

ACCELERATED DISTRIBUTED NESTEROV GRADIENT DESCENT *

GUANNAN QU[†] AND NA LI[†]

Abstract. This paper considers the distributed optimization problem over a network, where the objective is to optimize a global function formed by a sum of local functions, using only local computation and communication. We develop an Accelerated Distributed Nesterov Gradient Descent (Acc-DNGD) method. When the objective function is convex and L -smooth, we show that it achieves a $O(\frac{1}{t^{1.4-\epsilon}})$ convergence rate for all $\epsilon \in (0, 1.4)$. We also show the convergence rate can be improved to $O(\frac{1}{t^2})$ if the objective function is a composition of a linear map and a strongly-convex and smooth function. When the objective function is μ -strongly convex and L -smooth, we show that it achieves a linear convergence rate of $O([1 - C(\frac{\mu}{L})^{5/7}]^t)$, where $\frac{L}{\mu}$ is the condition number of the objective, and $C > 0$ is some constant that does not depend on $\frac{L}{\mu}$.

Key words. Consensus Optimization, distributed optimization, gradient method.

1. Introduction. Given a set of agents $\mathcal{N} = \{1, 2, \dots, n\}$, each of which has a local convex cost function $f_i(x) : \mathbb{R}^N \rightarrow \mathbb{R}$, the objective of distributed optimization is to find x that minimizes the average of all the functions,

$$\min_{x \in \mathbb{R}^N} f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f_i(x)$$

using local communication and local computation. The local communication is defined through an undirected connected communication graph. This problem has recently received much attention and has found various applications in multi-agent systems, sensor networks, machine learning, etc [2, 10, 14].

There exist many studies on developing distributed algorithms for this problem, e.g., [32, 4, 22, 15, 8, 30, 19, 20, 17, 26, 38, 16], most of which are distributed gradient descent algorithms based on a consensus scheme [24]. These methods have achieved sublinear convergence rates for convex functions. When the convex functions are nonsmooth, the sublinear convergence rate matches the Centralized Gradient Descent (CGD) method. More recent work [28, 31, 23, 21, 34, 35, 37], have improved these results to achieve linear convergence rates for strongly convex and smooth functions, or $O(\frac{1}{t})$ convergence rates for convex and smooth functions, which match the centralized gradient descent method as well.

It is known that among all centralized gradient based algorithms, centralized Nesterov Gradient Descent (CNGD) [24] achieves the optimal convergence speed in terms of first-order oracle complexity. Specifically, for L -smooth and convex problems, the convergence rate is $O(\frac{1}{t^2})$; for L -smooth and μ -strongly convex problems, the convergence rate is $O((1 - \sqrt{\mu\eta})^t)$ for step size $\eta \in (0, \frac{1}{L}]$. The nice convergence rates lead to the question of this paper: how to decentralize CNGD to achieve similar convergence rates? Paper [13] has developed Distributed Nesterov Gradient (D-NG) method and shown that for convex and L -smooth problems,¹ it has a convergence rate of $O(\frac{\log t}{t})$.² This convergence rate is worse than the $O(\frac{1}{t^2})$ rate of CNGD and is

*This work is supported under NSF ECCS 1608509 and NSF CAREER 1553407.

[†]John A. Paulson School of Engineering and Applied Sciences at Harvard University (gqu@g.harvard.edu, nali@seas.harvard.edu).

¹The convergence rate for strongly convex and smooth functions is not studied in [13]. From our simulation results in Section 5, the convergence rate for strongly convex and smooth functions is sublinear.

²Paper [13] also studies an algorithm that achieves $O(\frac{1}{t^2})$ convergence rate but it uses multiple

not better than the rate of CGD ($O(\frac{1}{t})$).

In this paper, we propose an Accelerated Distributed Nesterov Gradient Descent (Acc-DNGD) algorithm. Our algorithm has two versions. The first version Acc-DNGD-SC is designated for μ -strongly convex and L -smooth functions. It achieves a convergence rate of $O((1 - \sqrt{\mu\eta})^t)$ as CNGD, but with a more restricted range of step size $0 < \eta \leq C^2 \frac{1}{L} (\frac{\mu}{L})^{3/7}$ where C is a constant that does not depend on L or μ .³ This leads to a convergence rate of $O((1 - C(\frac{\mu}{L})^{5/7})^t)$. We emphasize here that the dependence on the condition number L/μ in our convergence rate ($[1 - C(\frac{\mu}{L})^{5/7}]^t$) is strictly better than that of CGD ($(1 - \frac{\mu}{L})^t$), and to the best of our knowledge, it is the best among all distributed algorithms proposed so far.⁴ The second version Acc-DNGD-NSC works for convex and smooth functions that are not necessarily strongly convex. It achieves a $O(\frac{1}{t^{1.4-\epsilon}})$ ($\forall \epsilon \in (0, 1.4)$) convergence rate if a vanishing step size is used. We further show that the convergence rate can be improved to $O(\frac{1}{t^2})$ when we use a fixed step size and the objective function is a composition of a linear map and a strongly-convex and smooth function. Both rates are faster than CGD ($O(\frac{1}{t})$). To the best of our knowledge, the $O(\frac{1}{t^{1.4-\epsilon}})$ rate is also the fastest among all distributed algorithms being proposed so far.⁴

Our algorithm is a combination of CNGD and a gradient estimation scheme. The gradient estimation scheme has recently been studied under various contexts in [36, 7, 33, 28, 23, 35, 21]. As [28] has pointed out, when combining the gradient estimation scheme with a centralized algorithm, the resulting distributed algorithm could potentially match the convergence rate of the centralized algorithm. The results in this paper show that, although combining the scheme with CNGD will not give a convergence rate ($O([1 - C(\frac{\mu}{L})^{5/7}]^t)$ or $O(1/t^{1.4-\epsilon})$) that matches CNGD ($O((1 - \sqrt{\frac{\mu}{L}})^t)$ or $O(1/t^2)$), it does give an improvement over CGD and CGD-based distributed algorithms ($O((1 - \frac{\mu}{L})^t)$ or $O(1/t)$).

The rest of the paper is organized as follows. Section 2 formally defines the problem and presents our algorithm and results. Section 3 and Section 4 prove the convergence of our algorithm. Lastly, Section 5 provides numerical simulations and Section 6 concludes the paper.

Notations. In this paper, n is the number of agents, and N is the dimension of the domain of the f_i 's. Notation $\|\cdot\|$ denotes 2-norm for vectors and Frobenius norm for matrices, while $\|\cdot\|_*$ denotes spectral norm for matrices, $\|\cdot\|_1$ denotes 1-norm for vectors, and $\langle \cdot, \cdot \rangle$ denotes inner product for vectors. Notation $\rho(\cdot)$ denotes spectral radius for square matrices, and $\mathbf{1}$ denotes a n -dimensional all one column vector. All vectors, when having dimension N (the dimension of the domain of the f_i 's), will all be regarded as row vectors. As a special case, all gradients, $\nabla f_i(x)$ and $\nabla f(x)$ are treated as N -dimensional row vectors. Notation " \leq ", when applied to vectors of the same dimension, denotes element wise "less than or equal to".

2. Problem and Algorithm.

2.1. Problem Formulation. Consider n agents, $\mathcal{N} = \{1, 2, \dots, n\}$, each of which has a convex function $f_i : \mathbb{R}^N \rightarrow \mathbb{R}$. The objective of distributed optimization

consensus steps per iteration. In the rest of the paper, we only focus on algorithms that use *one* consensus step per gradient iteration.

³The value of C will be clear in Theorem 5.

⁴We only include algorithms that are based on consensus averaging and gradient evaluation (without extra information like Hessian or solving sub optimization problems), and use one or a constant number of consensus steps per gradient evaluation.

is to find x to minimize the average of all the functions, i.e.

$$(1) \quad \min_{x \in \mathbb{R}^N} f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f_i(x)$$

using local communication and local computation. The local communication is defined through a *connected* and *undirected* communication graph $\mathcal{G} = (V, E)$, where the nodes $V = \mathcal{N}$ and edges $E \subset V \times V$. Agent i and j can send information to each other if and only if $(i, j) \in E$. The local computation means that each agent can only make its decision based on the local function f_i and the information obtained from its neighbors.

Global Assumptions. Throughout the paper, we make the following assumptions without explicitly stating them. We assume that each f_i is convex (hence f is also convex). We assume each f_i is L -smooth, that is, f_i is differentiable and the gradient is L -Lipschitz continuous, i.e., $\forall x, y \in \mathbb{R}^N$,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|.$$

As a result, f is also L -smooth. We assume f has at least one minimizer x^* with $f(x^*) = f^*$.

Other Assumptions. We will use the following assumptions in different parts of the paper, and those will be stated explicitly.

ASSUMPTION 1. $\forall i \in \mathcal{N}$, f_i is μ -strongly convex, i.e. $\forall x, y \in \mathbb{R}^N$, we have

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2.$$

As a result, f is also μ -strongly convex.

ASSUMPTION 2. The set of minimizers of f is compact.

2.2. Centralized Nesterov Gradient Descent (CNGD). We here briefly introduce two versions of centralized Nesterov Gradient Descent (CNGD) algorithm that is derived from [24, Scheme (2.2.6)].⁵ The first version, which we term ‘‘CNGD-SC’’, is designated for μ -strongly convex and L -smooth functions. Given step size η , let $\alpha = \sqrt{\mu\eta}$. CNGD-SC keeps updating three variables $x(t), v(t), y(t) \in \mathbb{R}^N$, starting from an initial point $x(0) = v(0) = y(0) \in \mathbb{R}^N$, and the updating rule is given by

$$(2a) \quad x(t+1) = y(t) - \eta \nabla f(y(t))$$

$$(2b) \quad v(t+1) = (1 - \alpha)v(t) + \alpha y(t) - \frac{\eta}{\alpha} \nabla f(y(t))$$

$$(2c) \quad y(t+1) = \frac{x(t+1) + \alpha v(t+1)}{1 + \alpha}.$$

The following theorem (adapted from [24, Theorem 2.2.1, Lemma 2.2.4]) gives the convergence rate of CNGD-SC.

THEOREM 3. Under Assumption 1, when $0 < \eta \leq \frac{1}{L}$, in CNGD-SC (2) we have $f(x(t)) - f^* = O((1 - \sqrt{\mu\eta})^t)$.

⁵The two versions are essentially the algorithm in [24, Scheme (2.2.6)] with two set of parameters. Describing the two versions in one form as in [24, Scheme (2.2.6)] will need more cumbersome notations. For this reason, we write down the two versions separately.

The second version, which we term ‘‘CNGD-NSC’’, is designated for convex (possibly not strongly-convex) and L -smooth functions. CNGD-NSC keeps updating three variables $x(t), v(t), y(t) \in \mathbb{R}^N$, starting from an initial point $x(0) = v(0) = y(0) \in \mathbb{R}^N$, and the updating rule is given by

$$(3a) \quad x(t+1) = y(t) - \eta \nabla f(y(t))$$

$$(3b) \quad v(t+1) = v(t) - \frac{\eta}{\alpha_t} \nabla f(y(t))$$

$$(3c) \quad y(t+1) = (1 - \alpha_{t+1})x(t+1) + \alpha_{t+1}v(t+1),$$

where $(\alpha_t)_{t=0}^\infty$ is defined by an arbitrarily chosen $\alpha_0 \in (0, 1)$ and the update equation $\alpha_{t+1}^2 = (1 - \alpha_{t+1})\alpha_t^2$, where α_{t+1} always takes the unique solution in $(0, 1)$. The following theorem (adapted from [24, Theorem 2.2.1, Lemma 2.2.4]) gives the convergence rate of CNGD-NSC.

THEOREM 4. *In CNGD-NSC (3), when $0 < \eta \leq \frac{1}{L}$, we have $f(x(t)) - f^* = O(\frac{1}{t^2})$.*

2.3. Our Algorithm: Accelerated Distributed Nesterov Gradient Descent (Acc-DNGD). We design our algorithm based on a consensus matrix $W = [w_{ij}] \in \mathbb{R}^{n \times n}$. Here w_{ij} stands for how much agent i weighs its neighbor j 's information. Matrix W satisfies the following conditions:

(a) For any $(i, j) \in E$, $w_{ij} > 0$. For any $i \in \mathcal{N}$, $w_{ii} > 0$. Elsewhere, $w_{ij} = 0$.

(b) Matrix W is doubly stochastic, i.e. $\sum_{i'} w_{i'j} = \sum_{j'} w_{ij'} = 1$ for all $i, j \in \mathcal{N}$.

As a result, $\exists \sigma \in (0, 1)$ being the second largest singular value of W , such that for any $\omega \in \mathbb{R}^{n \times 1}$, we have the ‘‘averaging property’’, $\|W\omega - \mathbf{1}\bar{\omega}\| \leq \sigma\|\omega - \mathbf{1}\bar{\omega}\|$ where $\bar{\omega} = \frac{1}{n}\mathbf{1}^T\omega$ (the average of the entries in ω) [27]. How to select a consensus matrix to satisfy these properties has been intensely studied, e.g. [25, 27].

Analogous to the centralized case, we present two versions of our algorithm. The first version, Acc-DNGD-SC, is designated for μ -strongly convex and L -smooth functions. Each agent keeps a copy of the three variables in CNGD-SC, $x_i(t), v_i(t), y_i(t)$ and in addition $s_i(t)$ which serves as a gradient estimator. The initial condition is $x_i(0) = v_i(0) = y_i(0) \in \mathbb{R}^{1 \times N}$ and $s_i(0) = \nabla f_i(y_i(0))$, and the algorithm updates as follows:

$$(4a) \quad x_i(t+1) = \sum_j w_{ij}y_j(t) - \eta s_i(t)$$

$$(4b) \quad v_i(t+1) = (1 - \alpha) \sum_j w_{ij}v_j(t) + \alpha \sum_j w_{ij}y_j(t) - \frac{\eta}{\alpha} s_i(t)$$

$$(4c) \quad y_i(t+1) = \frac{x_i(t+1) + \alpha v_i(t+1)}{1 + \alpha}$$

$$(4d) \quad s_i(t+1) = \sum_j w_{ij}s_j(t) + \nabla f_i(y_i(t+1)) - \nabla f_i(y_i(t))$$

where $[w_{ij}]_{n \times n}$ are the consensus weights, $\eta > 0$ is a fixed step size and $\alpha = \sqrt{\mu\eta}$.

The second version, Acc-DNGD-NSC, is designated for convex (not necessarily strongly convex) and L -smooth functions. Similarly as Acc-DNGD-SC, each agent keeps variable $x_i(t), v_i(t), y_i(t)$ and $s_i(t)$. The initial condition is $x_i(0) = v_i(0) = y_i(0) = 0$ and $s_i(0) = \nabla f(0)$,⁶ and the algorithm updates as follows:

⁶We note that the initial condition $s_i(0) = \nabla f(0) = \frac{1}{n} \sum_i \nabla f_i(0)$ requires the agents to conduct an initial run of consensus averaging. We impose this initial condition for technical reasons, while we expect the results of this paper to hold for a relaxed initial condition, $s_i(0) = \nabla f_i(y_i(0))$ which does not need initial coordination. We use the relaxed condition in numerical simulations.

$$\begin{aligned}
(5a) \quad x_i(t+1) &= \sum_j w_{ij} y_j(t) - \eta_t s_i(t) \\
(5b) \quad v_i(t+1) &= \sum_j w_{ij} v_j(t) - \frac{\eta_t}{\alpha_t} s_i(t) \\
(5c) \quad y_i(t+1) &= (1 - \alpha_{t+1}) x_i(t+1) + \alpha_{t+1} v_i(t+1) \\
(5d) \quad s_i(t+1) &= \sum_j w_{ij} s_j(t) + \nabla f_i(y_i(t+1)) - \nabla f_i(y_i(t))
\end{aligned}$$

where $[w_{ij}]_{n \times n}$ are the consensus weights and $\eta_t \in (0, \frac{1}{L})$ are the step sizes. Sequence $(\alpha_t)_{t \geq 0}$ is generated as follows. First we let $\alpha_0 = \sqrt{\eta_0 L} \in (0, 1)$. Then given $\alpha_t \in (0, 1)$, we select α_{t+1} to be the unique solution in $(0, 1)$ of the following equation,⁷

$$(6) \quad \alpha_{t+1}^2 = \frac{\eta_{t+1}}{\eta_t} (1 - \alpha_{t+1}) \alpha_t^2.$$

We will consider two variants of the algorithm with the following two step size rules.

- **Vanishing step size:** $\eta_t = \frac{1}{\eta(t+t_0)^\beta}$ for some $\eta \in (0, \frac{1}{L})$, $\beta \in (0, 2)$ and $t_0 \geq 1$.
- **Fixed step size:** $\eta_t = \eta > 0$.

In both versions (Acc-DNGD-SC (4) and Acc-DNGD-NSC (5)), because $w_{ij} = 0$ when $(i, j) \notin E$, each node i only needs to send $x_i(t)$, $v_i(t)$, $y_i(t)$ and $s_i(t)$ to its neighbors. Therefore, the algorithm can be operated in a fully distributed fashion with only local communication. The additional term $s_i(t)$ allows each agent to obtain an estimate on the global gradient $\frac{1}{n} \sum_i \nabla f_i(y_i(t))$. Compared with distributed algorithms without this estimation term, it helps improve the convergence speed. For more intuition behind how the gradient estimation term $s_i(t)$ works, we refer the readers to [28]. Because of the use of the gradient estimation term, we call this method as *Accelerated Distributed Nesterov Gradient Descent* (Acc-DNGD) method.

2.4. Convergence of the Algorithm. To state the convergence results, we need to define the following average sequence $\bar{x}(t) = \frac{1}{n} \sum_{i=1}^n x_i(t) \in \mathbb{R}^{1 \times N}$. We first provide the convergence result for Acc-DNGD-SC in Theorem 5.

THEOREM 5. *Consider algorithm Acc-DNGD-SC (4). Under the strongly convex assumption (Assumption 1), when $0 < \eta < \frac{\sigma^3(1-\sigma)^3}{61909L} (\frac{\mu}{L})^{3/7}$, we have (a) $f(\bar{x}(t)) - f^* = O((1 - \sqrt{\mu\eta})^t)$; (b) $\|y(t) - \mathbf{1}x^*\| = O((1 - \sqrt{\mu\eta})^{t/2})$.*

The step size in theorem 5 results in a convergence rate of $O([1 - C(\frac{\mu}{L})^{5/7}]^t)$ with $C = \frac{\sigma^{1.5}(1-\sigma)^{1.5}}{\sqrt{61909}}$. In this convergence rate, the dependence on the condition number $\frac{L}{\mu}$ is strictly better than that of CGD ($O((1 - \frac{\mu}{L})^t)$), and hence CGD based distributed algorithms.⁸ This is particularly beneficial because in many machine learning applications, the condition number $\frac{L}{\mu}$ can be as large as the sample size [5, Section 3.6].

We next provide the convergence result for Acc-DNGD-NSC in Theorem 6.

⁷Without causing any confusion with the α_t in (2), in the rest of the paper we abuse the notation of α_t .

⁸The quantity $\frac{L}{\mu}$ is called the condition number of the objective function. For a convergence rate of $\Theta((1 - (\frac{\mu}{L})^\delta)^t)$, it takes $\Theta((\frac{L}{\mu})^\delta \log \frac{1}{\epsilon})$ iterations to reach an accuracy level ϵ . Therefore, the smaller δ is, the better the convergence rate is.

THEOREM 6. Consider algorithm *Acc-DNGD-NSC* (5). Suppose Assumption 2 is true and without loss of generality we assume $\bar{v}(0) \neq x^*$. Let the step size be $\eta_t = \frac{\eta}{(t+t_0)^\beta}$ with $\beta = 0.6 + \epsilon$ where $\epsilon \in (0, 1.4)$. Suppose the following conditions are met.

(i)

$$t_0 > \frac{1}{\min\left(\left(\frac{\sigma+3}{\sigma+2}\frac{3}{4}\right)^{\sigma/(28\beta)}, \left(\frac{16}{15+\sigma}\right)^{\frac{1}{\beta}}\right) - 1}.$$

(ii)

$$\eta < \min\left(\frac{\sigma^2}{9^3 L}, \frac{(1-\sigma)^3}{6144 L}\right).$$

(iii)

$$\eta < \left(\frac{D(\beta, t_0)(\beta - 0.6)(1 - \sigma)^2}{9216(t_0 + 1)^{2-\beta} L^{2/3} [4 + R^2 / \|\bar{v}(0) - x^*\|^2]} \right)^{3/2}$$

where $D(\beta, t_0) = \frac{1}{(t_0+3)^2 e^{16+\frac{6}{2-\beta}}}$ and R is the diameter of the $(2f(\bar{x}(0)) - f^* + 2L\|\bar{v}(0) - x^*\|^2)$ -level set of f .⁹

Then, $f(\bar{x}(t)) - f^* = O\left(\frac{1}{t^{2-\beta}}\right) = O\left(\frac{1}{t^{1.4-\epsilon}}\right)$.

In Theorem 6, condition (i) is intended to make η_t/η_{t+1} close to 1 (as will be required in Lemma 17 (iii)), and condition (ii)(iii) are intended to make η_t close to 0 (as will be required in Lemma 17 (ii)). While the conditions are needed for the proof, we expect the same result will hold if we simply let $t_0 = 1$ and $\eta = \frac{1}{2L}$, which is what we choose in the simulations in Section 5. The reason is that, regardless of the value of η and t_0 , we have $\eta_t \rightarrow 0$ and $\eta_t/\eta_{t+1} \rightarrow 1$, and hence for large enough t , η_t and η_t/η_{t+1} will automatically be close to 0 and 1 respectively.

While in Theorem 6 we require $\beta > 0.6$, we conjecture that the algorithm will converge with rate $O\left(\frac{1}{t^{2-\beta}}\right)$ even if we choose $\beta \in [0, 0.6]$, with $\beta = 0$ corresponding to the case of fixed step size. In Section 5 we will use numerical methods to support this conjecture.

In the next theorem, we provide a $O\left(\frac{1}{t^2}\right)$ convergence result when a fixed step size is used and the objective functions belong to a special class.

THEOREM 7. Consider algorithm *Acc-DNGD-NSC* (5). Assume each $f_i(x)$ can be written as $f_i(x) = h_i(xA_i)$, where A_i is a non-zero $N \times M_i$ matrix, and $h_i(x) : \mathbb{R}^{1 \times M_i} \rightarrow \mathbb{R}$ is a μ_0 -strongly convex and L_0 -smooth function. Suppose we use the fixed step size rule $\eta_t = \eta$, with

$$0 < \eta < \min\left(\frac{\sigma^2}{9^3 L}, \frac{\mu^{1.5}(1-\sigma)^3}{L^{2.5} 3456^{1.5}}\right)$$

where $L = L_0\nu$ with $\nu = \max_i \|A_i\|_*^2$; and $\mu = \mu_0\gamma$ with γ being the smallest non-zero eigenvalue of matrix $A = \frac{1}{n} \sum_{i=1}^n A_i A_i^T$. Then, we have $f(\bar{x}(t)) - f^* = O\left(\frac{1}{t^2}\right)$.

An important example of the type of functions $f_i(x)$ in Theorem 7 is the square loss for linear regression (cf. Case I in Section 5) when the sample size is less than the parameter dimension.

Remark 8. All constants in the step size bounds provided in this section are conservative. The conservative constants are a result of the fact that, in order to simplify mathematical calculations, we have used coarse spectral bounds in the proofs (see Lemma 14, 20, 21, 22). In numerical simulations, we use much larger step sizes.

⁹Here we have used the fact that by Assumption 2, all level sets of f are bounded. See [3, Proposition B.9].

3. Convergence Analysis of Acc-DNGD-SC. In this section, we will provide the proof of Theorem 5. We will first provide a proof overview in Section 3.1 and then defer the detailed proof to the rest of the section.

3.1. Proof Overview. Without causing any confusion with notations in (2), in this section we abuse the use of notation $x(t), v(t), y(t)$. We introduce notations $x(t), v(t), y(t), s(t), \nabla(t) \in \mathbb{R}^{n \times N}$ as follows,

$$(7a) \quad x(t) = [x_1(t)^T, x_2(t)^T, \dots, x_n(t)^T]^T, \quad v(t) = [v_1(t)^T, v_2(t)^T, \dots, v_n(t)^T]^T$$

$$(7b) \quad y(t) = [y_1(t)^T, y_2(t)^T, \dots, y_n(t)^T]^T, \quad s(t) = [s_1(t)^T, s_2(t)^T, \dots, s_n(t)^T]^T$$

$$(7c) \quad \nabla(t) = [\nabla f_1(y_1(t))^T, \nabla f_2(y_2(t))^T, \dots, \nabla f_n(y_n(t))^T]^T.$$

Now the algorithm Acc-DNGD-SC (4) can be written as

$$(8a) \quad x(t+1) = Wy(t) - \eta s(t)$$

$$(8b) \quad v(t+1) = (1 - \alpha)Wv(t) + \alpha Wy(t) - \frac{\eta}{\alpha} s(t)$$

$$(8c) \quad y(t+1) = \frac{x(t+1) + \alpha v(t+1)}{1 + \alpha}$$

$$(8d) \quad s(t+1) = Ws(t) + \nabla(t+1) - \nabla(t).$$

Apart from the average sequence $\bar{x}(t) = \frac{1}{n} \sum_{i=1}^n x_i(t)$ that we have defined, we also define several other average sequences, $\bar{v}(t) = \frac{1}{n} \sum_{i=1}^n v_i(t)$, $\bar{y}(t) = \frac{1}{n} \sum_{i=1}^n y_i(t)$, $\bar{s}(t) = \frac{1}{n} \sum_{i=1}^n s_i(t)$, and $g(t) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(y_i(t))$.

Overview of the Proof. In our proof, we firstly derive the update formula for the average sequences (Lemma 9). Then, it turns out that the update rule for the average sequences is in fact CNGD-SC (4) with inexact gradients [6], and the inexactness is characterized by ‘‘consensus error’’ $\|y(t) - \mathbf{1}\bar{y}(t)\|$ (Lemma 10). The consensus error is then bounded in Lemma 11. With the bound on consensus error, we can roughly apply the same proof steps of CNGD-SC (see e.g. [24]) to the average sequence and finish the proof of Theorem 5 in Section 3.3.

LEMMA 9. *The following equalities hold.*

$$(9a) \quad \bar{x}(t+1) = \bar{y}(t) - \eta g(t)$$

$$(9b) \quad \bar{v}(t+1) = (1 - \alpha)\bar{v}(t) + \alpha \bar{y}(t) - \frac{\eta}{\alpha} g(t)$$

$$(9c) \quad \bar{y}(t+1) = \frac{\bar{x}(t+1) + \alpha \bar{v}(t+1)}{1 + \alpha}$$

$$(9d) \quad \bar{s}(t+1) = \bar{s}(t) + g(t+1) - g(t) = g(t+1)$$

Proof: We omit the proof since these can be easily derived using the fact that W is doubly stochastic. For (9d) we also need to use the fact that $\bar{s}(0) = g(0)$. \square

From (9a)-(9c) we see that the sequences $\bar{x}(t)$, $\bar{v}(t)$ and $\bar{y}(t)$ follow a update rule similar to the CNGD-SC in (2). The only difference is that the $g(t)$ in (9a)-(9c) is not the exact gradient $\nabla f(\bar{y}(t))$ in CNGD-SC. In the following Lemma, we show that $g(t)$ is an inexact gradient with error $O(\|y(t) - \mathbf{1}\bar{y}(t)\|^2)$.

LEMMA 10. *Under Assumption 1, $\forall t$, $g(t)$ is an inexact gradient of f at $\bar{y}(t)$ with error $O(\|y(t) - \mathbf{1}\bar{y}(t)\|^2)$ in the sense that if we let $\hat{f}(t) = \frac{1}{n} \sum_{i=1}^n [f_i(y_i(t)) + \langle \nabla f_i(y_i(t)), \bar{y}(t) - y_i(t) \rangle]$, then $\forall \omega \in \mathbb{R}^N$,*¹⁰

$$(10) \quad f(\omega) \geq \hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + \frac{\mu}{2} \|\omega - \bar{y}(t)\|^2$$

¹⁰The exact gradient $\nabla f(\bar{y}(t))$, satisfies $f(\bar{y}(t)) + \langle \nabla f(\bar{y}(t)), \omega - \bar{y}(t) \rangle + \frac{\mu}{2} \|\omega - \bar{y}(t)\|^2 \leq f(\omega) \leq f(\bar{y}(t)) + \langle \nabla f(\bar{y}(t)), \omega - \bar{y}(t) \rangle + \frac{\mu}{2} \|\omega - \bar{y}(t)\|^2$. This is why we call $g(t)$ inexact gradient.

$$(11) \quad f(\omega) \leq \hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + L\|\omega - \bar{y}(t)\|^2 + L\frac{1}{n}\|y(t) - \mathbf{1}\bar{y}(t)\|^2.$$

Proof: For any $\omega \in \mathbb{R}^N$, we have

$$\begin{aligned} f(\omega) &= \frac{1}{n} \sum_{i=1}^n f_i(\omega) \geq \frac{1}{n} \sum_{i=1}^n [f_i(y_i(t)) + \langle \nabla f_i(y_i(t)), \omega - y_i(t) \rangle + \frac{\mu}{2} \|\omega - y_i(t)\|^2] \\ &= \frac{1}{n} \sum_{i=1}^n [f_i(y_i(t)) + \langle \nabla f_i(y_i(t)), \bar{y}(t) - y_i(t) \rangle] + \frac{1}{n} \sum_{i=1}^n \langle \nabla f_i(y_i(t)), \omega - \bar{y}(t) \rangle \\ &\quad + \frac{1}{n} \sum_{i=1}^n \frac{\mu}{2} \|\omega - y_i(t)\|^2 \\ &\geq \hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + \frac{\mu}{2} \|\omega - \bar{y}(t)\|^2 \end{aligned}$$

which shows (10). For (11), similarly,

$$\begin{aligned} f(\omega) &\leq \frac{1}{n} \sum_{i=1}^n [f_i(y_i(t)) + \langle \nabla f_i(y_i(t)), \omega - y_i(t) \rangle + \frac{L}{2} \|\omega - y_i(t)\|^2] \\ &= \hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + \frac{L}{2} \frac{1}{n} \sum_{i=1}^n \|\omega - y_i(t)\|^2 \\ &\leq \hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + L\|\omega - \bar{y}(t)\|^2 + L\frac{1}{n} \sum_{i=1}^n \|\bar{y}(t) - y_i(t)\|^2 \end{aligned}$$

where in the last inequality we have used the elementary fact that $\|u+v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$ for all $u, v \in \mathbb{R}^N$. Then, (11) follows by $\sum_{i=1}^n \|\bar{y}(t) - y_i(t)\|^2 = \|y(t) - \mathbf{1}\bar{y}(t)\|^2$. \square

The consensus error $\|y(t) - \mathbf{1}\bar{y}(t)\|$ in the previous lemma is bounded by the following lemma whose proof is deferred to Section 3.2.

LEMMA 11. *When $0 < \eta < \min(\frac{1}{L} \frac{(1-\sigma)^3}{512}, \frac{1}{L} \frac{\sigma^3}{64})$, we have*

$$\|y(k) - \mathbf{1}\bar{y}(k)\| \leq A_1(\eta)\theta^k + A_2(\eta) \sum_{\ell=0}^{k-1} \theta^{k-1-\ell} a(\ell)$$

where $a(\ell) \triangleq \|\bar{y}(\ell) - \bar{x}(\ell)\| + 2\eta\|g(\ell)\|$, $\theta \triangleq \frac{1+\sigma}{2}$, and

$$A_1(\eta) \triangleq \frac{39[2\|y(0) - \mathbf{1}\bar{y}(0)\| + \frac{1}{L}\|s(0) - \mathbf{1}g(0)\|]}{(\sigma\eta L)^{2/3}}, A_2(\eta) \triangleq \frac{39\sqrt{n}(\eta L)^{1/3}}{(\sigma)^{2/3}}.$$

At last, we will finish the proof of Theorem 5 in Section 3.3.

3.2. Proof of the Bounded Consensus Error (Lemma 11). We now give the proof of Lemma 11. We will frequently use the following two lemmas, whose proofs are deferred to Appendix-A.1.

LEMMA 12. *The following inequality is true.*

$$(12) \quad \|\bar{y}(t+1) - \bar{y}(t)\| \leq \|\bar{y}(t) - \bar{x}(t)\| + 2\eta\|g(t)\|$$

LEMMA 13. *The following inequalities are true.*

$$(13) \quad \|\nabla(t+1) - \nabla(t)\| \leq L\|y(t+1) - y(t)\|$$

$$(14) \quad \|g(t) - \nabla f(\bar{y}(t))\| \leq \frac{L}{\sqrt{n}}\|y(t) - \mathbf{1}\bar{y}(t)\|$$

Proof of Lemma 11:

Overview of the proof. The proof is separated into two steps. In step 1, we treat the algorithm (8) as a linear system and derive a linear system inequality (15). In step 2, we analyze the state transition matrix in (15) and prove its spectral properties, from which the conclusion of the lemma follows.

Step 1: A Linear System Inequality. Define state $z(t) = [\|v(t) - \mathbf{1}\bar{v}(t)\|, \|y(t) - \mathbf{1}\bar{y}(t)\|, \frac{1}{L}\|s(t) - \mathbf{1}g(t)\|]^T \in \mathbb{R}^3$ and we will derive a linear system inequality that bounds the update of $z(t)$, given as follows

$$(15) \quad z(t+1) \leq G(\eta)z(t) + b(t).$$

Here $b(t) = [0, 0, \sqrt{na}(t)]^T \in \mathbb{R}^3$ is the input to the system with

$$a(t) \triangleq \|\bar{y}(t) - \bar{x}(t)\| + 2\eta\|g(t)\|.$$

The state transition matrix $G(\eta) \in \mathbb{R}^{3 \times 3}$ is given by

$$G(\eta) = \begin{bmatrix} (1-\alpha)\sigma & \alpha\sigma & \frac{\eta L}{\alpha} \\ \frac{1-\alpha}{1+\alpha}\alpha\sigma & \frac{1+\alpha^2}{1+\alpha}\sigma & 2\eta L \\ \alpha\sigma & 2 & \sigma + 2\eta L \end{bmatrix}.$$

We now prove (15). By (8a) and (9a), we have

$$(16) \quad \begin{aligned} \|x(t+1) - \mathbf{1}\bar{x}(t+1)\| &= \|[Wy(t) - \mathbf{1}\bar{y}(t)] - \eta[s(t) - \mathbf{1}g(t)]\| \\ &\leq \sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + \eta\|s(t) - \mathbf{1}g(t)\|. \end{aligned}$$

By (8b) and (9b), we have

$$(17) \quad \begin{aligned} \|v(t+1) - \mathbf{1}\bar{v}(t+1)\| &\leq \|(1-\alpha)[Wv(t) - \mathbf{1}\bar{v}(t)] + \alpha[Wy(t) - \mathbf{1}\bar{y}(t)] - \frac{\eta}{\alpha}[s(t) - \mathbf{1}g(t)]\| \\ &\leq (1-\alpha)\sigma\|v(t) - \mathbf{1}\bar{v}(t)\| + \alpha\sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + \frac{\eta}{\alpha}\|s(t) - \mathbf{1}g(t)\|. \end{aligned}$$

By (8c) and (9c), we have

$$(18) \quad \begin{aligned} &\|y(t+1) - \mathbf{1}\bar{y}(t+1)\| \\ &\leq \frac{1}{1+\alpha}\|x(t+1) - \mathbf{1}\bar{x}(t+1)\| + \frac{\alpha}{1+\alpha}\|v(t+1) - \mathbf{1}\bar{v}(t+1)\| \\ &\leq \frac{1}{1+\alpha} \left[\sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + \eta\|s(t) - \mathbf{1}g(t)\| \right] \\ &\quad + \frac{\alpha}{1+\alpha} \left[(1-\alpha)\sigma\|v(t) - \mathbf{1}\bar{v}(t)\| + \alpha\sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + \frac{\eta}{\alpha}\|s(t) - \mathbf{1}g(t)\| \right] \\ &\leq \frac{1-\alpha}{1+\alpha}\alpha\sigma\|v(t) - \mathbf{1}\bar{v}(t)\| + \frac{1+\alpha^2}{1+\alpha}\sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + 2\eta\|s(t) - \mathbf{1}g(t)\| \end{aligned}$$

where we have used (16) and (17) in the second inequality.

By (8d) and (9d), we have

$$\begin{aligned} &\|s(t+1) - \mathbf{1}g(t+1)\| \\ &= \|Ws(t) - \mathbf{1}g(t) + [\nabla(t+1) - \nabla(t) - \mathbf{1}(g(t+1) - g(t))]\| \\ &\stackrel{(a)}{\leq} \sigma\|s(t) - \mathbf{1}g(t)\| + \|\nabla(t+1) - \nabla(t)\| \end{aligned}$$

$$(19) \quad \stackrel{(b)}{\leq} \sigma \|s(t) - \mathbf{1}g(t)\| + L \|y(t+1) - y(t)\|$$

where in (a) we have used the fact that

$$\|[\nabla(t+1) - \nabla(t)] - [\mathbf{1}g(t+1) - \mathbf{1}g(t)]\|^2 = \|\nabla(t+1) - \nabla(t)\|^2 - n \|g(t+1) - g(t)\|^2$$

and in (b) we have used (13).

Now we expand $y(t+1) - y(t)$.

$$\begin{aligned} & \|y(t+1) - y(t)\| \\ & \leq \|y(t+1) - \mathbf{1}\bar{y}(t+1)\| + \|y(t) - \mathbf{1}\bar{y}(t)\| + \|\mathbf{1}\bar{y}(t+1) - \mathbf{1}\bar{y}(t)\| \\ & \leq \frac{1-\alpha}{1+\alpha} \alpha \sigma \|v(t) - \mathbf{1}\bar{v}(t)\| + \left[\frac{1+\alpha^2}{1+\alpha} \sigma + 1 \right] \|y(t) - \mathbf{1}\bar{y}(t)\| \\ & \quad + 2\eta \|s(t) - \mathbf{1}g(t)\| + \|\mathbf{1}\bar{y}(t+1) - \mathbf{1}\bar{y}(t)\| \\ & \leq \alpha \sigma \|v(t) - \mathbf{1}\bar{v}(t)\| + 2 \|y(t) - \mathbf{1}\bar{y}(t)\| \\ & \quad + 2\eta \|s(t) - \mathbf{1}g(t)\| + \sqrt{n} [\|\bar{y}(t) - \bar{x}(t)\| + 2\eta \|g(t)\|] \end{aligned}$$

where in the last inequality we have used (12) (Lemma 12). Combining the above with (19), we get

$$(20) \quad \begin{aligned} & \frac{1}{L} \|s(t+1) - \mathbf{1}g(t+1)\| \\ & \leq \sigma \frac{1}{L} \|s(t) - \mathbf{1}g(t)\| + \|y(t+1) - y(t)\| \\ & \leq \alpha \sigma \|v(t) - \mathbf{1}\bar{v}(t)\| + 2 \|y(t) - \mathbf{1}\bar{y}(t)\| \\ & \quad + (\sigma + 2\eta L) \frac{1}{L} \|s(t) - \mathbf{1}g(t)\| + \sqrt{n} \left[\|\bar{y}(t) - \bar{x}(t)\| + 2\eta \|g(t)\| \right]. \end{aligned}$$

Combining (17) (18) (20) gives the linear system inequality (15).

Step 2: Spectral Properties of $G(\eta)$. We give the following lemma regarding $G(\eta)$. We provide a proof-sketch here while the complete proof can be found in Appendix-A.2.

LEMMA 14. *When $0 < \eta < \min(\frac{1}{L}(\frac{1-\sigma}{8})^3, \frac{\sigma^3}{L64})$, the following holds.*

(a) *We have $\sigma + (\sigma\eta L)^{1/3} < \rho(G(\eta)) < \sigma + 4(\eta L)^{1/3} < \frac{1+\sigma}{2} = \theta$.*

(b) *The (2,3)th entry of $G(\eta)^t$ is upper bounded by $[G(\eta)^t]_{2,3} \leq \frac{39(\eta L)^{1/3}}{(\sigma)^{2/3}} \rho(G(\eta))^t$.*

(c) *The entries in the 2nd row of $G(\eta)^t$ are all upper bounded by $\frac{39}{(\sigma\eta L)^{2/3}} \rho(G(\eta))^t$.*

Proof-Sketch: Part (a) essentially says that when η is small, $\rho(G(\eta))$ should be close to σ . This follows from $\rho(G(0)) = \sigma$ and the fact that $\rho(G(\eta))$ is continuous in η . A detailed calculation leads to the bound in (a). Part (b)(c), essentially says that, the (p,q) th entry of $G(\eta)^t$ can be upper bounded by $C_{pq} \rho(G(\eta))^t$ for some constant $C_{pq} > 0$ that does not depend on t . This is a direct consequence of Perron-Frobenius Theorem [12, Theorem 8.2.11]. In the appendix we calculate detailed values of C_{21}, C_{22}, C_{23} which lead to (b)(c). \square

Notice in Lemma 14 (b), the constant $\frac{39(\eta L)^{1/3}}{(\sigma)^{2/3}}$ converges to 0 as $\eta \rightarrow 0$. This fact is crucial in the proof of Theorem 5 (cf. (29) and the argument following it).

The step size condition in Lemma 11 ensures the condition of Lemma 14 holds. By (15),

$$z(t) \leq G(\eta)^t z(0) + \sum_{k=0}^{t-1} G(\eta)^{t-1-k} b(k).$$

Recall the second entry of $z(t)$ is $\|y(t) - \mathbf{1}\bar{y}(t)\|$, and $b(t) = [0, 0, \sqrt{n}a(t)]^T$. Hence the above inequality implies,

$$\begin{aligned} & \|y(t) - \mathbf{1}\bar{y}(t)\| \\ & \leq \max([G(\eta)^t]_{2,1}, [G(\eta)^t]_{2,2}, [G(\eta)^t]_{2,3})\|z(0)\|_1 + \sum_{k=0}^{t-1} [G(\eta)^{t-1-k}]_{2,3}\sqrt{n}a(k) \\ & \leq \underbrace{\frac{39[2\|y(0) - \mathbf{1}\bar{y}(0)\| + \frac{1}{L}\|s(0) - \mathbf{1}g(0)\|]}{(\sigma\eta L)^{2/3}}}_{\triangleq A_1(\eta)} \theta^t + \underbrace{\frac{39\sqrt{n}(\eta L)^{1/3}}{(\sigma)^{2/3}} \sum_{k=0}^{t-1} \theta^{t-1-k} a(k)}_{\triangleq A_2(\eta)}. \end{aligned}$$

□

3.3. Proof of Theorem 5. We define a series of functions $\Phi_t(\omega)$. We define $\Phi_0(\omega) = f(\bar{x}(0)) + \frac{\mu}{2}\|\omega - \bar{v}(0)\|^2$ and given $\Phi_t(\cdot)$, we define $\Phi_{t+1}(\cdot)$ by,

$$(21) \quad \Phi_{t+1}(\omega) = (1 - \alpha)\Phi_t(\omega) + \alpha(\hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + \frac{\mu}{2}\|\omega - \bar{y}(t)\|^2).$$

Claim 1: $\Phi_t(\cdot)$ can be written as,

$$(22) \quad \Phi_t(\omega) = \phi_t^* + \frac{\mu}{2}\|\omega - \bar{v}(t)\|^2$$

where $\phi_0^* = f(\bar{x}(0))$, and given ϕ_t^* , we have

$$(23) \quad \begin{aligned} \phi_{t+1}^* &= (1 - \alpha)\phi_t^* + \frac{\mu}{2}(1 - \alpha)\alpha\|\bar{v}(t) - \bar{y}(t)\|^2 + \alpha\hat{f}(t) \\ &\quad - \frac{1}{2}\eta\|g(t)\|^2 + (1 - \alpha)\alpha\langle g(t), \bar{v}(t) - \bar{y}(t) \rangle. \end{aligned}$$

Claim 2: For any t , we have

$$(24) \quad \Phi_t(\omega) \leq f(\omega) + (1 - \alpha)^t(\Phi_0(\omega) - f(\omega)).$$

Proof of Claim 1: By (21), Φ_t is always a quadratic function. Since $\nabla^2\Phi_0(\omega) = \mu I$ and $\nabla^2\Phi_{t+1}(\omega) = (1 - \alpha)\nabla^2\Phi_t(\omega) + \alpha\mu I$, we get $\nabla^2\Phi_t(\omega) = \mu I$ for all t . We next show by induction that $\Phi_t(\cdot)$ achieves its minimum at $\bar{v}(t)$. Firstly, Φ_0 achieves its minimum at $\bar{v}(0)$. Assume $\Phi_t(\cdot)$ achieves minimum at $\bar{v}(t)$. Since Φ_t is a quadratic function with Hessian μI , we have $\nabla\Phi_t(\bar{v}(t+1)) = \mu(\bar{v}(t+1) - \bar{v}(t))$. Then by (21), we have

$$\begin{aligned} \nabla\Phi_{t+1}(\bar{v}(t+1)) &= (1 - \alpha)\mu(\bar{v}(t+1) - \bar{v}(t)) + \alpha(g(t) + \mu(\bar{v}(t+1) - \bar{y}(t))) \\ &= \mu[\bar{v}(t+1) - (1 - \alpha)\bar{v}(t) - \alpha\bar{y}(t) + \frac{\alpha}{\mu}g(t)] = 0. \end{aligned}$$

where the last equality follows from (9b) and the fact $\frac{\eta}{\alpha} = \frac{\alpha}{\mu}$. Hence Φ_{t+1} achieves its optimum at $\bar{v}(t+1)$. Now we have shown that $\Phi_t(\cdot)$ is a quadratic function that achieves minimum at $\bar{v}(t)$ with Hessian μI . This implies (22) is true. It remains to calculate ϕ_t^* . Clearly, $\phi_0^* = f(\bar{x}(0))$. Setting $\omega = \bar{y}(t)$ in (21), we get $\Phi_{t+1}(\bar{y}(t)) = (1 - \alpha)\Phi_t(\bar{y}(t)) + \alpha\hat{f}(t)$. Combining this with (22), we can get (23). □

Proof of Claim 2: Clearly (24) is true for $t = 0$. Assuming it's true for t , then for $t + 1$, we have by (21) and (10),

$$\Phi_{t+1}(\omega) = (1 - \alpha)\Phi_t(\omega) + \alpha(\hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + \frac{\mu}{2}\|\omega - \bar{y}(t)\|^2)$$

$$\leq (1 - \alpha)\Phi_t(\omega) + \alpha f(\omega) \leq f(\omega) + (1 - \alpha)^{t+1}(\Phi_0(\omega) - f(\omega)).$$

□

The major step of proving Theorem 5 is to show the following inequality,

$$(25) \quad f(\bar{x}(t)) \leq \phi_t^* + O((1 - \alpha)^t).$$

If (25) is true, then combining (25) with (24), we have

$$\begin{aligned} f(\bar{x}(t)) &\leq \phi_t^* + O((1 - \alpha)^t) \leq \Phi_t(x^*) + O((1 - \alpha)^t) \\ &\leq f^* + (1 - \alpha)^t(\Phi_0(x^*) - f(x^*)) + O((1 - \alpha)^t), \end{aligned}$$

which implies $f(\bar{x}(t)) - f^* = O((1 - \alpha)^t)$, the desired result of part (a) of the theorem. We will next prove (25). At last we will show part (b), which is an easy corollary.

Proof of (25): By (23),

$$\begin{aligned} &\phi_{t+1}^* - f(\bar{x}(t+1)) \\ &\stackrel{(a)}{\leq} (1 - \alpha)(\phi_t^* - f(\bar{x}(t))) + (1 - \alpha)f(\bar{x}(t)) + \frac{\mu(1 - \alpha)}{2\alpha} \|\bar{x}(t) - \bar{y}(t)\|^2 \\ &\quad + \alpha \hat{f}(t) - \frac{1}{2}\eta \|g(t)\|^2 + (1 - \alpha)\alpha \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle - f(\bar{x}(t+1)) \\ &\stackrel{(b)}{\geq} (1 - \alpha)(\phi_t^* - f(\bar{x}(t))) + \frac{\mu(1 - \alpha)}{2\alpha} \|\bar{x}(t) - \bar{y}(t)\|^2 \\ &\quad + (1 - \alpha)[\hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle] + \alpha \hat{f}(t) - \frac{1}{2}\eta \|g(t)\|^2 \\ &\quad + (1 - \alpha)\alpha \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle - f(\bar{x}(t+1)) \\ &\stackrel{(c)}{\leq} (1 - \alpha)(\phi_t^* - f(\bar{x}(t))) + \hat{f}(t) - \frac{1}{2}\eta \|g(t)\|^2 + \frac{\mu(1 - \alpha)}{2\alpha} \|\bar{x}(t) - \bar{y}(t)\|^2 - f(\bar{x}(t+1)) \\ (26) \quad &\stackrel{(d)}{\geq} (1 - \alpha)(\phi_t^* - f(\bar{x}(t))) + \left(\frac{1}{2}\eta - \eta^2 L\right) \|g(t)\|^2 + \frac{\mu(1 - \alpha)}{2\alpha} \|\bar{x}(t) - \bar{y}(t)\|^2 - \frac{L}{n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2 \end{aligned}$$

where (a) is due to $\bar{v}(t) - \bar{y}(t) = \frac{1}{\alpha}(\bar{y}(t) - \bar{x}(t))$ and (b) is due to (10) (for $\omega = \bar{x}(t)$). In (c), we have used $(1 - \alpha)\alpha(\bar{v}(t) - \bar{y}(t)) + (1 - \alpha)(\bar{x}(t) - \bar{y}(t)) = (1 - \alpha)[\alpha\bar{v}(t) + \bar{x}(t) - (1 + \alpha)\bar{y}(t)] = 0$. In (d), we have used $f(\bar{x}(t+1)) \leq \hat{f}(t) + (\eta^2 L - \eta) \|g(t)\|^2 + \frac{L}{n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2$, which follows from (11) (for $\omega = \bar{x}(t+1)$). We expand (26) recursively,

$$\begin{aligned} \phi_{t+1}^* - f(\bar{x}(t+1)) &\geq (1 - \alpha)^{t+1}(\phi_0^* - f(\bar{x}(0))) + \left(\frac{1}{2}\eta - \eta^2 L\right) \sum_{k=0}^t (1 - \alpha)^{t-k} \|g(k)\|^2 \\ (27) \quad &\quad + \frac{\mu}{2\alpha} \sum_{k=0}^t (1 - \alpha)^{t+1-k} \|\bar{x}(k) - \bar{y}(k)\|^2 - \frac{L}{n} \sum_{k=0}^t (1 - \alpha)^{t-k} \|y(k) - \mathbf{1}\bar{y}(k)\|^2. \end{aligned}$$

Now we bound $\sum_{k=0}^t (1 - \alpha)^{t-k} \|y(k) - \mathbf{1}\bar{y}(k)\|^2$. Fixing t , define vector $\nu, \pi_k \in \mathbb{R}^{t+1}$ (for $0 \leq k \leq t$),

$$\begin{aligned} \nu &= [A_1(\eta)(1 - \alpha)^{\frac{t}{2}}, A_2(\eta)a(0)(1 - \alpha)^{\frac{t-1}{2}}, A_2(\eta)a(1)(1 - \alpha)^{\frac{t-2}{2}}, \dots, A_2(\eta)a(t-1)]^T \\ \pi_k &= [\theta^k(1 - \alpha)^{-\frac{t}{2}}, \theta^{k-1}(1 - \alpha)^{-\frac{t-1}{2}}, \dots, \theta^0(1 - \alpha)^{-\frac{t-k}{2}}, 0, \dots, 0]^T. \end{aligned}$$

By Lemma 11 (the step size in Theorem 5 implies the condition of Lemma 11 holds), we have $\|y(k) - \mathbf{1}\bar{y}(k)\| \leq \nu^T \pi_k$, and hence $\|y(k) - \mathbf{1}\bar{y}(k)\|^2 \leq \nu^T \pi_k \pi_k^T \nu$. Therefore,

$$(28) \quad \sum_{k=0}^t (1 - \alpha)^{t-k} \|y(k) - \mathbf{1}\bar{y}(k)\|^2 \leq \nu^T \Pi \nu$$

where $\Pi = \sum_{k=0}^t (1-\alpha)^{t-k} \pi_k \pi_k^T \in \mathbb{R}^{(t+1) \times (t+1)}$ is a symmetric matrix. Let Π 's (p, q) 'th element be Π_{pq} . When $q \geq p$, the (p, q) 'th element of $\pi_k \pi_k^T$ is given by $[\pi_k \pi_k^T]_{p,q} = [\pi_k]_p [\pi_k]_q$, which equals $\theta^{2k+2-p-q} (1-\alpha)^{-t-1+\frac{p+q}{2}}$ if $k \geq q-1$, and 0 if $k < q-1$. Then, when $q \geq p$,

$$\Pi_{pq} = \sum_{k=q-1}^t (1-\alpha)^{t-k} \theta^{2k+2-p-q} (1-\alpha)^{-t-1+\frac{p+q}{2}} = \left(\frac{\theta}{\sqrt{1-\alpha}} \right)^{q-p} \frac{1 - \left(\frac{\theta^2}{1-\alpha} \right)^{t-q+2}}{1 - \frac{\theta^2}{1-\alpha}}.$$

By the step size in Theorem 5 we have $\eta < \frac{(1-\sigma)^2}{4\mu}$, and hence $\alpha < \frac{1-\sigma}{2} = 1 - \theta$. Therefore, $\frac{\theta}{\sqrt{1-\alpha}} < \sqrt{\theta} < 1$, and $1 - \frac{\theta}{\sqrt{1-\alpha}} > 1 - \sqrt{\theta} > \frac{1-\theta}{2} = \frac{1-\sigma}{4}$. Therefore,

$$\sum_{q=p+1}^{t+1} \Pi_{pq} < \frac{1}{1 - \frac{\theta^2}{1-\alpha}} \sum_{q=p+1}^{t+1} \left(\frac{\theta}{\sqrt{1-\alpha}} \right)^{q-p} = \frac{1}{1 - \frac{\theta^2}{1-\alpha}} \frac{\theta}{\sqrt{1-\alpha}} \frac{1 - \left(\frac{\theta}{\sqrt{1-\alpha}} \right)^{t-p+1}}{1 - \frac{\theta}{\sqrt{1-\alpha}}} < \frac{16}{(1-\sigma)^2}.$$

And similarly,

$$\sum_{q=1}^p \Pi_{pq} = \sum_{q=1}^p \Pi_{qp} < \frac{1}{1 - \frac{\theta^2}{1-\alpha}} \sum_{q=1}^p \left(\frac{\theta}{\sqrt{1-\alpha}} \right)^{p-q} = \frac{1}{1 - \frac{\theta^2}{1-\alpha}} \frac{1 - \left(\frac{\theta}{\sqrt{1-\alpha}} \right)^p}{1 - \frac{\theta}{\sqrt{1-\alpha}}} < \frac{16}{(1-\sigma)^2}.$$

Hence, by Gershgorin Disk Theorem[11], $\rho(\Pi) \leq \max_p (\sum_{q=1}^{t+1} \Pi_{pq}) \leq 32/(1-\sigma)^2$. Combining the above with (28),

$$\begin{aligned} \sum_{k=0}^t (1-\alpha)^{t-k} \|y(k) - \mathbf{1}\bar{y}(k)\|^2 &\leq \rho(\Pi) \|\nu\|^2 \\ &\leq \frac{32}{(1-\sigma)^2} \left[A_1(\eta)^2 (1-\alpha)^t + A_2(\eta)^2 \sum_{k=0}^{t-1} (1-\alpha)^{t-k-1} a(k)^2 \right] \\ &\leq \frac{32}{(1-\sigma)^2} \left[A_1(\eta)^2 (1-\alpha)^t + 2A_2(\eta)^2 \sum_{k=0}^{t-1} (1-\alpha)^{t-k-1} \|\bar{y}(k) - \bar{x}(k)\|^2 \right. \\ &\quad \left. + 8\eta^2 A_2(\eta)^2 \sum_{k=0}^{t-1} (1-\alpha)^{t-k-1} \|g(k)\|^2 \right] \end{aligned}$$

where in the last step, we have used by definition, $a(k)^2 = (\|\bar{y}(k) - \bar{x}(k)\| + 2\eta \|g(k)\|)^2 \leq 2\|\bar{y}(k) - \bar{x}(k)\|^2 + 8\eta^2 \|g(k)\|^2$. Now returning to (27), we get,

$$\begin{aligned} &\phi_{t+1}^* - f(\bar{x}(t+1)) \\ &\geq -\frac{L}{n} \frac{32A_1(\eta)^2}{(1-\sigma)^2} (1-\alpha)^t + \overbrace{\left(\frac{1}{2}\eta - \eta^2 L - \frac{256L\eta^2 A_2(\eta)^2}{n(1-\sigma)^2(1-\alpha)} \right)}^{\triangleq A_3(\eta)} \sum_{k=0}^t (1-\alpha)^{t-k} \|g(k)\|^2 \\ (29) \quad &+ \frac{1}{\alpha} \underbrace{\left[\frac{\mu}{2} - \frac{64LA_2(\eta)^2 \alpha}{n(1-\sigma)^2(1-\alpha)^2} \right]}_{\triangleq A_4(\eta)} \sum_{k=0}^t (1-\alpha)^{t+1-k} \|\bar{y}(k) - \bar{x}(k)\|^2. \end{aligned}$$

To prove (25), it remains to check $A_3(\eta)$ and $A_4(\eta)$ are positive. Plugging in $A_2(\eta) = \frac{39\sqrt{n}(\eta L)^{1/3}}{(\sigma)^{2/3}}$ into $A_3(\eta)$ and using $(1-\alpha) > \frac{1}{2}$ ($\iff \eta < \frac{1}{4\mu}$), we have

$$A_3(\eta) \geq \eta \left(\frac{1}{2} - \eta L - \frac{512 \times 39^2 (\eta L)^{5/3}}{(1-\sigma)^2 (\sigma)^{4/3}} \right) > 0$$

where in the second inequality we have used by the step size condition in Theorem 5, $\eta L < \frac{1}{4}$, and $\frac{512 \times 39^2 (\eta L)^{5/3}}{(1-\sigma)^2 (\sigma)^{4/3}} < \frac{1}{4}$ ($\Leftarrow \eta < \frac{(1-\sigma)^{1.2} \sigma^{0.8}}{7872L}$). For $A_4(\eta)$, similarly plugging in $A_2(\eta)$ and using $1 - \alpha > \frac{1}{\sqrt{2}}$ ($\Leftarrow \eta < \frac{1}{25\mu}$) and $\alpha = \sqrt{\mu\eta}$, we have

$$A_4(\eta) \geq \frac{\mu}{2} - \frac{128 \times 39^2}{(1-\sigma)^2 \sigma^{4/3}} \sqrt{L\mu} (\eta L)^{7/6} > 0$$

where in the last inequality we have used by the step size condition in Theorem 5, $\frac{\mu}{2} \geq \frac{128 \times 39^2}{(1-\sigma)^2 \sigma^{4/3}} \sqrt{L\mu} (\eta L)^{7/6}$ ($\Leftarrow \eta < \frac{\sigma^{8/7} (1-\sigma)^{12/7}}{61909L} (\frac{\mu}{L})^{3/7}$). \square

At last we will prove part (b) of Theorem 5. Using (24), $\phi_t^* \leq \Phi_t(x^*) \leq f^* + O((1-\alpha)^t)$. Hence $\phi_{t+1}^* - f(\bar{x}(t+1)) \leq \phi_{t+1}^* - f^* = O((1-\alpha)^t)$. Using (29), we have $A_3(\eta) \|g(t)\|^2 + \frac{1-\alpha}{\alpha} A_4(\eta) \|\bar{y}(t) - \bar{x}(t)\|^2 = O((1-\alpha)^t)$. Therefore, both $\|g(t)\|^2$ and $\|\bar{y}(t) - \bar{x}(t)\|^2$ are $O((1-\alpha)^t)$. Then, the $a(t)$ defined in Lemma 11 is $O((1-\alpha)^{t/2})$. By Lemma 11, we also have $\|y(t) - \mathbf{1}\bar{y}(t)\| = O((1-\alpha)^{t/2})$ (where we have used an easy-to-check fact: $\sqrt{1-\alpha} > \theta$). Since f is μ strongly convex, we have $f(\bar{x}(t)) - f^* \geq \frac{\mu}{2} \|\bar{x}(t) - x^*\|^2$ which implies $\|\bar{x}(t) - x^*\| = O((1-\alpha)^{t/2})$. Since $\|\bar{x}(t) - x^*\|$, $\|\bar{y}(t) - \bar{x}(t)\|$ and $\|y(t) - \mathbf{1}\bar{y}(t)\|$ are all $O((1-\alpha)^{t/2})$, by triangle inequality we have $\|y(t) - \mathbf{1}x^*\| = O((1-\alpha)^{t/2})$.

4. Convergence Analysis of Acc-DNGD-NSC. In this section, we will provide the proof of Theorem 6 and Theorem 7. We will first provide a proof overview in Section 4.1 and then defer the detailed proof to the rest of the section.

4.1. Proof Overview. Same as (7), we introduce matrix notations $x(t)$, $v(t)$, $y(t)$, $s(t)$, $\nabla(t) \in \mathbb{R}^{n \times N}$. We also define $\bar{x}(t)$, $\bar{v}(t)$, $\bar{y}(t)$, $\bar{s}(t)$ and $g(t)$ analogously. Then our algorithm in (5) can be written as

$$(30a) \quad x(t+1) = Wy(t) - \eta_t s(t)$$

$$(30b) \quad v(t+1) = Wv(t) - \frac{\eta_t}{\alpha_t} s(t)$$

$$(30c) \quad y(t+1) = (1 - \alpha_{t+1})x(t+1) + \alpha_{t+1}v(t+1)$$

$$(30d) \quad s(t+1) = Ws(t) + \nabla(t+1) - \nabla(t).$$

Overview of the Proof. We derive a series of lemmas (Lemma 15, 16, 17 and 18) that will work for both the vanishing and the fixed step size case. We firstly derive the update formula for the average sequences (Lemma 15). Then, we show that the update rule for the average sequences is in fact CNGD-NSC (3) with inexact gradients [6], and the inexactness is characterized by ‘‘consensus error’’ $\|y(t) - \mathbf{1}\bar{y}(t)\|$ (Lemma 16). The consensus error is bounded in Lemma 17. Then, we apply the proof of CNGD (see e.g. [24]) to the average sequences in spite of the consensus error, and derive an intermediate result in Lemma 18. Lastly, we finish the proof of Theorem 6 and Theorem 7 in Section 4.3 and Section 4.4 respectively.

As shown above, the proof is similar to that of Acc-DNGD-SC in Section 3. The main difference lies in how we bound the consensus error (Lemma 17) and how we apply the CNGD proof in Section 4.3. In what follows, we will mainly focus on the different parts while putting details for the parts that are similar to Acc-DNGD-SC into the Appendix.

LEMMA 15. *The following equalities hold.*

$$(31a) \quad \bar{x}(t+1) = \bar{y}(t) - \eta_t g(t)$$

$$(31b) \quad \bar{v}(t+1) = \bar{v}(t) - \frac{\eta_t}{\alpha_t} g(t)$$

$$(31c) \quad \bar{y}(t+1) = (1 - \alpha_{t+1})\bar{x}(t+1) + \alpha_{t+1}\bar{v}(t+1)$$

$$(31d) \quad \bar{s}(t+1) = \bar{s}(t) + g(t+1) - g(t) = g(t+1)$$

Proof: We omit the proof since these equalities can be easily derived using the fact that W is doubly stochastic and the fact that $\bar{s}(0) = g(0)$. \square

From (31a)-(31c) we see that the sequences $\bar{x}(t)$, $\bar{v}(t)$ and $\bar{y}(t)$ follow a update rule similar to the CNGD-NSC in (3). The only difference is that the $g(t)$ in (31a)-(31c) is not the exact gradient $\nabla f(\bar{y}(t))$ in CNGD-NSC. In the following Lemma, we show that $g(t)$ is an inexact gradient.

LEMMA 16. $\forall t$, $g(t)$ is an inexact gradient of f at $\bar{y}(t)$ with error $O(\|y(t) - \mathbf{1}\bar{y}(t)\|^2)$ in the sense that, $\forall \omega \in \mathbb{R}^N$,¹¹

$$(32) \quad f(\omega) \geq \hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle$$

$$(33) \quad f(\omega) \leq \hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle + L\|\omega - \bar{y}(t)\|^2 + L\frac{1}{n}\|y(t) - \mathbf{1}\bar{y}(t)\|^2,$$

where $\hat{f}(t) = \frac{1}{n} \sum_{i=1}^n [f_i(y_i(t)) + \langle \nabla f_i(y_i(t)), \bar{y}(t) - y_i(t) \rangle]$.

Proof: We omit the proof since it's almost identical as the proof of Lemma 10. \square

The consensus error $\|y(t) - \mathbf{1}\bar{y}(t)\|$ in the previous lemma is bounded by the following lemma whose proof is given in Section 4.2.

LEMMA 17. Suppose the step sizes satisfy

- (i) $\eta_t \geq \eta_{t+1} > 0$,
- (ii) $\eta_0 < \min(\frac{\sigma^2}{9^3 L}, \frac{(1-\sigma)^3}{6144 L})$,
- (iii) $\sup_{t \geq 0} \frac{\eta_t}{\eta_{t+1}} \leq \min((\frac{\sigma+3}{\sigma+2} \frac{3}{4})^{\sigma/28}, \frac{16}{15+\sigma})$.

Then, we have,

$$\|y(t) - \mathbf{1}\bar{y}(t)\| \leq \kappa \sqrt{n} \chi_2(\eta_t) \left[L\|\bar{y}(t) - \bar{x}(t)\| + \frac{8}{1-\sigma} L \eta_t \|g(t)\| \right]$$

where $\chi_2 : \mathbb{R} \rightarrow \mathbb{R}$ is a function satisfying $0 < \chi_2(\eta_t) \leq \frac{2}{L^{2/3}} \eta_t^{1/3}$, and $\kappa = \frac{6}{(1-\sigma)}$.

We next provide the following intermediate result.

LEMMA 18. Define $\gamma_0 = \frac{\alpha_0^2}{\eta_0(1-\alpha_0)} = \frac{L}{1-\alpha_0}$. We define a series of functions $(\Phi_t : \mathbb{R}^N \rightarrow \mathbb{R})_{t \geq 0}$, with $\Phi_0(\omega) = f(\bar{x}(0)) + \frac{\gamma_0}{2} \|\omega - \bar{v}(0)\|^2$ and

$$(34) \quad \Phi_{t+1}(\omega) = (1 - \alpha_t)\Phi_t(\omega) + \alpha_t[\hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle].$$

Then, we have,

$$(35) \quad \Phi_t(\omega) \leq f(\omega) + \lambda_t(\Phi_0(\omega) - f(\omega))$$

where λ_t is defined through $\lambda_0 = 1$, and $\lambda_{t+1} = (1 - \alpha_t)\lambda_t$. Further, we have function $\Phi_t(\omega)$ can be written as

$$(36) \quad \Phi_t(\omega) = \phi_t^* + \frac{\gamma_t}{2} \|\omega - \bar{v}(t)\|^2$$

¹¹The exact gradient $\nabla f(\bar{y}(t))$, satisfies $f(\bar{y}(t)) + \langle \nabla f(\bar{y}(t)), \omega - \bar{y}(t) \rangle \leq f(\omega) \leq f(\bar{y}(t)) + \langle \nabla f(\bar{y}(t)), \omega - \bar{y}(t) \rangle + \frac{L}{2} \|\omega - \bar{y}(t)\|^2$. This is why we call $g(t)$ inexact gradient.

where γ_t is defined through $\gamma_{t+1} = \gamma_t(1 - \alpha_t)$, and ϕ_t^* is some real number that satisfies $\phi_0^* = f(\bar{x}(0))$, and

$$(37) \quad \phi_{t+1}^* = (1 - \alpha_t)\phi_t^* + \alpha_t \hat{f}(t) - \frac{1}{2}\eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle.$$

Proof: Its proof is almost the same as the derivation of (22) (23) (24) in the proof of Theorem 5. For completeness we include a proof in Appendix-B.4. \square

4.2. Proof of the Bounded Consensus Error (Lemma 17). We will frequently use the following lemma, whose proofs can be found in Appendix-B.1. We will also use Lemma 13 in Section 3.2, which still holds under the setting of this section.

LEMMA 19. *The following equalities are true.*

$$(38) \quad \bar{y}(t+1) - \bar{y}(t) = \alpha_{t+1}(\bar{v}(t) - \bar{y}(t)) - \eta_t \left[\frac{\alpha_{t+1}}{\alpha_t} + 1 - \alpha_{t+1} \right] g(t)$$

$$(39) \quad \bar{v}(t+1) - \bar{y}(t+1) = (1 - \alpha_{t+1})(\bar{v}(t) - \bar{y}(t)) + \eta_t(1 - \alpha_{t+1})\left(1 - \frac{1}{\alpha_t}\right)g(t)$$

Proof of Lemma 17:

Overview of the proof. The proof is divided into three steps. In step 1, we treat the algorithm (30) as a linear system and derive a linear system inequality (40). In step 2, we analyze the state transition matrix in (40) and prove a few spectral properties. In step 3, we further analyze the linear system (40) and bound the state by the input, from which the conclusion of the lemma follows. Throughout the proof, we will frequently use an easy-to-check fact: α_t is a decreasing sequence.

Step 1: A Linear System Inequality. Define $z(t) = [\alpha_t \|v(t) - \mathbf{1}\bar{v}(t)\|, \|y(t) - \mathbf{1}\bar{y}(t)\|, \|s(t) - \mathbf{1}g(t)\|]^T \in \mathbb{R}^3$, $b(t) = [0, 0, \sqrt{n}a(t)]^T \in \mathbb{R}^3$ where

$$a(t) \triangleq \alpha_t L \|\bar{v}(t) - \bar{y}(t)\| + 2\lambda L \eta_t \|g(t)\|$$

in which $\lambda \triangleq \frac{4}{1-\sigma} > 1$. Then, we have the following linear system inequality holds.

$$(40) \quad z(t+1) \leq \overbrace{\begin{bmatrix} \sigma & 0 & \eta_t \\ \sigma & \sigma & 2\eta_t \\ L & 2L & \sigma + 2\eta_t L \end{bmatrix}}^{\triangleq G(\eta_t)} z(t) + b(t)$$

The derivation of (40) is almost the same as that of (15) (in the proof of Lemma 11). Due to space limit, we omit the derivation here. It can be found in Appendix-B.2.

Step 2: Spectral Properties of $G(\cdot)$. When η is positive, $G(\eta)$ is a nonnegative matrix and $G(\eta)^2$ is a positive matrix. By Perron-Frobenius Theorem ([12, Theorem 8.5.1]) $G(\eta)$ has a unique largest (in magnitude) eigenvalue that is a positive real with multiplicity 1, and the eigenvalue is associated with an eigenvector with positive entries. We let the unique largest eigenvalue be $\theta(\eta) = \rho(G(\eta))$ and let its eigenvector be $\chi(\eta) = [\chi_1(\eta), \chi_2(\eta), \chi_3(\eta)]^T$, normalized by $\chi_3(\eta) = 1$. We give bounds on the eigenvalue and the eigenvector in the following lemmas. Due to space limit, we only provide a proof-sketch, while the complete proof can be found in Appendix-B.3.

LEMMA 20. *When $0 < \eta L < 1$, we have $\sigma < \theta(\eta) < \sigma + 4(\eta L)^{1/3}$, and $\chi_2(\eta) \leq \frac{2}{L^{2/3}}\eta^{1/3}$.*

LEMMA 21. *When $\eta \in (0, \frac{\sqrt{\sigma}}{L2\sqrt{2}})$, $\theta(\eta) \geq \sigma + (\sigma\eta L)^{1/3}$ and $\chi_1(\eta) < \frac{\eta}{(\sigma\eta L)^{1/3}}$.*

LEMMA 22. When $0 < \zeta_2 < \zeta_1 < \frac{\sigma^2}{9^3 L}$, then $\frac{\chi_1(\zeta_1)}{\chi_1(\zeta_2)} \leq (\frac{\zeta_1}{\zeta_2})^{6/\sigma}$ and $\frac{\chi_2(\zeta_1)}{\chi_2(\zeta_2)} \leq (\frac{\zeta_1}{\zeta_2})^{28/\sigma}$.

Proof-Sketch: Lemma 20 and Lemma 21 essentially say that when η is small, $\theta(\eta) = \rho(G(\eta))$ should be close to σ , and $\chi_2(\eta)$, $\chi_1(\eta)$ should be close to 0. These can be seen through the fact that $G(0)$'s all three eigenvalues are σ , with eigenvector $[0, 0, 1]^T$. Therefore, when η is small, $G(\eta)$'s all eigenvalues should be close to σ , and eigenvectors "close" to $[0, 0, 1]^T$. Detailed calculations in the Appendix lead to the bound in Lemma 20 and Lemma 21. To prove Lemma 22, we essentially need to prove that both function $\chi_1(\eta)$ and $\chi_2(\eta)$ behave asymptotically similar to functions of the type $c\eta^\delta$ for some $c, \delta > 0$. Hence $\frac{\chi_1(\zeta_1)}{\chi_1(\zeta_2)} \approx (\frac{\zeta_1}{\zeta_2})^\delta$ and similarly for $\chi_2(\cdot)$. \square

It is easy to check that, under our step size condition (ii) in Lemma 17, all the conditions of Lemma 20, 21, 22 are satisfied.

Step 3: Bound the state by the input. With the above preparations, now we prove, by induction, the following statement,

$$(41) \quad z(t) \leq \sqrt{n}a(t)\kappa\chi(\eta_t)$$

where $\kappa = \frac{6}{1-\sigma}$. Eq. (41) is true for $t = 0$, since the left hand side is zero when $t = 0$.

Assume (41) holds for t . We now show (41) is true for $t + 1$. We divide the rest of the proof into two sub-steps. Briefly speaking, step 3.1 proves that the input to the system (40), $a(t+1)$ does not decrease too much compared to $a(t)$ ($a(t+1) \geq \frac{\sigma+3}{4}a(t)$); while step 3.2 shows that the state $z(t+1)$, compared to $z(t)$, decreases enough for (41) to hold for $t + 1$.

Step 3.1: We prove that $a(t+1) \geq \frac{\sigma+3}{4}a(t)$. By (39),

$$\begin{aligned} a(t+1) &= \alpha_{t+1}L\|\bar{v}(t+1) - \bar{y}(t+1)\| + 2\lambda\eta_{t+1}L\|g(t+1)\| \\ &= \alpha_{t+1}L\|(1 - \alpha_{t+1})(\bar{v}(t) - \bar{y}(t)) \\ &\quad + (1 - \alpha_{t+1})(1 - \frac{1}{\alpha_t})\eta_t g(t)\| + 2\lambda\eta_{t+1}L\|g(t+1)\| \\ &\geq \alpha_{t+1}(1 - \alpha_{t+1})L\|\bar{v}(t) - \bar{y}(t)\| - \frac{\alpha_{t+1}}{\alpha_t}(1 - \alpha_{t+1})(1 - \alpha_t)\eta_t L\|g(t)\| \\ &\quad + 2\lambda\eta_{t+1}L\|g(t)\| - 2\lambda\eta_{t+1}L\|g(t+1) - g(t)\|. \end{aligned}$$

Therefore, recalling $a(t) = \alpha_t L\|\bar{v}(t) - \bar{y}(t)\| + 2\lambda L\eta_t\|g(t)\|$, we have

$$\begin{aligned} a(t) - a(t+1) &\leq \left[\alpha_t - \alpha_{t+1}(1 - \alpha_{t+1}) \right] L\|\bar{v}(t) - \bar{y}(t)\| + \left[\frac{\alpha_{t+1}}{\alpha_t}(1 - \alpha_{t+1})(1 - \alpha_t)\eta_t L \right. \\ &\quad \left. + 2\lambda\eta_t L - 2\lambda\eta_{t+1}L \right] \|g(t)\| + 2\lambda\eta_{t+1}L\|g(t+1) - g(t)\| \\ &\leq \left[\alpha_t - \alpha_{t+1}(1 - \alpha_{t+1}) \right] L\|\bar{v}(t) - \bar{y}(t)\| + (\eta_t + 2\lambda(\eta_t - \eta_{t+1}))L\|g(t)\| \\ &\quad + 2\lambda\eta_{t+1}L\|g(t+1) - g(t)\| \\ (42) \quad &\leq \max\left(1 - \frac{\alpha_{t+1}}{\alpha_t} + \frac{\alpha_{t+1}^2}{\alpha_t}, \frac{1}{2\lambda} + \frac{\eta_t - \eta_{t+1}}{\eta_t}\right)a(t) + 2\lambda\eta_{t+1}L\|g(t+1) - g(t)\| \end{aligned}$$

where in the last inequality, we have used the elementary fact that for four positive numbers a_1, a_2, a_3, a_4 and $x, y \geq 0$, we have $a_1x + a_2y = \frac{a_1}{a_3}a_3x + \frac{a_2}{a_4}a_4y \leq \max(\frac{a_1}{a_3}, \frac{a_2}{a_4})(a_3x + a_4y)$.

Next, we expand $\|g(t+1) - g(t)\|$,

$$\begin{aligned} \|g(t+1) - g(t)\| &\leq \|g(t+1) - \nabla f(\bar{y}(t+1))\| + \|g(t) - \nabla f(\bar{y}(t))\| + \|\nabla f(\bar{y}(t+1)) - \nabla f(\bar{y}(t))\| \\ &\stackrel{(a)}{\leq} \frac{L}{\sqrt{n}}\|y(t+1) - \mathbf{1}\bar{y}(t+1)\| + \frac{L}{\sqrt{n}}\|y(t) - \mathbf{1}\bar{y}(t)\| + L\|\bar{y}(t+1) - \bar{y}(t)\| \end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{\leq} \frac{L}{\sqrt{n}} \sigma \alpha_t \|v(t) - \mathbf{1}\bar{v}(t)\| + \frac{L}{\sqrt{n}} 2\|y(t) - \mathbf{1}\bar{y}(t)\| + \frac{L}{\sqrt{n}} 2\eta_t \|s(t) - \mathbf{1}g(t)\| + a(t) \\
&\stackrel{(c)}{\leq} L\sigma\kappa\chi_1(\eta_t)a(t) + 2L\kappa\chi_2(\eta_t)a(t) + 2L\eta_t\kappa\chi_3(\eta_t)a(t) + a(t) \\
(43) \quad &\stackrel{(d)}{\leq} a(t) \left\{ L\sigma\kappa \frac{\eta_t}{(\sigma\eta_t L)^{1/3}} + 2L\kappa \frac{2\eta_t^{1/3}}{L^{2/3}} + 2L\eta_t\kappa + 1 \right\} \stackrel{(e)}{\leq} 8\kappa a(t).
\end{aligned}$$

Here (a) is due to (14); (b) is due to the second row of (40) and the fact that $a(t) \geq L\|\bar{y}(t+1) - \bar{y}(t)\|$ (cf. (38)); (c) is due to the induction assumption (41). In (d), we have used the bound on $\chi_1(\cdot)$ (Lemma 21), $\chi_2(\cdot)$ (Lemma 20), and $\chi_3(\eta_t) = 1$. In (e), we have used $\eta_t L < 1$, $\sigma < 1$ and $\kappa > 1$.

Combining (43) with (42) and recalling $\kappa = \frac{6}{1-\sigma}$, $\lambda = \frac{4}{1-\sigma}$, we have

$$\begin{aligned}
a(t) - a(t+1) &\leq \max\left(1 - \frac{\alpha_{t+1}}{\alpha_t} + \frac{\alpha_{t+1}^2}{\alpha_t}, \frac{1}{2\lambda} + \frac{\eta_t - \eta_{t+1}}{\eta_t}\right) a(t) + 16\kappa\lambda\eta_{t+1}La(t) \\
&\leq \left[\max\left(1 - \frac{\eta_{t+1}}{\eta_t} + 2\alpha_{t+1}, \frac{1-\sigma}{8} + \frac{\eta_t - \eta_{t+1}}{\eta_t}\right) + \frac{384}{(1-\sigma)^2}\eta_0L \right] a(t)
\end{aligned}$$

where in the last inequality, we have used the fact that

$$1 - \frac{\alpha_{t+1}}{\alpha_t} + \frac{\alpha_{t+1}^2}{\alpha_t} < 1 - \frac{\alpha_{t+1}^2}{\alpha_t^2} + \alpha_{t+1} = 1 - \frac{\eta_{t+1}}{\eta_t}(1 - \alpha_{t+1}) + \alpha_{t+1} < 1 - \frac{\eta_{t+1}}{\eta_t} + 2\alpha_{t+1}$$

where the equality follows from the update rule for α_t (6). By the step size condition (iii) in Lemma 17, $\frac{\eta_t}{\eta_{t+1}} \leq \frac{16}{15+\sigma}$, and hence $1 - \frac{\eta_{t+1}}{\eta_t} \leq \frac{1-\sigma}{16}$. By the step size condition (ii), $2\alpha_{t+1} \leq 2\alpha_0 = 2\sqrt{\eta_0 L} \leq \frac{1-\sigma}{16}$, and $\eta_0 L \frac{384}{(1-\sigma)^2} < \frac{1-\sigma}{16}$. Combining the above, we have $a(t) - a(t+1) \leq \frac{1-\sigma}{4}a(t)$. Hence $a(t+1) \geq \frac{3+\sigma}{4}a(t)$.

Step 3.2: Finishing the induction. We have,

$$\begin{aligned}
(44) \quad z(t+1) &\stackrel{(a)}{\leq} G(\eta_t)z(t) + b(t) \\
&\stackrel{(b)}{\leq} G(\eta_t)\sqrt{n}a(t)\kappa\chi(\eta_t) + \sqrt{n}a(t)\chi(\eta_t) \\
&\stackrel{(c)}{=} \theta(\eta_t)\sqrt{n}a(t)\kappa\chi(\eta_t) + \sqrt{n}a(t)\chi(\eta_t) \\
&= \sqrt{n}a(t)\chi(\eta_t)(\kappa\theta(\eta_t) + 1) \\
&\stackrel{(d)}{\leq} \sqrt{n}a(t+1)\chi(\eta_{t+1})\left(\kappa\frac{\sigma+1}{2} + 1\right)\frac{4}{3+\sigma} \times \max\left(\frac{\chi_1(\eta_t)}{\chi_1(\eta_{t+1})}, \frac{\chi_2(\eta_t)}{\chi_2(\eta_{t+1})}, 1\right) \\
&\stackrel{(e)}{=} \sqrt{n}a(t+1)\chi(\eta_{t+1})\frac{\sigma+2}{3}\kappa\frac{4}{\sigma+3} \times \max\left(\frac{\chi_1(\eta_t)}{\chi_1(\eta_{t+1})}, \frac{\chi_2(\eta_t)}{\chi_2(\eta_{t+1})}, 1\right) \\
&\stackrel{(f)}{\leq} \sqrt{n}a(t+1)\kappa\chi(\eta_{t+1})
\end{aligned}$$

where (a) is due to (40), and (b) is due to induction assumption (41), and (c) is because $\theta(\eta_t)$ is an eigenvalue of $G(\eta_t)$ with eigenvector $\chi(\eta_t)$, and (d) is due to step 3.1, and $\theta(\eta_t) < \sigma + 4(\eta_0 L)^{1/3} < \frac{1+\sigma}{2}$ (by Lemma 20 and step size condition (ii) in Lemma 17), and in (e), we have used the fact $\kappa\frac{\sigma+1}{2} + 1 = \frac{\sigma+2}{3}\kappa$ (since $\kappa = \frac{6}{1-\sigma}$). For (f), we have used that by Lemma 22 and step size condition (iii) (in Lemma 17),

$$\max\left(\frac{\chi_1(\eta_t)}{\chi_1(\eta_{t+1})}, \frac{\chi_2(\eta_t)}{\chi_2(\eta_{t+1})}, 1\right) \leq \left(\frac{\eta_t}{\eta_{t+1}}\right)^{28/\sigma} \leq \frac{\sigma+3}{\sigma+2} \frac{3}{4}.$$

Now, (41) is proven for $t+1$, and hence is true for all t . Therefore, we have

$$\|y(t) - \mathbf{1}\bar{y}(t)\| \leq \kappa\sqrt{n}\chi_2(\eta_t)a(t).$$

Notice that $a(t) = \alpha_t L\|\bar{v}(t) - \bar{y}(t)\| + 2\lambda L\eta_t\|g(t)\| \leq L\|\bar{y}(t) - \bar{x}(t)\| + \frac{8}{1-\sigma}L\eta_t\|g(t)\|$ (by $\alpha_t(\bar{v}(t) - \bar{y}(t)) = (1 - \alpha_t)(\bar{y}(t) - \bar{x}(t))$). The statement of the lemma follows. \square

4.3. Proof of Theorem 6. We first introduce Lemma 23 regarding the asymptotic behavior of α_t and λ_t .

LEMMA 23. *When the vanishing step size is used ($\eta_t = \frac{\eta}{(t+t_0)^\beta}$, $t_0 \geq 1$, $\beta \in (0, 2)$), and $\eta_0 < \frac{1}{4L}$ (equivalently $\alpha_0 < \frac{1}{2}$), we have*

- (i) $\alpha_t \leq \frac{2}{t+1}$.
- (ii) $\lambda_t = O(\frac{1}{t^{2-\beta}})$.
- (iii) $\lambda_t \geq \frac{D(\beta, t_0)}{(t+t_0)^{2-\beta}}$ where $D(\beta, t_0)$ is some constant that only depends on β and t_0 , given by $D(\beta, t_0) = \frac{1}{(t_0+3)^2 e^{16 + \frac{6}{2-\beta}}}$.

Proof-Sketch: Due to space limit, we provide a proof sketch here, while the complete proof can be found in Appendix-B.5. First through the update rule (6) it is not hard to check $\alpha_t \rightarrow 0$. Then, we can calculate $\frac{1}{\alpha_{t+1}} - \sqrt{\frac{\eta_t}{\eta_{t+1}}} \frac{1}{\alpha_t} = \frac{1}{2} + \frac{1}{4} \frac{1}{\frac{1}{2} \sqrt{1+4\frac{\eta_t}{\eta_{t+1}\alpha_t^2}} + \sqrt{\frac{\eta_t}{\eta_{t+1}}} \frac{1}{\alpha_t}} \rightarrow \frac{1}{2}$. This, coupled with the fact that $\sqrt{\frac{\eta_t}{\eta_{t+1}}} \approx (1 + \frac{1}{t})^{\beta/2} \approx 1 + \frac{\beta}{2t}$, would lead to $\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} - \frac{\beta}{2t} \frac{1}{\alpha_t} \approx \frac{1}{2}$. This implies $\frac{1}{\alpha_t} \approx t \frac{1}{2-\beta}$ and hence $\alpha_t \approx \frac{2-\beta}{t}$. So we have shown part (i). For (ii) and (iii), notice that $\log \lambda_t = \sum_{k=0}^{t-1} \log(1 - \alpha_k) \approx -\sum_{k=0}^{t-1} \alpha_k \approx -(2-\beta) \sum_{k=0}^{t-1} \frac{1}{k} \approx -(2-\beta) \log t$. Therefore $\lambda_t \approx \Theta(\frac{1}{t^{2-\beta}})$. \square

Now we proceed to prove Theorem 6.

Proof of Theorem 6: It is easy to check that with the step size condition in Theorem 6, all the conditions of Lemma 17 and 23 are satisfied, hence the conclusions of Lemma 17 and 23 hold. The major step of proving Theorem 6 is to show the following inequality,

$$(45) \quad \lambda_t(\Phi_0(x^*) - f^*) + \phi_t^* \geq f(\bar{x}(t)).$$

If (45) is true, by (45) and (35), we have

$$f(\bar{x}(t)) \leq \phi_t^* + \lambda_t(\Phi_0(x^*) - f^*) \leq \Phi_t(x^*) + \lambda_t(\Phi_0(x^*) - f^*) \leq f^* + 2\lambda_t(\Phi_0(x^*) - f^*).$$

Hence $f(\bar{x}(t)) - f^* = O(\lambda_t) = O(\frac{1}{t^{2-\beta}})$, i.e. the desired result of Theorem 6.

Now we use induction to prove (45). Firstly, (45) is true for $t = 0$, since $\phi_0^* = f(\bar{x}(0))$ and $\Phi_0(x^*) \geq f(\bar{x}(0)) \geq f^*$. Suppose it's true for $0, 1, 2, \dots, t$. For $0 \leq k \leq t$, by (35), $\Phi_k(x^*) \leq f^* + \lambda_k(\Phi_0(x^*) - f^*)$. Combining the above with (36),

$$\phi_k^* + \frac{\gamma_k}{2} \|x^* - \bar{v}(k)\|^2 \leq f^* + \lambda_k(\Phi_0(x^*) - f^*).$$

Using the induction assumption, we get

$$(46) \quad f(\bar{x}(k)) + \frac{\gamma_k}{2} \|x^* - \bar{v}(k)\|^2 \leq f^* + 2\lambda_k(\Phi_0(x^*) - f^*).$$

Since $f(\bar{x}(k)) \geq f^*$ and $\gamma_k = \lambda_k \gamma_0$, we have $\|x^* - \bar{v}(k)\|^2 \leq \frac{4}{\gamma_0} (\Phi_0(x^*) - f^*)$. Since $\bar{v}(k) = \frac{1}{\alpha_k} (\bar{y}(k) - \bar{x}(k)) + \bar{x}(k)$, we have $\|\bar{v}(k) - x^*\|^2 = \|\frac{1}{\alpha_k} (\bar{y}(k) - \bar{x}(k)) + \bar{x}(k) - x^*\|^2 \geq \frac{1}{2\alpha_k^2} \|\bar{y}(k) - \bar{x}(k)\|^2 - \|\bar{x}(k) - x^*\|^2$. By (46), $f(\bar{x}(k)) \leq 2\Phi_0(x^*) - f^* = 2f(\bar{x}(0)) - f^* + \gamma_0 \|\bar{v}(0) - x^*\|^2$. Also since $\gamma_0 = \frac{L}{1-\alpha_0} < 2L$, we have $\bar{x}(k)$ lies within the $(2f(\bar{x}(0)) - f^* + 2L\|\bar{v}(0) - x^*\|^2)$ -level set of f . By Assumption 2 and [3, Proposition B.9], we have the level set is compact. Hence we have $\|\bar{x}(k) - x^*\| \leq R$ where R is the diameter of that level set. Combining the above arguments, we get

$$\|\bar{y}(k) - \bar{x}(k)\|^2 \leq 2\alpha_k^2 (\|\bar{v}(k) - x^*\|^2 + \|\bar{x}(k) - x^*\|^2)$$

$$\begin{aligned}
&\leq 2\alpha_k^2 \left(\frac{4}{\gamma_0} (\Phi_0(x^*) - f^*) + R^2 \right) \\
&= 2\alpha_k^2 \left[\frac{4}{\gamma_0} (f(\bar{x}(0)) - f^*) + 2\|\bar{v}(0) - x^*\|^2 + R^2 \right] \\
(47) \quad &\leq 2\alpha_k^2 \underbrace{[4\|\bar{v}(0) - x^*\|^2 + R^2]}_{\triangleq C_1}
\end{aligned}$$

where C_1 is a constant that does *not* depend on η , and in the last inequality we have used by the L -smoothness of f , $f(\bar{x}(0)) - f^* \leq \frac{L}{2}\|\bar{x}(0) - x^*\|^2 \leq \frac{\gamma_0}{2}\|\bar{x}(0) - x^*\|^2$.

Next, we consider (37),

$$\begin{aligned}
\phi_{t+1}^* &= (1 - \alpha_t)\phi_t^* + \alpha_t \hat{f}(t) - \frac{1}{2}\eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle \\
&= (1 - \alpha_t)(\phi_t^* - f(\bar{x}(t))) + (1 - \alpha_t)f(\bar{x}(t)) + \alpha_t \hat{f}(t) \\
&\quad - \frac{1}{2}\eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle \\
&\stackrel{(a)}{\geq} (1 - \alpha_t)(\phi_t^* - f(\bar{x}(t))) + (1 - \alpha_t)\{ \hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle \} \\
&\quad + \alpha_t \hat{f}(t) - \frac{1}{2}\eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle \\
(48) \quad &\stackrel{(b)}{=} (1 - \alpha_t)(\phi_t^* - f(\bar{x}(t))) + \hat{f}(t) - \frac{1}{2}\eta_t \|g(t)\|^2
\end{aligned}$$

where (a) is due to (32) and (b) is due to $\alpha_t(\bar{v}(t) - \bar{y}(t)) + (1 - \alpha_t)(\bar{x}(t) - \bar{y}(t)) = 0$.

By (33) (setting $\omega = \bar{x}(t+1)$) and Lemma 17,

$$\begin{aligned}
f(\bar{x}(t+1)) &\leq \hat{f}(t) + \langle g(t), \bar{x}(t+1) - \bar{y}(t) \rangle + L\|\bar{x}(t+1) - \bar{y}(t)\|^2 + \frac{L}{n}\|y(t) - \mathbf{1}\bar{y}(t)\|^2 \\
(49) \quad &\leq \hat{f}(t) - (\eta_t - L\eta_t^2)\|g(t)\|^2 + 2L\kappa^2\chi_2(\eta_t)^2[L^2\|\bar{y}(t) - \bar{x}(t)\|^2 + \frac{64}{(1-\sigma)^2}L^2\eta_t^2\|g(t)\|^2].
\end{aligned}$$

Combining the above with (48) and recalling $\kappa = \frac{6}{1-\sigma}$, we get,

$$\begin{aligned}
\phi_{t+1}^* - f(\bar{x}(t+1)) &\geq (1 - \alpha_t)(\phi_t^* - f(\bar{x}(t))) + \left(\frac{1}{2}\eta_t - L\eta_t^2 - \frac{4608L^3\chi_2(\eta_t)^2\eta_t^2}{(1-\sigma)^4} \right) \|g(t)\|^2 \\
&\quad - 2\kappa^2\chi_2(\eta_t)^2L^3\|\bar{y}(t) - \bar{x}(t)\|^2 \\
(50) \quad &\geq (1 - \alpha_t)(\phi_t^* - f(\bar{x}(t))) - 2\kappa^2\chi_2(\eta_t)^2L^3\|\bar{y}(t) - \bar{x}(t)\|^2
\end{aligned}$$

where in the last inequality we have used that, recalling $\chi_2(\eta_t) \leq \frac{2}{L^{2/3}}\eta_t^{1/3}$,

$$\begin{aligned}
\frac{1}{2}\eta_t - L\eta_t^2 - \frac{4608L^3\chi_2(\eta_t)^2\eta_t^2}{(1-\sigma)^4} &\geq \frac{1}{2}\eta_t - L\eta_t^2 - \frac{4608L^3\eta_t^2}{(1-\sigma)^4} \frac{4\eta_t^{2/3}}{L^{4/3}} \\
&= \frac{1}{2}\eta_t - L\eta_t^2 - \frac{18432\eta_t(L\eta_t)^{5/3}}{(1-\sigma)^4} \\
&\geq \eta_t(1/2 - \eta L - \frac{18432(L\eta)^{5/3}}{(1-\sigma)^4}) > 0
\end{aligned}$$

where the last inequality follows from $\eta L < \frac{1}{4}$, and $\frac{18432(L\eta)^{5/3}}{(1-\sigma)^4} < \frac{1}{4}$ ($\Leftarrow \eta L < \frac{(1-\sigma)^{2.4}}{833}$) cf. step size condition (ii) in Theorem 6). Next, expanding (50) recursively, we get

$$\phi_{t+1}^* - f(\bar{x}(t+1)) \geq \prod_{k=0}^t (1 - \alpha_k)(\phi_0^* - f(\bar{x}(0))) - \sum_{k=0}^t 2\kappa^2\chi_2(\eta_k)^2L^3\|\bar{y}(k) - \bar{x}(k)\|^2 \prod_{\ell=k+1}^t (1 - \alpha_\ell)$$

$$= - \sum_{k=0}^t 2\kappa^2 \chi_2(\eta_k)^2 L^3 \|\bar{y}(k) - \bar{x}(k)\|^2 \prod_{\ell=k+1}^t (1 - \alpha_\ell).$$

Therefore to finish the induction, we need to show

$$\sum_{k=0}^t 2\kappa^2 \chi_2(\eta_k)^2 L^3 \|\bar{y}(k) - \bar{x}(k)\|^2 \prod_{\ell=k+1}^t (1 - \alpha_\ell) \leq (\Phi_0(x^*) - f^*) \lambda_{t+1}.$$

Notice that

$$\begin{aligned} \sum_{k=0}^t \frac{2\kappa^2 \chi_2(\eta_k)^2 L^3 \|\bar{y}(k) - \bar{x}(k)\|^2 \prod_{\ell=k+1}^t (1 - \alpha_\ell)}{(\Phi_0(x^*) - f^*) \lambda_{t+1}} &\stackrel{(a)}{\leq} \sum_{k=0}^t \frac{4\kappa^2 L^2}{\|\bar{v}(0) - x^*\|^2} \chi_2(\eta_k)^2 \|\bar{y}(k) - \bar{x}(k)\|^2 \frac{1}{\lambda_{k+1}} \\ &\stackrel{(b)}{\leq} \sum_{k=0}^t \frac{4(6/(1-\sigma))^2 L^2}{\|\bar{v}(0) - x^*\|^2} \left(\frac{2}{L^{2/3}} \eta_k^{1/3}\right)^2 2C_1 \alpha_k^2 \frac{1}{\lambda_{k+1}} \\ &= \sum_{k=0}^t \underbrace{\frac{1152L^{2/3}C_1}{(1-\sigma)^2 \|\bar{v}(0) - x^*\|^2}}_{\triangleq C_2} \eta_k^{2/3} \alpha_k^2 \frac{1}{\lambda_{k+1}} \end{aligned}$$

where C_2 is a constant that does *not* depend on η , and in (a) we have used $\Phi_0(x^*) - f^* \geq \frac{L}{2} \|\bar{v}(0) - x^*\|^2 > 0$, and $\prod_{\ell=k+1}^t (1 - \alpha_\ell) = \lambda_{t+1} / \lambda_{k+1}$; in (b), we have plugged in $\kappa = \frac{6}{1-\sigma}$, used $\chi_2(\eta_k) \leq \frac{2\eta_k^{1/3}}{L^{2/3}}$ (cf. Lemma 17) and the bound on $\|\bar{y}(k) - \bar{x}(k)\|^2$ (equation (47)). Now by Lemma 23, we get,

$$\begin{aligned} \sum_{k=0}^t \eta_k^{2/3} \alpha_k^2 \frac{1}{\lambda_{k+1}} &\leq \sum_{k=0}^t \frac{\eta^{2/3}}{(k+t_0)^{\frac{2}{3}\beta}} \frac{4}{(k+1)^2} \frac{(k+1+t_0)^{2-\beta}}{D(\beta, t_0)} \\ &\stackrel{(a)}{\leq} \eta^{2/3} \frac{4(t_0+1)^{2-\beta}}{D(\beta, t_0)} \sum_{k=0}^{\infty} \frac{1}{(k+1)^{\frac{5}{3}\beta}} \\ &\stackrel{(b)}{\leq} \eta^{2/3} \frac{4(t_0+1)^{2-\beta}}{D(\beta, t_0)} \times \frac{2}{\beta - 0.6} \end{aligned}$$

where in (a) we have used, $k+t_0 \geq k+1$, $k+1+t_0 \leq (t_0+1)(k+1)$; in (b) we have used $\frac{5}{3}\beta > 1$. So, we have

$$\sum_{k=0}^t \frac{2\kappa^2 \chi_2(\eta_k)^2 L^3 \|\bar{y}(k) - \bar{x}(k)\|^2 \prod_{\ell=k+1}^t (1 - \alpha_\ell)}{(\Phi_0(x^*) - f^*) \lambda_{t+1}} \leq C_2 \eta^{2/3} \frac{8(t_0+1)^{2-\beta}}{D(\beta, t_0)(\beta - 0.6)} < 1$$

where in the last inequality, we have simply required $\eta^{2/3} < \frac{D(\beta, t_0)(\beta - 0.6)}{8(t_0+1)^{2-\beta} C_2}$ (i.e. step size condition (iii) in Theorem 6), which is possible since the constants C_2 and $D(\beta, t_0)$ do not depend on η . So the induction is complete and we have (45) is true. \square

4.4. Proof-Sketch of Theorem 7. Due to space limit, we only provide a proof-sketch here. The full proof can be found in Appendix-B.6.

Proof-Sketch of Theorem 7: We first show that, since $\nabla f_i(y_i(t)) = \nabla h_i(y_i(t) A_i) A_i^T$, vector $\bar{x}(t) - \bar{y}(t)$ will always lie within the row space of matrix A . Then we show, the inequality (32) in Lemma 16, when evaluated at $\omega = \bar{x}(t)$, can be strengthened to,

$$(51) \quad f(\bar{x}(t)) \geq \hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle + \frac{\mu}{4} \|\bar{x}(t) - \bar{y}(t)\|^2 - \frac{L}{2n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2.$$

Next, we basically follow the same derivation as equation (48) (49) (50), but using the strengthened inequality (51) instead of (32) in Lemma 16. Because of the additional $\|\bar{x}(t) - \bar{y}(t)\|^2$ term in (51), we will get,

$$\phi_{t+1}^* - f(\bar{x}(t+1)) \geq (1 - \alpha_t)(\phi_t^* - f(\bar{x}(t))) + (E_1\eta - E_2\eta^2)\|g(t)\|^2 + (E_3 - E_4\eta^{2/3})\|\bar{x}(t) - \bar{y}(t)\|^2$$

where E_1, E_2, E_3, E_4 are constants that do not depend on η . By making η small enough, we have $\phi_{t+1}^* - f(\bar{x}(t+1)) \geq (1 - \alpha_t)(\phi_t^* - f(\bar{x}(t))) \geq 0$. Next we have, $f(\bar{x}(t)) \leq \phi_t^* \leq \Phi_t(x^*) \leq f^* + \lambda_t(\Phi_0(x^*) - f^*)$. Hence, $f(\bar{x}(t)) - f^* = O(\lambda_t)$. Lastly, it can be shown that, when using a fixed step size, $\lambda_t = O(\frac{1}{t^2})$.

5. Numerical Experiments. We simulate our algorithm on different objective functions and compare it with other algorithms. We choose $n = 100$ agents and the graph is generated using the Erdos-Renyi model [9] with connectivity probability 0.3. The weight matrix W is chosen using the Laplacian method [31]. In details, $W = I - \frac{1}{\max_{i=1}^n d_i + 1} \mathcal{L}$, where d_i is degree of node i in the graph \mathcal{G} , and $\mathcal{L} = [L_{ij}]$ is the Laplacian of the graph defined to be $L_{ij} = -1$ for $(i, j) \in E$, and $L_{ii} = d_i$ and $L_{ij} = 0$ for i, j not connected. We will compare our algorithm (Acc-DNGD-SC or Acc-DNGD-NSC) with Distributed Gradient Descent (DGD) in [22] with a vanishing step size, the ‘‘EXTRA’’ algorithm in [31] (with $\tilde{W} = \frac{W+I}{2}$), the algorithm studied in [36, 7, 33, 28, 23, 35, 21] (we name it ‘‘Acc-DGD’’), the ‘‘D-NG’’ method in [13]. We will also compare with two centralized methods that directly optimize f : Centralized Gradient Descent (CGD) and Centralized Nesterov Gradient Descent (CNGD-SC (2) or CNGD-NSC (3)). Each element of the initial point $x_i(0)$ is drawn from i.i.d. Gaussian with mean 0 and variance 25. For the functions f_i , we consider three cases.

Case I: The functions f_i are square losses for linear regression, i.e. $f_i(x) = \frac{1}{M_i} \sum_{m=1}^{M_i} (\langle u_{im}, x \rangle - v_{im})^2$ where $u_{im} \in \mathbb{R}^N$ ($N = 3$) are the features and $v_{im} \in \mathbb{R}$ are the observed outputs, and $\{(u_{im}, v_{im})\}_{m=1}^{M_i}$ are $M_i = 50$ data samples for agent i . We generate each data sample independently. We first fix a predefined parameter $\tilde{x} \in \mathbb{R}^N$ with each element drawn uniformly from $[0, 1]$. For each sample (u_{im}, v_{im}) , the last element of u_{im} is fixed to be 1, and the rest elements are drawn from i.i.d. Gaussian with mean 0 and variance 400. Then we generate $v_{im} = \langle \tilde{x}, u_{im} \rangle + \epsilon_{im}$ where ϵ_{im} are independent Gaussian noises with mean 0 and variance 100.

Case II: The functions f_i are the loss functions for logistic regression [1], i.e. $f_i(x) = \frac{1}{M_i} \sum_{m=1}^{M_i} [\ln(1 + e^{\langle u_{im}, x \rangle}) - v_{im} \langle u_{im}, x \rangle]$ where $u_{im} \in \mathbb{R}^N$ ($N = 3$) are the features and $v_{im} \in \{0, 1\}$ are the observed labels, and $\{(u_{im}, v_{im})\}_{m=1}^{M_i}$ are $M_i = 100$ data samples for agent i . The data samples are generated independently. We first fix a predefined parameter $\tilde{x} \in \mathbb{R}^N$ with each element drawn uniformly from $[0, 1]$. For each sample (u_{im}, v_{im}) , the last element of u_{im} is fixed to be 1, and the rest elements of u_{im} are drawn from i.i.d. Gaussian with mean 0 and variance 100. We then generate v_{im} from a Bernoulli distribution, with probability of $v_{im} = 1$ being $\frac{1}{1 + e^{-\langle \tilde{x}, u_{im} \rangle}}$.

For Case I and Case II, the functions are both strongly convex and smooth, so we test the strongly-convex variant of our algorithm Acc-DNGD-SC and the version of CNGD we compare with is CNGD-SC. We note here that the parameters of the f_i 's are chosen to have a large condition number $L/\mu \approx 400$ in order to verify that our algorithm has a better dependence on the condition number than CGD and CGD based distributed methods. The simulation results are shown in Figure 1, where the x -axis is iteration number t , and the y -axis is the average objective error $\frac{1}{n} \sum f(x_i(t)) - f^*$ for distributed methods, or objective error $f(x(t)) - f^*$ for centralized methods.

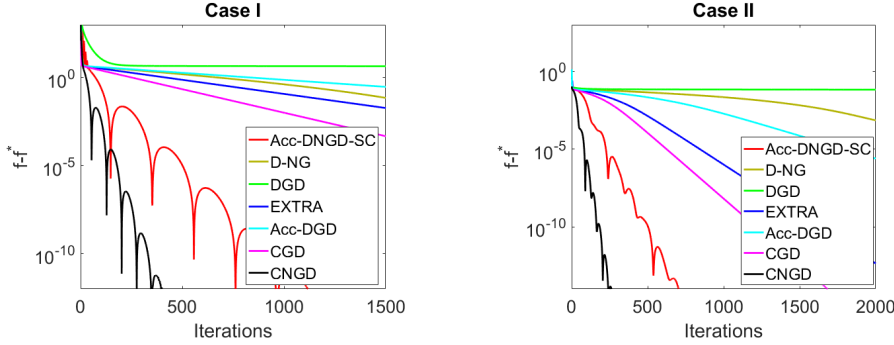


Fig. 1: Simulation results for case I and case II. Step sizes for Case I are, Acc-DNGD-SC: $\eta = 0.0002$, $\alpha = 0.0128$; D-NG: $\eta_t = \frac{0.0015}{t+1}$; DGD: $\eta_t = \frac{0.0002}{\sqrt{t}}$; EXTRA: $\eta = 0.0009$; Acc-DGD: $\eta_t = 0.0004$; CGD: $\eta = 0.0015$; CNGD: $\eta = 0.0015$, $\alpha = 0.0355$. Step sizes for Case II are, Acc-DNGD-SC: $\eta = 0.05$, $\alpha = 0.0216$; D-NG: $\eta_t = \frac{0.1633}{t+1}$; DGD: $\eta_t = \frac{0.05}{\sqrt{t}}$; EXTRA: $\eta = 0.2450$; Acc-DGD: $\eta = 0.1143$; CGD: $\eta = 0.3267$; CNGD: $\eta = 0.3267$, $\alpha = 0.0551$.

It can be seen that our algorithm Acc-DNGD-SC indeed performs significantly better than CGD, CGD-based distributed methods (DGD, EXTRA, Acc-DGD) and also D-NG.

Case III: the objective functions are given by,

$$f_i(x) = \begin{cases} \frac{1}{m} \langle a_i, x \rangle^m + \langle b_i, x \rangle & \text{if } |\langle a_i, x \rangle| \leq 1, \\ |\langle a_i, x \rangle| - \frac{m-1}{m} + \langle b_i, x \rangle & \text{if } |\langle a_i, x \rangle| > 1, \end{cases}$$

where $m = 12$, $a_i, b_i \in \mathbb{R}^N$ ($N = 4$) are vectors whose entries are i.i.d. Gaussian with mean 0 and variance 1, with the exception that b_n is set to be $b_n = -\sum_{i=1}^{n-1} b_i$ s.t. $\sum_i b_i = 0$. It is easy to check that f_i is convex and smooth, but not strongly convex (around the minimizer).

Case III is intended to test the sublinear convergence rate $\frac{1}{t^{2-\beta}}$ ($\beta > 0.6$) of the Acc-DNGD-NSC (5) and the conjecture that the $\frac{1}{t^{2-\beta}}$ rate still holds even if $\beta \in [0, 0.6]$ (cf. Theorem 6 and the comments following it). Therefore, we do two runs of Acc-DNGD-NSC, one with $\beta = 0.61$ and the other with $\beta = 0$. The results are shown in Figure 2, where the x -axis is the iteration t , and the y -axis is the (average) objective error. Notice that Figure 2 is a double log plot. It shows that Acc-DNGD-NSC with $\beta = 0.61$ performs faster than $1/t^{1.39}$, while D-NG, CGD and CGD-based distributed methods (DGD, Acc-DGD, EXTRA) are slower than $1/t^{1.39}$. Further, both Acc-DNGD-NSC with $\beta = 0$ and CNGD-NSC are faster than $\frac{1}{t^2}$.

6. Conclusion. In this paper we have proposed an Accelerated Distributed Nesterov Gradient Descent (Acc-DNGD) method. The first version works for convex and L -smooth functions and we show that it achieves a $O(\frac{1}{t^{1.4-\epsilon}})$ convergence rate for all $\epsilon \in (0, 1.4)$. We also show the convergence rate can be improved to $O(\frac{1}{t^2})$ if the objective function is a composition of a linear map and a strongly-convex and smooth function. The second version works for μ -strongly convex and L -smooth functions, and we show that it achieves a linear convergence rate of $O([1 - C(\frac{\mu}{L})^{5/7}]^t)$ for some constant C independent of L or μ . All the rates are better than CGD and CGD-based distributed methods. In the future, we plan to tighten our analysis to obtain

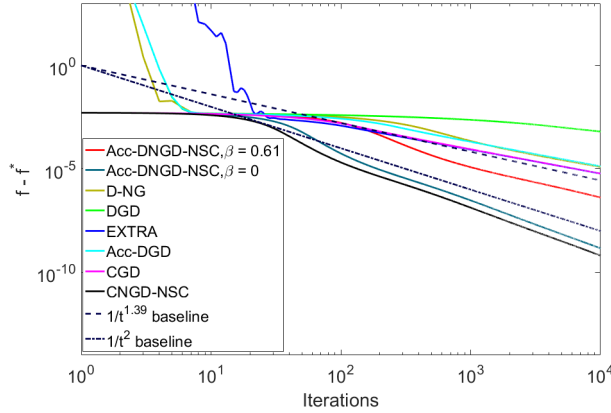


Fig. 2: Simulation results for case III. Steps sizes: Acc-DNGD-NSC with $\beta = 0.61$: $\eta_t = \frac{0.0045}{(t+1)^{0.61}}$, $\alpha_0 = 0.7071$; Acc-DNGD-NSC with $\beta = 0$: $\eta_t = 0.0045$, $\alpha_0 = 0.7071$; D-NG: $\eta_t = \frac{0.0091}{t+1}$; DGD: $\eta_t = \frac{0.0091}{\sqrt{t}}$; EXTRA: $\eta = 0.0091$; Acc-DGD: $\eta = 0.0045$; CGD: $\eta = 0.0091$; CNGD-NSC: $\eta = 0.0091$, $\alpha_0 = 0.5$.

better convergence rates. Specifically, we expect in the convex and L -smooth case, our method can achieve a $O(\frac{1}{t^{2-\beta}})$ rate for all $0 \leq \beta < 2$.

REFERENCES

- [1] *Logistic regression*, 2012, <http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf> (accessed 2016-09-23).
- [2] J. A. BAZERQUE AND G. B. GIANNAKIS, *Distributed spectrum sensing for cognitive radio networks by exploiting sparsity*, IEEE Transactions on Signal Processing, 58 (2010), pp. 1847–1862.
- [3] D. P. BERTSEKAS, *Nonlinear programming*, (1999).
- [4] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and distributed computation: numerical methods*, vol. 23, Prentice hall Englewood Cliffs, NJ, 1989.
- [5] S. BUBECK, *Convex optimization: Algorithms and complexity*, arXiv preprint arXiv:1405.4980, (2014).
- [6] O. DEVOLDER, F. GLINEUR, AND Y. NESTEROV, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming, 146 (2014), pp. 37–75.
- [7] P. DI LORENZO AND G. SCUTARI, *Distributed nonconvex optimization over networks*, in Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on, IEEE, 2015, pp. 229–232.
- [8] J. C. DUCHI, A. AGARWAL, AND M. J. WAINWRIGHT, *Dual averaging for distributed optimization: convergence analysis and network scaling*, Automatic control, IEEE Transactions on, 57 (2012), pp. 592–606.
- [9] P. ERDOS AND A. RENYI, *On random graphs i*, Publ. Math. Debrecen, 6 (1959), pp. 290–297.
- [10] P. A. FORERO, A. CANO, AND G. B. GIANNAKIS, *Consensus-based distributed support vector machines*, Journal of Machine Learning Research, 11 (2010), pp. 1663–1707.
- [11] S. A. GERSHGORIN, *über die abgrenzung der eigenwerte einer matrix*, Bulletin de l’Académie des Sciences de l’URSS. Classe des sciences mathématiques et na, (1931), pp. 749–754.
- [12] R. A. HORN AND C. R. JOHNSON, *Matrix analysis*, Cambridge university press, 2012.
- [13] D. JAKOVETIC, J. XAVIER, AND J. M. MOURA, *Fast distributed gradient methods*, Automatic Control, IEEE Transactions on, 59 (2014), pp. 1131–1146.
- [14] B. JOHANSSON, *On distributed optimization in networked systems*, (2008).
- [15] I. LOBEL AND A. OZDAGLAR, *Convergence analysis of distributed subgradient methods over random networks*, in Communication, Control, and Computing, 2008 46th Annual Allerton

- Conference on, IEEE, 2008, pp. 353–360.
- [16] I. LOBEL, A. OZDAGLAR, AND D. FEJER, *Distributed multi-agent optimization with state-dependent communication*, *Mathematical Programming*, 129 (2011), pp. 255–284.
 - [17] I. MATEI AND J. S. BARAS, *Performance evaluation of the consensus-based distributed subgradient method under random communication topologies*, *Selected Topics in Signal Processing*, IEEE Journal of, 5 (2011), pp. 754–771.
 - [18] C. D. MEYER, *Matrix analysis and applied linear algebra*, vol. 2, Siam, 2000.
 - [19] A. NEDIC AND A. OLSHEVSKY, *Stochastic gradient-push for strongly convex functions on time-varying directed graphs*, arXiv preprint arXiv:1406.2075, (2014).
 - [20] A. NEDIC AND A. OLSHEVSKY, *Distributed optimization over time-varying directed graphs*, *Automatic Control*, IEEE Transactions on, 60 (2015), pp. 601–615.
 - [21] A. NEDIC, A. OLSHEVSKY, W. SHI, AND C. A. URIBE, *Geometrically convergent distributed optimization with uncoordinated step-sizes*, arXiv preprint arXiv:1609.05877, (2016).
 - [22] A. NEDIĆ AND A. OZDAGLAR, *Distributed subgradient methods for multi-agent optimization*, *Automatic Control*, IEEE Transactions on, 54 (2009), pp. 48–61.
 - [23] A. NEDICH, A. OLSHEVSKY, AND W. SHI, *Achieving geometric convergence for distributed optimization over time-varying graphs*, arXiv preprint arXiv:1607.03218, (2016).
 - [24] Y. NESTEROV, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2013.
 - [25] R. OLFATI-SABER, J. A. FAX, AND R. M. MURRAY, *Consensus and cooperation in networked multi-agent systems*, *Proceedings of the IEEE*, 95 (2007), pp. 215–233, doi:10.1109/JPROC.2006.887293.
 - [26] A. OLSHEVSKY, *Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control*, arXiv preprint arXiv:1411.4186, (2014).
 - [27] A. OLSHEVSKY AND J. N. TSITSIKLIS, *Convergence speed in distributed consensus and averaging*, *SIAM Journal on Control and Optimization*, 48 (2009), pp. 33–55.
 - [28] G. QU AND N. LI, *Harnessing smoothness to accelerate distributed optimization*, arXiv preprint arXiv:1605.07112, (2016).
 - [29] G. QU AND N. LI, *Accelerated distributed nesterov gradient descent*, arXiv preprint arXiv:1705.07176, (2017).
 - [30] S. S. RAM, A. NEDIĆ, AND V. V. VEERAVALLI, *Distributed stochastic subgradient projection algorithms for convex optimization*, *Journal of optimization theory and applications*, 147 (2010), pp. 516–545.
 - [31] W. SHI, Q. LING, G. WU, AND W. YIN, *Extra: An exact first-order algorithm for decentralized consensus optimization*, *SIAM Journal on Optimization*, 25 (2015), pp. 944–966.
 - [32] J. N. TSITSIKLIS, D. P. BERTSEKAS, AND M. ATHANS, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, in 1984 American Control Conference, 1984, pp. 484–489.
 - [33] P. DI LORENZO AND G. SCUTARI, *Next: In-network nonconvex optimization*, *IEEE Transactions on Signal and Information Processing over Networks*, 2 (2016), pp. 120–136.
 - [34] C. XI AND U. A. KHAN, *On the linear convergence of distributed optimization over directed graphs*, arXiv preprint arXiv:1510.02149, (2015).
 - [35] C. XI AND U. A. KHAN, *Add-opt: Accelerated distributed directed optimization*, arXiv preprint arXiv:1607.04757, (2016).
 - [36] J. XU, S. ZHU, Y. C. SOH, AND L. XIE, *Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes*, in 2015 54th IEEE Conference on Decision and Control (CDC), IEEE, 2015, pp. 2055–2060.
 - [37] J. ZENG AND W. YIN, *Extrapush for convex smooth decentralized optimization over directed networks*, arXiv preprint arXiv:1511.02942, (2015).
 - [38] M. ZHU AND S. MARTÍNEZ, *On distributed convex optimization under inequality and equality constraints*, *Automatic Control*, IEEE Transactions on, 57 (2012), pp. 151–164.

Appendix A. Proofs of Auxiliary Lemmas in Section 3.

A.1. Proof of Lemma 12 and Lemma 13.

Proof of Lemma 12: We have,

$$\begin{aligned}
 & \|\bar{y}(t+1) - \bar{y}(t)\| \\
 &= \left\| \frac{1}{1+\alpha} [\bar{x}(t+1) - \bar{x}(t)] + \frac{\alpha}{1+\alpha} [\bar{v}(t+1) - \bar{v}(t)] \right\| \\
 &= \left\| \frac{1}{1+\alpha} [\bar{y}(t) - \bar{x}(t)] - \frac{\eta}{1+\alpha} g(t) + \frac{\alpha}{1+\alpha} \alpha (\bar{y}(t) - \bar{v}(t)) - \frac{\eta}{1+\alpha} g(t) \right\|
 \end{aligned}$$

$$\begin{aligned}
&= \left\| \frac{1}{1+\alpha} [\bar{y}(t) - \bar{x}(t)] + \frac{\alpha}{1+\alpha} \alpha [\bar{y}(t) - \frac{1+\alpha}{\alpha} \bar{y}(t) + \frac{1}{\alpha} \bar{x}(t)] - \frac{2\eta}{1+\alpha} g(t) \right\| \\
&\leq \frac{1-\alpha}{1+\alpha} \|\bar{y}(t) - \bar{x}(t)\| + \frac{2\eta}{1+\alpha} \|g(t)\| \\
&\leq \|\bar{y}(t) - \bar{x}(t)\| + 2\eta \|g(t)\|.
\end{aligned}$$

□

Proof of Lemma 13: For (13), we have

$$\begin{aligned}
\|\nabla(t+1) - \nabla(t)\| &= \sqrt{\sum_{i=1}^n \|\nabla f_i(y_i(t+1)) - \nabla f_i(y_i(t))\|^2} \\
&\leq \sqrt{\sum_{i=1}^n L^2 \|y_i(t+1) - y_i(t)\|^2} \\
&= L \|y(t+1) - y(t)\|.
\end{aligned}$$

For (14), we have

$$\begin{aligned}
\|g(t) - \nabla f(\bar{y}(t))\| &= \left\| \frac{\sum_{i=1}^n (\nabla f_i(y_i(t)) - \nabla f_i(\bar{y}(t)))}{n} \right\| \\
&\leq \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(y_i(t)) - \nabla f_i(\bar{y}(t))\| \\
&\leq \frac{L}{n} \sum_{i=1}^n \|y_i(t) - \bar{y}(t)\| \\
&\leq L \sqrt{\frac{1}{n} \sum_{i=1}^n \|y_i(t) - \bar{y}(t)\|^2} \\
&= \frac{L}{\sqrt{n}} \|y(t) - \mathbf{1}\bar{y}(t)\|.
\end{aligned}$$

□

A.2. Proof of Lemma 14. In this section, we provide the proof of Lemma 14.

Proof of Lemma 14: (a) We first calculate the characteristic polynomial of $G(\eta)$ as

$$p(\zeta) = \overbrace{(\zeta - \sigma) \left(\zeta - \frac{1-\alpha}{1+\alpha} \sigma \right) (\zeta - \sigma - 2\eta L)}^{\triangleq p_0(\zeta)} - \overbrace{[k_1(\zeta - \sigma - 2\eta L) + k_2]}^{\triangleq p_1(\zeta)}$$

where k_1 and k_2 are positive constants given by

$$k_1 = 4\eta L + \sigma\eta L < 5\eta L$$

and

$$\begin{aligned}
k_2 &= 2\eta^2 L^2 (4 + \sigma) + \eta L \sigma (2 + \alpha\sigma) + \frac{2\eta L \alpha^2 \sigma (2 + \alpha\sigma)}{1 + \alpha} \\
&< 10\eta^2 L^2 + 3\eta L + 6\eta L < 19\eta L.
\end{aligned}$$

where we have used $\eta L < 1$. Let $\zeta_0 = \sigma + 4(\eta L)^{1/3}$. Then $\zeta_0 > \sigma + 2\eta L$. Since $p(\sigma + 2\eta L) < 0$, and $p(\zeta)$ is a strictly increasing function on $[\zeta_0, +\infty)$ (since when $\zeta \geq \zeta_0$, $p'(\zeta) > (\zeta_0 - \sigma)(\zeta_0 - \frac{1-\alpha}{1+\alpha}\sigma) - k_1 > 16(\eta L)^{2/3} - 5\eta L > 0$), and also since $G(\eta)$'s largest eigenvalue in magnitude must be a positive real number (Perron-Frobenius Theorem [12, Theorem 8.2.11]), we have $\rho(G(\eta))$ is a real root of p on $[\sigma + 2\eta L, \infty)$. Then $\rho(G(\eta)) < \zeta_0$ will follow from $p(\zeta_0) > 0$, which is shown below.

$$\begin{aligned} p(\zeta_0) &= \left[4(\eta L)^{1/3}(4(\eta L)^{1/3} + \frac{2\alpha}{1+\alpha}\sigma) - k_1 \right] (4(\eta L)^{1/3} - 2\eta L) - k_2 \\ &> (16(\eta L)^{2/3} - 5\eta L)(4(\eta L)^{1/3} - 2\eta L) - k_2 \\ &> (16(\eta L)^{2/3} - 5(\eta L)^{2/3})(4(\eta L)^{1/3} - 2(\eta L)^{1/3}) - k_2 \\ &= 22\eta L - k_2 > 0. \end{aligned}$$

For the lower bound, notice that $k_2^{1/3} > 2(\eta L)^{2/3} > 4\eta L$ (which is equivalent to $\eta L < \frac{1}{8}$, cf. the step size condition in Lemma 14). Therefore, we have

$$\begin{aligned} p(\sigma + 0.84k_2^{1/3}) &= (0.84k_2^{1/3})(0.84k_2^{1/3} + \frac{2\alpha}{1+\alpha}\sigma) - k_1(0.84k_2^{1/3} - 2\eta L) - k_2 \\ &< (0.84k_2^{1/3})(0.84k_2^{1/3} + \frac{1}{2^{5/6}}k_2^{1/3})(0.84k_2^{1/3}) - k_2 < 0 \end{aligned}$$

where we have used that, noticing $k_2 > 2\eta L\sigma$ and $\alpha = \sqrt{\mu\eta} < \sqrt{\eta L} < \sqrt{1/8^3}$ (cf. the step size condition in Lemma 14),

$$\left[\frac{\alpha}{1+\alpha}\sigma \right]^3 < \alpha^3\sigma < \eta L\alpha\sigma < \frac{\alpha}{2}k_2 < \frac{k_2}{2}\sqrt{1/8^3} \Rightarrow \frac{\alpha}{1+\alpha}\sigma < \frac{k_2^{1/3}}{2 \cdot 2^{5/6}}.$$

Hence we have $\rho(G(\eta)) > \sigma + 0.84k_2^{1/3}$. Also noticing $k_2 > 2\eta L\sigma$, we have

$$\rho(G(\eta)) > \sigma + 0.84(2\eta L\sigma)^{1/3} > \sigma + (\sigma\eta L)^{1/3}.$$

Before we proceed to (b) and (c), we prove the following claim.

Claim: $G(\eta)$'s spectral gap is at least $\rho(G(\eta)) - \sigma$.

To prove the claim, we consider two cases. If $G(\eta)$ has three real eigenvalues, then notice that $p_0(\zeta)$ is nonpositive on $[\sigma, \sigma + 2\eta L]$, and $p_1(\zeta)$ is affine and positive on $[\sigma, \sigma + 2\eta L]$ (since $p_1(\sigma + 2\eta L) = k_2 > 0$, $p_1(\sigma) = k_2 - 2\eta Lk_1 = k_2 - 2\eta^2L^2(4 + \sigma) > 0$), therefore $p(\zeta)$ has no real roots on $[\sigma, \sigma + 2\eta L]$. Also notice $p(\zeta)$ has exactly one real root on $[\sigma + 2\eta L, \infty)$ (the leading eigenvalue; here we have used the fact that $p(\zeta)$ is strictly convex on $[\sigma + 2\eta L, \infty)$ and $p(\sigma + 2\eta L) < 0$), hence p 's other two real roots must be less than σ . We next show that the two other real roots must be nonnegative. Let the three eigenvalues be γ_1, γ_2 and γ_3 with γ_1 being the leading eigenvalue. Then,

$$\gamma_1 + \gamma_2 + \gamma_3 = \text{Trace}[G(\eta)] = (1 - \alpha)\sigma + \frac{1 + \alpha^2}{1 + \alpha}\sigma + \sigma + 2\eta L.$$

Then, using $\alpha = \sqrt{\mu\eta} < \frac{1}{2}$ (cf. the step size condition in Lemma 14), we have $1 - \alpha > \frac{1}{2}$ and $\frac{1 + \alpha^2}{1 + \alpha} > \frac{1}{2}$, and hence

$$\gamma_2 + \gamma_3 > 2\sigma - \gamma_1 > \sigma - 4(\eta L)^{1/3} > 0$$

where in the last inequality, we have used $\eta L < \frac{\sigma^3}{64}$ (cf. the step size condition in Lemma 14). Next, we calculate

$$\begin{aligned} \gamma_1\gamma_2\gamma_3 &= -p(0) = \sigma^2(\sigma + 2\eta L)\frac{1-\alpha}{1+\alpha} + k_2 - k_1(\sigma + 2\eta L) \\ &= \sigma^2(\sigma + 2\eta L)\frac{1-\alpha}{1+\alpha} - 2\eta L\sigma\frac{1+\alpha-2\alpha^2}{1+\alpha} - \sigma^2\eta L\frac{1-\alpha^2-2\alpha^3}{1+\alpha} \\ &= \sigma^2\eta L\frac{1-2\alpha+\alpha^2+2\alpha^3}{1+\alpha} + \sigma^3\frac{1-\alpha}{1+\alpha} - 2\eta L\sigma\frac{1+\alpha-2\alpha^2}{1+\alpha} \\ &> \frac{1}{2}\sigma^3 - 2\eta L\sigma. \end{aligned}$$

where in the last inequality, we have used $1-2\alpha > 0$, $\frac{1-\alpha}{1+\alpha} > \frac{1}{2}$ ($\Leftarrow \eta < \frac{1}{9\mu}$, implied by our step size selection in Lemma 14). At last, we use the fact that our step size bound in Lemma 14 implies $2\eta L\sigma < \frac{1}{2}\sigma^3$ ($\iff \eta L < \frac{1}{4}\sigma^2$). Then, $\gamma_1\gamma_2\gamma_3 > 0$. Since $\gamma_1 > 0$, we have $\gamma_2\gamma_3 > 0$. We have already shown $\gamma_2 + \gamma_3 > 0$. Hence, both γ_2 and γ_3 are positive. Now that we have already shown that, γ_2 and γ_3 are positive reals less than σ . This implies the spectral gap is at least $\rho(G(\eta)) - \sigma$.

On the other hand, if $G(\eta)$ has one real eigenvalue (the leading one) and two conjugate complex eigenvalues, then let the modulus of the two conjugate eigenvalues be τ and we have

$$\tau^2\rho(G(\eta)) = -p(0) = \sigma^2(\sigma + 2\eta L)\frac{1-\alpha}{1+\alpha} + k_2 - k_1(\sigma + 2\eta L)$$

We calculate

$$k_2 - k_1(\sigma + 2\eta L) = -2\eta L\sigma\frac{1+\alpha-2\alpha^2}{1+\alpha} - \sigma^2\eta L\frac{1-\alpha^2-2\alpha^3}{1+\alpha} < 0,$$

where we have used $2\alpha^2 < 1$ and $\alpha^2 + 2\alpha^3 < 1$ (since $\alpha < \sqrt{\eta L} < \sqrt{1/8^3}$). Therefore, $\tau^2\rho(G(\eta)) < \sigma^2(\sigma + 2\eta L)$. Noticing that $\rho(G(\eta)) > \sigma + 2\eta L$, we have $\tau < \sigma$. Therefore, we can conclude that the spectral gap is at least $\rho(G(\eta)) - \sigma$.

Now we proceed to prove (b) and (c).

(b) Same as before, we let the three eigenvalues of $G(\eta)$ be γ_1 , γ_2 and γ_3 with $\gamma_1 = \rho(G(\eta))$ being the leading eigenvalue and $\gamma_1 > |\gamma_2| \geq |\gamma_3|$. We assume $\gamma_2 \neq \gamma_3$.¹² Then by the claim on spectral gap, we have $\gamma_1 - |\gamma_3| \geq \gamma_1 - |\gamma_2| \geq (\sigma\eta L)^{1/3}$. We will use the following lemma.

LEMMA 24. *If a series $(B_t)_{t \geq 0}$ can be written as $B_t = \alpha_1\gamma_1^t + \alpha_2\gamma_2^t + \alpha_3\gamma_3^t$ for some $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{C}$, then we have*

$$|B_t| \leq \frac{(5|B_0| + 6|B_1| + 3|B_2|)\gamma_1^t}{(\sigma\eta L)^{2/3}}.$$

We know that through diagonalization, all the entries of $G(\eta)^t$ can be written as the form described in Lemma 24. We know that $[G(\eta)^0]_{23} = 0$, $[G(\eta)^1]_{23} = 2\eta L$,

¹²There will be some values of η and α for which $G(\eta)$ will have only two eigenvalues, one of which has multiplicity 2. This case can be dealt with by taking the $\gamma_2 \rightarrow \gamma_3$ limit and won't affect our result.

$[G(\eta)^2]_{23} = 2\eta L(\sigma + 2\eta L + \sigma \frac{1+\alpha^2}{1+\alpha}) + \eta L \sigma \frac{1-\alpha}{1+\alpha} < 9\eta L$. As a result,

$$[G(\eta)^t]_{23} \leq \frac{39\eta L}{(\sigma\eta L)^{2/3}} \gamma_1^t = \frac{39(\eta L)^{1/3}}{(\sigma)^{2/3}} \gamma_1^t.$$

(c) Similarly to (c), we can bound $[G(\eta)^t]_{21}$ and $[G(\eta)^t]_{22}$. We calculate, $[G(\eta)^0]_{21} = 0$, $[G(\eta)^1]_{21} < 1$, $[G(\eta)^2]_{21} < 4$. Hence

$$[G(\eta)^t]_{21} < \frac{18}{(\sigma\eta L)^{2/3}} \gamma_1^t.$$

Similarly, $[G(\eta)^0]_{22} = 1$, $[G(\eta)^1]_{22} < 1$, $[G(\eta)^2]_{22} < 6$. Hence

$$[G(\eta)^t]_{22} < \frac{29}{(\sigma\eta L)^{2/3}} \gamma_1^t.$$

As a result,

$$\max([G(\eta)^t]_{21}, [G(\eta)^t]_{22}, [G(\eta)^t]_{23}) < \frac{39}{(\sigma\eta L)^{2/3}} \gamma_1^t.$$

□

At last, we provide the derivation of Lemma 24.

Proof of Lemma 24: We have,

$$\begin{bmatrix} 1 & 1 & 1 \\ \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_1^2 & \gamma_2^2 & \gamma_3^2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}.$$

Hence we can solve for $\alpha_1, \alpha_2, \alpha_3$, getting

$$\begin{aligned} \alpha_1 &= \frac{1}{\Delta} \left[\gamma_2 \gamma_3 (\gamma_3 - \gamma_2) B_0 + (\gamma_2^2 - \gamma_3^2) B_1 + (\gamma_3 - \gamma_2) B_2 \right] \\ \alpha_2 &= \frac{1}{\Delta} \left[\gamma_3 \gamma_1 (\gamma_1 - \gamma_3) B_0 + (\gamma_3^2 - \gamma_1^2) B_1 + (\gamma_1 - \gamma_3) B_2 \right] \\ \alpha_3 &= \frac{1}{\Delta} \left[\gamma_1 \gamma_2 (\gamma_2 - \gamma_1) B_0 + (\gamma_1^2 - \gamma_2^2) B_1 + (\gamma_2 - \gamma_1) B_2 \right] \end{aligned}$$

where $\Delta = (\gamma_1 - \gamma_2)(\gamma_2 - \gamma_3)(\gamma_3 - \gamma_1)$. We now calculate,

$$\begin{aligned} |(\gamma_1 - \gamma_3)\gamma_2^t + (\gamma_2 - \gamma_1)\gamma_3^t| &= |(\gamma_2 - \gamma_3)\gamma_2^t + (\gamma_2 - \gamma_1)(\gamma_3^t - \gamma_2^t)| \\ &= |(\gamma_2 - \gamma_3)\gamma_2^t + (\gamma_2 - \gamma_1)(\gamma_3 - \gamma_2) \sum_{k=0}^{t-1} \gamma_3^{t-1-k} \gamma_2^k| \\ &\leq |\gamma_2 - \gamma_3| \left[|\gamma_2|^t + t|\gamma_1 - \gamma_2| |\gamma_2|^{t-1} \right]. \end{aligned}$$

Similarly,

$$|(\gamma_3^2 - \gamma_1^2)\gamma_2^t + (\gamma_1^2 - \gamma_2^2)\gamma_3^t| = |(\gamma_3^2 - \gamma_2^2)\gamma_2^t + (\gamma_1^2 - \gamma_2^2)(\gamma_3^t - \gamma_2^t)|$$

$$\begin{aligned}
&\leq 2|\gamma_2 - \gamma_3||\gamma_2|^t + 2|\gamma_1 - \gamma_2||\gamma_2 - \gamma_3| \left| \sum_{k=0}^{t-1} \gamma_2^{t-1-k} \gamma_3^k \right| \\
&\leq |\gamma_2 - \gamma_3| \left[2|\gamma_2|^t + 2t|\gamma_1 - \gamma_2||\gamma_2|^{t-1} \right].
\end{aligned}$$

$$\begin{aligned}
|\gamma_3\gamma_1(\gamma_1 - \gamma_3)\gamma_2^t + \gamma_1\gamma_2(\gamma_2 - \gamma_1)\gamma_3^t| &= |(\gamma_3 - \gamma_2)(\gamma_1^2 - \gamma_1(\gamma_2 + \gamma_3))\gamma_2^t + \gamma_1\gamma_2(\gamma_2 - \gamma_1)(\gamma_3^t - \gamma_2^t)| \\
&\leq 3|\gamma_2 - \gamma_3||\gamma_2|^t + |\gamma_1 - \gamma_2||\gamma_2 - \gamma_3| \left| \sum_{k=0}^{t-1} \gamma_2^{t-1-k} \gamma_3^k \right| \\
&\leq |\gamma_2 - \gamma_3| \left[3|\gamma_2|^t + t|\gamma_1 - \gamma_2||\gamma_2|^{t-1} \right].
\end{aligned}$$

Hence,

$$\begin{aligned}
&|\alpha_2\gamma_2^t + \alpha_3\gamma_3^t| \\
&\leq \frac{1}{|\Delta|} \left[|B_0||\gamma_3\gamma_1(\gamma_1 - \gamma_3)\gamma_2^t + \gamma_1\gamma_2(\gamma_2 - \gamma_1)\gamma_3^t| + |B_1||(\gamma_3^2 - \gamma_1^2)\gamma_2^t + (\gamma_1^2 - \gamma_2^2)\gamma_3^t| \right. \\
&\quad \left. + |B_2||(\gamma_1 - \gamma_3)\gamma_2^t + (\gamma_2 - \gamma_1)\gamma_3^t| \right] \\
&\leq \frac{|\gamma_2 - \gamma_3|}{|\Delta|} \left[(3|B_0| + 2|B_1| + |B_2|)|\gamma_2|^t + (|B_0| + 2|B_1| + |B_2|)t|\gamma_1 - \gamma_2||\gamma_2|^{t-1} \right].
\end{aligned}$$

Now let $\min(|\gamma_1 - \gamma_2|, |\gamma_1 - \gamma_3|) = \beta \geq (\sigma\eta L)^{1/3}$. Notice that

$$\frac{\gamma_1^t}{|\gamma_2|^t} = \left(1 + \frac{|\gamma_1 - \gamma_2|}{|\gamma_2|}\right)^t > t \frac{|\gamma_1 - \gamma_2|}{|\gamma_2|} \geq \frac{\beta}{|\gamma_2|} t.$$

Therefore, $t|\gamma_2|^{t-1} \leq \frac{1}{\beta}\gamma_1^t$. Hence,

$$\begin{aligned}
&|\alpha_2\gamma_2^t + \alpha_3\gamma_3^t| \\
&\leq \frac{|\gamma_2 - \gamma_3|}{|\Delta|} (3|B_0| + 2|B_1| + |B_2|)|\gamma_2|^t + \frac{|\gamma_2 - \gamma_3|}{|\Delta|} (|B_0| + 2|B_1| + |B_2|)t|\gamma_1 - \gamma_2||\gamma_2|^{t-1} \\
&\leq \frac{(3|B_0| + 2|B_1| + |B_2|)|\gamma_2|^t}{|\gamma_1 - \gamma_2||\gamma_1 - \gamma_3|} + \frac{(|B_0| + 2|B_1| + |B_2|)t|\gamma_2|^{t-1}}{|\gamma_1 - \gamma_3|} \\
&\leq \frac{(3|B_0| + 2|B_1| + |B_2|)\gamma_1^t}{\beta^2} + \frac{(|B_0| + 2|B_1| + |B_2|)\gamma_1^t}{\beta^2} \\
&= \frac{(4|B_0| + 4|B_1| + 2|B_2|)\gamma_1^t}{\beta^2}.
\end{aligned}$$

At last, it is easy to check

$$|\alpha_1\gamma_1^t| \leq \frac{(|B_0| + 2|B_1| + |B_2|)\gamma_1^t}{\beta^2}.$$

Therefore,

$$|B_t| \leq \frac{(5|B_0| + 6|B_1| + 3|B_2|)\gamma_1^t}{\beta^2} \leq \frac{(5|B_0| + 6|B_1| + 3|B_2|)\gamma_1^t}{(\sigma\eta L)^{2/3}}.$$

□

Appendix B. Proofs of Auxiliary Lemmas in Section 4.

B.1. Proof of Lemma 19. The derivations are shown below.

$$\begin{aligned}\bar{y}(t+1) - \bar{y}(t) &= (1 - \alpha_{t+1})(\bar{y}(t) - \eta_t g(t)) + \alpha_{t+1}(\bar{v}(t) - \frac{\eta_t}{\alpha_t} g(t)) - \bar{y}(t) \\ &= \alpha_{t+1}(\bar{v}(t) - \bar{y}(t)) - \eta_t \left[\frac{\alpha_{t+1}}{\alpha_t} + 1 - \alpha_{t+1} \right] g(t)\end{aligned}$$

$$\begin{aligned}\bar{v}(t+1) - \bar{y}(t+1) &= \bar{v}(t) - \frac{\eta_t}{\alpha_t} g(t) - (1 - \alpha_{t+1})(\bar{y}(t) - \eta_t g(t)) - \alpha_{t+1}(\bar{v}(t) - \frac{\eta_t}{\alpha_t} g(t)) \\ &= (1 - \alpha_{t+1})(\bar{v}(t) - \bar{y}(t)) + \eta_t(1 - \alpha_{t+1})\left(1 - \frac{1}{\alpha_t}\right)g(t)\end{aligned}$$

□

B.2. Derivation of (40) in the Proof of Lemma 17. By (30a) and (31a), we have

$$(52) \quad \begin{aligned}\|x(t+1) - \mathbf{1}\bar{x}(t+1)\| &= \|[Wy(t) - \mathbf{1}\bar{y}(t)] - \eta_t[s(t) - \mathbf{1}g(t)]\| \\ &\leq \sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + \eta_t\|s(t) - \mathbf{1}g(t)\|.\end{aligned}$$

By (30b) and (31b), we have

$$(53) \quad \begin{aligned}\|v(t+1) - \mathbf{1}\bar{v}(t+1)\| &\leq \|[Wv(t) - \mathbf{1}\bar{v}(t)] - \frac{\eta_t}{\alpha_t}[s(t) - \mathbf{1}g(t)]\| \\ &\leq \sigma\|v(t) - \mathbf{1}\bar{v}(t)\| + \frac{\eta_t}{\alpha_t}\|s(t) - \mathbf{1}g(t)\|.\end{aligned}$$

Hence,

$$(54) \quad \alpha_{t+1}\|v(t+1) - \mathbf{1}\bar{v}(t+1)\| \leq \sigma\alpha_t\|v(t) - \mathbf{1}\bar{v}(t)\| + \eta_t\|s(t) - \mathbf{1}g(t)\|.$$

By (30c) and (31c), we have

$$(55) \quad \begin{aligned}\|y(t+1) - \mathbf{1}\bar{y}(t+1)\| &\leq (1 - \alpha_{t+1})\|x(t+1) - \mathbf{1}\bar{x}(t+1)\| + \alpha_{t+1}\|v(t+1) - \mathbf{1}\bar{v}(t+1)\| \\ &\leq (1 - \alpha_{t+1}) \left[\sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + \eta_t\|s(t) - \mathbf{1}g(t)\| \right] \\ &\quad + \alpha_{t+1} \left[\sigma\|v(t) - \mathbf{1}\bar{v}(t)\| + \frac{\eta_t}{\alpha_t}\|s(t) - \mathbf{1}g(t)\| \right] \\ &\leq \alpha_t\sigma\|v(t) - \mathbf{1}\bar{v}(t)\| + \sigma\|y(t) - \mathbf{1}\bar{y}(t)\| + 2\eta_t\|s(t) - \mathbf{1}g(t)\|\end{aligned}$$

where we have used (52) and (53) in the second inequality.

By (30d) and (31d), we have

$$(56) \quad \begin{aligned}\|s(t+1) - \mathbf{1}g(t+1)\| &= \|Ws(t) - \mathbf{1}g(t) + [\nabla(t+1) - \nabla(t) - \mathbf{1}(g(t+1) - g(t))]\| \\ &\stackrel{(a)}{\leq} \sigma\|s(t) - \mathbf{1}g(t)\| + \|\nabla(t+1) - \nabla(t)\| \\ &\stackrel{(b)}{\leq} \sigma\|s(t) - \mathbf{1}g(t)\| + L\|y(t+1) - y(t)\|\end{aligned}$$

where in (a) we have used the fact that

$$\left\| [\nabla(t+1) - \nabla(t)] - [\mathbf{1}g(t+1) - \mathbf{1}g(t)] \right\|^2$$

$$\begin{aligned}
&= \|\nabla(t+1) - \nabla(t)\|^2 - n\|g(t+1) - g(t)\|^2 \\
&\leq \|\nabla(t+1) - \nabla(t)\|^2
\end{aligned}$$

and in (b) we have used (13).

Now we expand $y(t+1) - y(t)$,

$$\begin{aligned}
\|y(t+1) - y(t)\| &\leq \|y(t+1) - \mathbf{1}\bar{y}(t+1)\| + \|y(t) - \mathbf{1}\bar{y}(t)\| + \|\mathbf{1}\bar{y}(t+1) - \mathbf{1}\bar{y}(t)\| \\
&\leq \alpha_t \|v(t) - \mathbf{1}\bar{v}(t)\| + 2\|y(t) - \mathbf{1}\bar{y}(t)\| \\
(57) \quad &+ 2\eta_t \|s(t) - \mathbf{1}g(t)\| + \|\mathbf{1}\bar{y}(t+1) - \mathbf{1}\bar{y}(t)\|.
\end{aligned}$$

Combining the above with (56), we get,

$$\begin{aligned}
\|s(t+1) - \mathbf{1}g(t+1)\| &\leq \alpha_t L \|v(t) - \mathbf{1}\bar{v}(t)\| + 2L \|y(t) - \mathbf{1}\bar{y}(t)\| \\
(58) \quad &+ \left[\sigma + 2\eta_t L \right] \|s(t) - \mathbf{1}g(t)\| + L\sqrt{n} \|\bar{y}(t+1) - \bar{y}(t)\|.
\end{aligned}$$

By (38) we have $a(t) \geq L\|\bar{y}(t+1) - \bar{y}(t)\|$. Now we combine (54) (55) (58) and get the desired linear system inequality (40). \square

B.3. Proof of Lemma 20, Lemma 21 and Lemma 22. We first prove Lemma 20.

Proof of Lemma 20: We write down the characteristic polynomial of $G(\eta)$ as

$$(59) \quad p(\zeta) = (\zeta - \sigma)^2(\zeta - \sigma - 2\eta L) - 5\eta L(\zeta - \sigma - 2\eta L) - 2\eta L\sigma - 10\eta^2 L^2.$$

We evaluate $p(\cdot)$ on $\sigma + 4(\eta L)^{1/3}$ and get,

$$\begin{aligned}
p(\sigma + 4(\eta L)^{1/3}) &= (16(\eta L)^{2/3} - 5\eta L)(4(\eta L)^{1/3} - 2\eta L) - 2\eta L\sigma - 10\eta^2 L^2 \\
&> 22\eta L - 2\eta L\sigma - 10\eta^2 L^2 > 0
\end{aligned}$$

where we have used $(\eta L)^2 < (\eta L) < (\eta L)^{2/3} < (\eta L)^{1/3}$. It's easy to check $p(\sigma) < 0$. Also, $p'(\zeta) = 2(\zeta - \sigma)(\zeta - \sigma - 2\eta L) + (\zeta - \sigma)^2 - 5\eta L > 0$ on $[\sigma + 4(\eta L)^{1/3}, \infty)$, so p 's largest real root must lie within $(\sigma, \sigma + 4(\eta L)^{1/3})$. Therefore, $\sigma < \theta(\eta) < \sigma + 4(\eta L)^{1/3}$. For the eigenvectors, notice that

$$L\chi_1(\eta) + 2L\chi_2(\eta) + (\sigma + 2\eta L)\chi_3(\eta) = \theta(\eta)\chi_3(\eta).$$

Hence, $\theta(\eta)\chi_3(\eta) \geq \sigma\chi_3(\eta) + 2L\chi_2(\eta)$, and $\chi_2(\eta)/\chi_3(\eta) \leq \frac{\theta(\eta) - \sigma}{2L} \leq \frac{2}{L^{2/3}}\eta^{1/3}$. \square

For the rest of this section, we will use the following formula for $\chi_1(\cdot)$ and $\chi_2(\cdot)$.

LEMMA 25. *We have*

$$\begin{aligned}
\chi_1(\eta) &= \frac{\eta}{\theta(\eta) - \sigma}, \\
\chi_2(\eta) &= \frac{\theta(\eta) - \sigma}{2L} - \eta - \frac{\eta}{2(\theta(\eta) - \sigma)}.
\end{aligned}$$

Proof: Since $G(\eta)\chi(\eta) = \theta(\eta)\chi(\eta)$, we have, writing down the first line,

$$\sigma\chi_1(\eta) + \eta\chi_3(\eta) = \theta(\eta)\chi_1(\eta)$$

from which we can get the formula for $\chi_1(\eta)$. Writing the third line of $G(\eta)\chi(\eta) = \theta(\eta)\chi(\eta)$, we get

$$L\chi_1(\eta) + 2L\chi_2(\eta) + (\sigma + 2\eta L)\chi_3(\eta) = \theta(\eta)\chi_3(\eta)$$

from which we can derive the formula for $\chi_2(\eta)$. \square

Now we proceed to prove Lemma 21.

Proof of Lemma 21: Since $\eta L < \frac{\sqrt{\sigma}}{\sqrt{8}}$, we have $(\sigma\eta L)^{1/3} > 2\eta L$. Hence by (59), $p(\sigma + (\sigma\eta L)^{1/3}) \leq (\sigma\eta L)^{2/3}((\sigma\eta L)^{1/3} - 2\eta L) - 2\sigma\eta L < 0$. Hence $\theta(\eta) > \sigma + (\sigma\eta L)^{1/3}$. As a result, $\chi_1(\eta) < \frac{\eta}{(\sigma\eta L)^{1/3}}$. \square

The rest of the section will be devoted to Lemma 22. Before proving Lemma 22, we prove an axillary lemma first.

LEMMA 26. *We have*

$$\theta'(\eta) = \frac{2L(\theta(\eta) - \sigma)^2 + 5L(\theta(\eta) - \sigma) + 2\sigma L}{3(\theta(\eta) - \sigma)^2 - 4\eta L(\theta(\eta) - \sigma) - 5\eta L}.$$

Further, when $0 < \eta < \frac{\sigma^2}{L9^3}$, we have

$$0 < \theta'(\eta) < 5 \frac{L^{1/3}}{\eta^{2/3}}.$$

Proof: Since $p(\theta(\eta)) = 0$ (where p is the characteristic polynomial of $G(\eta)$ as defined in (59), we take derivative w.r.t. η on both sides of $p(\theta(\eta)) = 0$, and get

$$\begin{aligned} 0 = & 2(\theta(\eta) - \sigma)\theta'(\eta)(\theta(\eta) - \sigma - 2\eta L) \\ & + (\theta(\eta) - \sigma)^2(\theta'(\eta) - 2L) - 5L(\theta(\eta) - \sigma - 2\eta L) \\ & - 5\eta L(\theta'(\eta) - 2L) - 2\sigma L - 20\eta L^2. \end{aligned}$$

From the above equation we obtain

$$(60) \quad \theta'(\eta) = \frac{2L(\theta(\eta) - \sigma)^2 + 5L(\theta(\eta) - \sigma) + 2\sigma L}{3(\theta(\eta) - \sigma)^2 - 4\eta L(\theta(\eta) - \sigma) - 5\eta L}.$$

By Lemma 20, the step size $\eta < \frac{\sigma^2}{L9^3}$ implies that

$$(61) \quad \theta(\eta) - \sigma < 4(\eta L)^{1/3} < \sigma^{2/3}.$$

Hence, the nominator part of $\theta'(\eta)$ satisfy,

$$2L(\theta(\eta) - \sigma)^2 + 5L(\theta(\eta) - \sigma) + 2\sigma L < 9L\sigma^{2/3}.$$

Also notice,

$$4\eta L(\theta(\eta) - \sigma) + 5\eta L < 9\eta L < (\sigma\eta L)^{2/3}$$

where the last inequality comes from the step size condition $\eta < \frac{\sigma^2}{L9^3}$. The step size condition also implies $\eta < \frac{\sqrt{\sigma}}{L2\sqrt{2}}$ and hence we can use Lemma 21 to get $(\theta(\eta) - \sigma)^2 \geq (\sigma\eta L)^{2/3}$. Hence, the denominator part of $\theta'(\eta)$ satisfy

$$3(\theta(\eta) - \sigma)^2 - 4\eta L(\theta(\eta) - \sigma) - 5\eta L \geq 3(\sigma\eta L)^{2/3} - (\sigma\eta L)^{2/3} = 2(\sigma\eta L)^{2/3}.$$

Combining the bound on the nominator and the denominator, we get

$$0 < \theta'(\eta) \leq \frac{9L\sigma^{2/3}}{2(\sigma\eta L)^{2/3}} < 5 \frac{L^{1/3}}{\eta^{2/3}}.$$

□

Now we proceed to prove Lemma 22.

Proof of Lemma 22: Define $\xi_1, \xi_2 : \mathbb{R} \rightarrow \mathbb{R}$ to be $\xi_1(y) = \log \chi_1(e^y)$ and $\xi_2(y) = \log \chi_2(e^y)$. The statement of the lemma can be rephrased as,

$$\begin{aligned} \log \chi_1(e^{\log \zeta_1}) - \log \chi_1(e^{\log \zeta_2}) &\leq \frac{6}{\sigma} |\log \zeta_1 - \log \zeta_2|, \\ \log \chi_2(e^{\log \zeta_1}) - \log \chi_2(e^{\log \zeta_2}) &\leq \frac{28}{\sigma} |\log \zeta_1 - \log \zeta_2|. \end{aligned}$$

So we only need to show that ξ_1 is $\frac{6}{\sigma}$ -Lipschitz continuous and ξ_2 is $\frac{28}{\sigma}$ -Lipschitz continuous, on interval $y \in (-\infty, \log(\frac{\sigma^2}{9^3 L}))$. We calculate the derivative of ξ_1 and let $\eta = e^y \in (0, \sigma^2/(9^3 L))$,

$$\xi_1'(y) = \frac{\chi_1'(\eta)\eta}{\chi_1(\eta)} = \frac{\eta}{\chi_1(\eta)} \left[\frac{(\theta(\eta) - \sigma) - \eta\theta'(\eta)}{(\theta(\eta) - \sigma)^2} \right] = 1 - \frac{\eta\theta'(\eta)}{\theta(\eta) - \sigma}.$$

Notice that η satisfies all the step size conditions of Lemma 20,21,26. We have

$$\begin{aligned} |\xi_1'(y)| &\leq 1 + \frac{\eta\theta'(\eta)}{\theta(\eta) - \sigma} \\ &\leq 1 + \frac{\eta \cdot 5L^{1/3}/\eta^{2/3}}{(\sigma\eta L)^{1/3}} \\ &\leq 6/\sigma^{1/3} < 6/\sigma. \end{aligned}$$

This implies ξ_1 is $6/\sigma$ -Lipschitz continuous. Similarly, for ξ_2 , we have

$$\xi_2'(y) = \frac{\chi_2'(\eta)\eta}{\chi_2(\eta)}.$$

By Lemma 25, we have

$$\begin{aligned} \chi_2(\eta) &\geq \frac{(\sigma\eta L)^{1/3}}{2L} - \eta - \frac{\eta}{2(\sigma\eta L)^{1/3}} \\ &\stackrel{(a)}{\geq} \frac{(\sigma\eta L)^{1/3}}{2L} - 2\frac{\eta}{(\sigma\eta L)^{1/3}} \\ &\stackrel{(b)}{\geq} \frac{(\sigma\eta L)^{1/3}}{4L} \end{aligned}$$

where in (a), we have used $(\sigma\eta L)^{1/3} < \sigma/9 < 1$, and in (b), we have used $\frac{(\sigma\eta L)^{1/3}}{4L} \geq 2\frac{\eta}{(\sigma\eta L)^{1/3}}$, which is equivalent to $\eta L \leq \sigma^2/8^3$ and follows from our step size condition.

Now, we calculate $\chi_2'(\eta)$,

$$\begin{aligned} |\chi_2'(\eta)| &= \left| \frac{\theta'(\eta)}{2L} - 1 - \frac{(\theta(\eta) - \sigma) - \eta\theta'(\eta)}{2(\theta(\eta) - \sigma)^2} \right| \\ &\leq \frac{\theta'(\eta)}{2L} + 1 + \frac{1}{2(\theta(\eta) - \sigma)} + \frac{\eta\theta'(\eta)}{2(\theta(\eta) - \sigma)^2} \\ &\leq \frac{5}{2} \frac{1}{(\eta L)^{2/3}} + 1 + \frac{1}{2(\sigma\eta L)^{1/3}} + \frac{1}{2} \frac{\eta 5L^{1/3}/\eta^{2/3}}{(\sigma\eta L)^{2/3}} \end{aligned}$$

$$\begin{aligned}
&= \frac{5}{2} \frac{1}{(\eta L)^{2/3}} + 1 + \frac{1}{2(\sigma \eta L)^{1/3}} + \frac{5}{2} \frac{1}{\sigma^{2/3}(\eta L)^{1/3}} \\
&\leq \frac{7}{(\sigma \eta L)^{2/3}}.
\end{aligned}$$

Therefore,

$$|\xi_2'(y)| \leq \frac{\eta \cdot 7 / (\sigma \eta L)^{2/3}}{(\sigma \eta L)^{1/3} / (4L)} = \frac{28}{\sigma}.$$

Hence, ξ_2 is $\frac{28}{\sigma}$ -Lipschitz continuous. \square

B.4. Proof of the Intermediate Result (Lemma 18). The proof essentially follows the same argument as the CNGD (see e.g. Section 2.2 in [24]).

First, we prove (35). Notice (35) is true for $t = 0$. Then, assume it's true for t , then for $t + 1$, we have

$$\begin{aligned}
\Phi_{t+1}(\omega) &= (1 - \alpha_t)\Phi_t(\omega) + \alpha_t(\hat{f}(t) + \langle g(t), \omega - \bar{y}(t) \rangle) \\
&\leq (1 - \alpha_t)\Phi_t(\omega) + \alpha_t f(\omega) \\
&\leq f(\omega) + \lambda_{t+1}(\Phi_0(\omega) - f(\omega)).
\end{aligned}$$

where in the first inequality we have used (32) and in the second inequality we have used the induction assumption.

Next, we prove (36). It's clear that Φ_t is always a quadratic function. Notice that $\nabla^2 \Phi_0(\omega) = \gamma_0 I$ and

$$\nabla^2 \Phi_{t+1}(\omega) = (1 - \alpha_t) \nabla^2 \Phi_t(\omega).$$

We get $\nabla^2 \Phi_t(\omega) = \gamma_t I$ for all t , by the definition of γ_t .

We next claim, Φ_t , as a quadratic function, achieves minimum at $\bar{v}(t)$. We prove the claim by induction. Firstly, Φ_0 achieves minimum at $\bar{v}(0)$. Assume Φ_t achieves minimum at $\bar{v}(t)$. Then, $\nabla \Phi_t(\omega) = \gamma_t(\omega - \bar{v}(t))$. Notice that by (34),

$$\begin{aligned}
\nabla \Phi_{t+1}(\omega) &= (1 - \alpha_t) \nabla \Phi_t(\omega) + \alpha_t g(t) \\
&= (1 - \alpha_t) \gamma_t(\omega - \bar{v}(t)) + \alpha_t g(t).
\end{aligned}$$

Set $\omega = \bar{v}(t + 1)$, we get

$$\begin{aligned}
\nabla \Phi_{t+1}(\bar{v}(t + 1)) &= (1 - \alpha_t) \gamma_t(\bar{v}(t + 1) - \bar{v}(t)) + \alpha_t g(t) \\
&= [-(1 - \alpha_t) \gamma_t \frac{\eta t}{\alpha_t} + \alpha_t] g(t) \\
&= 0
\end{aligned}$$

where the last equality follows from the fact that $\alpha_t^2 = \eta_t(1 - \alpha_t)\gamma_t$, which can be proved recursively. It's true for $t = 0$ by definition of γ_0 . And note $\alpha_{t+1}^2 = \frac{\eta_{t+1}}{\eta_t}(1 - \alpha_{t+1})\alpha_t^2 = (1 - \alpha_{t+1})\eta_{t+1}(1 - \alpha_t)\gamma_t = \eta_{t+1}(1 - \alpha_{t+1})\gamma_{t+1}$. Hence Φ_{t+1} achieves optimum at $\bar{v}(t + 1)$, and hence the claim.

We have proven $\Phi_t(\omega)$ is a quadratic function that achieves minimum at $\bar{v}(t)$ and satisfies $\nabla^2 \Phi_t = \gamma_t I$. This implies Φ_t can be written in the form of (36) for some unique $\phi_t^* \in \mathbb{R}$.

We next show ϕ_t^* satisfies (37). Clearly, $\phi_0^* = f(\bar{x}(0))$. We now derive a recursive formula for ϕ_t^* . Since by (34)

$$\Phi_{t+1}(\bar{y}(t)) = (1 - \alpha_t)\Phi_t(\bar{y}(t)) + \alpha_t \hat{f}(t).$$

Also by (36), we get

$$\begin{aligned} & \phi_{t+1}^* + \frac{\gamma_{t+1}}{2} \|\bar{y}(t) - \bar{v}(t+1)\|^2 \\ &= (1 - \alpha_t) \left[\phi_t^* + \frac{\gamma_t}{2} \|\bar{y}(t) - \bar{v}(t)\|^2 \right] + \alpha_t \hat{f}(t). \end{aligned}$$

Since $\bar{v}(t+1) - \bar{y}(t) = (\bar{v}(t) - \bar{y}(t)) - \frac{\eta_t}{\alpha_t} g(t)$, therefore

$$\|\bar{v}(t+1) - \bar{y}(t)\|^2 = \|\bar{v}(t) - \bar{y}(t)\|^2 - 2 \frac{\eta_t}{\alpha_t} \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle + \frac{\eta_t^2}{\alpha_t^2} \|g(t)\|^2.$$

Hence

$$\begin{aligned} \phi_{t+1}^* &= (1 - \alpha_t) \phi_t^* + \alpha_t \hat{f}(t) - \frac{1}{2} \frac{\eta_t^2 \gamma_{t+1}}{\alpha_t^2} \|g(t)\|^2 + \frac{\gamma_{t+1}}{\alpha_t} \eta_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle \\ &= (1 - \alpha_t) \phi_t^* + \alpha_t \hat{f}(t) - \frac{1}{2} \eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle. \end{aligned}$$

□

B.5. Proof of Lemma 23. We provide a comparison lemma that will be helpful for the analysis.

LEMMA 27. *Given two step size sequences $(\eta_t)_t$ and $(\eta'_t)_t$ that satisfy, $\eta_0 \leq \eta'_0$, and $\forall t$, $\frac{\eta_{t+1}}{\eta_t} \leq \frac{\eta'_{t+1}}{\eta'_t}$, then $\alpha_t \leq \alpha'_t$.*

Proof: We prove the statement by induction. First, $\alpha_0 = \sqrt{\eta_0 L} \leq \sqrt{\eta'_0 L} = \alpha'_0$. Next, assume $\alpha_t \leq \alpha'_t$, then $\frac{\eta_{t+1}}{\eta_t} \alpha_t^2 \leq \frac{\eta'_{t+1}}{\eta'_t} \alpha_t'^2$. Define function $\xi : (0, 1) \rightarrow (0, \infty)$ with $\xi(y) = y^2/(1-y)$. It's easy to check that ξ is a strictly increasing function and is a bijection. Notice that $\alpha_{t+1} = \xi^{-1}(\frac{\eta_{t+1}}{\eta_t} \alpha_t^2) \leq \xi^{-1}(\frac{\eta'_{t+1}}{\eta'_t} \alpha_t'^2) = \alpha'_{t+1}$. So we are done. □

Proof of Lemma 23: Since $\eta_{t+1} < \eta_t$, we have $\alpha_{t+1}^2 < \alpha_t^2$ and hence α_t is decreasing. Now we derive the asymptotic convergence rate of α_t .

Proof of (i). Firstly, we have,

$$\alpha_{t+1}^2 + \frac{\eta_{t+1}}{\eta_t} \alpha_t^2 \alpha_{t+1} - \frac{\eta_{t+1}}{\eta_t} \alpha_t^2 = 0$$

Hence

$$\begin{aligned} \alpha_{t+1} &= \frac{1}{2} \left(-\frac{\eta_{t+1}}{\eta_t} \alpha_t^2 + \sqrt{\frac{\eta_{t+1}^2}{\eta_t^2} \alpha_t^4 + 4 \frac{\eta_{t+1}}{\eta_t} \alpha_t^2} \right) \\ &= \frac{2 \frac{\eta_{t+1}}{\eta_t} \alpha_t^2}{\frac{\eta_{t+1}}{\eta_t} \alpha_t^2 + \sqrt{\frac{\eta_{t+1}^2}{\eta_t^2} \alpha_t^4 + 4 \frac{\eta_{t+1}}{\eta_t} \alpha_t^2}} \\ &= \frac{2}{1 + \sqrt{1 + 4 \frac{\eta_t}{\eta_{t+1} \alpha_t^2}}}. \end{aligned}$$

Hence

$$\frac{1}{\alpha_{t+1}} - \sqrt{\frac{\eta_t}{\eta_{t+1}}} \frac{1}{\alpha_t} = \frac{1}{2} + \frac{1}{2} \sqrt{1 + 4 \frac{\eta_t}{\eta_{t+1} \alpha_t^2}} - \sqrt{\frac{\eta_t}{\eta_{t+1}}} \frac{1}{\alpha_t}$$

$$\begin{aligned}
(62) \quad &= \frac{1}{2} + \frac{1}{4} \frac{1}{\frac{1}{2} \sqrt{1 + 4 \frac{\eta_t}{\eta_{t+1} \alpha_t^2}} + \sqrt{\frac{\eta_t}{\eta_{t+1}} \frac{1}{\alpha_t}}} \\
&> \frac{1}{2}.
\end{aligned}$$

Hence, $\frac{\sqrt{\eta_t}}{\alpha_t} - \frac{\sqrt{\eta_{t-1}}}{\alpha_{t-1}} > \frac{1}{2} \sqrt{\eta_t}$. Therefore, for $t \geq 1$,

$$(63) \quad \frac{\sqrt{\eta_t}}{\alpha_t} > \frac{\sqrt{\eta_0}}{\alpha_0} + \frac{1}{2} \sum_{k=1}^t \sqrt{\eta_k}.$$

Now we use the comparison lemma (Lemma 27) with a fixed step size $\eta'_t = \eta_0$. Notice that $\eta_{t+1}/\eta_t < \eta'_{t+1}/\eta'_t$, we have $\alpha_t \leq \alpha'_t$. Repeating the argument for (63), we get (63) is also true if we replace η_t with η'_t and α_t with α'_t . Hence, for $t \geq 1$,

$$\frac{\sqrt{\eta_0}}{\alpha'_t} > \frac{\sqrt{\eta_0}}{\alpha'_0} + \frac{1}{2} t \sqrt{\eta_0} > \frac{1}{2} (t+1) \sqrt{\eta_0}.$$

This implies that for $t \geq 0$, $\alpha'_t \leq \frac{2}{t+1}$. Hence,

$$(64) \quad \alpha_t \leq \alpha'_t \leq \frac{2}{t+1}.$$

This gives part (i) of the lemma. Now we derive a tighter upper bound for α_t which will be used later. Returning to (63), we have

$$\begin{aligned}
\frac{\sqrt{\eta_t}}{\alpha_t} &> \frac{\sqrt{\eta_0}}{\alpha_0} + \frac{1}{2} \sum_{k=1}^t \sqrt{\eta_k} \\
&> \frac{1}{2} \sum_{k=1}^t \frac{\sqrt{\eta}}{(k+t_0)^{\beta/2}} \\
&> \frac{1}{2} \sqrt{\eta} \int_1^{t+1} \frac{1}{(y+t_0)^{\beta/2}} dy \\
&= \frac{1}{2} \sqrt{\eta} \frac{2}{2-\beta} [(t+1+t_0)^{1-\frac{\beta}{2}} - (t_0+1)^{1-\frac{\beta}{2}}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{\alpha_t} &> \frac{1}{2} \sqrt{\eta} \frac{2}{2-\beta} [(t+1+t_0)^{1-\frac{\beta}{2}} - (t_0+1)^{1-\frac{\beta}{2}}] \times \frac{(t+t_0)^{\beta/2}}{\sqrt{\eta}} \\
&> \frac{1}{2-\beta} (t+t_0 - (t+t_0)^{\beta/2} (t_0+1)^{1-\frac{\beta}{2}}).
\end{aligned}$$

Notice that when $t \geq 2$, the right hand side of the above formula is positive. Hence,

$$(65) \quad \alpha_t < \frac{2-\beta}{t+t_0 - (t+t_0)^{\beta/2} (t_0+1)^{1-\frac{\beta}{2}}}, \forall t \geq 2.$$

Proof of (ii). We return to (62), and use (64) to get,

$$\frac{1}{\alpha_{t+1}} - \sqrt{\frac{\eta_t}{\eta_{t+1}}} \frac{1}{\alpha_t} < \frac{1}{2} + \frac{1}{8} \alpha_t \sqrt{\frac{\eta_{t+1}}{\eta_t}} \leq \frac{1}{2} + \frac{1}{4(t+1)}.$$

Hence, for $t \geq 1$,

$$\frac{\sqrt{\eta_t}}{\alpha_t} - \frac{\sqrt{\eta_{t-1}}}{\alpha_{t-1}} \leq \frac{1}{2}\sqrt{\eta_t} + \frac{\sqrt{\eta_t}}{4t}.$$

Therefore, for $t \geq 1$,

$$\begin{aligned} \frac{\sqrt{\eta_t}}{\alpha_t} &\leq \frac{\sqrt{\eta_0}}{\alpha_0} + \frac{1}{2} \sum_{k=1}^t \sqrt{\eta_k} + \sum_{k=1}^t \frac{\sqrt{\eta_k}}{4k} \\ &< \frac{\sqrt{\eta_0}}{\alpha_0} + \frac{1}{2} \sum_{k=1}^t \sqrt{\eta} \frac{1}{(k+t_0)^{\beta/2}} + \sum_{k=1}^t \sqrt{\eta} \frac{1}{4k^{\beta/2+1}} \\ &\leq \frac{\sqrt{\eta_0}}{\alpha_0} + \frac{\sqrt{\eta}}{2} \int_0^t \frac{1}{(y+t_0)^{\beta/2}} dy + \frac{\sqrt{\eta}}{4} [1 + \int_1^t \frac{1}{y^{\beta/2+1}} dy] \\ &= \frac{\sqrt{\eta_0}}{\alpha_0} + \frac{\sqrt{\eta}}{2} \frac{2}{2-\beta} [(t+t_0)^{1-\beta/2} - t_0^{1-\beta/2}] + \frac{\sqrt{\eta}}{4} [1 + \frac{2}{\beta} - \frac{2}{\beta} \frac{1}{t^{\beta/2}}] \\ &\leq \frac{\sqrt{\eta_0}}{\alpha_0} + \frac{\sqrt{\eta}}{2-\beta} (t+t_0)^{1-\beta/2} + \frac{\sqrt{\eta}}{4} [1 + \frac{2}{\beta}]. \end{aligned}$$

Therefore, for $t \geq 1$,

$$\frac{1}{\alpha_t} \leq \frac{1}{2-\beta} (t+t_0) + (\frac{1}{\alpha_0 t_0^{\beta/2}} + \frac{1}{2\beta} + \frac{1}{4})(t+t_0)^{\beta/2}.$$

And then, for $t \geq 1$,

$$\begin{aligned} \frac{2-\beta}{t+t_0} - \alpha_t &\leq \frac{2-\beta}{t+t_0} - \frac{2-\beta}{(t+t_0) + (2-\beta)(\frac{1}{\alpha_0 t_0^{\beta/2}} + \frac{1}{2\beta} + \frac{1}{4})(t+t_0)^{\beta/2}} \\ &= \frac{(2-\beta)^2 (\frac{1}{\alpha_0 t_0^{\beta/2}} + \frac{1}{2\beta} + \frac{1}{4})}{(t+t_0)[(t+t_0)^{1-\frac{\beta}{2}} + (2-\beta)(\frac{1}{\alpha_0 t_0^{\beta/2}} + \frac{1}{2\beta} + \frac{1}{4})]} \\ &= O(\frac{1}{(t+t_0)^{2-\frac{\beta}{2}}}). \end{aligned}$$

Now we consider $\lambda_t = \prod_{k=0}^{t-1} (1 - \alpha_k)$. We have

$$\begin{aligned} \log \lambda_t &\leq - \sum_{k=0}^{t-1} \alpha_k \\ &= -\alpha_0 + \sum_{k=1}^{t-1} (\frac{2-\beta}{k+t_0} - \alpha_k) - \sum_{k=1}^{t-1} \frac{2-\beta}{k+t_0} \\ &\leq -\alpha_0 + O(\sum_{k=1}^{\infty} \frac{1}{(k+t_0)^{2-\frac{\beta}{2}}}) - (2-\beta) \log \frac{t+t_0}{1+t_0} \\ &= -(2-\beta) \log(t+t_0) + O(1) \end{aligned}$$

where we have used $\sum_{k=1}^{\infty} \frac{1}{(k+t_0)^{2-\frac{\beta}{2}}} < \infty$ since $2-\frac{\beta}{2} > 1$. Hence, $\lambda_t = O(\frac{1}{(t+t_0)^{2-\beta}}) = O(\frac{1}{t^{2-\beta}})$, i.e. the part (ii) of this lemma.

Proof of (iii). It is easy to check that $\forall y \in (-1, \infty)$, $\log(1+y) \geq \frac{y}{1+y}$. Therefore,

$$(66) \quad \log \lambda_t \geq -\sum_{k=0}^{t-1} \frac{\alpha_k}{1-\alpha_k} = -\sum_{k=0}^{t-1} \alpha_k - \sum_{k=0}^{t-1} \frac{\alpha_k^2}{1-\alpha_k}.$$

By (64) and the fact $\alpha_k \leq \alpha_0 < \frac{1}{2}$, we have

$$(67) \quad \sum_{k=0}^{t-1} \frac{\alpha_k^2}{1-\alpha_k} \leq 2 \sum_{k=0}^{t-1} \alpha_k^2 \leq 8 \sum_{k=0}^{t-1} \frac{1}{(k+1)^2} \leq 8 \cdot \frac{\pi^2}{6} < 14.$$

We next bound $\sum_{k=0}^{t-1} \alpha_k$. Notice that when $k \geq t_0 + 2$, we have

$$(k+t_0) \geq 2^{1-\frac{\beta}{2}}(k+t_0)^{\beta/2}(1+t_0)^{1-\frac{\beta}{2}}.$$

Hence, by (65), we have when $k \geq t_0 + 2$,

$$\begin{aligned} \alpha_k &< \frac{2-\beta}{k+t_0 - (k+t_0)^{\beta/2}(t_0+1)^{1-\frac{\beta}{2}}} - \frac{2-\beta}{k+t_0} + \frac{2-\beta}{k+t_0} \\ &= \frac{(2-\beta)(k+t_0)^{\beta/2}(t_0+1)^{1-\frac{\beta}{2}}}{(k+t_0)(k+t_0 - (k+t_0)^{\beta/2}(t_0+1)^{1-\frac{\beta}{2}})} + \frac{2-\beta}{k+t_0} \\ &\leq \frac{(2-\beta)(t_0+1)^{1-\frac{\beta}{2}}}{(1-2^{\frac{\beta}{2}-1})(k+t_0)^{2-\frac{\beta}{2}}} + \frac{2-\beta}{k+t_0}. \end{aligned}$$

Hence, for $t > t_0 + 3$,

$$\begin{aligned} \sum_{k=t_0+3}^{t-1} \alpha_k &\leq \sum_{k=t_0+3}^{t-1} \frac{(2-\beta)(t_0+1)^{1-\frac{\beta}{2}}}{(1-2^{\frac{\beta}{2}-1})(k+t_0)^{2-\frac{\beta}{2}}} + \sum_{k=t_0+3}^{t-1} \frac{2-\beta}{k+t_0} \\ &\leq \frac{2(t_0+1)^{1-\frac{\beta}{2}}}{(1-2^{\frac{\beta}{2}-1})(2t_0+2)^{1-\frac{\beta}{2}}} + (2-\beta) \log \frac{t-1+t_0}{2t_0+2} \\ &= \frac{2}{2^{1-\frac{\beta}{2}}-1} + (2-\beta) \log \frac{t-1+t_0}{2t_0+2} \\ &< \frac{4}{(2-\beta) \log 2} + (2-\beta) \log(t-1+t_0). \end{aligned}$$

Notice that, when $t \leq t_0 + 3$, the above inequality is still true since the left hand side is 0 while the right hand side is positive. Hence,

$$\begin{aligned} \sum_{k=0}^{t-1} \alpha_k &\leq \sum_{k=0}^{t_0+2} \alpha_k + \sum_{k=t_0+3}^{t-1} \alpha_k \\ &\leq \sum_{k=0}^{t_0+2} \frac{2}{k+1} + \sum_{k=t_0+3}^{t-1} \alpha_k \\ &\leq 2 + 2 \log(t_0+3) + \frac{4}{(2-\beta) \log 2} + (2-\beta) \log(t-1+t_0) \\ &\leq (2-\beta) \log(t-1+t_0) + 2 \log(t_0+3) + 2 + \frac{6}{2-\beta}. \end{aligned}$$

Therefore, combining the above with (67) and (66), we get

$$\log \lambda_t \geq -(2 - \beta) \log(t - 1 + t_0) - 2 \log(t_0 + 3) - \frac{6}{2 - \beta} - 16.$$

Hence,

$$\lambda_t \geq \frac{1}{(t + t_0)^{2-\beta} (t_0 + 3)^2 e^{16 + \frac{6}{2-\beta}}}.$$

□

B.6. Proof of Theorem 7. In this section, we provide a detailed proof for Theorem 7.

LEMMA 28. *Under the conditions of Theorem 7, we have inequality (32) in Lemma 16, when evaluated at $\omega = \bar{x}(t)$, can be strengthened to,*

$$f(\bar{x}(t)) \geq \hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle + \frac{\mu}{4} \|\bar{x}(t) - \bar{y}(t)\|^2 - \frac{L}{2n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2$$

where $\mu = \mu_0 \gamma$, and γ is the smallest non-zero eigenvalue of the positive semidefinite matrix $A = \frac{1}{n} \sum_{i=1}^n A_i A_i^T$ (Matrix A has at least one nonzero eigenvalue since otherwise, all A_i would be zero.); $L = L_0 \nu$ and $\nu = \max_i \|A_i\|_*^2$.

Proof: It is easy to check that A is a symmetric and positive semidefinite matrix. Let $y \in \mathbb{R}^{1 \times N}$ be any vector in the null space of A . Notice,

$$0 = \langle yA, y \rangle = \frac{1}{n} \sum_{i=1}^n \|yA_i\|^2 \geq 0.$$

This implies $\forall i, yA_i = 0$. Then,

$$\langle g(t), y \rangle = \left\langle \frac{1}{n} \sum_{i=1}^n \nabla h_i(y_i(t)A_i)A_i^T, y \right\rangle = \frac{1}{n} \sum_{i=1}^n \nabla h_i(y_i(t)A_i)A_i^T y^T = 0.$$

This implies that, $g(t)$ lies in the space spanned by the rows of A . By (39) and the fact that $\bar{v}(0) - \bar{y}(0) = 0$, we have $\bar{v}(t) - \bar{y}(t)$ lies in the row space of A . Hence $\bar{x}(t) - \bar{y}(t) = \frac{\alpha_t}{1-\alpha_t} (\bar{y}(t) - \bar{v}(t))$ also lies in the row space of A . Therefore,

$$\langle (\bar{x}(t) - \bar{y}(t))A, \bar{x}(t) - \bar{y}(t) \rangle \geq \gamma \|\bar{x}(t) - \bar{y}(t)\|^2.$$

Therefore,

$$\begin{aligned} f(\bar{x}(t)) &= \frac{1}{n} \sum_{i=1}^n h_i(\bar{x}(t)A_i) \\ &\stackrel{(a)}{\geq} \frac{1}{n} \sum_{i=1}^n [h_i(y_i(t)A_i) + \langle \nabla h_i(y_i(t)A_i), \bar{x}(t)A_i - y_i(t)A_i \rangle + \frac{\mu_0}{2} \|(\bar{x}(t) - y_i(t))A_i\|^2] \\ &= \frac{1}{n} \sum_{i=1}^n [f_i(y_i(t)) + \langle \nabla f_i(y_i(t)), \bar{x}(t) - y_i(t) \rangle + \frac{\mu_0}{2} \|(\bar{x}(t) - y_i(t))A_i\|^2] \\ &\geq \frac{1}{n} \sum_{i=1}^n [f_i(y_i(t)) + \langle \nabla f_i(y_i(t)), \bar{y}(t) - y_i(t) \rangle] + \frac{1}{n} \sum_{i=1}^n \langle \nabla f_i(y_i(t)), \bar{x}(t) - \bar{y}(t) \rangle \end{aligned}$$

$$\begin{aligned}
& + \frac{\mu_0}{2n} \sum_{i=1}^n \left[\frac{1}{2} \|(\bar{x}(t) - \bar{y}(t))A_i\|^2 - \|(\bar{y}(t) - y_i(t))A_i\|^2 \right] \\
\stackrel{(b)}{\geq} & \hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle + \frac{\mu_0}{4n} \sum_{i=1}^n \langle (\bar{x}(t) - \bar{y}(t))A_i A_i^T, \bar{x}(t) - \bar{y}(t) \rangle \\
& - \frac{\mu_0 \nu}{2n} \sum_{i=1}^n \|y_i(t) - \bar{y}(t)\|^2 \\
= & \hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle + \frac{\mu_0}{4} \langle (\bar{x}(t) - \bar{y}(t))A, \bar{x}(t) - \bar{y}(t) \rangle - \frac{\mu_0 \nu}{2n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2 \\
\geq & \hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle + \frac{1}{4} \mu_0 \gamma \|\bar{x}(t) - \bar{y}(t)\|^2 - \frac{L}{2} \frac{1}{n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2
\end{aligned}$$

where in (a) we have used the fact that $h_i(\cdot)$ is μ_0 -strongly convex, and in (b), we have used the definition of $\hat{f}(t)$, and the fact that $\|(\bar{y}(t) - y_i(t))A_i\| \leq \sqrt{\nu} \|\bar{y}(t) - y_i(t)\|$ (since A_i 's spectral norm is upper bounded by $\sqrt{\nu}$). \square

We now finish the proof of Theorem 7.

Proof of Theorem 7: It is easy to check that f_i is convex and L -smooth. It is easy to check under the step size conditions, Lemma 17 holds. We will prove by induction that,

$$(68) \quad \phi_t^* \geq f(\bar{x}(t)).$$

Equation (68) is true for $t = 0$. Next, by (37),

$$\begin{aligned}
\phi_{t+1}^* & = (1 - \alpha_t) \phi_t^* + \alpha_t \hat{f}(t) - \frac{1}{2} \eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle \\
& \stackrel{(a)}{\geq} (1 - \alpha_t) f(\bar{x}(t)) + \alpha_t \hat{f}(t) - \frac{1}{2} \eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle \\
& \stackrel{(b)}{\geq} (1 - \alpha_t) \{ \hat{f}(t) + \langle g(t), \bar{x}(t) - \bar{y}(t) \rangle + \frac{\mu}{4} \|\bar{x}(t) - \bar{y}(t)\|^2 \\
& \quad - \frac{L}{2n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2 \} + \alpha_t \hat{f}(t) - \frac{1}{2} \eta_t \|g(t)\|^2 + \alpha_t \langle g(t), \bar{v}(t) - \bar{y}(t) \rangle \\
(69) \quad & \stackrel{(c)}{\geq} \hat{f}(t) - \frac{1}{2} \eta_t \|g(t)\|^2 + (1 - \alpha_t) \frac{\mu}{4} \|\bar{x}(t) - \bar{y}(t)\|^2 - \frac{L}{2n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2.
\end{aligned}$$

where (a) is due to the induction assumption (68), (b) is due to Lemma 28 and (c) is due to $\alpha_t (\bar{v}(t) - \bar{y}(t)) + (1 - \alpha_t) (\bar{x}(t) - \bar{y}(t)) = 0$. By (33) (Lemma 16),

$$\begin{aligned}
f(\bar{x}(t+1)) & \leq \hat{f}(t) + \langle g(t), \bar{x}(t+1) - \bar{y}(t) \rangle + L \|\bar{x}(t+1) - \bar{y}(t)\|^2 + \frac{L}{n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2 \\
& \leq \hat{f}(t) - (\eta_t - L\eta_t^2) \|g(t)\|^2 + \frac{L}{n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2.
\end{aligned}$$

Combining the above with (69) and using Lemma 17, we have,

$$\begin{aligned}
\phi_{t+1}^* - f(\bar{x}(t+1)) & \geq \left(\frac{1}{2} \eta_t - L\eta_t^2 \right) \|g(t)\|^2 + (1 - \alpha_t) \frac{\mu}{4} \|\bar{x}(t) - \bar{y}(t)\|^2 - \frac{3}{2} \frac{L}{n} \|y(t) - \mathbf{1}\bar{y}(t)\|^2 \\
& \geq \left(\frac{1}{2} \eta_t - L\eta_t^2 \right) \|g(t)\|^2 + (1 - \alpha_t) \frac{\mu}{4} \|\bar{x}(t) - \bar{y}(t)\|^2 \\
& \quad - 3L\kappa^2 \chi_2(\eta_t)^2 (L^2 \|\bar{x}(t) - \bar{y}(t)\|^2 + \frac{64}{(1 - \sigma)^2} L^2 \eta_t^2 \|g(t)\|^2)
\end{aligned}$$

$$\begin{aligned}
&= \left(\frac{1}{2}\eta_t - L\eta_t^2 - \frac{192L^3\kappa^2\chi_2(\eta_t)^2\eta_t^2}{(1-\sigma)^2} \right) \|g(t)\|^2 \\
(70) \quad &+ \left((1-\alpha_t)\frac{\mu}{4} - 3\kappa^2\chi_2(\eta_t)^2L^3 \right) \|\bar{x}(t) - \bar{y}(t)\|^2.
\end{aligned}$$

Since $\eta_t = \eta$, and $\chi_2(\eta) < \frac{2\eta^{1/3}}{L^{2/3}}$, and recalling $\kappa = \frac{6}{1-\sigma}$, we have

$$\begin{aligned}
\frac{1}{2}\eta_t - L\eta_t^2 - \frac{192L^3\kappa^2\chi_2(\eta_t)^2\eta_t^2}{(1-\sigma)^2} &\geq \frac{1}{2}\eta - L\eta^2 - \frac{27648}{(1-\sigma)^4}\eta \cdot (\eta L)^{5/3} \\
&= \eta \left(\frac{1}{2} - L\eta - \frac{27648}{(1-\sigma)^4}(\eta L)^{5/3} \right) \geq 0
\end{aligned}$$

where in the last inequality, we have used $\eta L < \frac{1}{4}$, and $\frac{27648}{(1-\sigma)^4}(\eta L)^{5/3} < \frac{1}{4}$ ($\Leftarrow \eta L < \frac{(1-\sigma)^{2.4}}{1063}$), all following from our step size condition.

Next, since $\alpha_t < \alpha_0 \leq \frac{1}{2}$ and $\chi_2(\eta) < \frac{2\eta^{1/3}}{L^{2/3}}$, we have,

$$(1-\alpha_t)\frac{\mu}{4} - 3\kappa^2\chi_2(\eta_t)^2L^3 \geq \frac{\mu}{8} - \frac{432}{(1-\sigma)^2}\eta^{2/3}L^{5/3} \geq 0$$

where the last inequality (equivalent to $\eta^{2/3} < \frac{\mu}{L^{5/3}} \frac{(1-\sigma)^2}{3456}$) follows from the step size condition. Hence, returning to (70), we get $\phi_{t+1}^* \geq f(\bar{x}(t+1))$. Therefore the induction is finished and (68) is true for all t . Hence, by (35),

$$f(\bar{x}(t)) \leq \phi_t^* \leq \Phi_t(x^*) \leq f^* + \lambda_t(\Phi_0(x^*) - f^*).$$

Therefore $f(\bar{x}(t)) - f^* = O(\lambda_t)$. Following an argument similar to Lemma 23 (or simply let $\beta \rightarrow 0$ in Lemma 23), we will have $\lambda_t = O(1/t^2)$. As a result, $f(\bar{x}(t)) - f^* = O(\frac{1}{t^2})$.