Jordan, Martin J. Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Transactions on Information Theory, 61(5):2788–2806, 2015.

[31] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. Journal of Machine Learning Research, 18(52):1–11, 2017.

[32] Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization, 2018. arXiv preprint arXiv:1810.03233.

[33] Davood Hajinezhad, Mingyi Hong, and Alfredo Garcia. Zeroth order nonconvex multi-agent optimization over networks, 2017. arXiv preprint arXiv:1710.09997.

[34] Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, volume 84 of Proceedings of Machine Learning Research, pages 288–297. PMLR, 2018.

[35] Anit Kumar Sahu, Dusan Jakovetic, Dragana Bajovic, and Soummya Kar. Distributed zeroth order optimization over random networks: A kiefer-wolfowitz stochastic approximation approach. In 2018 IEEE Conference on Decision and Control (CDC), pages 4951–4958. IEEE, 2018.

[36] Zhan Yu, Daniel W. C. Ho, and Deming Yuan. Distributed randomized gradient-free mirror descent algorithm for constrained optimization, 2019. arXiv preprint arXiv:1903.04157.

[37] Angelia Nedić and Alex Olshevsky. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. IEEE Transactions on Automatic Control, 61(12):3936–3947, 2016.
[38] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 795–811, 2016.

[39] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. USSR Computational Mathematics and Mathematical Physics, 3(4):864–878, 1963.

[40] Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1467–1476. PMLR, 2018.

[41] Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks, 2018. arXiv preprint arXiv:1809.08587.

[42] Jack Kiefer, Jacob Wolfowitz, et al. Stochastic estimation of the maximum of a regression function. The Annals of Mathematical Statistics, 23(3):462–466, 1952.

[43] Ali H Sayed. Diffusion adaptation over networks. In Academic Press Library in Signal Processing, volume 3, pages 323–453. Elsevier, 2014.

[44] Jorge Nocedal and Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.

[45] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In Conference on Learning Theory, pages 797–842, 2015.

[46] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points, 2017. arXiv preprint arXiv:1710.07406.

[47] Lin Xiao, Stephen Boyd, and Sanjay Lall. A scheme for robust distributed sensor fusion based on average consensus. In 2005 Fourth International Symposium on Information Processing in Sensor Networks, pages 63–70, 2005.

[48] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pages 1724–1732. JMLR.org, 2017.

[49] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In Optimizing methods in statistics, pages 233–257. Elsevier, 1971.
Appendix

The set of positive integers will be denoted by \( \mathbb{Z}_+ \). We use \( I_d \) to denote the \( d \)-dimensional identity matrix, and use \( 1_n \) to denote the \( n \)-dimensional vector whose entries are all 1.

For two matrices \( A = [a_{ij}] \in \mathbb{R}^{p \times q} \) and \( B = [b_{ij}] \in \mathbb{R}^{r \times s} \), their tensor product \( A \otimes B \) is defined by

\[
A \otimes B = \begin{bmatrix}
    a_{11}B & a_{12}B & \cdots & a_{1q}B \\
    a_{21}B & a_{22}B & \cdots & a_{2q}B \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{p1}B & a_{p2}B & \cdots & a_{pq}B
\end{bmatrix} \in \mathbb{R}^{pr \times qs}.
\]

Let \( W \in \mathbb{R}^{n \times n} \) be a consensus matrix that satisfies Assumption 1 in the main text, and denote \( n \)-dimensional identity matrix, and use \( 1_n \) to denote the \( n \)-dimensional vector whose entries are all 1. We have

Lemma 1. We have \( \rho < 1 \) when \( G \) is a connected undirected graph. Moreover, for any \( x^1, \ldots, x^n \in \mathbb{R}^d \), we have

\[
\| (W \otimes I_d)(x - 1_n \otimes x) \| \leq \rho \| x - 1_n \otimes x \|
\]

where we denote.

\[
x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \bar{x} = \frac{1}{n} (1_n^T \otimes I_d)x = \frac{1}{n} \sum_{i=1}^n x_i.
\]

The following lemma provides a useful property of smooth functions.

Lemma 2. Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) is L-smooth and attains its global minimum at \( x^* \in \mathbb{R}^d \). Then

\[
\| \nabla f(x) \|^2 \leq 2L(f(x) - f(x^*)).
\]

Proof. The L-smoothness of \( f \) implies

\[
f(x^*) \leq f(x - L^{-1} \nabla f(x)) \leq f(x) - \frac{1}{2L} \| \nabla f(x) \|^2.
\]

The following lemma will be used to establish convergence of the proposed algorithms.

Lemma 3 ([19]). Let \( (\Omega, \mathcal{F}, \mathbb{P}) \) be a probability space and \( (\mathcal{F}_t)_{t \in \mathbb{Z}_+} \) be a filtration. Let \( U(t), \xi(t) \) and \( \zeta(t) \) be nonnegative \( \mathcal{F}_t \)-measurable random variables for \( t \in \mathbb{Z}_+ \) such that

\[
\mathbb{E}[U(t + 1)|\mathcal{F}_t] \leq U(t) + \xi(t) - \zeta(t), \quad \forall t = 1, 2, \ldots
\]

Then almost surely on the event \( \{ \sum_{t=1}^\infty \zeta(t) < +\infty \} \), \( U(t) \) converges to a random variable and \( \sum_{t=1}^\infty \zeta(t) < +\infty \).

As a special case, let \( U_t, \xi_t \) and \( \zeta_t \) be (deterministic) nonnegative sequences for \( t \in \mathbb{Z}_+ \) such that

\[
U_{t+1} \leq U_t + \xi_t - \zeta_t,
\]

with \( \sum_{t=1}^\infty \xi_t < +\infty \). Then \( U_t \) converges and \( \sum_{t=1}^\infty \zeta_t < +\infty \).
We will also use the following inequalities:

\[
\sum_{t=t_1}^{t_2} \frac{1}{t} \geq \int_{t_1}^{t_2+1} \frac{ds}{s^\epsilon} = \frac{(t_2 + 1)^{1-\epsilon} - t_1^{1-\epsilon}}{1-\epsilon},
\]

and

\[
\sum_{t=t_1}^{t_2} \frac{1}{t} \leq \begin{cases} 
1 + \int_{3/2}^{t_2+1/2} \frac{ds}{s^\epsilon} = 1 + \frac{(t_2 + 1/2)^{1-\epsilon} - (3/2)^{1-\epsilon}}{1-\epsilon}, & t_1 = 1, \\
\int_{t_1-1/2}^{t_2+1/2} \frac{ds}{s^\epsilon} = \frac{(t_2 + 1/2)^{1-\epsilon} - (t_1 - 1/2)^{1-\epsilon}}{1-\epsilon}, & t_1 > 1.
\end{cases}
\]

where \(\epsilon > 0\) and \(\epsilon \neq 1\), and

\[
\ln \frac{t_2 + 1}{t_1} = \int_{t_1}^{t_2+1} \frac{ds}{s} \leq \sum_{t=t_1}^{t_2} \frac{1}{t} \leq \int_{t_1-1/2}^{t_2+1/2} \frac{ds}{s} = \ln \frac{2t_2 + 1}{2t_1 - 1}.
\]

Especially, when \(\epsilon > 1\), we have

\[
\sum_{t=1}^{\infty} \frac{1}{t^\epsilon} \leq 1 + \int_{3/2}^{\infty} \frac{ds}{s^\epsilon} = 1 + \frac{1}{(\epsilon - 1)(3/2)^{\epsilon - 1}} \leq \frac{\epsilon}{\epsilon - 1}.
\]

**Analysis of Algorithm 1**

Let \((\mathcal{F}_t)_{t\in\mathbb{Z}_+}\) be a filtration to which \((z^i(t), x^i(t) : i \in \mathcal{N})_{t\in\mathbb{Z}_+}\) is adapted. We will extensively use the following properties of the distribution \(U(\mathbb{S}_{d-1})\):

\[
E_{z \sim U(\mathbb{S}_{d-1})}[d \cdot \langle g, z \rangle z] = g, \quad E_{z \sim U(\mathbb{S}_{d-1})}[d \cdot \langle g, z \rangle^2] = \|g\|^2
\]

for any (deterministic) \(g \in \mathbb{R}^d\).

**Lemma 4.** 1. Let \(u > 0\) be arbitrary, and suppose \(f : \mathbb{R}^d \to \mathbb{R}\) is differentiable. Then

\[
\nabla f^u(x) = \mathbb{E}_{z \sim U(\mathbb{S}_{d-1})}\left[G_f^{(2)}(x; u, z)\right],
\]

where \(f^u(x) := \mathbb{E}_{y \sim U(\mathbb{B}_d)}[f(x + uy)]\) is the smoothed version of \(f\). Moreover, if \(f\) is \(L\)-smooth, then \(f^u\) is also \(L\)-smooth.

2. [31] Lemma 10. Suppose \(f : \mathbb{R}^d \to \mathbb{R}\) is \(G\)-Lipschitz. Then for any \(x \in \mathbb{R}^d\) and \(u \geq 0\),

\[
E_{z \sim U(\mathbb{S}_{d-1})}\left[\left\|G_f^{(2)}(x; u, z)\right\|^2\right] \leq \kappa^2 G^2 d
\]

where \(\kappa > 0\) is some numerical constant.

3. Suppose \(f : \mathbb{R}^d \to \mathbb{R}\) is \(L\)-Lipschitz, and let \(u\) be positive. Then for any \(x \in \mathbb{R}^d\) and \(h \in \mathbb{R}^d\), we have

\[
\left| \frac{f(x + uh) - f(x - uh)}{2u} - \langle \nabla f(x), h \rangle \right| \leq \frac{1}{2} u L \|h\|^2.
\]

In addition,

\[
\|\nabla f(x) - \nabla f^u(x)\| \leq u L.
\]
Proof. 1. The equality [27] follows from [26] Lemma 1 and the fact that the distribution \( U(S_{d-1}) \) has zero mean. When \( f \) is \( L \)-smooth, we have

\[
\|\nabla f(x_1) - \nabla f(x_2)\| = \left\| \frac{1}{J_{S_d}} \int_{J_{S_d}} (\nabla f(x_1 + uy) - \nabla f(x_2 + uy)) \, dy \right\| \\
\leq \frac{1}{J_{S_d}} \int_{J_{S_d}} \|\nabla f(x_1 + uy) - \nabla f(x_2 + uy)\| \, dy \leq L \|x_1 - x_2\|
\]

for any \( x_1, x_2 \in \mathbb{R}^d \).

3. We have

\[
\left| \frac{f(x + uh) - f(x - uh)}{2u} - \langle \nabla f(x), h \rangle \right| = \left| \frac{1}{2u} \int_{-1}^{1} \langle \nabla f(x + uh), uh \rangle \, ds - \langle \nabla f(x), h \rangle \right| \\
= \frac{1}{2} \left| \int_{-1}^{1} \langle \nabla f(x + uh) - \nabla f(x), h \rangle \, ds \right| \leq \frac{1}{2} \int_{-1}^{1} Lu|s||h|^2 \, ds = \frac{1}{2} uL \|h\|^2,
\]

and

\[
\|\nabla f(x) - \nabla f^n(x)\| = \left\| \frac{1}{J_{S_d}} \int_{J_{S_d}} (\nabla f(x) - \nabla f(x + uy)) \, dy \right\| \\
\leq \frac{uL}{J_{S_d}} \int_{J_{S_d}} \|y\| \, dy \leq uL.
\]

\( \square \)

Now we introduce the following quantities:

\[
x(t) = \begin{bmatrix} x^1(t) \\ \vdots \\ x^n(t) \end{bmatrix}, \quad g(t) = \begin{bmatrix} g^1(t) \\ \vdots \\ g^n(t) \end{bmatrix}, \quad \bar{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x(i)(t), \quad \bar{g}(t) = \frac{1}{n} \sum_{i=1}^{n} g(i)(t).
\]

We can see that

\[
x(t) = (W \otimes I_d)(x(t - 1) - \eta t g(t)), \quad \bar{x}(t) = \bar{x}(t - 1) - \eta \bar{g}(t).
\]

**Lemma 5.** Suppose each \( f_i \) is \( G \)-Lipschitz and \( L \)-smooth. Then

\[
\|x(t) - 1_n \otimes \bar{x}(t)\|^2 \leq \left( \frac{1 + \rho^2}{2} \right)^t \|x(0) - 1_n \otimes \bar{x}(0)\|^2 + \frac{8n\rho^2}{1 - \rho^2} G^2 d^2 \sum_{\tau=0}^{t-1} \left( \frac{1 + \rho^2}{2} \right)^\tau \eta_{t-\tau}^2 (31)
\]

almost surely, and

\[
\mathbb{E} \left[ \|x(t) - 1_n \otimes \bar{x}(t)\|^2 \right] \leq \left( \frac{1 + \rho^2}{2} \right)^t \|x(0) - 1_n \otimes \bar{x}(0)\|^2 + \frac{8n\rho^2 \kappa^2}{1 - \rho^2} G^2 d \sum_{\tau=0}^{t-1} \left( \frac{1 + \rho^2}{2} \right)^\tau \eta_{t-\tau}^2. (32)
\]

18
Proof. We have
\[ x(t) - 1_n \otimes \bar{x}(t) = (W \otimes I_d) (x(t-1) - 1_n \otimes \bar{x}(t-1) - \eta_t g(t) - 1_n \otimes \bar{g}(t)), \]
and therefore
\[
\begin{align*}
\|x(t) - 1_n \otimes \bar{x}(t)\|^2 &\leq \rho^2 (\|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2 + \eta_t^2 \|g(t) - 1_n \otimes \bar{g}(t)\|^2) \\
&\quad + 1 - \frac{\rho^2}{2\rho^2} \cdot \rho^2 \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2 + \frac{2\rho^2}{1 - \rho^2} \cdot \eta_t^2 \rho^2 \|g(t) - 1_n \otimes \bar{g}(t)\|^2 \\
&= 1 + \frac{\rho^2}{2} \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2 + \eta_t^2 \frac{\rho^2(1 + \rho^2)}{1 - \rho^2} \|g(t) - 1_n \otimes \bar{g}(t)\|^2,
\end{align*}
\]
where we used Lemma 1. Since each \( f_i \) is \( G \)-Lipschitz, we have \( \|g^i(t)\| \leq Gd\|z\|^2 \), and so
\[
\|g(t) - 1_n \otimes \bar{g}(t)\|^2 = \sum_{i=1}^n \left( \frac{n-1}{n} g^i(t) - \frac{1}{n} \sum_{j \neq i} g^j(t) \right)^2 \\
\leq \sum_{i=1}^n \left( \frac{n-1}{n} \|g^i(t)\| + \frac{1}{n} \sum_{j \neq i} \|g^j(t)\| \right)^2 \\
\leq \sum_{i=1}^n \left( \frac{n-1}{n} Gd\|z^i(t)\|^2 + \frac{1}{n} \sum_{j \neq i} Gd\|z^j(t)\|^2 \right)^2 \\
= \frac{4(n-1)^2}{n} G^2 d^2 \leq 4nG^2 d^2,
\]
and by \[28\] of Lemma 1 we have
\[
\begin{align*}
\mathbb{E} \left[ \|g(t) - 1_n \otimes \bar{g}(t)\|^2 | F_{t-1} \right] &= \sum_{i=1}^n \mathbb{E} \left[ \left\| \frac{n-1}{n} g^i(t) - \frac{1}{n} \sum_{j \neq i} g^j(t) \right\|^2 | F_{t-1} \right] \\
&\leq 2 \sum_{i=1}^n \mathbb{E} \left[ \left\| \frac{n-1}{n} g^i(t) \right\|^2 + \left\| \frac{1}{n} \sum_{j \neq i} g^j(t) \right\|^2 | F_{t-1} \right] \\
&\leq \frac{4(n-1)^2}{n} \kappa^2 G^2 d \leq 4nk^2 G^2 d.
\end{align*}
\]
By plugging these bounds into \[33\] and noting that \( \rho < 1 \), we get \[31\] and \[32\] by mathematical induction.

Corollary 1. 1. Let \( \eta_t \) be a non-increasing sequence that converges to zero. Then
\[
\lim_{t \to \infty} \|x(t) - 1_n \otimes \bar{x}(t)\|^2 = 0.
\]
Furthermore, if \( \sum_{t=1}^\infty \eta_t^2 < +\infty \), then
\[
\sum_{t=1}^\infty \eta_t \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2 < +\infty
\]
almost surely.
2. Suppose $\eta_t = \eta_1/t^3$ for $\beta > 1/3$. Then

$$\sum_{t=1}^{\infty} \eta_t \mathbb{E} \left[ \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2 \right] \leq \frac{2\eta_1 \|x(0) - 1_n \otimes \bar{x}(0)\|^2}{1 - \rho^2} + \eta_1^3 \frac{48\beta n\kappa^2\rho^2}{(3\beta - 1)(1 - \rho^2)^2} G^2 d. \quad (34)$$

\textbf{Proof.} 1. By the monotonicity of $\eta_t$ and $((1 + \rho^2)/2)^t$, we have

$$\sum_{t=1}^{t-1} \left( \frac{1 + \rho^2}{2} \right)^\tau \eta_t^2 \eta_{t-\tau}^2 = \sum_{t=1}^t \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} \eta_t \eta_{t-\tau} \leq \sum_{t=1}^t \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} \cdot \frac{t}{t} \sum_{t=1}^t \eta_t^2 \rightarrow 0$$
as $t \to \infty$.

For the summability of $\eta_t \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2$, we have

$$\sum_{t=1}^{\infty} \eta_t \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2 \leq \|x(0) - 1_n \otimes \bar{x}(0)\|^2 \sum_{t=2}^{\infty} \eta_t \left( \frac{1 + \rho^2}{2} \right)^{t-1} + \frac{8n\rho^2G^2d^2}{1 - \rho^2} \sum_{t=2}^{\infty} \sum_{\tau=0}^{t-2} \eta_t \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} \eta_{t-\tau-1}^2$$

The first term on the right-hand side obviously converges. For the second term, we have

$$\sum_{t=2}^{\infty} \sum_{\tau=0}^{t-2} \eta_t \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} \eta_{t-\tau-1}^2 \leq \sum_{t=2}^{\infty} \sum_{\tau=0}^{t-2} \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} \eta_{t-\tau-1} \eta_{t-\tau-1} = \sum_{t=2}^{\infty} \sum_{\tau=0}^{t-2} \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} \eta_{t-\tau-1}^3$$

Therefore we can conclude that $\eta_t \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2$ is summable almost surely.

2. We have

$$\sum_{t=1}^{\infty} \eta_t \mathbb{E} \left[ \|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2 \right]$$

\leq \eta_1 \|x(0) - 1_n \otimes \bar{x}(0)\|^2 \sum_{t=1}^{\infty} \left( \frac{1 + \rho^2}{2} \right)^{t-1} + \eta_1^3 \frac{8n\rho^2\kappa^2}{1 - \rho^2} G^2 d \sum_{t=2}^{\infty} \sum_{\tau=0}^{t-2} \frac{1}{(t-1-\tau)^{3\beta}} \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} \eta_t^2$$

Then since

$$\sum_{t=2}^{\infty} \sum_{\tau=0}^{t-2} \frac{1}{(t-1-\tau)^{3\beta}} \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} = \sum_{t=2}^{\infty} \sum_{\tau=2}^{t} \frac{1}{(t-1)^{3\beta}} \left( \frac{1 + \rho^2}{2} \right)^{t-\tau} = \frac{2}{1 - \rho^2} \sum_{t=2}^{\infty} \sum_{\tau=0}^{t} \frac{1}{(t-1)^{3\beta}} \leq \frac{6\beta}{(3\beta - 1)(1 - \rho^2)},$$
we get the inequality (21).

**Lemma 6.** We have

\[
\mathbb{E} \left[ \| \bar{g}(t) \|^2 | \mathcal{F}_{t-1} \right] \leq \frac{4G^2(d - 1)}{3n^2} + 2\| \nabla f(\bar{x}(t - 1)) \|^2 + \frac{4L^2}{n} \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2 + u_t^2 L^2 d^2.
\]

**Proof.** Since

\[
\| \bar{g}(t) \|^2 = \frac{d}{n} \sum_{i=1}^{n} \left[ \langle \nabla f_i(x^i(t - 1)), z^i(t) \rangle z^i(t) + \left( \frac{f_i(x^i(t-1)+u_t z^i(t)) - f_i(x^i(t-1)-u_t z^i(t))}{2u_t} - \langle \nabla f_i(x^i(t-1)), z^i(t) \rangle \right) z^i(t) \right]^2,
\]

by (29) of Lemma 4, we see that

\[
\mathbb{E} \left[ \| \bar{g}(t) \|^2 | \mathcal{F}_{t-1} \right] \leq \mathbb{E} \left[ \left( 1 + \frac{1}{3} \right) \left( \frac{d}{n} \sum_{i=1}^{n} \langle \nabla f_i(x^i(t - 1)), z^i(t) \rangle z^i(t) \right)^2 + (1 + 3) \left( \frac{d}{n} \sum_{i=1}^{n} \frac{1}{2} u_t L \right)^2 \right] \mathcal{F}_{t-1} \]

\[
\leq \frac{4}{3} \left( \frac{d}{n^2} \sum_{i=1}^{n} \| \nabla f_i(x^i(t - 1)) \|^2 + \frac{1}{n^2} \sum_{i \neq j} \langle \nabla f_i(x^i(t - 1)), \nabla f_j(x^j(t - 1)) \rangle \right) + u_t^2 L^2 d^2,
\]

where we used (29) and the fact that \( \langle \nabla f_i(x^i(t - 1)), z^i(t) \rangle z^i(t) \) and \( \langle \nabla f_j(x^j(t - 1)), z^j(t) \rangle z^j(t) \) are independent for \( j \neq i \) conditioned on \( \mathcal{F}_{t-1} \). Then since

\[
\frac{d}{n^2} \sum_{i=1}^{n} \| \nabla f_i(x^i(t - 1)) \|^2 + \frac{1}{n^2} \sum_{i \neq j} \langle \nabla f_i(x^i(t - 1)), \nabla f_j(x^j(t - 1)) \rangle
\]

\[
= \frac{d - 1}{n^2} \sum_{i=1}^{n} \| \nabla f_i(x^i(t - 1)) \|^2 + \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^i(t - 1)) \right\|^2,
\]

and

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^i(t - 1)) \right\|^2 \leq \left( 1 + \frac{1}{2} \right) \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\bar{x}(t - 1)) \right\|^2 + (1 + 2) \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla f_i(x^i(t - 1)) - \nabla f_i(\bar{x}(t - 1))) \right\|^2
\]

\[
\leq \frac{3}{2} \| \nabla f(\bar{x}(t - 1)) \|^2 + 3 \cdot \frac{1}{n} \sum_{i=1}^{n} \| \nabla f_i(x^i(t - 1)) - \nabla f_i(\bar{x}(t - 1)) \|^2
\]

\[
\leq \frac{3}{2} \| \nabla f(\bar{x}(t - 1)) \|^2 + \frac{3L^2}{n} \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2
\]

\[\text{21}\]
we get
\[ \mathbb{E} [\|g(t)\|^2 | \mathcal{F}_{t-1}] \leq \frac{4(d-1)}{3n^2} \sum_{i=1}^{n} \| \nabla f_i(x^i(t-1)) \|^2 + \frac{4}{3} \left( \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x^i(t-1)) \right)^2 + u_t^2 L^2 d^2 \]
\[ \leq \frac{4G^2(d-1)}{3n^2} + 2\| \nabla f(\bar{x}(t-1)) \|^2 + \frac{4L^2}{n} \| x(t-1) - \mathbf{1}_n \otimes \bar{x}(t-1) \|^2 + u_t^2 L^2 d^2. \]

**Lemma 7.** Suppose \( \eta_t L \leq 1/4 \). Then
\[
\mathbb{E} [f(\bar{x}(t)) | \mathcal{F}_{t-1}] \leq f(\bar{x}(t-1)) - \frac{\eta_t}{4} \| \nabla f(\bar{x}(t-1)) \|^2 + \frac{3\eta_t L^2}{2n} \| x(t-1) - \mathbf{1}_n \otimes \bar{x}(t-1) \|^2 \\
+ \frac{2\eta_t^2 L^2 G^2 (d-1)}{3n^2} + \eta_t^2 L^2 \left( 1 + \frac{1}{2} d^2 \eta_t L \right). \tag{35}
\]

**Proof.** Since \( \bar{x}(t) = \bar{x}(t-1) - \eta_t g(t), \) by the \( L \)-smoothness of the function \( f, \) we get
\[ f(\bar{x}(t)) \leq f(\bar{x}(t-1)) - \eta_t \langle \nabla f(\bar{x}(t-1)), g(t) \rangle + \eta_t^2 \frac{L}{2} \| g(t) \|^2. \]

Note that by (27) of Lemma 4 we have
\[ \mathbb{E} [g(t) | \mathcal{F}_{t-1}] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i^u(x^i(t-1)). \]

By taking the expectation conditioned on \( \mathcal{F}_{t-1}, \) we get
\[
\mathbb{E} [f(\bar{x}(t)) | \mathcal{F}_{t-1}] \leq f(\bar{x}(t-1)) - \eta_t \| \nabla f(\bar{x}(t-1)) \|^2 + \frac{3\eta_t L^2}{2n} \| x(t-1) - \mathbf{1}_n \otimes \bar{x}(t-1) \|^2 \\
- \eta_t \left( \nabla f(\bar{x}(t-1)), \frac{1}{n} \sum_{i=1}^{n} \left( \nabla f_i^u(x^i(t-1)) - \nabla f_i^u(\bar{x}(t-1)) \right) \right) \\
- \eta_t \langle \nabla f(\bar{x}(t-1)), \nabla f^u(\bar{x}(t-1)) - \nabla f(\bar{x}(t-1)) \rangle.
\]

Since each \( f_i^u \) is \( L \)-smooth (see Part 1 of Lemma 3), we have
\[
- \left( \nabla f(\bar{x}(t-1)), \frac{1}{n} \sum_{i=1}^{n} \left( \nabla f_i^u(x^i(t-1)) - \nabla f_i^u(\bar{x}(t-1)) \right) \right) \\
\leq \frac{1}{2} \left( \frac{1}{2} \| \nabla f(\bar{x}(t-1)) \|^2 + 2 \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \nabla f_i^u(x^i(t-1)) - \nabla f_i^u(\bar{x}(t-1)) \right) \right\|^2 \right) \\
\leq \frac{1}{4} \| \nabla f(\bar{x}(t-1)) \|^2 + \left( \frac{1}{2} \sum_{i=1}^{n} \| x^i(t-1) - \bar{x}(t-1) \| \right)^2 \\
\leq \frac{1}{4} \| \nabla f(\bar{x}(t-1)) \|^2 + \frac{L^2}{n} \| x(t-1) - \mathbf{1}_n \otimes \bar{x}(t-1) \|^2,
\]

\[ 22 \]
and by (30), we have
\[
- \langle \nabla f(\bar{x}(t - 1)), \nabla f_{\eta}(\bar{x}(t - 1)) - \nabla f(\bar{x}(t - 1)) \rangle \\
\leq \frac{1}{2} \left( \frac{1}{2} \| \nabla f(\bar{x}(t - 1)) \|^2 + 2 \| \nabla f_{\eta}(\bar{x}(t - 1)) - \nabla f(\bar{x}(t - 1)) \|^2 \right) \\
\leq \frac{1}{4} \| \nabla f(\bar{x}(t - 1)) \|^2 + u^2 L^2.
\]

Therefore
\[
\mathbb{E} [f(\bar{x}(t)) | F_{t-1}] \leq f(\bar{x}(t - 1)) - \frac{\eta}{2} \| \nabla f(\bar{x}(t - 1)) \|^2 + \eta^2 L^2 \left( 1 + \frac{1}{2} d^2 \eta L \right) \\
+ \frac{\eta L^2}{n} \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2 + \eta u^2 L^2.
\]

Finally, by plugging the bound of Lemma 3, we get
\[
\mathbb{E} [f(\bar{x}(t)) | F_{t-1}] \leq f(\bar{x}(t - 1)) - \frac{\eta}{2} (1 - 2\eta L) \| \nabla f(\bar{x}(t - 1)) \|^2 + \eta^2 u^2 L^2 \left( 1 + \frac{1}{2} d^2 \eta L \right) \\
+ \frac{\eta L^2}{n} \frac{2LG^2(d - 1)}{3n^2} + \frac{\eta L^2}{n} (1 + 2\eta L) \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2,
\]

and the inequality follows from the assumption $\eta L \leq 1/4$. \hfill \Box

Now we are ready to prove Theorem 1 and Theorem 2 in the main text.

**Proof of Theorem 1.** Without loss of generality we assume that $f^* = \inf_{x \in \mathbb{R}^d} f(x) = 0$.

1. Suppose $\eta_t$ is non-increasing and $\sum_{t=1}^{\infty} \eta_t = +\infty$, $\sum_{t=1}^{\infty} \eta^2_t < +\infty$, and $\sum_{t=1}^{\infty} \eta_t u^2 < +\infty$. The convergence of $x^t(t)$ to $\bar{x}(t)$ is already shown by Corollary 4. Moreover, the random variable
\[
\frac{3\eta L^2}{2n} \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2 + \frac{2\eta^2 L G^2 (d - 1)}{3n^2} + \eta u^2 L^2 \left( 1 + \frac{1}{2} d^2 \eta L \right)
\]
is summable almost surely. By Lemma 3 we see that $f(\bar{x}(t))$ converges and
\[
\sum_{t=1}^{\infty} \eta_t \| \nabla f(t - 1) \|^2 < +\infty
\]
almost surely, which implies that $\liminf_{t \to \infty} \| \nabla f(\bar{x}(t)) \| = 0$.

Now let $\delta > 0$ be arbitrary, and consider the event
\[
A_\delta := \left\{ \limsup_{t \to \infty} \| \nabla f(\bar{x}(t)) \| \geq \delta \right\}.
\]

On the event $A_\delta$, we can always find a (random) subsequence of $\| \nabla f(\bar{x}(t)) \|$, which we denote by $(\| \nabla f(\bar{x}(t_k)) \|)_{k \in \mathbb{Z}^+}$, such that $\| \nabla f(\bar{x}(t_k)) \| \geq \frac{\delta}{2}$ for all $k$. It’s not hard to verify that
\[
M := \sup_{t \in \mathbb{Z}^+} \| \hat{g}(t) \| < +\infty.
\]
Then for any $s \in \mathbb{Z}_+$, we have

$$
\|\nabla f(\bar{x}(t_k + s))\| \geq \|\nabla f(\bar{x}(t_k))\| - \sum_{\tau = 1}^{s} \|\nabla f(\bar{x}(t_k + \tau)) - \nabla f(\bar{x}(t_k + \tau - 1))\|
$$

$$
\geq \frac{2\delta}{3} - \sum_{\tau = 1}^{\hat{s}(k)+1} L \cdot \eta_{t_k + \tau} M
$$

Let $\hat{s}(k)$ be the smallest positive integer such that

$$
\frac{2\delta}{3} - \sum_{\tau = 1}^{\hat{s}(k)+1} L \cdot \eta_{t_k + \tau} M < \frac{\delta}{3}
$$

(such $\hat{s}(k)$ exists as $\sum_{t=1}^{\infty} \eta_t = +\infty$). This implies that

$$
\sum_{\tau = 1}^{\hat{s}(k)+1} \eta_{t_k + \tau} \geq \frac{\delta}{3LM},
$$

and $\|\nabla f(\bar{x}(t_k + s))\| \geq \delta/3$ for all $s = 0, \ldots, \hat{s}(k)$. Therefore

$$
\sum_{\tau = 1}^{\hat{s}(k)+1} \eta_{t_k + \tau} \|\nabla f(\bar{x}(t_k + \tau - 1))\| \geq \frac{\delta^2}{9} \geq \frac{\delta^3}{27LM}
$$

Since $t_k \to \infty$ as $k \to \infty$, we can find a subsequence of $(t_{k_p})_{p \in \mathbb{Z}_+}$ satisfying $t_{k_{p+1}} - t_{k_p} > \hat{s}(k_p)$ by induction, and then

$$
\sum_{t=1}^{\infty} \eta_t \|\nabla f(\bar{x}(t-1))\|^2 \geq \sum_{p=1}^{\infty} \frac{\delta^3}{27LM} = +\infty.
$$

In other words, on $A_3$ the series $\sum_{t=1}^{\infty} \eta_t \|\nabla f(t-1)\|^2$ diverges. Since $\sum_{t=1}^{\infty} \eta_t \|\nabla f(t-1)\|^2 < +\infty$ converges almost surely, we have $P(A_3) = 0$, and consequently

$$
P\left(\limsup_{t \to \infty} \|\nabla f(\bar{x}(t))\| > 0\right) = P\left(\bigcup_{k \in \mathbb{Z}_+} A_{1/k}\right) = \lim_{k \to \infty} P(A_{1/k}) = 0,
$$

and we see that $\|\nabla f(\bar{x}(t))\|$ converges almost surely.

2. When $\eta_t = \eta_1/t^\beta$ and $u_t = u_1/t^{(\gamma-\beta)/2}$, by taking the telescoping sum of (34) and noting that $f$ is nonnegative, we get

$$
\sum_{\tau = 1}^{t} \eta_{\tau} E\left[\|\nabla f(\bar{x}(t-1))\|^2\right] \leq 4f(\bar{x}(0)) + \frac{6L^2}{n} \sum_{\tau = 1}^{t} \eta_{\tau} E\left[\|x(t-1) - 1_n \otimes \bar{x}(t-1)\|^2\right]
$$

$$
+ \eta_{\tau}^2 \frac{8LG^2(d-1)}{3n^2} \sum_{\tau = 1}^{t} \frac{1}{\tau^{2\beta}} + 4\eta_1 u_1^2 L^2 \sum_{\tau = 1}^{t} \left(\frac{1}{\tau^{\gamma}} + \frac{1}{2} d^2 \eta_1 L \frac{1}{\tau^{\gamma+\beta}}\right)
$$

$$
\leq 4f(\bar{x}(0)) + \eta_{\tau}^2 \frac{8LG^2(d-1)}{3n^2} \frac{2}{2\beta - 1} + 4\eta_1 u_1^2 L^2 \left(\frac{1}{\gamma} + \frac{1}{2} d^2 \eta_1 L / 2\right)\gamma - 1
$$

$$
+ 6\eta_1^2 L^2 \left(\frac{2\|x(0) - 1_n \otimes \bar{x}(0)\|^2}{1 - \rho^2} \right) + \eta_{\tau}^2 \frac{48n^2 \rho^2 \kappa^2}{(1 - \rho^2)^2 G^2 d},
$$

24
where we used (34), (25) and \( \beta \in (1/2, 1) \). Now since \( \eta_1 = \alpha_\eta/(4L\sqrt{d}) \) and \( u_1 \leq \alpha_u G/(L\sqrt{d}) \), and notice that when \( \beta < 1 \),

\[
\sum_{\tau=1}^{t} \eta_{\tau} \geq \eta_1 \sum_{\tau=1}^{t} \frac{1}{\tau} \geq \eta_1 \int_{1}^{t+1} \frac{ds}{s^\beta} = \frac{\eta_1}{1-\beta} \left( (1 + t)^{1-\beta} - 1 \right)
\]

we have

\[
\sum_{\tau=1}^{t} \eta_{\tau} \mathbb{E} \left[ \| \nabla f(\bar{x}(t-1)) \|^2 \right] \leq \frac{(1-\beta)}{(t+1)^{1-\beta} - 1} \left( \frac{16\sqrt{d}L f(\bar{x}(0))}{\alpha_\eta} + \frac{12L^2 \| x(0) - 1_n \otimes \bar{x}(0) \|^2}{n(1-\rho^2)} + \frac{\alpha_u^2 \gamma}{2(\gamma - 1)} G^2 (\sqrt{d} + 8d^{-1}) \right)
\]

\[
+ \left( \frac{4}{(6\beta - 3)n^2} + \frac{18\alpha_u \gamma \rho^2}{\sqrt{d}(1-\rho^2)^2} \right) \frac{(1-\beta)\alpha_u G^2 \sqrt{d}}{(t+1)^{1-\beta} - 1},
\]

and we get the convergence rate of \( \mathbb{E}[\| \nabla f(\bar{x}(t)) \|^2] \) stated in the theorem.

The convergence rate of the consensus error follows from Lemma 5 and the fact that

\[
\sum_{\tau=1}^{t-1} \frac{\lambda^\tau}{(t-\tau)\epsilon} = \frac{1}{(1-\lambda)t^\epsilon} + o(t^\epsilon)
\]

for any \( \lambda \in (0, 1) \) and \( \epsilon > 0 \).

3. When \( \eta_1 = \eta_1/\sqrt{t} \) and \( u_t = u_1/t^{\gamma/2-1/4} \), by (35) we have

\[
\mathbb{E} \left[ \frac{f(\bar{x}(t))}{(t+1)^{\epsilon}} \mathcal{F}_{t-1} \right] \leq \frac{1}{t^\epsilon} f(\bar{x}(t-1)) - \frac{\eta_1}{4t^{1/2+\epsilon}} \| \nabla f(\bar{x}(t-1)) \|^2
\]

\[
+ \frac{3\eta_1 L^2}{2nt^\epsilon} \| x(t-1) - 1_n \otimes \bar{x}(t-1) \|^2
\]

\[
+ \frac{2\eta_1^2 L G^2(d-1)}{3n^2 t^{1+\epsilon}} + \frac{\eta_1 u_t^2 L^2}{t^\epsilon} \left( 1 + \frac{1}{2} d^2 \eta_t L \right),
\]

where \( \epsilon > 0 \) is arbitrary. Since

\[
\frac{3\eta_1 L^2}{2nt^\epsilon} \| x(t-1) - 1_n \otimes \bar{x}(t-1) \|^2 + \frac{2\eta_1^2 L G^2(d-1)}{3n^2 t^{1+\epsilon}} + \frac{\eta_1 u_t^2 L^2}{t^\epsilon} \left( 1 + \frac{1}{2} d^2 \eta_t L \right)
\]

is summable, we see that

\[
\sum_{t=1}^{\infty} \frac{\eta_1}{t^{1/2+\epsilon}} \| \nabla f(\bar{x}(t-1)) \| < +\infty,
\]

which implies that

\[
\liminf_{t \to \infty} \| \nabla f(\bar{x}(t)) \| = 0.
\]
Now by taking the telescoping sum of \( (35) \) and using \( \eta_1 = \alpha_\eta/(4L\sqrt{d}) \) and \( u_1 \leq \alpha_\eta G/(L\sqrt{d}) \), we can show that

\[
\sum_{t=1}^{\tau} \eta_t \mathbb{E} \left[ \| \nabla f(\bar{x}(t-1)) \|^2 \right] \\
\leq 4f(\bar{x}(0)) + \eta_1 \frac{2\alpha_\eta G^2 \sqrt{d}}{3n^2} \ln(2t + 1) + \eta_1 \cdot \frac{\alpha_\eta^2 \gamma}{2(\gamma - 1)} G^2 (\sqrt{d} + 8d^{-1}) \\
+ \frac{12\eta L^2}{n} \| x(0) - 1_n \otimes \bar{x}(0) \|^2 \\
+ \eta_1 18\alpha_\eta^2 \kappa^2 \rho^2 G^2,
\]

where we used \( \sum_{t=1}^{t} \tau^{-1} \leq \ln(2t + 1) \). This further leads to

\[
\sum_{t=1}^{t} \eta_t \mathbb{E} \left[ \| \nabla f(\bar{x}(t-1)) \|^2 \right] \\
\leq \frac{1}{\sqrt{t+1}} \left( 8\sqrt{d}Lf(\bar{x}(0)) \right) + \frac{6L^2 \| x(0) - 1_n \otimes \bar{x}(0) \|^2}{n(1 - \rho^2)} \\
+ \frac{\alpha_\eta^2 \gamma}{4(\gamma - 1)} G^2 (\sqrt{d} + 8d^{-1}) \\
+ \frac{\alpha_\eta G^2 \sqrt{d} \ln(2t + 1)}{3n^2} + \frac{9\kappa^2 \rho^2}{(1 - \rho^2)^2} \frac{\alpha_\eta^2 G^2}{\sqrt{t+1} - 1}.
\]

We now get the convergence rate of \( \mathbb{E} \left[ \| \nabla f(\bar{x}(t)) \|^2 \right] \) stated in the theorem. The convergence rate of the consensus error follows from the same argument in the previous part.

\[\square\]

**Proof of Theorem 2.** Denote \( \delta(t) := f(\bar{x}(t)) - f^* \). It’s easy to check that \( \eta_\tau L \leq 1/4 \). By \( (35) \) and using the fact that \( f \) is \( \mu \)-gradient dominated, we have

\[
\mathbb{E} \left[ \delta(t) | \mathcal{F}_{t-1} \right] \leq \left( 1 - \frac{\eta_\tau \mu}{2} \right) \delta(t-1) + \frac{3\eta L^2}{2n} \| x(t-1) - 1_n \otimes \bar{x}(t-1) \|^2 \\
+ \frac{2\eta L^2 G^2 (d-1)}{3n^2} + \eta_\tau u_\tau^2 L^2 \left( 1 + \frac{1}{2} d^2 \eta_\tau L \right),
\]

(36)

By induction we see that

\[
\mathbb{E} \left[ \delta(t) \right] \leq \delta(0) \prod_{\tau=1}^{t} \left( 1 - \frac{\eta_\tau \mu}{2} \right) \\
+ \frac{3L^2}{2n} \sum_{\tau=0}^{t-1} \eta_{\tau-\tau} \mathbb{E} \left[ \| x(\tau-\tau) - 1_n \otimes \bar{x}(\tau-\tau) \|^2 \right] \prod_{s=\tau+s+1}^{t} \left( 1 - \frac{\eta_\tau \mu}{2} \right) \\
+ \sum_{\tau=0}^{t-1} \left( \frac{2\eta L^2 G^2 (d-1)}{3n^2} + \eta_{\tau-\tau} u_\tau^2 L^2 \left( 1 + \frac{1}{2} d^2 \eta_{\tau-\tau} L \right) \right) \prod_{s=\tau+s+1}^{t} \left( 1 - \frac{\eta_\tau \mu}{2} \right).
\]

26
Since
\[
\prod_{\tau=t_1}^{t_2} \left(1 - \frac{\eta \tau \mu}{2}\right) \leq \exp \left(-\sum_{\tau=t_1}^{t_2} \frac{\eta \tau \mu}{2}\right) = \exp \left(-\alpha \eta \sum_{\tau=t_1}^{t_2} \frac{1}{\tau}\right)
\]
we get
\[
E[\delta(t)] \leq \delta(0) \cdot \left(\frac{1}{t+1}\right)^{\alpha \eta}
\]
and
\[
+ \frac{3L^2}{2n} \sum_{\tau=0}^{t-1} \left[ \eta_{t-\tau} E \left[\|x(t-\tau) - 1_n \otimes \bar{x}(t-\tau)\|^2\right] \right] \left(\frac{t-\tau+1}{t+1}\right)^{\alpha \eta}
\]
and
\[
+ \sum_{\tau=0}^{t-1} \left(2\eta^2 \tau G^2 d - 1\right) + \eta_{t-\tau} g^2 \left(1 + \frac{1}{2} \eta^2 \tau L\right) \left(\frac{t-\tau+1}{t+1}\right)^{\alpha \eta}.
\]
By Lemma 5 we have
\[
\sum_{\tau=1}^{t} \eta_{t-\tau} E \left[\|x(\tau) - 1_n \otimes \bar{x}(\tau)\|^2\right] \left(\frac{t+1}{t+1}\right)^{\alpha \eta}
\]
\[
\leq \frac{2\alpha}{\mu(t+1)^{\alpha \eta}} \sum_{\tau=1}^{t} \left(\frac{\tau+1}{\tau + t_0}\right)^{\alpha \eta} \left(\frac{1 + \rho^2}{2}\right)^{\tau-1} \|x(0) - 1_n \otimes \bar{x}(0)\|^2
\]
\[
+ \frac{8\alpha^3}{\mu^3 (t+1)^{\alpha \eta}} \frac{8n \kappa^2 \rho^2 G^2 d}{1 - \rho^2} \sum_{\tau=2}^{t} \sum_{\tau=0}^{\tau-1} \left(\frac{1 + \rho^2}{2}\right)^s \frac{1}{(\tau - s - 1 + t_0)^2}
\]
and
\[
\leq \frac{2\alpha}{\mu(t+1)^{\alpha \eta}} \|x(0) - 1_n \otimes \bar{x}(0)\|^2 \sum_{\tau=1}^{t} \frac{(\tau+1)^{\alpha \eta-1}}{s} \left(\frac{1 + \rho^2}{2}\right)^{\tau-1} (s-1+t_0)^2.
\]
Since
\[
\frac{1}{(t+1)^{\alpha \eta}} \sum_{\tau=2}^{t} \sum_{s=2}^{\tau} \left(\frac{1 + \rho^2}{2}\right)^{\tau-s} \frac{1}{(s-1+t_0)^2}
\]
\[
= \frac{1}{(t+1)^{\alpha \eta}} \sum_{\tau=2}^{t} (\tau+1)^{\alpha \eta-1} \left(\frac{2}{1 - \rho^2 (\tau+1)^2} + o(\tau^{-2})\right)
\]
\[
= \left\{\begin{array}{ll}
\frac{2}{(\alpha \eta - 2)(1 - \rho^2)(t+1)^2} + o(t^{-2}), & \alpha \eta > 2, \\
\frac{2 \ln t}{(1 - \rho^2)(t+1)^2} + o\left(\frac{\ln t}{(t+1)^2}\right), & \alpha \eta = 2, \\
\frac{C_1(\alpha \eta, \mu/L, \rho)}{(t+1)^{\alpha \eta}}, & \alpha \eta \in (1, 2),
\end{array}\right.
\]
and
\[
\begin{array}{ll}
2 \ln t & \frac{\ln t}{(t+1)^2}, \\
\frac{\ln t}{(t+1)^2} & \frac{C_1(\alpha \eta, \mu/L, \rho)}{(t+1)^{\alpha \eta}}.
\end{array}
\]
where $C_1(\alpha, \mu/L, \rho)$ denotes some positive quantity that depends only on $\alpha$, $\mu/L$ and $\rho$, and

$$
\sum_{\tau=0}^{t-1} 2\eta^2_{t-\tau}L^2d(1-1) \left( \frac{t-\tau+1}{t+1} \right)^{\alpha_n} \leq \frac{8\alpha_n^2 L^2G^2(d-1)}{3\eta^2} \frac{1}{(t+1)^{\alpha_n}} \sum_{\tau=0}^{t-1} (t-\tau + 1)^{\alpha_n-2} \\
\leq \frac{8\alpha_n^2 L^2G^2(d-1)}{3(\alpha-1)\mu^2n^2} \frac{1}{(t+1)^{\alpha_n}} \sum_{\tau=0}^{t-1} (t-\tau + 1)^{\alpha_n-2} \\
\leq \frac{2\alpha_n^2 G^2}{\mu(\alpha-1)} \frac{1}{t+1},
$$

$$
\sum_{\tau=0}^{t-1} \eta^2_{t-\tau} L^2 \left( \frac{t-\tau+1}{t+1} \right)^{\alpha_n} \leq \frac{2\alpha_n^2 G^2}{\mu^2} \frac{1}{(t+1)^{\alpha_n}} \sum_{\tau=0}^{t-1} (t-\tau + 1)^{\alpha_n-3} \\
\leq \frac{2\alpha_n^2 G^2 Ld^2}{\mu^2} \cdot \left\{ \begin{array}{ll}
\frac{(\alpha-2)(t+1)^2}{\ln(t+1)} & , \quad \alpha > 2, \\
\frac{1}{(t+1)^2} & , \quad \alpha = 2, \\
\frac{1}{(2-\alpha)(t+1)^{\alpha_n}} & , \quad \alpha \in (1, 2).
\end{array} \right.
$$

We see that

$$
\mathbb{E}[\delta(t)] \leq \frac{\alpha_n G^2d}{\mu(\alpha-1)(t+1)} \left( \frac{8\alpha_n^2 L}{3\eta^2} \cdot \frac{L}{\mu} + 2\alpha_n^2 \right) + o(t^{-1}).
$$

The bound on the consensus error follows from Lemma 5.

**Analysis of Algorithm 2**

**Lemma 8.** Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be $L$-smooth. Then for any $x \in \mathbb{R}^d$,

$$
\left\| \sum_{k=1}^{d} \frac{f(x + u e_k) - f(x - u e_k)}{2u} e_k - \nabla f(x) \right\| \leq \frac{1}{2} u L \sqrt{d}.
$$

**Proof.** We have

$$
\left\| \sum_{k=1}^{d} \frac{f(x + u e_k) - f(x - u e_k)}{2u} e_k - \nabla f(x) \right\| = \left\| \sum_{k=1}^{d} \left( \frac{f(x + u e_k) - f(x - u e_k)}{2u} - \langle \nabla f(x), e_k \rangle \right) e_k \right\| \\
= \left( \sum_{k=1}^{d} \left( \frac{f(x + u e_k) - f(x - u e_k)}{2u} - \langle \nabla f(x), e_k \rangle \right)^2 \right)^{1/2} \\
\leq \left( \sum_{k=1}^{d} \left( \frac{1}{2u} \right)^2 \right)^{1/2} = \frac{1}{2} u L \sqrt{d},
$$

where we used \[\text{29}\] of Lemma 4.

28
We introduce the following quantities:

\[
x(t) = \begin{bmatrix} x^1(t) \\ \vdots \\ x^n(t) \end{bmatrix}, \quad g(t) = \begin{bmatrix} g^1(t) \\ \vdots \\ g^n(t) \end{bmatrix}, \quad s(t) = \begin{bmatrix} s^1(t) \\ \vdots \\ s^n(t) \end{bmatrix}, \quad \bar{x}(t) = \frac{1}{n} \sum_{i=1}^{n} x^i(t), \quad \bar{g}(t) = \frac{1}{n} \sum_{i=1}^{n} g^i(t).
\]

It’s not hard to see that

\[
s(t) = (W \otimes I_d)(s(t - 1) + g(t) - g(t - 1)),
\]

\[
x(t) = (W \otimes I_d)(x(t - 1) - \eta s(t)),
\]

and

\[
\frac{1}{n} \sum_{i=1}^{n} s^i(t) = \bar{g}(t), \quad \bar{x}(t) = \bar{x}(t - 1) - \eta \bar{g}(t).
\]

**Lemma 9.** Suppose \( \eta L \leq 1/6 \). Then

\[
f(\bar{x}(t)) \leq f(\bar{x}(t - 1)) - \frac{n}{3} \| \nabla f(\bar{x}(t - 1)) \|^2 + \frac{4 \eta L^2}{3n} \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2 + \frac{\eta u^2 L^2 d}{3}.
\]  

**(37)**

**Proof.** By \( \bar{x}(t) = \bar{x}(t - 1) - \eta \bar{g}(t) \) and the \( L \)-smoothness of the function \( f \), we have

\[
f(\bar{x}(t)) \leq f(\bar{x}(t - 1)) - \eta \langle \nabla f(\bar{x}(t - 1)), \bar{g}(t) \rangle + \frac{\eta^2 L^2}{2} \| \bar{g}(t) \|^2
\]

\[
= f(\bar{x}(t - 1)) - \eta \| \nabla f(\bar{x}(t - 1)) \|^2 + \frac{\eta^2 L^2}{2} \| \bar{g}(t) \|^2
\]

\[
- \eta \left( \nabla f(\bar{x}(t - 1)), \frac{1}{n} \sum_{i=1}^{n} (g^i(t) - f_i(\bar{x}(t - 1)) \right)
\]

\[
\leq f(\bar{x}(t - 1)) - \frac{\eta}{2} \| \nabla f(\bar{x}(t - 1)) \|^2 + \frac{\eta^2 L^2}{2} \| \bar{g}(t) \|^2 + \frac{\eta}{2} \left( \frac{1}{n} \sum_{i=1}^{n} (g^i(t) - \nabla f_i(\bar{x}(t - 1)) \right)^2.
\]

Then, by Lemma 8

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} (g^i(t) - \nabla f_i(\bar{x}(t - 1)) \right\|^2
\]

\[
\leq 2 \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla f_i(x^i(t - 1)) - \nabla f_i(\bar{x}(t - 1))) \right\|^2 + 2 \left( \frac{1}{n} \sum_{i=1}^{n} \| g^i(t) - \nabla f_i(x^i(t - 1)) \|^2\right)^2
\]

\[
\leq 2 \left( \frac{1}{n} \sum_{i=1}^{n} L \| x^i(t - 1) - \bar{x}(t - 1) \|^2 \right)^2 + \frac{1}{2} u^2 L^2 d
\]

\[
\leq \frac{2L^2}{n} \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2 + \frac{1}{2} u^2 L^2 d,
\]

we see that

\[
f(\bar{x}(t)) \leq f(\bar{x}(t - 1)) - \frac{\eta}{2} \| \nabla f(\bar{x}(t - 1)) \|^2 + \frac{\eta^2 L^2}{2} \| \bar{g}(t) \|^2
\]

\[
+ \frac{\eta L^2}{n} \| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2 + \frac{\eta u^2 L^2 d}{4}.
\]

29
Next, we bound the term $\|\bar{g}(t)\|^2$:

$$
\|\bar{g}(t)\|^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} g^i(t) \right\|^2 
\leq 2 \|\nabla f(\bar{x}(t-1))\|^2 + 2 \left\| \frac{1}{n} \sum_{i=1}^{n} (g^i(t) - \nabla f_i(\bar{x}(t-1))) \right\|^2
\leq 2 \|\nabla f(\bar{x}(t-1))\|^2 + \frac{4L^2}{n} \|x(t) - 1_n \otimes \bar{x}(t)\|^2 + u_1^2 L^2 d.
$$

Then we see that

$$
f(\bar{x}(t)) \leq f(\bar{x}(t-1)) - \frac{\eta}{2} (1 - 2\eta L) \|\nabla f(\bar{x}(t-1))\|^2
\quad + \frac{\eta L^2}{n} (1 + 2\eta L) \|x(t) - 1_n \otimes \bar{x}(t-1)\|^2 + \frac{\eta u_1^2 L^2 d}{4} (1 + 2\eta L).
$$

Finally, by using $\eta L \leq 1/6$, we get the desired result.

\[\square\]

**Lemma 10.** We have

$$
\|s(1) - 1_n \otimes \bar{g}(1)\|^2 \leq \rho^2 \left( \frac{4}{3} \sum_{i=1}^{n} \|\nabla f_i(x^i(0))\|^2 + nu_1^2 L^2 d \right).
$$

**Proof.** Since $s(0) = g(0) = 0$, we have

$$
\|s(1) - 1_n \otimes \bar{g}(1)\|^2 = \|(W \otimes I_d)(g(1) - 1_n \otimes \bar{g}(1))\|^2 \leq \rho^2 \|g(1) - 1_n \otimes \bar{g}(1)\|^2.
$$

Then since

$$
\|g(1) - 1_n \otimes \bar{g}(1)\|^2 = \|g(1)\|^2 + n\|\bar{g}(1)\|^2 - 2 \sum_{i=1}^{n} \left\langle g^i(1), \frac{1}{n} \sum_{j=1}^{n} g^j(1) \right\rangle
\quad = \|g(1)\|^2 - n\|\bar{g}(1)\|^2 \leq \|g(1)\|^2,
$$

and by Lemma \[\square\]

$$
\|g(1)\|^2 \leq \sum_{i=1}^{n} \left( \frac{4}{3} \|\nabla f^i(x(0))\|^2 + 4\|g^i(1) - \nabla f^i(x(0))\|^2 \right)
\quad \leq \frac{4}{3} \sum_{i=1}^{n} \|\nabla f^i(x(0))\|^2 + 4 \sum_{i=1}^{n} \left( \frac{1}{2} u_1 L \sqrt{d} \right)^2
\quad = \frac{4}{3} \sum_{i=1}^{n} \|\nabla f^i(x(0))\|^2 + nu_1^2 L^2 d,
$$

we get the desired result. \[\square\]

**Lemma 11.** If

$$
\eta L \leq \min \left\{ \frac{1}{6}, \frac{(1 - \rho^2)^2}{4\rho^2(3 + 4\rho^2)} \right\},
$$

then

$$
\max \left\{ \|x(t) - 1_n \otimes \bar{x}(t)\|^2 - \frac{3\eta}{10L} \|s(t) - 1_n \otimes \bar{g}(t)\|^2 \right\}
\leq \lambda^t n R_0 + \frac{4\eta n^2 L \rho^2(1 + 2\rho^2)}{3(1 - \rho^2)} \sum_{\tau=0}^{t-1} \lambda^{\tau} \|\nabla f(\bar{x}(\tau - t - 1))\|^2 + \frac{5n\eta Ld\rho^2(1 + 2\rho^2)}{6(1 - \rho^2)} \sum_{\tau=0}^{t-1} \lambda^{\tau} u_1^2 L^2 d - \tau
$$

(41)
where
\[ \lambda := \frac{2 + \rho^2}{3} < 1. \]

Consequently
\[
\max \left\{ \frac{t-1}{n} \sum_{\tau=0}^{t-1} \| x(\tau) - 1_n \otimes x(\tau) \|^2, \frac{3\eta}{10L} \sum_{\tau=1}^{t} \| s(\tau) - 1_n \otimes \bar{g}(\tau) \|^2 \right\} 
\leq \frac{3nR_0}{1 - \rho^2} + \frac{4n\eta^2 L^2(1 + 2\rho^2)}{(1 - \rho^2)^2} \sum_{\tau=0}^{t-2} \| \nabla f(\bar{x}(\tau)) \|^2 + \frac{5n\eta L \rho^2(1 + 2\rho^2)}{2(1 - \rho^2)^2} \sum_{\tau=1}^{t-1} u^2_{\tau}. \quad (42)
\]

**Proof.** We have
\[
s(t) - 1_n \otimes \bar{g}(t) = (W \otimes I_d)(s(t-1) - 1_n \otimes \bar{g}(t-1) + g(t) - g(t-1) - 1_n \otimes \bar{g}(t) + 1_n \otimes \bar{g}(t-1)).
\]
Then since
\[
\| g(t) - g(t-1) - 1_n \otimes \bar{g}(t) + 1_n \otimes \bar{g}(t-1) \|^2 
= \| g(t) - g(t-1) \|^2 + n\| \bar{g}(t) - \bar{g}(t-1) \|^2 - 2 \sum_{i=1}^{n} \langle g^i(t) - g^i(t-1), \bar{g}(t) - \bar{g}(t-1) \rangle 
= \| g(t) - g(t-1) \|^2 - n\| \bar{g}(t) - \bar{g}(t-1) \|^2 \leq \| g(t) - g(t-1) \|^2,
\]
we have
\[
\| s(t) - 1_n \otimes \bar{g}(t) \|^2 
\leq \rho^2 (\| s(t-1) - 1_n \otimes \bar{g}(t-1) \|^2 + \| g(t) - g(t-1) \|^2) 
\leq \frac{1 + 2\rho^2}{3} \| s(t-1) - 1_n \otimes \bar{g}(t-1) \|^2 + \frac{\rho^2(1 + 2\rho^2)}{1 - \rho^2} \sum_{i=1}^{n} \| g^i(t) - g^i(t-1) \|^2.
\]

Now since
\[
\| g^i(t) - g^i(t-1) \|^2 \leq 2 \| \nabla f_i(x^i(t-1)) - \nabla f_i(x^i(t-2)) \|^2 + 2 \left( \frac{u_i + u_{i-1}}{2} - 1 \right)^2
\]
we get
\[
\| s(t) - 1_n \otimes \bar{g}(t) \|^2 \leq \frac{1 + 2\rho^2}{3} \| s(t-1) - 1_n \otimes \bar{g}(t-1) \|^2 
+ \frac{2\rho^2(1 + 2\rho^2)L^2}{1 - \rho^2} \| x(t-1) - x(t-2) \|^2 + \frac{2\rho^2(1 + 2\rho^2)L^2}{1 - \rho^2} \| x(t-2) \|^2
\]
Then
\[
x(t-1) - x(t-2) = ((W \otimes I_d) - I)x(t-2) - \eta(W \otimes I_d)s(t-1)
= ((W \otimes I_d) - I)(x(t-2) - 1_n \otimes \bar{x}(t-2)) - \eta(W \otimes I_d)(s(t-1) - 1_n \otimes \bar{g}(t-1)) 
- \eta(1_n \otimes (\bar{g}(t-1) - \nabla f(\bar{x}(t-2))) - \eta(1_n \otimes \nabla f(\bar{x}(t-2))).
\]
We notice that for any \( u_1, \ldots, u_n \in \mathbb{R}^d \) and \( v \in \mathbb{R}^d \), we have
\[
\sum_{i=1}^{n} \left( u_i - \frac{1}{n} \sum_{j=1}^{n} u_j, v \right) = 0, \quad \text{and} \quad \sum_{i=1}^{n} \left( \sum_{j=1}^{n} W_{ij} u_j - \frac{1}{n} \sum_{j=1}^{n} u_j, v \right) = 0. \tag{43}
\]
In addition, similar to (38), we have
\[
\| \bar{g}(t - 1) - \nabla f(\bar{x}(t - 2)) \|^2 \leq \frac{18L^2}{17n} \| x(t - 2) - 1_n \otimes \bar{x}(t - 2) \|^2 + \frac{9}{2} u_{t-1}^{2} L^2 d.
\]
Therefore we get
\[
\| x(t - 1) - x(t - 2) \|^2 \\
= \| ((W \otimes I_d) - 1)(x(t - 2) - 1_n \otimes \bar{x}(t - 2)) - \eta \| W \otimes I_d \| s(t - 1) - 1_n \otimes \bar{g}(t - 1) \| \|^2 \\
+ \eta^2 n \| \bar{g}(t - 1) - \nabla f(\bar{x}(t - 2)) + \nabla f(\bar{x}(t - 2)) \|^2 \\
\leq \frac{9}{2} \| x(t - 2) - 1_n \otimes \bar{x}(t - 2) \|^2 + 9\eta^2 \rho^2 \| s(t - 1) - 1_n \otimes \bar{g}(t - 1) \|^2 \\
+ 2\eta^2 n \| \bar{g}(t - 1) - \nabla f(\bar{x}(t - 2)) \|^2 + \eta^2 n \| \nabla f(\bar{x}(t - 2)) \|^2 \\
\leq \left( \frac{9}{2} + \frac{36}{17} \eta^2 L^2 \right) \| x(t - 2) - 1_n \otimes \bar{x}(t - 2) \|^2 \\
+ 9\eta^2 \rho^2 \| s(t - 1) - 1_n \otimes \bar{g}(t - 1) \|^2 + 2\eta^2 n \| \nabla f(\bar{x}(t - 2)) \|^2 + 9\eta^2 n u_{t-1}^{2} L^2 d \\
\leq \frac{155}{34} \| x(t - 2) - 1_n \otimes \bar{x}(t - 2) \|^2 + 9\eta^2 \rho^2 \| s(t - 1) - 1_n \otimes \bar{g}(t - 1) \|^2 \\
+ 2\eta^2 n \| \nabla f(\bar{x}(t - 2)) \|^2 + \frac{1}{4} n u_{t-1}^{2} d,
\]
where the first inequality follows from \( \| W \otimes I_d - I \| \leq 2 \) and that \( \| u + v \|^2 \leq (1 + \epsilon) \| u \|^2 + (1 + \epsilon) \| v \|^2 \) for any vectors \( u, v \) and \( \epsilon > 0 \), and the third inequality follows from \( \eta L \leq 1/6 \). Consequently
\[
\| s(t) - 1_n \otimes \bar{g}(t) \|^2 \leq \left( \frac{1 + 2\rho^2}{3} + \frac{18\rho^4(1 + 2\rho^2)}{1 - \rho^2} \eta^2 L^2 \right) \| s(t - 1) - 1_n \otimes \bar{g}(t - 1) \|^2 \\
+ \frac{228\rho^2(1 + 2\rho^2)}{25(1 - \rho^2)} L^2 \| x(t - 2) - 1_n \otimes \bar{x}(t - 2) \|^2 \\
+ \frac{2\rho^2(1 + 2\rho^2)}{1 - \rho^2} \left( 2\eta^2 L^2 n \| \nabla f(\bar{x}(t - 2)) \|^2 + \frac{5}{4} n L^2 u_{t-1}^{2} d \right),
\]
where we used \( 155/34 < 114/25 \). On the other hand,
\[
\| x(t - 1) - 1_n \otimes \bar{x}(t - 1) \|^2 \\
= \| (W \otimes I_d)[x(t - 2) - 1_n \otimes \bar{x}(t - 2) - \eta(s(t - 1) - 1_n \otimes \bar{g}(t - 1)))] \| \|^2 \\
\leq \frac{1 + 2\rho^2}{3} \| x(t - 2) - 1_n \otimes \bar{x}(t - 2) \|^2 + \frac{\rho^2(1 + 2\rho^2)}{1 - \rho^2} \eta^2 \| s(t - 1) - 1_n \otimes \bar{g}(t - 1) \|^2.
\]
32
Therefore
\[
\begin{align*}
\frac{5n}{2\sqrt{57L}}\|s(t) - 1_n \otimes \bar{g}(t)\|^2 \\
\|x(t - 1) - 1_n \otimes x(t - 1)\|^2
\end{align*}
\leq A \left[ \frac{5n}{2\sqrt{57L}}\|s(t - 1) - 1_n \otimes \bar{g}(t - 1)\|^2 \right] + \frac{5n\eta^2 L\rho^2(1 + 2\rho^2)}{\sqrt{57(1 - \rho^2)}} \left[ 2\|\nabla f(\bar{x}(t - 2))\|^2 + \frac{5n^2\eta^2 u_{t-1}^2}{\eta L} \right],
\]
where
\[
A = \left[ \frac{1 + 2\rho^2 + \sqrt{1 + 2\rho^2}}{2\sqrt{57(1 - \rho^2)}} \eta^2 L^2 \right].
\]

This leads to
\[
\begin{align*}
\frac{5n}{2\sqrt{57L}}\|s(t) - 1_n \otimes \bar{g}(t)\|^2 \\
\|x(t) - 1_n \otimes x(t)\|^2
\end{align*}
\leq A' \left[ \frac{5n}{2\sqrt{57L}}\|s(0) - 1_n \otimes \bar{g}(0)\|^2 \right] + \frac{5n\eta^2 L\rho^2(1 + 2\rho^2)}{\sqrt{57(1 - \rho^2)}} \sum_{\tau = 0}^{t-1} A' \left[ 2\|\nabla f(\bar{x}(t - \tau - 1))\|^2 + \frac{5n^2\eta^2 u_{t-\tau}^2}{\eta L} \right],
\]
and consequently
\[
\max \left\{ \frac{\|x(t) - 1_n \otimes x(t)\|^2}{\frac{5n}{2\sqrt{57L}}\|s(t + 1) - 1_n \otimes \bar{g}(t + 1)\|^2} \right\}
\leq \|A\| \left( \frac{5n}{2\sqrt{57L}}\|s(0) - 1_n \otimes \bar{g}(0)\|^2 + \|x(0) - 1_n \otimes \bar{x}(0)\|^2 \right)
+ \frac{10n\eta^2 L\rho^2(1 + 2\rho^2)}{\sqrt{57(1 - \rho^2)}} \sum_{\tau = 0}^{t-1} \|A\|\|\nabla f(\bar{x}(t - \tau - 1))\|^2 + \frac{25n\eta L\rho^4(1 + 2\rho^2)}{4\sqrt{57(1 - \rho^2)}} \sum_{\tau = 0}^{t-1} \|A\|\|u_{t-\tau}\|^2.
\]
Now, since \(A\) is symmetric, straightforward calculation shows that
\[
\|A\| = \frac{1 + 2\rho^2}{3(1 - \rho^2)} \left( 1 - \rho^2 + 2\rho^4(\eta L)^2 + \frac{3\sqrt{3}}{5} \sqrt{76\rho^2(\eta L)^2 + 675\rho^6(\eta L)^4} \right).
\]

By solving the inequality \(\|A\| \leq (2 + \rho^2)/3\), we get
\[
(\eta L)^2 \leq \frac{25(1 - \rho^2)^4}{\rho^4(3402 + 8208\rho^2 + 4158\rho^4 + 2700\rho^6)}.
\]

It can be checked that
\[
\frac{1}{25} (3402 + 8208\rho^2 + 4158\rho^4 + 2700\rho^6) \leq \left[ 4(3 + 4\rho^2) \right]^2, \quad \forall \rho \in [0, 1).
\]

Therefore if \(\eta L\) satisfies \(\eta L \leq 10(3 + 4\rho^2)\), we have \(\|A\| \leq (2 + \rho^2)/3\). By Lemma \(\#10\) and that \(5/(2\sqrt{57}) < 1/3\), we get \(\#11\). The bound \(\#12\) follows by taking the sum of \(\#11\) and using
\[
\sum_{\tau = 1}^{t-1} \sum_{s=0}^{t-1} \lambda^\tau a_{t-\tau} = \sum_{\tau = 1}^{t-1} \sum_{s=0}^{t-1} \lambda^\tau s a_s = \sum_{s=1}^{t-1} a_s \sum_{\tau = s}^{t-1} \lambda^\tau s \leq \frac{1}{1 - \lambda} \sum_{s=1}^{t-1} a_s
\]
for any nonnegative sequence \((a_s)_{s \in \mathbb{Z}_+}\).
Now we are ready to prove Theorems 3 and 4 in the main text.

**Proof of Theorem 3.** Let \( t \in \mathbb{Z}_+ \) be arbitrary. By Lemma 9 and 12, we see that

\[
0 \leq f(\bar{x}(0)) - f^* - \frac{\eta}{3} \sum_{\tau=0}^{t-1} \|\nabla f(\bar{x}(\tau))\|^2 + \frac{\eta L^2 d}{3} \sum_{\tau=1}^{t} u^2_{\tau} + \frac{4\eta L^2}{3n} \left( \frac{3n R_0}{1 - \rho^2} + \frac{4\eta^3 L \rho^2 (1 + 2\rho^2)}{(1 - \rho^2)^2} \sum_{\tau=0}^{t-2} \|\nabla f(\bar{x}(\tau))\|^2 + \frac{5\eta L d \rho^2 (1 + 2\rho^2)}{2(1 - \rho^2)^2} \sum_{\tau=1}^{t-1} u^2_{\tau} \right)
\]

\[
\leq f(\bar{x}(0)) - f^* + \frac{4\eta L^2 R_0}{1 - \rho^2} \cdot \frac{\eta}{3} \left( 1 - \frac{16\eta^3 L^3 \rho^2 (1 + 2\rho^2)}{(1 - \rho^2)^2} \right) \sum_{\tau=0}^{t-1} \|\nabla f(\bar{x}(\tau))\|^2 + \left( \frac{10\eta L \rho^2 (1 + 2\rho^2)}{3(1 - \rho^2)^2} + \frac{1}{3} \right) \eta L^2 d \sum_{\tau=1}^{t} u^2_{\tau}.
\]

Then since

\[
\frac{16\eta^3 L^3 \rho^2 (1 + 2\rho^2)}{(1 - \rho^2)^2} \leq \frac{16}{36} \cdot \frac{(1 - \rho^2)^2}{4\rho^2(3 + 4\rho^2)} \rho^2 (1 + 2\rho^2) = \frac{4}{9} \frac{(1 + 2\rho^2)}{4\rho^2(3 + 4\rho^2)} = \frac{1}{18},
\]

and \( \frac{1}{3}(1 - 1/18) = \frac{17}{54} \geq \frac{5}{16} \), and

\[
\frac{10\eta L \rho^2 (1 + 2\rho^2)}{3(1 - \rho^2)^2} \leq \frac{10}{3} \cdot \frac{(1 - \rho^2)^2}{4\rho^2(3 + 4\rho^2)} \rho^2 (1 + 2\rho^2) \leq \frac{5}{12},
\]

we get

\[
0 \leq f(\bar{x}(0)) + \frac{4\eta L^2 R_0}{1 - \rho^2} - \frac{5\eta}{16} \sum_{\tau=0}^{t-1} \|\nabla f(\bar{x}(\tau))\|^2 + \frac{3}{4} \eta L^2 d \sum_{\tau=1}^{t} u^2_{\tau}.
\]

Since \( u^2_{\tau} \) is summable, this implies that \( \|\nabla f(\bar{x}(t))\| \) converges to zero, and we have

\[
\frac{1}{t} \sum_{\tau=0}^{t-1} \|\nabla f(\bar{x}(\tau))\|^2 \leq \frac{1}{t} \left[ \frac{3.2(f(\bar{x}(0)) - f^*)}{\eta} + \frac{12.8L^2 R_0}{1 - \rho^2} + 2.4L^2 d \sum_{\tau=1}^{t} u^2_{\tau} \right] \quad \text{(15)}
\]

Now by (12) and (13), we see that \( \|x(t) - 1_n \otimes \bar{x}(t)\|^2 \) is summable, and

\[
\frac{1}{n} \sum_{\tau=0}^{\infty} \|x(\tau) - 1_n \otimes \bar{x}(\tau)\|^2
\]

\[
\leq \frac{3R_0}{1 - \rho^2} + \frac{4\eta^3 L \rho^2 (1 + 2\rho^2)}{(1 - \rho^2)^2} \left[ \frac{3.2(f(\bar{x}(0)) - f^*)}{\eta} + \frac{12.8L^2 R_0}{1 - \rho^2} + 2.4L^2 d \sum_{\tau=1}^{\infty} u^2_{\tau} \right]
\]

\[
+ \frac{5\eta L \rho^2 (1 + 2\rho^2)}{2(1 - \rho^2)^2} \sum_{\tau=1}^{t-1} u^2_{\tau}
\]

\[
\leq 1.6\eta(f(\bar{x}(0)) - f^*) + 6.4(\eta L)^2 R_0 + 0.35d \sum_{\tau=1}^{\infty} u^2_{\tau}.
\]
For the convergence of \( s(t) \), we have
\[
\frac{1}{n} \sum_{\tau=1}^{\infty} \| s(\tau) - 1_n \otimes \nabla f(\bar{x}(\tau - 1)) \|^2 \\
\leq \frac{3}{2n} \sum_{\tau=1}^{\infty} \| s(\tau) - 1_n \otimes \bar{g}(\tau) \|^2 + 3 \sum_{\tau=0}^{\infty} \| \bar{g}(\tau) - \nabla f(\bar{x}(\tau - 1)) \|^2 \\
\leq \frac{3}{2n} \sum_{\tau=1}^{\infty} \| s(\tau) - 1_n \otimes \bar{g}(\tau) \|^2 + 3 \sum_{\tau=1}^{\infty} \left( \frac{2L^2}{n} \| x(\tau - 1) - 1_n \otimes \bar{x}(\tau - 1) \|^2 + \frac{1}{2} u_\tau^2L^2 d \right) \\
\leq \left( \frac{5L}{\eta} + 6L^2 \right) \left( 1.6\eta (f(\bar{x}(0)) - f^*) + 6.4(\eta L)^2 R_0 + 0.35d \sum_{\tau=1}^{\infty} u_\tau^2 \right) + \frac{3}{2} \sum_{\tau=1}^{\infty} u_\tau^2 L^2 d \\
\leq 9.6L(f(\bar{x}(0)) - f^*) + 38.4\eta L^2 R_0 + \frac{2.35}{\eta} L d \sum_{\tau=1}^{\infty} u_\tau^2,
\]
where we used (38), (34) and \( \eta L \leq 1/6 \). Finally, since \( u_\tau^2 \) is also summable, by Lemma 3 and the deterministic version of Lemma 3 we see that \( f(\bar{x}(t)) \) converges. \( \square \)

**Proof of Theorem 4.** Denote \( \kappa := \mu/L \) and
\[
\delta_t := f(\bar{x}(t)) - f(x^*).
\]
By Lemma 2 we see that \( \mu \leq L \). Notice that the condition on the step size
\[
\eta L = \alpha \cdot \left( \frac{\mu}{L} \right)^\beta \left( 1 - \rho^2 \right)^2
\]
implies \( \eta L \leq 1/6 \). By (14) and Lemma 2 we get
\[
\left[ \frac{25n}{\eta L} \left\| s(t) - 1_n \otimes \bar{g}(t) \right\|^2 \right] \leq A \left[ \frac{25n}{\eta L} \left\| s(t - 1) - 1_n \otimes \bar{g}(t - 1) \right\|^2 \right] \\
+ \frac{2n\eta L \rho^2 (1 + \rho^2)}{3(1 - \rho^2)} \left[ 4\eta^2 L^2 \delta_{t-2} + \frac{5}{4} u_{t-1}^2 \right],
\]
where \( A \) is given by (15) and the norm of \( A \) is given by (40). By solving the inequality \( \| A \| \leq 1 - (1 - \rho^2)^2/21 \), we get
\[
(\eta L)^2 \leq \frac{25(1 - \rho^2)^4 (13 + \rho^2)^2}{\rho^4(223398 + 411642\rho^2 + 33642\rho^4 + 217350\rho^6 + 18900\rho^8)}.
\]
It can be verified that
\[
\frac{25(13 + \rho^2)^2}{\rho^4(223398 + 411642\rho^2 + 33642\rho^4 + 217350\rho^6 + 18900\rho^8)} \geq \frac{1}{14^2}
\]
for all \( \rho \in [0, 1) \). By the condition (47) on \( \eta L \), we see that \( \| A \| \leq 1 - (1 - \rho^2)^2/21 \). Then, since
\[
\frac{8\eta^3 L^2 \rho^2 (1 + \rho^2)}{3(1 - \rho^2)} = \frac{8n\eta^2 (1 + \rho^2)}{3} \cdot \frac{\alpha^2 \kappa^2/3 (1 - \rho^2)^3}{196} \\
\leq \frac{2n\alpha^2 \kappa^{2/3} \eta}{147} \max_{\rho \in [0,1]} \frac{\rho^2 (1 + \rho^2)(1 - \rho^2)^3}{6} (1 - \chi),
\]
35
where we denote
\[
\chi := 1 - \frac{4}{49} \max_{\rho \in [0, 1]} \rho^2 (1 + 2\rho^2)(1 - \rho^2)^3 \approx 0.9865,
\]
we get
\[
\left\| \frac{5n}{2\sqrt{\alpha}L} \| s(t) - 1_n \otimes \bar{y}(t) \| ^2 \right\| \leq \left( 1 - \frac{1 - \rho^2)^2}{21} \right) \left\| \frac{5n}{2\sqrt{\alpha}L} \| s(t - 1) - 1_n \otimes \bar{y}(t - 1) \| ^2 \right\| + \frac{n\alpha^2 \kappa^{2/3} \eta}{6} \cdot (1 - \chi) \delta_{t-2} + \frac{5n\eta L^2 (1 + 2\rho^2)(1 - \rho^2)}{6(1 - \rho^2)} u_{t-1}^2 d,
\]
where the condition (47) was used. Consequently, if we denote
\[
\sigma_{t-1} := \frac{2\sqrt{2}L}{n\alpha\kappa^{1/3} \sqrt{1 - \chi} \left\| \frac{5n}{2\sqrt{\alpha}L} \| s(t) - 1_n \otimes \bar{y}(t) \| ^2 \right\| ^2},
\]
we get
\[
\sigma_{t-1} \leq \left( 1 - \frac{(1 - \rho^2)^2}{21} \right) \sigma_{t-2} + \sqrt{2\alpha\kappa^{1/3} \sqrt{1 - \chi} \eta L \cdot \delta_{t-2} + \frac{5\sqrt{2} \rho^2 (1 + 2\rho^2)(1 - \rho^2)}{42 \sqrt{1 - \chi} \sigma_{t-2}} u_{t-1}^2 L d.
\]
On the other hand, by Lemma 9 and the fact that \( f \) is \( \mu \)-gradient dominated, we have
\[
\delta_{t-1} \leq \left( 1 - \frac{2\eta \mu}{3} \right) \delta_{t-2} + \frac{4\eta L^2}{3n} \| x(t - 2) - 1_n \otimes \bar{x}(t - 2) \|^2 + \frac{\eta L^2 u_{t-1}^2 d}{3}
\]
\[
\leq \left( 1 - \frac{2\eta \mu}{3} \right) \delta_{t-2} + \sqrt{2\alpha \kappa^{1/3} \sqrt{1 - \chi} \eta L \cdot \sigma_{t-2} + \frac{\eta L^2 u_{t-1}^2 d}{3}}.
\]
Therefore
\[
\left[ \begin{array}{c}
\sigma_{t-1} \\
\delta_{t-1}
\end{array} \right] \leq B \left[ \begin{array}{c}
\sigma_{t-2} \\
\delta_{t-2}
\end{array} \right] + \left[ \begin{array}{c}
\frac{5\sqrt{2} \rho^2 (1 + 2\rho^2)(1 - \rho^2)}{42 \sqrt{1 - \chi} \sigma_{t-2}}
\frac{\eta L^2 u_{t-1}^2 d}{3}
\end{array} \right],
\]
where
\[
B := \left[ \begin{array}{cc}
1 - \frac{2\eta \mu}{3}(1 - \rho^2)^2 & \frac{1}{3} \sqrt{2(1 - \chi)\alpha^{1/3} \eta L} \\
\frac{1}{3} \sqrt{2(1 - \chi) \alpha^{1/3} \eta L} & \frac{1}{2} \frac{\eta L^2 u_{t-1}^2 d}{3}
\end{array} \right].
\]
Straightforward calculation shows that
\[
\| B \| = 1 - \frac{(1 - \rho^2)^2}{42} \left( 1 + \alpha \kappa^{4/3} - \sqrt{(1 - \alpha \kappa^{4/3})^2 + 2(1 - \chi) \alpha^4 \kappa^{4/3}} \right)
\leq 1 - \frac{(1 - \rho^2)^2}{42} \left( 1 + \alpha \kappa^{4/3} - \sqrt{(1 - \alpha \kappa^{4/3})^2 + 2(1 - \chi) \alpha^4 \kappa^{4/3}} \right)
\leq 1 - \frac{(1 - \rho^2)^2}{42} \left( 1 + \alpha \kappa^{4/3} - \sqrt{(1 - \chi) \alpha \kappa^{4/3})^2 + (1 - \chi^2) \alpha^2 \kappa^{8/3}} \right)
\]
Since \( x \mapsto \sqrt{(1 - \chi x)^2 + (1 - \chi^2)x^2} \) is a convex function over \( x \in [0, 1] \), it can be shown that
\[
\sqrt{(1 - \chi x)^2 + (1 - \chi^2)x^2} \leq 1 + (\sqrt{2(1 - \chi)} - 1)x,
\]
and so
\[ \|B\| \leq 1 - \frac{(1 - \rho^2)^2}{42} \left( 2 - \sqrt{2(1 - \chi)} \right) \alpha \kappa^{4/3} \leq 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3}, \]
where we used the fact that \( 2 - \sqrt{2(1 - \chi)} > \frac{42}{25} \). By (48), we then have
\[
\left\| \begin{bmatrix} \sigma_{t-1} \\ \delta_{t-1} \end{bmatrix} \right\| \leq \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right) \left\| \begin{bmatrix} \sigma_{t-2} \\ \delta_{t-2} \end{bmatrix} \right\| + \frac{\sqrt{2} \sqrt{3} \left( 1 + 2 \rho^2 \right)(1 - \rho^2)}{\eta L} \frac{u_{t-1}^2 Ld}{3}
\leq \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right) \left\| \begin{bmatrix} \sigma_{t-2} \\ \delta_{t-2} \end{bmatrix} \right\| + 5(1 - \rho^2)u_{t-1}^2 Ld,
\]
where we used \( \sqrt{1 - \chi} > 1/9 \) and
\[
\left\| \begin{bmatrix} \sigma_{t-1} \\ \delta_{t-1} \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \sigma_{t-2} \\ \delta_{t-2} \end{bmatrix} \right\| + \frac{\sqrt{3} \sqrt{2} \sqrt{3} \left( 1 + 2 \rho^2 \right)(1 - \rho^2)}{\eta L} \frac{u_{t-1}^2 Ld}{3}
\leq \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right) \left\| \begin{bmatrix} \sigma_{t-2} \\ \delta_{t-2} \end{bmatrix} \right\| + \frac{\sqrt{3} \sqrt{2} \sqrt{3} \left( 1 + 2 \rho^2 \right)(1 - \rho^2)}{\eta L} \frac{u_{t-1}^2 Ld}{3}
\leq \frac{\sqrt{3} \sqrt{2} \sqrt{3} \left( 1 + 2 \rho^2 \right)(1 - \rho^2)}{\eta L} \frac{u_{t-1}^2 Ld}{3} \leq \frac{50(1 + 2 \rho^2)}{16(1 - \rho^2)} + \frac{1}{14^2} < 15(1 - \rho^2).
\]
By induction we get
\[
\left\| \begin{bmatrix} \sigma_t \\ \delta_t \end{bmatrix} \right\| \leq \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right)^t \left\| \begin{bmatrix} \sigma_0 \\ \delta_0 \end{bmatrix} \right\| + 5(1 - \rho^2) Ld \sum_{\tau=0}^{t-1} \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right)^\tau u_{t-\tau}^2,
\]
which implies the bound on \( f(\bar{x}(t)) - f(x^*) \). The bound on \( \frac{1}{n} \sum_{i=1}^{n} \| x(t) - \bar{x}(t) \|^2 \) follows from
\[
\frac{n \alpha \kappa^{1/3} \sqrt{1 - \chi}}{2 \sqrt{2L}} \cdot 5(1 - \rho^2) Ld < \frac{3n \alpha \kappa^{1/3}}{10 \sqrt{2}} (1 - \rho^2)d
\]
as \( \sqrt{1 - \chi} < 3/25 \). The bound on \( \frac{1}{n} \sum_{i=1}^{n} \| s(t) - \nabla f(\bar{x}(t)) \|^2 \) follows from
\[
\frac{1}{n} \| s(t+1) - \mathbf{1}_n \otimes \nabla f(\bar{x}(t)) \|^2 \\
\leq \frac{3}{2n} \| s(t+1) - \mathbf{1}_n \otimes \bar{g}(t+1) \|^2 + 3 \| \bar{g}(t+1) - \nabla f(\bar{x}(t)) \|^2 \\
\leq \frac{3}{2n} \| s(t+1) - \mathbf{1}_n \otimes \bar{g}(t+1) \|^2 + 3 \left( \frac{2L^2}{n} \| x(t) - \mathbf{1}_n \otimes \bar{x}(t) \| \right)^2 + \frac{1}{2} u_{t+1}^2 L^2 d \\
\leq \left( \frac{3}{2n} \frac{10L}{3\eta} + \frac{6L^2}{n} \right) \cdot \frac{n \alpha \kappa^{1/3} \sqrt{1 - \chi}}{2 \sqrt{2L}} \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right)^t \left\| \begin{bmatrix} \sigma_0 \\ \delta_0 \end{bmatrix} \right\| \\
\quad + \frac{3}{2} u_{t+1}^2 L^2 d + \left( \frac{2L^2}{n} \frac{10L}{3\eta} + \frac{6L^2}{n} \right) \cdot \frac{3n \alpha \kappa^{1/3}}{10 \sqrt{2}} (1 - \rho^2)d \sum_{\tau=0}^{t-1} \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right)^\tau u_{t-\tau}^2 \\
\leq \frac{18L}{5(1 - \rho^2)^2} \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right)^t \left\| \begin{bmatrix} \sigma_0 \\ \delta_0 \end{bmatrix} \right\| + \frac{7 \sqrt{2L^2 d}}{5(1 - \rho^2)^2} \sum_{\tau=0}^{t-1} \left( 1 - \frac{(1 - \rho^2)^2}{25} \alpha \kappa^{4/3} \right)^\tau u_{t+1-\tau}^2.
\]
Details on the Numerical Example

In this part we present more details on the numerical example.

We have tested Algorithm 1, Algorithm 2, decentralized gradient descent (DGD), and distributed gradient descent with gradient tracking [6] on the phase retrieval problem (21), starting from initial points randomly generated from the distribution $\mathcal{N}(0, \frac{1}{d} I_{nd})$. We only present numerical results on 4 cases of these initial points here for legibility. Since Algorithm 1 utilizes random vectors in its iterations, we run Algorithm 1 for 10 times to get 10 random instances for each initial point. In the following illustrations, we’ll use light blue curves to represent the individual results of the 10 instances, and dark blue curves to represent their average for Algorithm 1. The parameters of the algorithms are set as follows:

1. Algorithm 1: $\eta_t = 0.05/\sqrt{t + 24}$, $u_t = 0.2/\sqrt{t + 24}$.
2. Algorithm 2: $\eta = 0.03$, $u_t = 0.1/t^{3/4}$.
3. DGD: $\eta_t = 0.05/\sqrt{t + 24}$.
4. Gradient tracking: $\eta = 0.03$.

Note that the step sizes used for Algorithm 1 has a slightly different form from the one studied in Theorem 1. We point out that this does not affect the applicability of the theoretical results much.

We first compare Algorithm 1 with DGD, and Figure 2 show the associated illustrations of the two algorithms. We see that for all the 4 cases, both algorithms converge relatively fast during the initial stage and then stabilizes at a sublinear convergence rate, meaning that they exhibit similar qualitative convergence behavior. On the other hand, Algorithm 1 converges slower and enters the sublinear convergence stage with higher sub-optimality and consensus errors compared to DGD. Especially, the squared magnitudes of the gradient $\|\nabla f(\bar{x}(t))\|^2$ for Algorithm 1 is about 100 times larger than those for DGD when $t \gtrsim 7.5 \times 10^3$. We speculate that this results from the randomization of the 2-point gradient estimator employed in Algorithm 1.

Next we compare Algorithm 2 with the gradient tracking method in [6]. The results are shown in Figure 3. We can see that the two algorithms exhibit almost identical behavior. We point out that this might be a consequence of the sufficient smoothness of the objective functions that could lead to very accurate estimation of the gradient even if $u_t$ is not very small (the smallest $u_t$ used in the simulation is $0.1/500^{3/4} \approx 1 \times 10^{-3}$). We are currently working on testing the performance of the proposed algorithm on other problems, and it is likely that different phenomena can occur for different problems.

Finally, we compare the convergence of Algorithm 1 and Algorithm 2 in terms of the number of function value queries as they are both zero-order methods. We have already presented and discussed one case of the initial points in the main text. Figure 4 illustrates the results for all the 4 cases, and also include the graphs of the consensus behavior. We see that observations made in the main text are still valid for the rest of the cases.
Figure 2: Comparison of Algorithm 1 and DGD for 4 cases of initial points. Figures on the left-hand side show convergence of $\|\nabla f(\bar{x}(t))\|^2$, and Figures on the right-hand side show the corresponding consensus errors $\frac{1}{n} \sum_{i=1}^{n} \|x^i(t) - \bar{x}(t)\|^2$. 
Figure 3: Comparison of Algorithm 2 and the distributed gradient descent with gradient tracking in $\mathbf{6}$ for 4 cases of initial points. Figures on the left-hand side show convergence of $\|\nabla f(\bar{x}(t))\|^2$, figures in the middle show the corresponding consensus errors $\frac{1}{n} \sum_{i=1}^{n} \|x^i(t) - \bar{x}(t)\|^2$, and figures on the right-hand side illustrate the corresponding gradient tracking errors $\frac{1}{n} \sum_{i=1}^{n} \|s^i(t) - \nabla f(\bar{x}(t-1))\|^2$. 
Figure 4: Comparison of Algorithm 1 and Algorithm 2 for 4 cases of initial points. Figures on the left-hand side show convergence of $\|\nabla f(\bar{x}(t))\|^2$, and Figures on the right-hand side show the corresponding consensus errors $\frac{1}{n} \sum_{i=1}^{n} \|x^i(t) - \bar{x}(t)\|^2$. 