Journal of Machine Learning Research 25 (2024) 1-52 Submitted 5/21; Revised 12/23; Published 9/24

Accelerated Gradient Tracking over Time-varying Graphs for Decentralized Optimization

Huan Li LIHUANSS@NANKAI.EDU.CN
Institute of Robotics and Automatic Information Systems
College of Artificial Intelligence
Nankai University
Tianjin 300071, China

Zhouchen Lin ✉ ZLIN@PKU.EDU.CN
State Key Lab of General AI, School of Intelligence Science and Technology, Peking University
Institute for Artificial Intelligence, Peking University, Beijing 100871, China
Pazhou Laboratory (Huangpu), Guangzhou 510555, China

Editor: Pradeep Ravikumar

Abstract

Decentralized optimization over time-varying graphs has been increasingly common in modern machine learning with massive data stored on millions of mobile devices, such as in federated learning. This paper revisits the widely used accelerated gradient tracking and extends it to time-varying graphs. We prove that the practical single loop accelerated gradient tracking needs $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^2\sqrt{\frac{L}{\epsilon}})$ and $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ iterations to reach an ϵ-optimal solution over time-varying graphs when the problems are nonstrongly convex and strongly convex, respectively, where γ and σ_γ are two common constants charactering the network connectivity, L and μ are the smoothness and strong convexity constants, respectively, and one iteration corresponds to one gradient oracle call and one communication round. Our convergence rates improve significantly over the ones of $\mathcal{O}(\frac{1}{\epsilon^{5/7}})$ and $\mathcal{O}((\frac{L}{\mu})^{5/7}\frac{1}{(1-\sigma)^{1.5}}\log\frac{1}{\epsilon})$, respectively, which were proved in the original literature of accelerated gradient tracking only for static graphs, where $\frac{\gamma}{1-\sigma_\gamma}$ equals $\frac{1}{1-\sigma}$ when the network is time-invariant. When combining with a multiple consensus subroutine, the dependence on the network connectivity constants can be further improved to $\mathcal{O}(1)$ and $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma})$ for the gradient oracle and communication round complexities, respectively. When the network is static, by employing the Chebyshev acceleration, our complexities exactly match the lower bounds without hiding any poly-logarithmic factor for both nonstrongly convex and strongly convex problems.

Keywords: decentralized optimization, accelerated gradient tracking, time-varying graphs

1. Introduction

Distributed optimization has emerged as a promising framework in machine learning motivated by large-scale data being produced or stored in a network of nodes. Due to the popularity of smartphones and their growing computational power, time-varying graphs are increasingly common in modern distributed optimization, where the communication links in the network may vary with time, and the devices may not be active all the time such that the network may be even unconnected at each time. A typical example is federated

learning (Li et al., 2020b; Kairouz et al., 2021), which involves training a global statistical model from data stored on millions of mobile devices. The physical constraints on each device typically result in only a small fraction of the devices being active at once, and it is possible for an active device to drop out at a given time (Bonawitz et al., 2019). Although centralized network is the predominant topology in most machine learning systems, such as TensorFlow, decentralized network has been a potential alternative because it reduces the high communication cost on the central server (Lian et al., 2017). This motivates us to study decentralized optimization over time-varying graphs. In this paper, we consider the following convex optimization problem:

$$\min_{x\in\mathbb{R}^p} F(x) = \frac{1}{m}\sum_{i=1}^{m} f_{(i)}(x), \tag{1}$$

where the local objective functions $f_{(i)}$ are distributed separately over a network of nodes. The network is mathematically represented as a sequence of time-varying graphs $\{\mathcal{G}^0, \mathcal{G}^1, ...\}$, and each graph instance $\mathcal{G}^k$ consists of a fixed set of agents $\mathcal{V} = \{1, ..., m\}$ and a set of time-varying edges $\mathcal{E}^k$. Agents i and j can exchange information at time k if and only if $(i, j) \in \mathcal{E}^k$. Each agent i privately holds a local objective $f_{(i)}$, and makes its decision only based on the local computations on $f_{(i)}$ and the local information received from its neighbors. The local objective functions are assumed to be smooth. We consider both strongly convex and nonstrongly convex objectives in this paper.

Although decentralized optimization over static graphs has been well studied, for example, lower bounds on the number of communication rounds and gradient or stochastic gradient oracle calls for strongly convex and smooth problems are well-known (Scaman et al., 2017, 2019; Hendrikx et al., 2021), and optimal accelerated algorithms with upper bounds exactly matching the lower bounds are developed (Kovalev et al., 2020; Li et al., 2022), for the time-varying graphs, the problem is more challenging. It is unclear how to design practical accelerated methods with the optimal dependence on the precision ϵ and the condition number of the objectives, exactly matching that of the classical centralized accelerated gradient descent. In this paper, we aim to address this question.

1.1 Notations and Assumptions

Throughout this article, we denote $x_{(i)}$ to be the local variable for agent i. We use the subscript (i) to distinguish the ith element of vector x. To write the algorithm in a compact form, we introduce the aggregate objective function $f(\mathbf{x})$ with its aggregate variable $\mathbf{x} \in \mathbb{R}^{m\times p}$ and aggregate gradient $\nabla f(\mathbf{x}) \in \mathbb{R}^{m\times p}$ as

$$f(\mathbf{x}) = \sum_{i=1}^{m} f_{(i)}(x_{(i)}), \quad \mathbf{x} = \begin{pmatrix} x_{(1)}^\top \\ \vdots \\ x_{(m)}^\top \end{pmatrix}, \quad \nabla f(\mathbf{x}) = \begin{pmatrix} \nabla f_{(1)}(x_{(1)})^\top \\ \vdots \\ \nabla f_{(m)}(x_{(m)})^\top \end{pmatrix}. \tag{2}$$

Denote $\mathbf{x}^k$ to be the value at iteration k. For scalars, for example, θ, we use θ_k instead of θ^k to denote the value at iteration k, while the latter represents its kth power. Specially, $x^\top$ means the transpose of x. We denote $\|\cdot\|$ to be the Frobenius norm for matrices and the ℓ_2 Euclidean norm for vectors uniformly, and $\|\cdot\|_2$ as the spectral norm of matrices. Denote I

as the identity matrix and $\mathbf{1}$ as the column vector of m ones. Assume that problem (1) has a solution, and let x^* be any one of them. Define the average variable across all the local variables as

$$\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_{(i)}, \quad \overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_{(i)}, \quad \overline{z} = \frac{1}{m}\sum_{i=1}^{m} z_{(i)}, \quad \overline{s} = \frac{1}{m}\sum_{i=1}^{m} s_{(i)}, \tag{3}$$

where x, y, z, and s will be used later in the development of the algorithm. Define operator $\Pi = I - \frac{\mathbf{1}\mathbf{1}^\top}{m}$ to measure the consensus violation such that

$$\Pi\mathbf{x} = \begin{pmatrix} x_{(1)}^\top - \overline{x}^\top \\ \cdots \\ x_{(m)}^\top - \overline{x}^\top \end{pmatrix}. \tag{4}$$

We make the following assumptions for each local objective function in problem (1).

Assumption 1

1. *Each $f_{(i)}(x)$ is μ-strongly convex: $f_{(i)}(y) \geq f_{(i)}(x) + \langle \nabla f_{(i)}(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$. Especially, we allow μ to be zero throughout this paper, and in this case we say $f_{(i)}(x)$ is convex.*

2. *Each $f_{(i)}(x)$ is L-smooth, that is, $f_{(i)}(x)$ is differentiable and its gradient is L-Lipschitz continuous: $\|\nabla f_{(i)}(y) - \nabla f_{(i)}(x)\| \leq L\|y - x\|$.*

A direct consequence of the smoothness and convexity assumptions is the following property (Nesterov, 2004):

$$\frac{1}{2L}\|\nabla f_{(i)}(y) - \nabla f_{(i)}(x)\|^2 \leq f_{(i)}(y) - f_{(i)}(x) - \langle \nabla f_{(i)}(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2. \tag{5}$$

The information exchange between different agents in the network is realized through a gossip matrix such that communication can be represented as a matrix multiplication with the gossip matrix. When the network is static, we make the following standard assumptions for the gossip matrix $W \in \mathbb{R}^{m\times m}$ (Qu and Li, 2018):

Assumption 2

1. *(Decentralized property) $W_{i,j} > 0$ if and only if $(i,j) \in \mathcal{E}$ or $i = j$. Otherwise, $W_{i,j} = 0$.*

2. *(Double stochasticity) $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top W = \mathbf{1}^\top$.*

Note that we do not assume that W is symmetric. If the network is connected, Assumption 2 implies that the second largest singular value σ of W is less than 1 (its largest one equals 1), that is, $\sigma = \|W - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\|_2 < 1$. Moreover, we have the following classical consensus contraction:

$$\|\Pi W\mathbf{x}\| = \left\|\left(W - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\right)\mathbf{x}\right\| = \left\|\left(W - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\right)\left(I - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\right)\mathbf{x}\right\| \leq \sigma\|\Pi\mathbf{x}\|. \tag{6}$$

We often use $\frac{1}{1-\sigma}$ as the condition number of the communication network.

When the network is time-varying, each graph instance $\mathcal{G}^k$ associates with a gossip matrix W^k. Denote

$$W^{k,\gamma} = W^k W^{k-1} \cdots W^{k-\gamma+1}, \quad \text{for any } k \geq \gamma - 1, \tag{7}$$

$W^{k,0} = I$, and we follow (Nedić et al., 2017) to make the following standard assumptions for the sequence of gossip matrices $\{W^k\}_{k=0}^{\infty}$.

Assumption 3

1. *(Decentralized property)* $W_{i,j}^k > 0$ *if and only if* $(i,j) \in \mathcal{E}^k$ *or* $i = j$. *Otherwise,* $W_{i,j}^k = 0$.
2. *(Double stochasticity)* $W^k \mathbf{1} = \mathbf{1}$ *and* $\mathbf{1}^\top W^k = \mathbf{1}^\top$.
3. *(Joint spectrum property) There exists a constant integer* γ *such that*

$$\sigma_\gamma < 1, \quad \textit{where} \quad \sigma_\gamma = \sup_{k \geq \gamma - 1} \left\| W^{k,\gamma} - \frac{1}{m} \mathbf{1}\mathbf{1}^\top \right\|_2.$$

Assumption 3 is weaker than the assumption that every graph $\mathcal{G}^k$ is connected. A typical example of the gossip matrix satisfying Assumption 3 is the Metropolis weight over γ-connected graphs. The former is defined as

$$W_{ij}^k = \begin{cases} 1/(1 + \max\{d_i^k, d_j^k\}), & \text{if } (i,j) \in \mathcal{E}^k, \\ 0, & \text{if } (i,j) \neq \mathcal{E}^k \text{ and } i \neq j, \\ 1 - \sum_{l \in \mathcal{N}_{(i)}^k} W_{il}^k, & \text{if } i = j, \end{cases} \tag{8}$$

where $\mathcal{N}_{(i)}^k$ is the set of neighbors of agent i at time k, and $d_i^k = |\mathcal{N}_{(i)}^k|$ is the degree. The γ-connected graph sequence is defined as follows (Nedić et al., 2017).

Definition 1 *The time-varying undirected graph sequence* $\{\mathcal{V}, \mathcal{E}^k\}_{k=0}^{\infty}$ *is* γ*-connected if there exists some positive integer* γ *such that the union of these* γ *consecutive undirected graphs* $\{\mathcal{V}, \cup_{r=k}^{k+\gamma-1} \mathcal{E}^r\}$ *is connected for all* $k = 0, 1, ...$

When Assumption 3 holds, we have the following γ-step consensus contraction:

$$\|\Pi W^{k,\gamma} \mathbf{x}\| \leq \sigma_\gamma \|\Pi \mathbf{x}\|, \qquad \text{for any } k \geq \gamma - 1. \tag{9}$$

When the algorithm proceeds less than γ steps, we only have

$$\|\Pi W^{k,t} \mathbf{x}\| \leq \|\Pi \mathbf{x}\|, \qquad \text{for any } 0 \leq t < \gamma \text{ and } k \geq t - 1. \tag{10}$$

In decentralized optimization, people often use communication round complexity and gradient oracle complexity to measure the convergence speed. The former means the number of communication rounds to reach an ϵ-optimal solution with $F(x) - F(x^*) \leq \epsilon$, while the latter means the number of gradient oracle calls. In one communication round, all the agents can receive $\mathcal{O}(1)$ vectors, such as $x_{(j)}^k$, from each of its neighbors in parallel, which can be represented as $W^k \mathbf{x}^k$ mathematically. In one gradient oracle call, all the agents call the oracle to compute their local gradients $\nabla f_{(i)}(x_{(i)}^k)$ in parallel.

1.2 Literature Review

In this section, we briefly review the decentralized algorithms over static graphs and time-varying graphs, mainly focusing on the accelerated methods. We emphasize gradient tracking (Nedić et al., 2017) and its acceleration (Qu and Li, 2020), which are mostly relevant for our work. Tables 1 and 2 sum up the complexity comparisons of the state-of-the-art methods.

1.2.1 Decentralized Optimization over Static Graphs

Decentralized optimization has been studied for a long time (Bertsekas, 1983; Tsitsiklis et al., 1986). The representative decentralized algorithms include distributed gradient/subgradient descent (DGD) (Nedić and Ozdaglar, 2009; Nedić, 2011; Ram et al., 2010; Yuan et al., 2016), EXTRA (Shi et al., 2015b,a), gradient tracking (Nedić et al., 2017; Qu and Li, 2018; Xu et al., 2015; Xin et al., 2018), NIDS (Li et al., 2019), as well as the dual based methods, such as dual ascent (Terelius et al., 2011), dual averaging (Duchi et al., 2012), ADMM (Wei and Ozdaglar, 2013; Iutzeler et al., 2016; Makhdoumi and Ozdaglar, 2017), and the primal-dual method (Lan et al., 2020; Scaman et al., 2018; Hong et al., 2017; Jakovetić, 2019). Among these methods, gradient tracking has the $\mathcal{O}((\frac{L}{\mu}+\frac{1}{(1-\sigma)^2})\log\frac{1}{\epsilon})$ communication round and gradient oracle complexities for strongly convex problems and the $\mathcal{O}(\frac{L}{\epsilon(1-\sigma)^2})$ complexities for nonstrongly convex ones. Recently, accelerated decentralized methods have gained significant attention due to their provable faster convergence rates.

Accelerated Methods for Strongly Convex and Smooth Decentralized Optimization. The accelerated methods which can be applied to this scenario include the accelerated distributed Nesterov gradient descent (Acc-DNGD) (Qu and Li, 2020), the robust distributed accelerated stochastic gradient method (Fallah et al., 2022), the multi-step dual accelerated method (Scaman et al., 2017, 2019), accelerated penalty method (APM) (Li et al., 2020a; Dvinskikh and Gasnikov, 2021), the multi-consensus decentralized accelerated gradient descent (Mudag) (Ye et al., 2023, 2020), accelerated EXTRA (Li and Lin, 2020; Li et al., 2022), the decentralized accelerated augmented Lagrangian method (Arjevani et al., 2020), and the accelerated proximal alternating predictor-corrector method (APAPC) (Kovalev et al., 2020). Scaman et al. (2017, 2019) proved the $\Omega(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon})$ and $\Omega(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ lower bounds for communication rounds and gradient oracle calls, respectively. To the best of our knowledge, APAPC combined with the Chebyshev acceleration (CA) (Arioli and Scott, 2014) is the first to exactly achieve these lower bounds without hiding any poly-logarithmic factor. Although gradient tracking has been widely used in practice, its accelerated variant, Acc-DNGD, only has the $\mathcal{O}((\frac{L}{\mu})^{5/7}\frac{1}{(1-\sigma)^{1.5}}\log\frac{1}{\epsilon})$ communication round and gradient oracle complexities (Qu and Li, 2020).

Accelerated Methods for Nonstrongly Convex and Smooth Decentralized Optimization. The accelerated methods for this scenario are much scarcer. Examples include the distributed Nesterov gradient with consensus (Jakovetić et al., 2014a), Acc-DNGD (Qu and Li, 2020), APM (Li et al., 2020a; Dvinskikh and Gasnikov, 2021), accelerated EXTRA (Li and Lin, 2020), and the accelerated dual ascent (Uribe et al., 2021), where the last one adds a small regularizer to translate the problem to a strongly convex and smooth one. Scaman et al. (2019) proved the $\Omega(\sqrt{\frac{L}{\epsilon(1-\sigma)}})$ communication round complexity lower bound and

Table 1: Comparisons among the state-of-the-art complexities of decentralized methods over static graphs, as well as those of gradient tracking and its accelerated variant Acc-DNGD. Double loop means the method needs to call a subroutine with multiple steps at each iteration, such as the Chebyshev acceleration, the multiple consensus, the gradient evaluation of Fenchel conjugate, or the minimization of a subproblem.

Methods	gradient oracle complexity	communication round complexity	single or double loop
Nonstrongly convex and smooth functions			
Gradient tracking (Qu and Li, 2018)	$\mathcal{O}\left(\frac{L}{\epsilon(1-\sigma)^2}\right)$	$\mathcal{O}\left(\frac{L}{\epsilon(1-\sigma)^2}\right)$	single
Acc-DNGD (Qu and Li, 2020)	$\mathcal{O}\left(\frac{1}{\epsilon^{5/7}}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^{5/7}}\right)$	single
APM (Li et al., 2020a) (Dvinskikh and Gasnikov, 2021)	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	double
Acc-EXTRA (Li and Lin, 2020)	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	double
Our results for Acc-GT	$\mathcal{O}\left(\frac{1}{(1-\sigma)^2}\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\frac{1}{(1-\sigma)^2}\sqrt{\frac{L}{\epsilon}}\right)$	single
Our results for Acc-GT+CA	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon(1-\sigma)}}\right)$	double
Lower bounds (Scaman et al., 2019)	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon(1-\sigma)}}\right)$	\
Strongly convex and smooth functions			
Gradient tracking (Alghunaim et al., 2021)	$\mathcal{O}\left(\left(\frac{L}{\mu}+\frac{1}{(1-\sigma)^2}\right)\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\left(\frac{L}{\mu}+\frac{1}{(1-\sigma)^2}\right)\log\frac{1}{\epsilon}\right)$	single
Acc-DNGD (Qu and Li, 2020)	$\mathcal{O}\left(\left(\frac{L}{\mu}\right)^{5/7}\frac{1}{(1-\sigma)^{1.5}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\left(\frac{L}{\mu}\right)^{5/7}\frac{1}{(1-\sigma)^{1.5}}\log\frac{1}{\epsilon}\right)$	single
APAPC+CA (Kovalev et al., 2020)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	double
Our results for Acc-GT	$\mathcal{O}\left(\sqrt{\frac{L}{\mu(1-\sigma)^3}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu(1-\sigma)^3}}\log\frac{1}{\epsilon}\right)$	single
Our results for Acc-GT+CA	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	double
Lower bounds (Scaman et al., 2019)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	\

the $\Omega(\sqrt{\frac{L}{\epsilon}})$ gradient oracle complexity lower bound. To the best of our knowledge, there is no method matching these lower bounds exactly without hiding any poly-logarithmic factor. APM comes close to this target, but with an additional $\mathcal{O}(\log\frac{1}{\epsilon})$ factor in the communication round complexity. Acc-DNGD, acceleration of gradient tracking, only has the $\mathcal{O}(\frac{1}{\epsilon^{5/7}})$ complexities of communication rounds and gradient oracles, originally proved in (Qu and Li, 2020). Note that the dependence on $1-\sigma$, a small constant charactering

the network connectivity, was not explicitly given in (Qu and Li, 2020). Xu et al. (2020) proposed an accelerated primal dual method, however, their complexities remain $\mathcal{O}(\frac{1}{\epsilon})$.

1.2.2 Decentralized Optimization over Time-varying Graphs

We review the decentralized algorithms over time-varying graphs in two scenarios. In the first scenario, the network may not be connected at every time, but it is assumed to be γ-connected. In the second scenario, the network is assumed to be connected at every time.

Not Connected at Every Time but γ-connected. In this scenario, DIGing (that is, gradient tracking over time-varying graphs) (Nedić et al., 2017), PANDA (Maros and Jalden, 2018, 2019), the time-varying $\mathcal{AB}$/push-pull method (Saadatniaki et al., 2020), the decentralized stochastic gradient descent (SGD) (Koloskova et al., 2020), and the push-sum based methods (Nedić and Olshevsky, 2016, 2015; Nedić et al., 2017) are the representative non-accelerated methods over time-varying graphs for convex problems, as well as NEXT (Lorenzo and Scutari, 2016) and SONATA (Scutari and Sun, 2019) for nonconvex problems. When combing with Nesterov's acceleration, to the best of our knowledge, the decentralized accelerated gradient descent with consensus subroutine (DAGD-C) (Rogozin et al., 2021b,a) is the only accelerated method for strongly convex and smooth objectives with explicit complexities in this general time-varying setting. However, the communication round complexity of DAGD-C has an additional $\mathcal{O}(\log \frac{1}{\epsilon})$ factor compared with the classical centralized accelerated gradient method. For nonstrongly convex and smooth problems, no literature studies the accelerated methods over time-varying graphs. While APM (Li et al., 2020a) was originally designed for static graphs, it can be easily extended to the time-varying case. However, as introduced in the previous section, APM also has an additional $\mathcal{O}(\log \frac{1}{\epsilon})$ factor in the communication round complexity. Both DAGD-C and APM are double-loop methods, where one gradient is computed at each iteration of the outer loop, and multiple rounds of consensus communications follow up in the inner loop. The multiple consensus double loop may limit the applications of DAGD-C and APM. See the discussions in Remark 16.

Connected at Every Time. In this scenario, the literature is rich and many distributed methods originally designed over static graphs, such as Acc-DNGD, can be directly used. Kovalev et al. (2021b,a) proposed a dual based method named ADOM and its primal-only extension ADOM+, where the latter has the state-of-the-art $\mathcal{O}(\frac{1}{1-\sigma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ communication round complexity and the $\mathcal{O}(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ gradient oracle complexity for strongly convex and smooth problems. Kovalev et al. (2021a) also established the lower bounds showing that their ADOM+ is optimal. Rogozin et al. (2020) gave the complexity of $\mathcal{O}(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon})$ under a stronger assumption that the network changes slowly in the sense that the number of network changes cannot exceed some percentage of the number of total iterations. Nguyen et al. (2024) studied the accelerated AB/push-pull method over directed graphs, but no accelerated rate is proved.

1.3 Contributions

In this paper, we study accelerated gradient tracking over time-varying graphs with sharper complexities. We give our analysis over static graphs and time-varying graphs in a uni-

Table 2: Comparisons among the state-of-the-art complexities of decentralized methods over time-varying graphs. We only compare with the methods working over γ-connected graphs.

Methods	gradient oracle complexity	communication round complexity	single or double loop
Nonstrongly convex and smooth functions			
DGD[1] (Koloskova et al., 2020)	$\mathcal{O}\left(\frac{\gamma\bar{\zeta}\sqrt{L}}{(1-\sigma_\gamma)\epsilon^{3/2}}+\frac{\gamma}{1-\sigma_\gamma}\frac{L}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\gamma\bar{\zeta}\sqrt{L}}{(1-\sigma_\gamma)\epsilon^{3/2}}+\frac{\gamma}{1-\sigma_\gamma}\frac{L}{\epsilon}\right)$	single
APM (Li et al., 2020a)	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\epsilon}}\log\frac{1}{\epsilon}\right)$	double
Our results for Acc-GT	$\mathcal{O}\left(\frac{\gamma^2}{(1-\sigma_\gamma)^2}\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\frac{\gamma^2}{(1-\sigma_\gamma)^2}\sqrt{\frac{L}{\epsilon}}\right)$	single
Our results for Acc-GT+ multiple consensus	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\epsilon}}\right)$	double
Strongly convex and smooth functions			
DGD (Koloskova et al., 2020)	$\mathcal{O}\left(\frac{\gamma\bar{\zeta}\sqrt{L}}{\mu(1-\sigma_\gamma)\sqrt{\epsilon}}+\frac{\gamma}{1-\sigma_\gamma}\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\gamma\bar{\zeta}\sqrt{L}}{\mu(1-\sigma_\gamma)\sqrt{\epsilon}}+\frac{\gamma}{1-\sigma_\gamma}\frac{L}{\mu}\log\frac{1}{\epsilon}\right)$	single
DIGing (Nedić et al., 2017)	$\mathcal{O}\left(\sqrt{m}\left(\frac{L}{\mu}\right)^{1.5}\frac{\gamma^3}{(1-\sigma_\gamma)^2}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{m}\left(\frac{L}{\mu}\right)^{1.5}\frac{\gamma^3}{(1-\sigma_\gamma)^2}\log\frac{1}{\epsilon}\right)$	single
DAGD-C (Rogozin et al., 2021b)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\left(\log\frac{1}{\epsilon}\right)^2\right)$	double
Our results for Acc-GT	$\mathcal{O}\left(\left(\frac{\gamma}{1-\sigma_\gamma}\right)^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\left(\frac{\gamma}{1-\sigma_\gamma}\right)^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	single
Our results for Acc-GT+ multiple consensus	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	double

fied framework. The former scenario provides the basis and insights for the latter. Our contributions are summarized as follows:

1. For time-varying graphs, our contributions include:

   (a) When the local objective functions are nonstrongly convex and smooth, we prove the $\mathcal{O}(\frac{\gamma^2}{(1-\sigma_\gamma)^2}\sqrt{\frac{L}{\epsilon}})$ complexities of communication rounds and gradient oracle calls for the practical single loop accelerated gradient tracking (Acc-GT). When combing with a multiple consensus subroutine, our complexities can be improved to $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\epsilon}})$ for communication rounds and $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ for gradient oracles. The number of our communication rounds is less than that of the state-of-the-art APM (Li et al., 2020a) by a $\mathcal{O}(\log\frac{1}{\epsilon})$ factor, while that of our gradient oracle calls is the same as that of APM.

1. The method in (Koloskova et al., 2020) was designed for stochastic decentralized optimization. We recover the complexities for deterministic optimization by setting the variance of the stochastic gradient to be zero. On the other hand, $\bar{\zeta}=\frac{1}{m}\sum_{i=1}^m\|\nabla f_{(i)}(x^*)\|^2$.

(b) When the local objective functions are strongly convex and smooth, we prove the $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ communication round and gradient oracle complexities for the practical single loop Acc-GT. When combing with the multiple consensus subroutine, we can improve the communication round complexity to $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ and the gradient oracle complexity to $\mathcal{O}(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$. The number of our communication rounds is less than that of the state-of-the-art DAGD-C (Rogozin et al., 2021b) by a $\mathcal{O}(\log\frac{1}{\epsilon})$ factor, while our gradient oracle calls remain the same as that of DAGD-C.

(c) To the best of our knowledge, this is the first time that the communication round upper bound with the optimal dependence on the precision ϵ and condition number L/μ is given for both nonstrongly convex and strongly convex problems. More importantly, they are established for a practical single loop algorithm.

2. For static graphs as a special case, our contributions include:

(a) When the local objective functions are nonstrongly convex and smooth, we prove the $\mathcal{O}(\frac{1}{(1-\sigma)^2}\sqrt{\frac{L}{\epsilon}})$ complexities of communication rounds and gradient oracles for the practical single loop Acc-GT, which significantly improve over the existing $\mathcal{O}(\frac{1}{\epsilon^{5/7}})$ ones originally proved in (Qu and Li, 2020). When combing with the Chebyshev acceleration, we can improve the complexities to $\mathcal{O}(\sqrt{\frac{L}{\epsilon(1-\sigma)}})$ for communication rounds and $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ for gradient oracles, which exactly match the complexity lower bounds. As far as we know, we are the first to establish the optimal upper bounds for nonstrongly convex and smooth problems, which exactly match the corresponding lower bounds without hiding any poly-logarithmic factor.

(b) When the local objective functions are strongly convex and smooth, we prove the $\mathcal{O}(\sqrt{\frac{L}{\mu(1-\sigma)^3}}\log\frac{1}{\epsilon})$ communication round and gradient oracle complexities for the practical single loop Acc-GT, which improves over the existing $\mathcal{O}((\frac{L}{\mu})^{5/7}\frac{1}{(1-\sigma)^{1.5}}\log\frac{1}{\epsilon})$ ones originally given in (Qu and Li, 2020). When combing with the Chebyshev acceleration, the complexities can be further improved to match the corresponding lower bounds and existing optimal upper bounds.

2. Accelerated Gradient Tracking over Time-varying Graphs

We first review the gradient tracking and its accelerated variant, where the latter was only designed over static graphs, and then give our extensions of the accelerated gradient tracking to time-varying graphs with sharper complexities.

2.1 Review of Gradient Tracking and Its Acceleration

Gradient tracking (Nedić et al., 2017; Qu and Li, 2018; Xu et al., 2015; Xin et al., 2018) keeps an auxiliary variable $s_{(i)}^k$ at each iteration for each agent i to track the average of the gradients $\nabla f_{(j)}(x_{(j)}^k)$ for all $j = 1, ..., m$, such that if $x_{(i)}^k$ converges to some point x^∞, $s_{(i)}^k$

converges to $\frac{1}{m}\sum_{i=1}^{m}\nabla f_{(i)}(x^{\infty})$. The auxiliary variable is updated recursively as follows:

$$s_{(i)}^{k}=\sum_{j\in\mathcal{N}_{(i)}}W_{ij}s_{(j)}^{k-1}+\nabla f_{(i)}(x_{(i)}^{k})-\nabla f_{(i)}(x_{(i)}^{k-1}),$$

and each agent uses this auxiliary variable as the descent direction in the general distributed gradient descent framework:

$$x_{(i)}^{k+1}=\sum_{j\in\mathcal{N}_{(i)}}W_{ij}x_{(j)}^{k}-\alpha s_{(i)}^{k},$$

where α is the step size. Writing gradient tracking in the compact form, it reads as follows:

$$\begin{aligned}\mathbf{s}^{k}&=W\mathbf{s}^{k-1}+\nabla f(\mathbf{x}^{k})-\nabla f(\mathbf{x}^{k-1}),\\ \mathbf{x}^{k+1}&=W\mathbf{x}^{k}-\alpha\mathbf{s}^{k}.\end{aligned}$$

Gradient tracking can be used over both static graphs and time-varying graphs (Nedić et al., 2017).

To further accelerate gradient tracking, Qu and Li (2020) employed Nesterov's acceleration technique (Nesterov, 2004) and proposed the following accelerated distributed Nesterov gradient descent for nonstrongly convex problems:

$$\mathbf{y}^{k}=\theta_k\mathbf{z}^{k}+(1-\theta_k)\mathbf{x}^{k}, \tag{12a}$$
$$\mathbf{s}^{k}=W\mathbf{s}^{k-1}+\nabla f(\mathbf{y}^{k})-\nabla f(\mathbf{y}^{k-1}), \tag{12b}$$
$$\mathbf{x}^{k+1}=W\mathbf{y}^{k}-\alpha\mathbf{s}^{k}, \tag{12c}$$
$$\mathbf{z}^{k+1}=W\mathbf{z}^{k}-\frac{\alpha}{\theta_k}\mathbf{s}^{k}. \tag{12d}$$

It can be checked that step (12c) is equivalent to the following one:

$$\mathbf{x}^{k+1}=\theta_k\mathbf{z}^{k+1}+(1-\theta_k)W\mathbf{x}^{k}.$$

When strong convexity is assumed, Qu and Li (2020) fixed θ_k at each iteration and replaced steps (12a) and (12d) by the following two steps:

$$\mathbf{y}^{k}=\frac{\mathbf{x}^{k}+\theta\mathbf{z}^{k}}{1+\theta},\qquad \mathbf{z}^{k+1}=(1-\theta)W\mathbf{z}^{k}+\theta W\mathbf{y}^{k}-\frac{\alpha}{\theta}\mathbf{s}^{k}.$$

The main idea behind the development of the above accelerated algorithms is to relate it to the inexact accelerated gradient descent (Devolder et al., 2014) by taking average of the local variables over all $i=1,...,m$. See Section 3.1 for the details. Tables 1 and 2 list the complexities of gradient tracking and its accelerated variant.

2.2 Extension of Accelerated Gradient Tracking to Time-varying Graphs

Algorithm 1 Accelerated Gradient Tracking (Acc-GT)

Initialize: $x_{(i)}^0 = y_{(i)}^0 = z_{(i)}^0 = x_{int}$, $s_{(i)}^0 = \nabla f_{(i)}(y_{(i)}^0)$, $z_{(i)}^1 = \sum_{j\in\mathcal{N}_{(i)}} W_{ij}^0 z_{(j)}^0 - \frac{\alpha}{\theta_0+\mu\alpha} s_{(i)}^0$, and $x_{(i)}^1 = \theta_0 z_{(i)}^1 + (1-\theta_0)\sum_{j\in\mathcal{N}_{(i)}} W_{ij}^0 x_{(j)}^0$.

for $k = 1, 2, \ldots$ **do**

$$
\begin{aligned}
y_{(i)}^k &= \theta_k z_{(i)}^k + (1-\theta_k) x_{(i)}^k, \\
s_{(i)}^k &= \sum_{j\in\mathcal{N}_{(i)}} W_{ij}^k s_{(j)}^{k-1} + \nabla f_{(i)}(y_{(i)}^k) - \nabla f_{(i)}(y_{(i)}^{k-1}), \\
z_{(i)}^{k+1} &= \frac{1}{1+\frac{\mu\alpha}{\theta_k}} \left(\sum_{j\in\mathcal{N}_{(i)}} W_{ij}^k \left(\frac{\mu\alpha}{\theta_k} y_{(j)}^k + z_{(j)}^k \right) - \frac{\alpha}{\theta_k} s_{(i)}^k \right), \\
x_{(i)}^{k+1} &= \theta_k z_{(i)}^{k+1} + (1-\theta_k) \sum_{j\in\mathcal{N}_{(i)}} W_{ij}^k x_{(j)}^k.
\end{aligned}
$$

end for

In this paper, we study the following accelerated gradient tracking with time-varying gossip matrices:

$$
\begin{aligned}
\mathbf{y}^k &= \theta_k \mathbf{z}^k + (1-\theta_k)\mathbf{x}^k, && \text{(13a)} \\
\mathbf{s}^k &= W^k \mathbf{s}^{k-1} + \nabla f(\mathbf{y}^k) - \nabla f(\mathbf{y}^{k-1}), && \text{(13b)} \\
\mathbf{z}^{k+1} &= \frac{1}{1+\frac{\mu\alpha}{\theta_k}} \left(W^k \left(\frac{\mu\alpha}{\theta_k} \mathbf{y}^k + \mathbf{z}^k \right) - \frac{\alpha}{\theta_k} \mathbf{s}^k \right), && \text{(13c)} \\
\mathbf{x}^{k+1} &= \theta_k \mathbf{z}^{k+1} + (1-\theta_k) W^k \mathbf{x}^k, && \text{(13d)}
\end{aligned}
$$

where we initialize $\mathbf{x}^0$ such that $\Pi\mathbf{x}^0 = 0$. We give the specific descriptions of the method in Algorithm 1 in a distributed way. Step (13b) is the standard gradient tracking, while steps (13a), (13c), and (13d) come from Nesterov's classical accelerated gradient descent (Nesterov, 2004), except that one round of consensus communication is performed by multiplying the aggregate variables with a gossip matrix. We see that algorithm (13a)-(13d) is equivalent to (12a)-(12d) when the gossip matrix is fixed and $\mu = 0$. However, when $\mu > 0$, it is not equivalent to the method proposed in (Qu and Li, 2020). In fact, Nesterov's accelerated gradient methods have several variants, and we choose the one in the form of (13a)-(13d) due to its simple convergence proof.

We follow the proof idea in (Jakovetić et al., 2014a; Qu and Li, 2020) to rewrite the distributed algorithm in the form of inexact accelerated gradient descent. However, we use a different proof framework from (Qu and Li, 2020) with much simpler proofs, and give sharper complexities. See Remark 23 for the differences and the reasons of the convergence rates improvement. On the other hand, for time-varying graphs, unlike the classical analysis relying on the small gain theorem (Nedić et al., 2017), we construct a different way to bound the consensus errors such that the proof framework over static graphs can be extended to time-varying graphs. See the proof of Lemma 25 and the remark following it. Our proof technique may shed new light to decentralized optimization over time-varying graphs, and

gives an alternative to the small gain theorem. There are two advantages of our proof technique: it can be embedded into many algorithm frameworks from the perspective of error analysis, and it can be applied to both strongly convex and nonstrongly convex problems, while the small gain theorem only applies to strongly convex ones.

Our main technical results concerning the convergence rates of the accelerated gradient tracking are summarized in the following two theorems for nonstrongly convex and strongly convex problems, respectively.

Theorem 2 *Suppose that Assumption 1 holds with $\mu = 0$ and Assumption 3 holds for the sequence $\{W^k\}_{k=0}^{T\gamma}$. Let the sequence $\{\theta_k\}_{k=0}^{T\gamma}$ satisfy $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ with $\theta_0 = 1$, let $\alpha \leq \frac{(1-\sigma_\gamma)^4}{21675L\gamma^4}$. Then for algorithm (13a)-(13d), we have for any $T \geq 1$,*

$$F(\overline{x}^{T\gamma+1}) - F(x^*) \leq \frac{2C}{\alpha(T\gamma+1)^2},$$

and

$$\frac{1}{m}\|\Pi \mathbf{x}^{T\gamma}\|^2 \leq \frac{9C}{\alpha L(T\gamma+1)^2},$$

where $C = \|\overline{z}^0 - x^*\|^2 + \frac{\alpha(1-\sigma_\gamma)}{10mL\gamma}\max_{r=0,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2$.

Theorem 3 *Suppose that Assumption 1 holds with $\mu > 0$ and Assumption 3 holds for the sequences $\{W^k\}_{k=0}^{T\gamma}$. Let $\alpha \leq \frac{(1-\sigma_\gamma)^3}{4244L\gamma^3}$ and $\theta_k \equiv \theta = \frac{\sqrt{\mu\alpha}}{2}$. Then for algorithm (13a)-(13d), we have for any $T \geq 1$,*

$$F(\overline{x}^{T\gamma+1}) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^{T\gamma+1} - x^*\|^2 \leq (1-\theta)^{T\gamma+1}C,$$

and

$$\frac{1}{m}\|\Pi \mathbf{x}^{T\gamma}\|^2 \leq (1-\theta)^{T\gamma+1}\frac{4C}{L},$$

where $C = F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{1-\sigma_\gamma}{49mL\gamma(1-\theta)}\mathcal{M}_{\mathbf{s}}^{\gamma,\gamma} + \frac{1459L\gamma^3}{m(1-\theta)(1-\sigma_\gamma)^3}\mathcal{M}_{\mathbf{z}}^{\gamma,\gamma} + \frac{6.6L\gamma}{m(1-\theta)(1-\sigma_\gamma)}\mathcal{M}_{\mathbf{x}}^{\gamma,\gamma}$, $\mathcal{M}_{\mathbf{s}}^{\gamma,\gamma} = \max_{r=1,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2$, and similarly for $\mathcal{M}_{\mathbf{z}}^{\gamma,\gamma}$ and $\mathcal{M}_{\mathbf{x}}^{\gamma,\gamma}$.

When the local objectives are nonstrongly convex, we see from Theorem 2 that algorithm (13a)-(13d) needs $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^2\sqrt{\frac{LC}{\epsilon}})$ communication rounds and gradient oracle calls to find an ϵ-optimal averaged solution (see Remark 31). When strong convexity is assumed, we see from Theorem 3 that both the communication round and gradient oracle complexities are $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$. Our complexities have the optimal dependence on the precision ϵ and the condition number L/μ, matching that of the classical centralized accelerated gradient method. As illustrated in Table 2, our communication round complexities improve over the state-of-the-art APM (Li et al., 2020a) and DAGD-C (Rogozin et al., 2021b) on the dependence of ϵ since they have an additional $\mathcal{O}(\log\frac{1}{\epsilon})$ factor. However, our dependence on $\frac{\gamma}{1-\sigma_\gamma}$ is not state-of-the-art. We will improve it in Section 2.4.

Remark 4 *In order to establish the proof, we use very small stepsizes with huge constants in the theorems, which is impractical. We suggest to tune the best stepsize in practice, rather than the ones used in the theorems.*

Remark 5 *We measure the convergence rates at the averaged solution, which can be obtained by an additional consensus average routine* $\mathbf{u}^{t+1} = W^t\mathbf{u}^t$ *initialized at* $\mathbf{u}^0 = \mathbf{x}^{T\gamma+1}$*, and* $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\log\frac{1}{\epsilon})$ *rounds of communications are enough. So the total complexities are* $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^2\sqrt{\frac{LC}{\epsilon}}) + \mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\log\frac{1}{\epsilon})$ *and* $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}) + \mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\log\frac{1}{\epsilon})$ *for nonstrongly convex and strongly convex problems, respectively, and they are dominated by the first parts.*

Remark 6 *For nonstrongly convex problems, we can also prove the convergence rate measured at the point* $x_{(i)}^{T\gamma+1}$ *for any* i*:*

$$F(x_{(i)}^{T\gamma+1}) - F(x^*) \leq \frac{2C}{\alpha(T\gamma+1)^2}\max\left\{\frac{\sqrt{m}(1-\sigma_\gamma)}{L\alpha\gamma}, 8m\right\}.$$

However, the complexities increase to $\mathcal{O}(\max\{\sqrt{m}, \sqrt[4]{m}(\frac{\gamma}{1-\sigma_\gamma})^{1.5}\}(\frac{\gamma}{1-\sigma_\gamma})^2\sqrt{\frac{L}{\epsilon}})$*. For strongly convex problems, the complexities stay the same no matter measured at* $\overline{x}^{T\gamma+1}$ *or* $x_{(i)}^{T\gamma+1}$ *because the additional terms, such as* $\max\{\sqrt{m}, \sqrt[4]{m}(\frac{\gamma}{1-\sigma_\gamma})^{1.5}\}$ *in the nonstrongly convex case, appear in the constant* C' *in* $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{C'}{\epsilon})$*.*

Remark 7 *In Theorems 2 and 3, we measure the convergence rates at the* $(T\gamma+1)$*th iteration for simplicity. In fact, the same rates hold for any* $K = T\gamma + r$ *with* $1 \leq r \leq \gamma$ *by regarding the* $(r-1)$*th iteration as the virtual initialization, which only influences the constant* C *in Theorems 2 and 3. In addition, since* $\theta_{r-1} < 1$*, the constant* C *in Theorem 2 contains an additional term* $\frac{\alpha(1-\theta_{r-1})}{\theta_{r-1}^2}(F(\overline{x}^{r-1}) - F(x^*))$*.*

Remark 8 *Due to the physical constraints such as the battery dies, the device shuts down, or the WiFi network is unavailable, the agents may not be active all the time. Most literature let the agents wait and use the old iterates when rejoining the network. Alternatively, we can formulate this case by local updates (Stich, 2019; Koloskova et al., 2020) and use our analysis framework to ensure the convergence. Mathematically, letting* $W_{ii}^k = 1$ *and* $W_{ij}^k = 0$ *for all* $j \neq i$ *and* $k = t+1, t+2, ..., t'$*, which means that agent* i *drops out from the communication network during the time* $[t+1, t']$*, algorithm (13a)-(13d) reduces to the following steps for agent* i *at iterations* $k = t+1, t+2, ..., t'$*:*

$$y_{(i)}^k = \theta_k z_{(i)}^k + (1-\theta_k)x_{(i)}^k, \tag{14a}$$

$$s_{(i)}^k = s_{(i)}^{k-1} + \nabla f_{(i)}(y_{(i)}^k) - \nabla f_{(i)}(y_{(i)}^{k-1}), \tag{14b}$$

$$z_{(i)}^{k+1} = \frac{1}{1+\frac{\mu\alpha}{\theta_k}}\left(\left(\frac{\mu\alpha}{\theta_k}y_{(i)}^k + z_{(i)}^k\right) - \frac{\alpha}{\theta_k}s_{(i)}^k\right), \tag{14c}$$

$$x_{(i)}^{k+1} = \theta_k z_{(i)}^{k+1} + (1-\theta_k)x_{(i)}^k, \tag{14d}$$

which are a serious of local updates without communications. When joining the network again, we require agent i *to make up the delayed computations by performing (14a)-(14d)*

for $t'-t$ iterations. Note that (14a)-(14d) has much lower cost than the same number of iterations (13a)-(13d) because the CPU speed is much faster than the communication speed over TCP/IP or the slow WiFi (Lan et al., 2020).

2.3 Special Cases over Static Graphs

When we fix $W^k = W$, algorithm (13a)-(13d) can be applied to static graphs. As a special case of Theorems 2 and 3, we have the following theorems over static graphs.

Theorem 9 *Suppose that Assumptions 1 and 2 hold with connected graphs and $\mu = 0$. Let the sequence $\{\theta_k\}_{k=0}^K$ satisfy $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ with $\theta_0 = 1$, let $\alpha \le \frac{(1-\sigma)^4}{537L}$. Then for algorithm (13a)-(13d) with fixed gossip matrix W, we have for any $K \ge 1$*

$$F(\overline{x}^{K+1}) - F(x^*) \le \frac{1}{\alpha(K+1)^2}\left(2\|\overline{z}^0 - x^*\|^2 + \frac{\alpha(1-\sigma)}{2mL}\|\Pi \mathbf{s}^0\|^2\right),$$

and

$$\frac{1}{m}\|\Pi \mathbf{x}^K\|^2 \le \frac{1}{\alpha L(K+1)^2}\left(5\|\overline{z}^0 - x^*\|^2 + \frac{9\alpha(1-\sigma)}{4mL}\|\Pi \mathbf{s}^0\|^2\right).$$

Theorem 10 *Suppose that Assumptions 1 and 2 hold with connected graphs and $\mu > 0$. Let $\alpha \le \frac{(1-\sigma)^3}{119L}$ and $\theta_k \equiv \theta = \frac{\sqrt{\mu\alpha}}{2}$. Then for algorithm (13a)-(13d) with fixed gossip matrix W, we have for any $K \ge 1$*

$$F(\overline{x}^{K+1}) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^{K+1} - x^*\|^2 \le (1-\theta)^{K+1}C,$$

and

$$\frac{1}{m}\|\Pi \mathbf{x}^K\|^2 \le (1-\theta)^{K+1}\frac{4C}{L},$$

where $C = F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{4(1-\sigma)}{59mL(1-\theta)}\|\Pi \mathbf{s}^0\|^2$.

The above theorems give the $\mathcal{O}(\frac{1}{(1-\sigma)^2}\sqrt{\frac{L}{\epsilon}})$ and $\mathcal{O}(\sqrt{\frac{L}{\mu(1-\sigma)^3}}\log\frac{1}{\epsilon})$ convergence rates for nonstrongly convex and strongly convex problems, respectively. As illustrated in Table 1, our convergence rates significantly improve over the ones of $\mathcal{O}(\frac{1}{\epsilon^{5/7}})$ and $\mathcal{O}((\frac{L}{\mu})^{5/7}\frac{1}{(1-\sigma)^{1.5}}\log\frac{1}{\epsilon})$, respectively, which were originally proved in (Qu and Li, 2020).

Remark 11 *In our Theorem 10, we require each $f_{(i)}(x)$ to be strongly convex. Some literatures study the weaker assumptions where only $F(x)$ is required to be strongly convex and each $f_{(i)}$ can be convex and smooth. Sun et al. (2022) established the $\mathcal{O}((\frac{L}{\mu(1-\sigma)})^2\log\frac{1}{\epsilon})$ complexity for gradient tracking over general undirected graphs. As a comparison, when each $f_{(i)}$ is strongly convex, the state-of-the-art complexity of gradient tracking is $\mathcal{O}((\frac{L}{\mu} + \frac{1}{(1-\sigma)^2})\log\frac{1}{\epsilon})$ (Alghunaim et al., 2021). Currently, it is unclear how to combine our techniques with those in (Sun et al., 2022) and we conjecture that the complexity would be higher than the one given in Theorem 10. On the other hand, for some algorithms relying on multi-consensus (Ye et al., 2023, 2020), the weaker assumptions have no influence on the complexity.*

2.4 Improved Dependence on the Network Connectivity Constants

As shown in Tables 1 and 2, the dependence on the network connectivity constants in our complexities is not optimal. We improve it over static graphs and time-varying graphs in the next two sections, respectively.

2.4.1 Chebyshev Acceleration over Static Graphs

Chebyshev acceleration was first used to accelerate distributed algorithms by Scaman et al. (2017), and it has become a standard technique now. Define the Chebyshev polynomials as $T_0(x) = 1$, $T_1(x) = x$, and $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$ for all $k \geq 1$. For symmetric W, define $A = I - W$ with $2 \geq \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{m-1} > \lambda_m = 0$ being its eigenvalues. We know $\lambda_{m-1} = 1 - \sigma$. Define $\nu = \frac{\lambda_{m-1}}{\lambda_1}$, $c_1 = \frac{1-\sqrt{\nu}}{1+\sqrt{\nu}}$, $c_2 = \frac{1+\nu}{1-\nu}$, $c_3 = \frac{2}{\lambda_1+\lambda_{m-1}}$, and $P_t(x) = 1-\frac{T_t(c_2(1-x))}{T_t(c_2)}$. Then, $P_t(c_3A)$ is a symmetric matrix satisfying $P_t(c_3A)\mathbf{1} = 0$ with its spectrum in $[1 - \frac{2c_1^t}{1+c_1^{2t}}, 1 + \frac{2c_1^t}{1+c_1^{2t}}] \cup 0$ (Auzinger and Melenk, 2017). Let $t = \frac{1}{\sqrt{\nu}}$ so to have $c_1^t \leq \frac{1}{e}$ and $[1-\frac{2c_1^t}{1+c_1^{2t}}, 1+\frac{2c_1^t}{1+c_1^{2t}}] \subseteq [0.35, 1.65]$. Thus, we can replace the fixed gossip matrix W in algorithm (13a)-(13d) by $I - P_t(c_3A)$ because its second largest singular value σ' satisfies $\sigma' \leq 0.65$, which is independent of $1 - \sigma$. From Theorems 9 and 10 with σ replaced by σ', we see that the algorithm needs $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ iterations for nonstrongly convex problems and $\mathcal{O}(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ iterations for strongly convex ones to find an ϵ-optimal solution, which corresponds to the gradient oracle complexity. On the other hand, we can compute $(I - P_t(c_3A))\mathbf{x}$ by the following procedure (Scaman et al., 2017):

Input: $\mathbf{x}$. Initialize: $a_0 = 1$, $a_1 = c_2$, $\mathbf{z}^0 = \mathbf{x}$, $\mathbf{z}^1 = c_2(I - c_3A)\mathbf{x}$.
for $s = 1, 2, ..., t-1$ **do**
  $a_{s+1} = 2c_2a_s - a_{s-1}$,
  $\mathbf{z}^{s+1} = 2c_2(I - c_3A)\mathbf{z}^s - \mathbf{z}^{s-1}$.
end for
Output: $(I - P_t(c_3A))\mathbf{x} = \frac{\mathbf{z}^t}{a_t}$.

Thus, the communication round complexities for nonstrongly convex and strongly convex problems are $\mathcal{O}(\sqrt{\frac{L}{\epsilon(1-\sigma)}})$ and $\mathcal{O}(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon})$, respectively.

Corollary 12 *Under the settings of Theorem 9 with symmetric and fixed gossip matrix W, algorithm (13a)-(13d) with Chebyshev acceleration requires time of $\mathcal{O}(\sqrt{\frac{L}{\epsilon(1-\sigma)}})$ communication rounds and $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ gradient oracle calls to find an ϵ-optimal averaged solution such that $F(\overline{x}) - F(x^*) \leq \epsilon$.*

Corollary 13 *Under the settings of Theorem 10 with symmetric and fixed gossip matrix W, algorithm (13a)-(13d) with Chebyshev acceleration requires time of $\mathcal{O}(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon})$ communication rounds and $\mathcal{O}(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ gradient oracle calls to find an ϵ-optimal averaged solution such that $F(\overline{x}) - F(x^*) \leq \epsilon$.*

2.4.2 Multiple Consensus over Time-varying Graphs

Although Chebyshev acceleration has been widely used in decentralized optimization, it is unclear how to extend it to time-varying graphs. In this section, we use a multiple consensus subroutine as an alternative to improve the dependence on the network connectivity constants. Motivated by Chebyshev acceleration, our idea is to replace W^k in (13a)-(13d) by virtual gossip matrices $W^{k,\zeta}$ with carefully designed ζ such that

$$\|\Pi W^{k,\zeta}\mathbf{x}\| \leq \frac{1}{e}\|\Pi\mathbf{x}\|, \quad r = 1, 2, 3.$$

Here, $\frac{1}{e}$ can be replaced by any constant not close to 1. Then, it can be regarded as running the resultant algorithm over time-varying graphs with each graph instance being connected at every time, and moreover, $\sigma = \frac{1}{e}$. Note that we do not require the symmetry of the gossip matrices in Assumptions 2 and 3, thus our theorems apply to the virtual gossip matrices $W^{k,\zeta}$. From Theorems 2 and 3 with $\gamma = 1$ and $\sigma_\gamma = \frac{1}{e}$, we see that $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ iterations for nonstrongly convex problems and $\mathcal{O}(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ for strongly convex problems suffice to find an ϵ-optimal solution, which correspond to the gradient oracle complexity. Next, we consider the communication round complexity. Letting $\zeta = \lceil\frac{\gamma}{1-\sigma_\gamma}\rceil$, it follows from (9) that

$$\|\Pi W^{k,\zeta}\mathbf{x}\| \leq \sigma_\gamma^{\frac{1}{1-\sigma_\gamma}}\|\Pi\mathbf{x}\| = (1-(1-\sigma_\gamma))^{\frac{1}{1-\sigma_\gamma}}\|\Pi\mathbf{x}\| \leq \frac{1}{e}\|\Pi\mathbf{x}\|,$$

where we use the fact that $(1-x)^{1/x} \leq 1/e$ for any $x \in (0,1)$. Since $W^{k,\zeta}\mathbf{x}$ can be implemented by the multiple consensus subroutine

$$\mathbf{u}^{t+1} = W^t\mathbf{u}^t$$

with ζ rounds of communications initialized at $\mathbf{u}^0 = \mathbf{x}$, the communication round complexity is $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\epsilon}})$ for nonstrongly convex problems and $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ for strongly convex ones, respectively.

Corollary 14 *Under the settings of Theorem 2, algorithm (13a)-(13d) combined with the multiple consensus subroutine requires time of $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\epsilon}})$ communication rounds and $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ gradient oracle calls to find an ϵ-optimal averaged solution such that $F(\overline{x}) - F(x^*) \leq \epsilon$.*

Corollary 15 *Under the settings of Theorem 3, algorithm (13a)-(13d) combined with the multiple consensus subroutine requires time of $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ communication rounds and $\mathcal{O}(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ gradient oracle calls to find an ϵ-optimal averaged solution such that $F(\overline{x}) - F(x^*) \leq \epsilon$.*

Remark 16 *The multiple consensus subroutine is only for the theoretical purpose. It may be impractical in a realistic time-varying network because communication has been recognized as the major bottleneck in distributed optimization. The multiple consensus may place a*

larger communication burden in practice, although it gives theoretically lower communication round complexities. A similar issue also happens in APM (Li et al., 2020a) and DAGD-C (Rogozin et al., 2021b), which also need a multiple consensus subroutine.

On the other hand, decentralized optimization over time-varying graphs is important because of two reasons. Firstly, in many applications, the communication network varies with time, and algorithms for this scenario are needed. Secondly, many other scenarios can be reformulated as optimization over time-varying graphs, such as asynchrony (Spiridonoff et al., 2020), local SGD (Koloskova et al., 2020), and sparsification (Chen et al., 2022). In these scenarios, the real network may be fixed, and the time-varying graphs are only used for analysis. So the single loop methods are much more favored.

Remark 17 *Unlike the scenario over static graphs, the communication round complexity lower bounds over γ-connected time-varying graphs have not been established, and it is unclear whether the $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\epsilon}})$ and $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ communication round complexities can be further improved. We leave it as an open problem. On the other hand, Kovalev et al. (2021a) established the $\mathcal{O}(\frac{1}{1-\sigma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ communication round complexity lower bound and the $\mathcal{O}(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ gradient oracle complexity lower bound for the special scenario of connected graphs at every time. That is, $\gamma = 1$ in our scenario.*

3. Proofs of Theorems

In this section, we prove the theorems in Sections 2.2 and 2.3. We first reformulate algorithm (13a)-(13d) as the inexact accelerated gradient descent and give its convergence rates in Section 3.1, and then bound the consensus errors. To help the readers get a quick start on our proof framework, we first bound the consensus errors over static graphs in Sections 3.2, and then extend it to the time-varying graphs in Section 3.3. The former scenario provides some basis and insights for the complex proofs of the latter.

3.1 Convergence Rates of the Inexact Accelerated Gradient Descent

Following the proof framework in (Jakovetić et al., 2014a; Qu and Li, 2020), we multiply both sides of (13a)-(13d) by $\frac{1}{m}\mathbf{1}^\top$ and use the definitions in (3) and (2) to yield

$$\overline{y}^k = \theta_k\overline{z}^k + (1-\theta_k)\overline{x}^k, \tag{15a}$$

$$\overline{s}^k - \frac{1}{m}\sum_{i=1}^m \nabla f_{(i)}(y_{(i)}^k) = \overline{s}^{k-1} - \frac{1}{m}\sum_{i=1}^m \nabla f_{(i)}(y_{(i)}^{k-1}), \tag{15b}$$

$$\overline{z}^{k+1} = \frac{1}{1+\frac{\mu\alpha}{\theta_k}}\left(\frac{\mu\alpha}{\theta_k}\overline{y}^k + \overline{z}^k - \frac{\alpha}{\theta_k}\overline{s}^k\right), \tag{15c}$$

$$\overline{x}^{k+1} = \theta_k\overline{z}^{k+1} + (1-\theta_k)\overline{x}^k, \tag{15d}$$

where we use the column stochasticity of $\mathbf{1}^\top W^k = \mathbf{1}^\top$. From the initialization $\mathbf{s}^0 = \nabla f(\mathbf{y}^0)$ and (15b), we have the following standard but important property in gradient tracking:

$$\overline{s}^k = \frac{1}{m}\sum_{i=1}^m \nabla f_{(i)}(y_{(i)}^k). \tag{16}$$

Iterations (15a)-(15d) can be regarded as the inexact accelerated gradient descent (Devolder et al., 2014) in the sense that we use $\frac{1}{m}\sum_{i=1}^m \nabla f_{(i)}(y_{(i)}^k)$ as the descent direction, rather than the true gradient $\frac{1}{m}\sum_{i=1}^m \nabla f_{(i)}(\overline{y}^k)$. In fact, when we replace $\overline{s}^k$ in step (15c) by the true gradient, steps (15a), (15c), and (15d) reduce to the updates of the standard accelerated gradient descent, see (Nesterov, 2004; Lin et al., 2020) for example.

The next lemma demonstrates the analogy properties of convexity and smoothness with the inexact gradients. The proof can be found in (Jakovetić et al., 2014a; Qu and Li, 2020). For the completeness and the readers' convenience, we give the proof in the appendix.

Lemma 18 *Define*

$$f(\overline{y}^k, \mathbf{y}^k) = \frac{1}{m}\sum_{i=1}^m \left(f_{(i)}(y_{(i)}^k) + \left\langle \nabla f_{(i)}(y_{(i)}^k), \overline{y}^k - y_{(i)}^k \right\rangle \right). \tag{17}$$

Suppose that Assumption 1 holds. Then, we have for any w,

$$F(w) \geq f(\overline{y}^k, \mathbf{y}^k) + \left\langle \overline{s}^k, w - \overline{y}^k \right\rangle + \frac{\mu}{2}\|w - \overline{y}^k\|^2, \tag{18}$$

$$F(w) \leq f(\overline{y}^k, \mathbf{y}^k) + \left\langle \overline{s}^k, w - \overline{y}^k \right\rangle + \frac{L}{2}\|w - \overline{y}^k\|^2 + \frac{L}{2m}\|\Pi \mathbf{y}^k\|^2. \tag{19}$$

Especially, we allow μ to be zero.

Define the Bregman divergence as follows:

$$D_f(x, \mathbf{y}^k) = \frac{1}{m}\sum_{i=1}^m \left(f_{(i)}(x) - f_{(i)}(y_{(i)}^k) - \left\langle \nabla f_{(i)}(y_{(i)}^k), x - y_{(i)}^k \right\rangle \right). \tag{20}$$

The next lemma gives the convergence rates of the inexact accelerated gradient descent. The techniques in this proof are standard, see (Lin et al., 2020) for example. The crucial difference is that we keep the Bregman divergence term $D_f(\overline{x}^k, \mathbf{y}^k)$ in (21) and (22), which is motivated by (Tseng, 2008).

Compared with the standard accelerated gradient descent, for example, see (Nesterov, 2004; Lin et al., 2020), there are two additional error terms (a) and (c) in our lemma due to the inexact gradients. In the next two sections, we bound the two terms carefully by (b) and (d), respectively, such that the convergence rates of the accelerated gradient tracking match those of the classical centralized accelerated gradient descent, which is the main technical contribution of this paper compared with the existing work on accelerated gradient tracking in (Qu and Li, 2020).

Lemma 19 *Suppose that Assumption 1 with $\mu = 0$ and part 2 of Assumption 3 hold. Let the sequence $\{\theta_k\}_{k=0}^K$ satisfy $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ with $\theta_0 = 1$. Then for algorithm (13a)-(13d), we have*

$$\begin{aligned}\frac{F(\overline{x}^{K+1}) - F(x^*)}{\theta_K^2} + \frac{1}{2\alpha}\|\overline{z}^{K+1} - x^*\|^2 \leq \frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 \\ + \underbrace{\sum_{k=0}^K \frac{L}{2m\theta_k^2}\|\Pi \mathbf{y}^k\|^2}_{\text{term (a)}} - \underbrace{\sum_{k=0}^K \left(\left(\frac{1}{2\alpha} - \frac{L}{2} \right) \|\overline{z}^{k+1} - \overline{z}^k\|^2 + \frac{1}{\theta_{k-1}^2} D_f(\overline{x}^k, \mathbf{y}^k) \right)}_{\text{term (b)}}.\end{aligned} \tag{21}$$

Suppose that Assumption 1 with $\mu > 0$ and part 2 of Assumption 3 hold. Let $\theta_k = \theta = \frac{\sqrt{\alpha\mu}}{2}$ for all k and assume that $\alpha\mu \leq 1$. Then for algorithm (13a)-(13d), we have

$$\begin{aligned}
&\frac{1}{(1-\theta)^{K+1}}\left(F(\overline{x}^{K+1}) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^{K+1} - x^*\|^2\right) \\
&\quad \leq F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \underbrace{\sum_{k=0}^{K}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2}_{\text{term (c)}} \\
&\quad \underbrace{- \sum_{k=0}^{K}\left(\frac{1}{(1-\theta)^{k+1}}\left(\frac{\theta^2}{2\alpha} - \frac{L\theta^2}{2}\right)\|\overline{z}^{k+1} - \overline{z}^k\|^2 + \frac{1}{(1-\theta)^k}D_f(\overline{x}^k, \mathbf{y}^k)\right)}_{\text{term (d)}}.
\end{aligned} \tag{22}$$

Proof From the inexact smoothness (19), we have

$$\begin{aligned}
F(\overline{x}^{k+1}) \leq& f(\overline{y}^k, \mathbf{y}^k) + \left\langle \overline{s}^k, \overline{x}^{k+1} - \overline{y}^k\right\rangle + \frac{L}{2}\|\overline{x}^{k+1} - \overline{y}^k\|^2 + \frac{L}{2m}\|\Pi\mathbf{y}^k\|^2 \\
\overset{a}{=}& f(\overline{y}^k, \mathbf{y}^k) + \theta_k\left\langle \overline{s}^k, \overline{z}^{k+1} - \overline{z}^k\right\rangle + \frac{L\theta_k^2}{2}\|\overline{z}^{k+1} - \overline{z}^k\|^2 + \frac{L}{2m}\|\Pi\mathbf{y}^k\|^2 \\
=& f(\overline{y}^k, \mathbf{y}^k) + \theta_k\left\langle \overline{s}^k, x^* - \overline{z}^k\right\rangle + \theta_k\left\langle \overline{s}^k, \overline{z}^{k+1} - x^*\right\rangle \\
&+ \frac{L\theta_k^2}{2}\|\overline{z}^{k+1} - \overline{z}^k\|^2 + \frac{L}{2m}\|\Pi\mathbf{y}^k\|^2,
\end{aligned} \tag{23}$$

where we use (15a) and (15d) in $\overset{a}{=}$. Next, we bound the two inner product terms. For the first inner product, we have

$$\begin{aligned}
&f(\overline{y}^k, \mathbf{y}^k) + \theta_k\left\langle \overline{s}^k, x^* - \overline{z}^k\right\rangle \\
&\overset{b}{=} f(\overline{y}^k, \mathbf{y}^k) + \left\langle \overline{s}^k, \theta_k x^* + (1-\theta_k)\overline{x}^k - \overline{y}^k\right\rangle \\
&= \theta_k\left(f(\overline{y}^k, \mathbf{y}^k) + \left\langle \overline{s}^k, x^* - \overline{y}^k\right\rangle\right) + (1-\theta_k)\left(f(\overline{y}^k, \mathbf{y}^k) + \left\langle \overline{s}^k, \overline{x}^k - \overline{y}^k\right\rangle\right) \\
&\overset{c}{\leq} \theta_k F(x^*) - \frac{\mu\theta_k}{2}\|\overline{y}^k - x^*\|^2 + \frac{1-\theta_k}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k) + \left\langle \nabla f_{(i)}(y_{(i)}^k), \overline{x}^k - y_{(i)}^k\right\rangle\right) \\
&= \theta_k F(x^*) - \frac{\mu\theta_k}{2}\|\overline{y}^k - x^*\|^2 + (1-\theta_k)F(\overline{x}^k) \\
&\quad - \frac{1-\theta_k}{m}\sum_{i=1}^{m}\left(f_{(i)}(\overline{x}^k) - f_{(i)}(y_{(i)}^k) - \left\langle \nabla f_{(i)}(y_{(i)}^k), \overline{x}^k - y_{(i)}^k\right\rangle\right) \\
&= \theta_k F(x^*) - \frac{\mu\theta_k}{2}\|\overline{y}^k - x^*\|^2 + (1-\theta_k)F(\overline{x}^k) - (1-\theta_k)D_f(\overline{x}^k, \mathbf{y}^k),
\end{aligned}$$

where we use (15a) in $\overset{b}{=}$, (18), (17), and (16) in $\overset{c}{\leq}$. For the second inner product, we have

$$\begin{aligned}\theta_k \left\langle \overline{s}^k, \overline{z}^{k+1} - x^* \right\rangle \overset{d}{=} & -\frac{\theta_k^2}{\alpha} \left\langle \overline{z}^{k+1} - \overline{z}^k + \frac{\mu\alpha}{\theta_k}(\overline{z}^{k+1} - \overline{y}^k), \overline{z}^{k+1} - x^* \right\rangle \\ = & \frac{\theta_k^2}{2\alpha}\left(\|\overline{z}^k - x^*\|^2 - \|\overline{z}^{k+1} - x^*\|^2 - \|\overline{z}^{k+1} - \overline{z}^k\|^2\right) \\ & + \frac{\mu\theta_k}{2}\left(\|\overline{y}^k - x^*\|^2 - \|\overline{z}^{k+1} - x^*\|^2 - \|\overline{z}^{k+1} - \overline{y}^k\|^2\right),\end{aligned}$$

where we use (15c) in $\overset{d}{=}$. Plugging into (23) and rearranging the terms, it gives

$$\begin{aligned}& F(\overline{x}^{k+1}) - F(x^*) + \left(\frac{\theta_k^2}{2\alpha} + \frac{\mu\theta_k}{2}\right)\|\overline{z}^{k+1} - x^*\|^2 \\ & \leq (1-\theta_k)(F(\overline{x}^k) - F(x^*)) + \frac{\theta_k^2}{2\alpha}\|\overline{z}^k - x^*\|^2 \\ & \quad - \left(\frac{\theta_k^2}{2\alpha} - \frac{L\theta_k^2}{2}\right)\|\overline{z}^{k+1} - \overline{z}^k\|^2 - (1-\theta_k)D_f(\overline{x}^k, \mathbf{y}^k) + \frac{L}{2m}\|\Pi\mathbf{y}^k\|^2.\end{aligned} \tag{24}$$

Case 1: Each $f_{(i)}$ is nonstrongly convex. In this case, (24) holds with $\mu = 0$. Dividing both sides of (24) by θ_k^2, summing over $k = 0, 1, ..., K$, using $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ and $\theta_0 = 1$, we have (21).

Case 2: Each $f_{(i)}$ is μ-strongly convex. Letting $\theta_k = \theta = \frac{\sqrt{\alpha\mu}}{2}$ for all k, we know $\frac{\theta^2}{2\alpha} \leq \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)(1-\theta)$ holds if $\alpha\mu \leq 1$. Dividing both sides of (24) by $(1-\theta)^{k+1}$ and summing over $k = 0, 1, ..., K$, it gives (22). ■

3.2 Bounding the Consensus Errors over Static Graphs

In this section, we bound the term (a) by (b) appeared in (21), and the term (c) by (d) in (22) over static graphs. We first bound $\|\Pi\mathbf{y}^k\|^2$ in the next lemma. The crucial trick is to construct a linear combination of the consensus errors with carefully designed weights such that it shrinks geometrically with an additional error term. Moreover, the step size α remains to be a constant of the order $\mathcal{O}(\frac{1}{L})$ as large as possible. Another trick is that we use a constant τ to balance $D_f(\overline{x}^{r+1}, \mathbf{y}^{r+1})$ and $\|\overline{z}^{r+1} - \overline{z}^r\|^2$ in Φ^r, which is generated by Young's inequality and will be specified later.

Lemma 20 *Suppose that Assumptions 1 and 2 hold with $\mu \geq 0$. Let $\alpha \leq \frac{(1-\sigma)^3}{80L\sqrt{1+\frac{1}{\tau}}}$ and the sequence $\{\theta_k\}_{k=0}^K$ satisfy $\theta_{k+1} \leq \theta_k \leq 1$. Then for algorithm (13a)-(13d) with fixed gossip matrix W, we have*

$$\max\left\{\|\Pi\mathbf{y}^{k+1}\|^2, \|\Pi\mathbf{x}^{k+1}\|^2\right\} \leq C_1\rho^{k+1} + C_2\sum_{r=0}^{k}\rho^{k-r}\theta_r^2\Phi^r, \tag{25}$$

where $\rho = 1 - \frac{1-\sigma}{4}$, $C_1 = \frac{(1-\sigma)^2}{18(1+\frac{1}{\tau})L^2}\|\Pi\mathbf{s}^0\|^2$, $C_2 = \frac{1-\sigma}{9(1+\frac{1}{\tau})L^2}$,

$$\Phi^r = \frac{2mL(1+\tau)}{\theta_r^2} D_f(\overline{x}^{r+1}, \mathbf{y}^{r+1}) + 2mL^2\left(1+\frac{1}{\tau}\right)\|\overline{z}^{r+1} - \overline{z}^r\|^2, \tag{26}$$

and τ *can be any positive constant.*

Proof Multiplying both sides of (13a)-(13d) by Π, using (6) and $\|\Pi\mathbf{x}\| \le \|\mathbf{x}\|$, we have

$$\|\Pi\mathbf{y}^k\| \le \theta_k\|\Pi\mathbf{z}^k\| + (1-\theta_k)\|\Pi\mathbf{x}^k\| \le \theta_k\|\Pi\mathbf{z}^k\| + \|\Pi\mathbf{x}^k\|, \tag{27}$$

$$\|\Pi\mathbf{s}^{k+1}\| \le \sigma\|\Pi\mathbf{s}^k\| + \|\nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k)\|, \tag{28}$$

$$\begin{aligned}\|\Pi\mathbf{z}^{k+1}\| \le& \sigma\left(\frac{\mu\alpha}{\theta_k+\mu\alpha}\|\Pi\mathbf{y}^k\| + \frac{\theta_k}{\theta_k+\mu\alpha}\|\Pi\mathbf{z}^k\|\right) + \frac{\alpha}{\theta_k+\mu\alpha}\|\Pi\mathbf{s}^k\| \\ \overset{a}{\le}& \frac{\sigma(\mu\alpha+1)\theta_k}{\theta_k+\mu\alpha}\|\Pi\mathbf{z}^k\| + \frac{\sigma\mu\alpha}{\theta_k+\mu\alpha}\|\Pi\mathbf{x}^k\| + \frac{\alpha}{\theta_k+\mu\alpha}\|\Pi\mathbf{s}^k\| \\ \overset{b}{\le}& \sigma\|\Pi\mathbf{z}^k\| + \frac{\mu\alpha}{\theta_k}\|\Pi\mathbf{x}^k\| + \frac{\alpha}{\theta_k}\|\Pi\mathbf{s}^k\|,\end{aligned} \tag{29}$$

$$\|\Pi\mathbf{x}^{k+1}\| \le \theta_k\|\Pi\mathbf{z}^{k+1}\| + \sigma\|\Pi\mathbf{x}^k\|, \tag{30}$$

where $\overset{a}{\le}$ uses (27), $\overset{b}{\le}$ uses $\sigma < 1$ and $\frac{(\mu\alpha+1)\theta_k}{\theta_k+\mu\alpha} \le 1$ with $\theta_k \le 1$. Next, we bound $\|\nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k)\|$.

$$\begin{aligned}&\|\nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k)\|^2 = \sum_{i=1}^m \|\nabla f_{(i)}(y_{(i)}^{k+1}) - \nabla f_{(i)}(y_{(i)}^k)\|^2 \\ &\overset{c}{\le} \sum_{i=1}^m (1+\tau)\|\nabla f_{(i)}(y_{(i)}^{k+1}) - \nabla f_{(i)}(\overline{x}^{k+1})\|^2 \\ &\quad + \sum_{i=1}^m \left(1+\frac{1}{\tau}\right) 2\left(\|\nabla f_{(i)}(\overline{x}^{k+1}) - \nabla f_{(i)}(\overline{y}^k)\|^2 + \|\nabla f_{(i)}(\overline{y}^k) - \nabla f_{(i)}(y_{(i)}^k)\|^2\right) \\ &\overset{d}{\le} 2mL(1+\tau)D_f(\overline{x}^{k+1}, \mathbf{y}^{k+1}) + 2L^2\left(1+\frac{1}{\tau}\right)\sum_{i=1}^m\left(\|\overline{x}^{k+1} - \overline{y}^k\|^2 + \|\overline{y}^k - y_{(i)}^k\|^2\right) \\ &\overset{e}{=} 2mL(1+\tau)D_f(\overline{x}^{k+1}, \mathbf{y}^{k+1}) + 2L^2\left(1+\frac{1}{\tau}\right)\left(m\theta_k^2\|\overline{z}^{k+1} - \overline{z}^k\|^2 + \|\Pi\mathbf{y}^k\|^2\right) \\ &= \theta_k^2\Phi^k + 2L^2\left(1+\frac{1}{\tau}\right)\|\Pi\mathbf{y}^k\|^2 \\ &\overset{f}{\le} \theta_k^2\Phi^k + 4L^2\left(1+\frac{1}{\tau}\right)\left(\theta_k^2\|\Pi\mathbf{z}^k\|^2 + \|\Pi\mathbf{x}^k\|^2\right),\end{aligned} \tag{31}$$

where $\overset{c}{\le}$ uses Young's inequality of $\|a-b\|^2 \le (1+\tau)\|a\|^2 + (1+\frac{1}{\tau})\|b\|^2$ for any $\tau > 0$, $\overset{d}{\le}$ uses (5), the smoothness of $f_{(i)}$, and the definition of D_f in (20), $\overset{e}{=}$ uses (15a), (15d), and the definition of $\Pi\mathbf{y}$ in (4), $\overset{f}{\le}$ uses (27). Denote $c_0 = 4L^2\left(1+\frac{1}{\tau}\right)$ for simplicity in the remaining proof of this lemma.

Squaring both sides of (28), it follows that

$$\begin{aligned}\|\Pi\mathbf{s}^{k+1}\|^2 \leq& \left(1+\frac{1-\sigma}{2\sigma}\right)\sigma^2\|\Pi\mathbf{s}^k\|^2+\left(1+\frac{2\sigma}{1-\sigma}\right)\|\nabla f(\mathbf{y}^{k+1})-\nabla f(\mathbf{y}^k)\|^2\\ =&\frac{\sigma+\sigma^2}{2}\|\Pi\mathbf{s}^k\|^2+\frac{1+\sigma}{1-\sigma}\|\nabla f(\mathbf{y}^{k+1})-\nabla f(\mathbf{y}^k)\|^2 \qquad (32)\\ \overset{g}{\leq}&\frac{1+\sigma}{2}\|\Pi\mathbf{s}^k\|^2+\frac{2}{1-\sigma}\left(\theta_k^2\Phi^k+c_0\theta_k^2\|\Pi\mathbf{z}^k\|^2+c_0\|\Pi\mathbf{x}^k\|^2\right),\end{aligned}$$

where we use $\sigma<1$ and (31) in $\overset{g}{\leq}$. Similarly, for (29) and (30), we also have

$$\|\Pi\mathbf{z}^{k+1}\|^2 \leq \frac{1+\sigma}{2}\|\Pi\mathbf{z}^k\|^2+\frac{4}{1-\sigma}\left(\frac{\mu^2\alpha^2}{\theta_k^2}\|\Pi\mathbf{x}^k\|^2+\frac{\alpha^2}{\theta_k^2}\|\Pi\mathbf{s}^k\|^2\right), \qquad (33)$$

$$\|\Pi\mathbf{x}^{k+1}\|^2 \leq \frac{1+\sigma}{2}\|\Pi\mathbf{x}^k\|^2+\frac{2\theta_k^2}{1-\sigma}\|\Pi\mathbf{z}^{k+1}\|^2. \qquad (34)$$

Adding (32), (33), and (34) together with the weights c_1, $c_2\theta_{k+1}^2$, and c_3, respectively, we have

$$\begin{aligned}&c_1\|\Pi\mathbf{s}^{k+1}\|^2+c_2\theta_{k+1}^2\|\Pi\mathbf{z}^{k+1}\|^2+c_3\|\Pi\mathbf{x}^{k+1}\|^2\\ &\overset{h}{\leq} c_1\|\Pi\mathbf{s}^{k+1}\|^2+\left(c_2\theta_k^2+\frac{2c_3\theta_k^2}{1-\sigma}\right)\|\Pi\mathbf{z}^{k+1}\|^2+\frac{c_3(1+\sigma)}{2}\|\Pi\mathbf{x}^k\|^2\\ &\leq \left(\frac{c_1(1+\sigma)}{2}+\left(c_2+\frac{2c_3}{1-\sigma}\right)\frac{4\alpha^2}{1-\sigma}\right)\|\Pi\mathbf{s}^k\|^2\\ &\quad+\left(\frac{2c_0c_1}{1-\sigma}+\left(c_2+\frac{2c_3}{1-\sigma}\right)\frac{1+\sigma}{2}\right)\theta_k^2\|\Pi\mathbf{z}^k\|^2\\ &\quad+\left(\frac{2c_0c_1}{1-\sigma}+\left(c_2+\frac{2c_3}{1-\sigma}\right)\frac{4\mu^2\alpha^2}{1-\sigma}+\frac{c_3(1+\sigma)}{2}\right)\|\Pi\mathbf{x}^k\|^2+\frac{2c_1}{1-\sigma}\theta_k^2\Phi^k,\end{aligned}$$

where we use $\theta_{k+1}\leq\theta_k$ and (34) in $\overset{h}{\leq}$. Letting $c_3=\frac{9c_0c_1}{(1-\sigma)^2}$, $c_2=\frac{80c_0c_1}{(1-\sigma)^4}\geq\frac{8c_0c_1}{(1-\sigma)^2}+\frac{8c_3}{(1-\sigma)^2}$, and $\alpha^2\leq\min\left\{\frac{(1-\sigma)^6}{1600c_0},\frac{(1-\sigma)^4}{1600\mu^2}\right\}$ such that

$$\begin{aligned}&\frac{c_1(1+\sigma)}{2}+\left(c_2+\frac{2c_3}{1-\sigma}\right)\frac{4\alpha^2}{1-\sigma}\leq\frac{c_1(1+\sigma)}{2}+\frac{400c_0c_1\alpha^2}{(1-\sigma)^5}\leq\frac{c_1(3+\sigma)}{4},\\ &\frac{2c_0c_1}{1-\sigma}+\left(c_2+\frac{2c_3}{1-\sigma}\right)\frac{1+\sigma}{2}\leq\frac{2c_0c_1}{1-\sigma}+\frac{c_2(1+\sigma)}{2}+\frac{2c_3}{1-\sigma}\leq\frac{c_2(3+\sigma)}{4},\\ &\frac{2c_0c_1}{1-\sigma}+\left(c_2+\frac{2c_3}{1-\sigma}\right)\frac{4\mu^2\alpha^2}{1-\sigma}+\frac{c_3(1+\sigma)}{2}\leq\frac{2c_0c_1}{1-\sigma}+\frac{400c_0c_1\mu^2\alpha^2}{(1-\sigma)^5}+\frac{c_3(1+\sigma)}{2}\leq\frac{c_3(3+\sigma)}{4},\end{aligned}$$

we have

$$\begin{aligned}&c_1\|\Pi\mathbf{s}^{k+1}\|^2+c_2\theta_{k+1}^2\|\Pi\mathbf{z}^{k+1}\|^2+c_3\|\Pi\mathbf{x}^{k+1}\|^2\\ &\leq\frac{3+\sigma}{4}\left(c_1\|\Pi\mathbf{s}^k\|^2+c_2\theta_k^2\|\Pi\mathbf{z}^k\|^2+c_3\|\Pi\mathbf{x}^k\|^2\right)+\frac{2c_1}{1-\sigma}\theta_k^2\Phi^k\\ &\leq\left(\frac{3+\sigma}{4}\right)^{k+1}\left(c_1\|\Pi\mathbf{s}^0\|^2+c_2\theta_0^2\|\Pi\mathbf{z}^0\|^2+c_3\|\Pi\mathbf{x}^0\|^2\right)+\frac{2c_1}{1-\sigma}\sum_{r=0}^{k}\left(\frac{3+\sigma}{4}\right)^{k-r}\theta_r^2\Phi^r.\end{aligned}$$

From (27), $c_2 > c_3$, and the initialization such that $\Pi \mathbf{x}^0 = \Pi \mathbf{y}^0 = \Pi \mathbf{z}^0 = 0$, we have

$$\begin{aligned}\|\Pi \mathbf{y}^{k+1}\|^2 \leq & \frac{2}{c_3}\left(c_1\|\Pi \mathbf{s}^{k+1}\|^2 + c_2\theta_{k+1}^2\|\Pi \mathbf{z}^{k+1}\|^2 + c_3\|\Pi \mathbf{x}^{k+1}\|^2\right) \\ \leq & \left(\frac{3+\sigma}{4}\right)^{k+1}\frac{2c_1}{c_3}\|\Pi \mathbf{s}^0\|^2 + \frac{4c_1}{c_3(1-\sigma)}\sum_{r=0}^{k}\left(\frac{3+\sigma}{4}\right)^{k-r}\theta_r^2\Phi^r \\ = & \left(\frac{3+\sigma}{4}\right)^{k+1}\frac{(1-\sigma)^2}{18(1+\frac{1}{\tau})L^2}\|\Pi \mathbf{s}^0\|^2 + \frac{1-\sigma}{9(1+\frac{1}{\tau})L^2}\sum_{r=0}^{k}\left(\frac{3+\sigma}{4}\right)^{k-r}\theta_r^2\Phi^r,\end{aligned}$$

which is exactly (25). ■

Having (25) at hand, we are ready to bound the term (a) by (b) appeared in (21). The remaining challenge is to upper bound the weighted cumulative consensus errors.

Lemma 21 *Suppose that Assumptions 1 and 2 hold with $\mu = 0$. Let the sequence $\{\theta_k\}_{k=0}^K$ satisfy $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ with $\theta_0 = 1$, let $\alpha \leq \frac{(1-\sigma)^3}{80L\sqrt{1+\frac{1}{\tau}}}$. Then for algorithm (13a)-(13d) with fixed gossip matrix W, we have*

$$\begin{aligned}&\max\left\{\sum_{k=0}^{K}\frac{L}{2m\theta_k^2}\|\Pi \mathbf{y}^k\|^2, \sum_{k=0}^{K}\frac{L}{2m\theta_k^2}\|\Pi \mathbf{x}^k\|^2\right\} \\ &\leq \frac{16}{3mL(1+\frac{1}{\tau})(1-\sigma)}\|\Pi \mathbf{s}^0\|^2 + \frac{11}{mL(1+\frac{1}{\tau})(1-\sigma)^2}\sum_{r=0}^{K-1}\Phi^r,\end{aligned} \tag{35}$$

where τ and Φ^r are defined in Lemma 20.

Proof We first give some properties of the sequence $\{\theta_k\}_{k=0}^K$. From $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ and $\theta_0 = 1$, we have $\theta_k \leq \theta_{k-1}$, $\frac{1}{\theta_k} - 1 \leq \frac{1}{\theta_{k-1}}$, and $\frac{1}{\theta_k} - \frac{1}{2} \geq \frac{1}{\theta_{k-1}}$, which further give

$$\frac{1}{\theta_k} - \frac{1}{\theta_{k-1}} \leq 1, \qquad \frac{1}{k+1} \leq \theta_k \leq \frac{2}{k+1}. \tag{36}$$

From (25), we get

$$\begin{aligned}\sum_{k=1}^{K}\frac{L}{2m\theta_k^2}\|\Pi \mathbf{y}^k\|^2 \leq & \sum_{k=1}^{K}\frac{C_1L\rho^k}{2m\theta_k^2} + \sum_{k=1}^{K}\frac{C_2L}{2m\theta_k^2}\sum_{r=0}^{k-1}\rho^{k-1-r}\theta_r^2\Phi^r \\ = & \frac{C_1L}{2m}\sum_{k=1}^{K}\frac{\rho^k}{\theta_k^2} + \frac{C_2L}{2m\rho}\sum_{k=1}^{K}\frac{\rho^k}{\theta_k^2}\sum_{r=0}^{k-1}\frac{\theta_r^2}{\rho^r}\Phi^r \\ = & \frac{C_1L}{2m}\sum_{k=1}^{K}\frac{\rho^k}{\theta_k^2} + \frac{C_2L}{2m\rho}\sum_{r=0}^{K-1}\frac{\theta_r^2}{\rho^r}\Phi^r\sum_{k=r+1}^{K}\frac{\rho^k}{\theta_k^2}.\end{aligned} \tag{37}$$

Recall that for scalars, θ_k means the value at iteration k, while ρ^k is its kth power. Next, we compute $\sum_{k=r+1}^{K} \frac{\rho^k}{\theta_k^2}$ for any $r \geq 0$. Denote $S = \sum_{k=r+1}^{K} \frac{\rho^k}{\theta_k^2}$ for simplicity. We have

$$\rho S = \sum_{k=r+1}^{K} \frac{\rho^{k+1}}{\theta_k^2} = \sum_{k=r+1}^{K} \frac{\rho^k}{\theta_{k-1}^2} - \frac{\rho^{r+1}}{\theta_r^2} + \frac{\rho^{K+1}}{\theta_K^2},$$

and

$$S - \rho S = \sum_{k=r+1}^{K} \rho^k \left(\frac{1}{\theta_k^2} - \frac{1}{\theta_{k-1}^2} \right) + \frac{\rho^{r+1}}{\theta_r^2} - \frac{\rho^{K+1}}{\theta_K^2} \overset{a}{=} \sum_{k=r+1}^{K} \frac{\rho^k}{\theta_k} + \frac{\rho^{r+1}}{\theta_r^2} - \frac{\rho^{K+1}}{\theta_K^2},$$

where we use $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ in $\overset{a}{=}$. It further gives

$$\rho(1-\rho)S = \sum_{k=r+1}^{K} \frac{\rho^{k+1}}{\theta_k} + \frac{\rho^{r+2}}{\theta_r^2} - \frac{\rho^{K+2}}{\theta_K^2} = \sum_{k=r+1}^{K} \frac{\rho^k}{\theta_{k-1}} - \frac{\rho^{r+1}}{\theta_r} + \frac{\rho^{K+1}}{\theta_K} + \frac{\rho^{r+2}}{\theta_r^2} - \frac{\rho^{K+2}}{\theta_K^2},$$

and

$$\begin{aligned}
(1-\rho)^2 S =& (1-\rho)S - \rho(1-\rho)S \\
=& \sum_{k=r+1}^{K} \rho^k \left(\frac{1}{\theta_k} - \frac{1}{\theta_{k-1}} \right) + \frac{\rho^{r+1}}{\theta_r^2} - \frac{\rho^{K+1}}{\theta_K^2} + \frac{\rho^{r+1}}{\theta_r} - \frac{\rho^{K+1}}{\theta_K} - \frac{\rho^{r+2}}{\theta_r^2} + \frac{\rho^{K+2}}{\theta_K^2} \\
=& \sum_{k=r+1}^{K} \rho^k \left(\frac{1}{\theta_k} - \frac{1}{\theta_{k-1}} \right) + \frac{(1-\rho)\rho^{r+1}}{\theta_r^2} - \frac{(1-\rho)\rho^{K+1}}{\theta_K^2} + \frac{\rho^{r+1}}{\theta_r} - \frac{\rho^{K+1}}{\theta_K} \\
\overset{b}{\leq}& \sum_{k=r+1}^{K} \rho^k + \frac{(1-\rho)\rho^{r+1}}{\theta_r^2} + \frac{\rho^{r+1}}{\theta_r} \leq \frac{\rho^{r+1}}{1-\rho} + \frac{2\rho^{r+1}}{\theta_r^2},
\end{aligned}$$

where we use (36) in $\overset{b}{\leq}$. Thus, we get

$$\sum_{k=r+1}^{K} \frac{\rho^k}{\theta_k^2} \leq \frac{1}{(1-\rho)^2} \left(\frac{\rho^{r+1}}{1-\rho} + \frac{2\rho^{r+1}}{\theta_r^2} \right) \leq \frac{3\rho^{r+1}}{(1-\rho)^3 \theta_r^2}. \tag{38}$$

Plugging into (37), it follows from $\Pi \mathbf{y}^0 = 0$ and $\theta_0 = 1$ that

$$\begin{aligned}
\sum_{k=0}^{K} \frac{L}{2m\theta_k^2} \|\Pi \mathbf{y}^k\|^2 \leq& \frac{3C_1 L\rho}{2m(1-\rho)^3} + \frac{3C_2 L}{2m(1-\rho)^3} \sum_{r=0}^{K-1} \Phi^r \\
\leq& \frac{16}{3mL(1+\frac{1}{\tau})(1-\sigma)} \|\Pi \mathbf{s}^0\|^2 + \frac{11}{mL(1+\frac{1}{\tau})(1-\sigma)^2} \sum_{r=0}^{K-1} \Phi^r,
\end{aligned}$$

where the last inequality uses the definitions of C_1, C_2, and ρ given in Lemma 20. Replacing $\|\Pi \mathbf{y}^k\|$ by $\|\Pi \mathbf{x}^k\|$ in the above analysis, we have the same bound for $\sum_{k=0}^{K} \frac{L}{2m\theta_k^2} \|\Pi \mathbf{x}^k\|^2$. ■

In the next lemma, we bound the term (c) by (d) appeared in (22) in a similar way to the proof of Lemma 21.

Lemma 22 *Suppose that Assumptions 1 and 2 hold with $\mu > 0$. Let $\alpha \leq \frac{(1-\sigma)^3}{80L\sqrt{1+\frac{1}{\tau}}}$ and $\theta_k \equiv \theta = \frac{\sqrt{\mu\alpha}}{2}$. Then for algorithm (13a)-(13d) with fixed gossip matrix W, we have*

$$\begin{aligned}&\max\left\{\sum_{k=0}^{K}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2,\sum_{k=0}^{K}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{x}^k\|^2\right\}\\&\leq\frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)}\|\Pi\mathbf{s}^0\|^2+\frac{8\theta^2}{27mL(1+\frac{1}{\tau})}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}},\end{aligned}\tag{39}$$

where τ and Φ^r are defined in Lemma 20.

Proof From (25), we get

$$\begin{aligned}&\sum_{k=1}^{K}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2\\&\leq\sum_{k=1}^{K}\frac{C_1L\rho^k}{2m(1-\theta)^{k+1}}+\sum_{k=1}^{K}\frac{C_2L}{2m(1-\theta)^{k+1}}\sum_{r=0}^{k-1}\rho^{k-1-r}\theta^2\Phi^r\\&=\frac{C_1L}{2m(1-\theta)}\sum_{k=1}^{K}\left(\frac{\rho}{1-\theta}\right)^k+\frac{\theta^2C_2L}{2m\rho(1-\theta)}\sum_{k=1}^{K}\left(\frac{\rho}{1-\theta}\right)^k\sum_{r=0}^{k-1}\frac{\Phi^r}{\rho^r}\\&=\frac{C_1L}{2m(1-\theta)}\sum_{k=1}^{K}\left(\frac{\rho}{1-\theta}\right)^k+\frac{\theta^2C_2L}{2m\rho(1-\theta)}\sum_{r=0}^{K-1}\frac{\Phi^r}{\rho^r}\sum_{k=r+1}^{K}\left(\frac{\rho}{1-\theta}\right)^k.\end{aligned}$$

From the settings of θ and α, we know $\theta \leq \frac{1-\sigma}{16}$. Thus we have $\frac{\rho}{1-\theta} < 1$, $1-\theta-\rho \geq \frac{3(1-\sigma)}{16}$, and $\sum_{k=r+1}^{K}\left(\frac{\rho}{1-\theta}\right)^k \leq \left(\frac{\rho}{1-\theta}\right)^r\frac{\rho}{1-\theta-\rho}$ with $\rho = 1-\frac{1-\sigma}{4}$. It follows from $\Pi\mathbf{y}^0 = 0$ that

$$\begin{aligned}\sum_{k=0}^{K}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2\leq&\frac{C_1L}{2m(1-\theta)}\frac{\rho}{1-\theta-\rho}+\frac{\theta^2C_2L}{2m(1-\theta-\rho)}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}}\\\leq&\frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)}\|\Pi\mathbf{s}^0\|^2+\frac{8\theta^2}{27mL(1+\frac{1}{\tau})}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}},\end{aligned}$$

where the last inequality uses the definitions of C_1 and C_2 given in Lemma 20 and $1-\theta-\rho \geq \frac{3(1-\sigma)}{16}$. Replacing $\|\Pi\mathbf{y}^k\|$ by $\|\Pi\mathbf{x}^k\|$ in the above analysis, we have the same bound for $\Pi\mathbf{x}^k$. ■

Now, we are ready to prove Theorems 9 and 10. We first prove Theorem 9. The crucial trick in this proof is to make the constant before $D_f(\overline{x}^k, \mathbf{y}^k)$ positive by setting the constant τ small, and make the constant before $\|\overline{z}^{t+1}-\overline{z}^t\|^2$ positive by setting the step size α small. This is the reason why we introduce the constant τ in the definition of Ψ^r in (26).

Proof Plugging (35) into (21) and using the definition of Φ^r in (26), we obtain

$$\begin{aligned}
&\frac{F(\overline{x}^{K+1})-F(x^*)}{\theta_K^2}+\frac{1}{2\alpha}\|\overline{z}^{K+1}-x^*\|^2\\
&\le \frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{16}{3mL(1+\frac{1}{\tau})(1-\sigma)}\|\Pi\mathbf{s}^0\|^2\\
&\quad-\sum_{k=0}^{K}\left(\left(\frac{1}{2\alpha}-\frac{L}{2}-\frac{22L}{(1-\sigma)^2}\right)\|\overline{z}^{t+1}-\overline{z}^t\|^2+\frac{1}{\theta_{k-1}^2}\left(1-\frac{22(1+\tau)}{(1+\frac{1}{\tau})(1-\sigma)^2}\right)D_f(\overline{x}^k,\mathbf{y}^k)\right)\\
&\overset{a}{\le} \frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{16}{3mL(1+\frac{1}{\tau})(1-\sigma)}\|\Pi\mathbf{s}^0\|^2-\sum_{k=0}^{K}\left(\frac{1}{4\alpha}\|\overline{z}^{t+1}-\overline{z}^t\|^2+\frac{1}{2\theta_{k-1}^2}D_f(\overline{x}^k,\mathbf{y}^k)\right)\\
&\le \frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{1-\sigma}{8mL}\|\Pi\mathbf{s}^0\|^2-\frac{1}{5mL}\sum_{r=0}^{K-1}\Phi^r,
\end{aligned}$$

where in $\overset{a}{\le}$ we let $\tau=\frac{(1-\sigma)^2}{44}$ so to have $\frac{22(1+\tau)}{(1+\frac{1}{\tau})(1-\sigma)^2}=\frac{1}{2}$, $\alpha\le\frac{(1-\sigma)^4}{537L}\le\frac{(1-\sigma)^3}{80L\sqrt{1+\frac{1}{\tau}}}$, and $\frac{1}{4\alpha}\ge\frac{L}{2}+\frac{22L}{(1-\sigma)^2}$. So we have

$$\begin{aligned}
&F(\overline{x}^{K+1})-F(x^*)\le\theta_K^2\left(\frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{1-\sigma}{8mL}\|\Pi\mathbf{s}^0\|^2\right),\\
&\frac{1}{5mL}\sum_{r=0}^{K-1}\Phi^r\le\frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{1-\sigma}{8mL}\|\Pi\mathbf{s}^0\|^2.
\end{aligned}$$

It follows from (35) that

$$\begin{aligned}
&\max\left\{\sum_{k=0}^{K}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{y}^k\|^2,\sum_{k=0}^{K}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{x}^k\|^2\right\}\\
&\le\frac{16}{3mL(1+\frac{1}{\tau})(1-\sigma)}\|\Pi\mathbf{s}^0\|^2+\frac{11}{mL(1+\frac{1}{\tau})(1-\sigma)^2}\sum_{r=0}^{K-1}\Phi^r\\
&\le\frac{1-\sigma}{8mL}\|\Pi\mathbf{s}^0\|^2+\frac{1}{4mL}\sum_{r=0}^{K-1}\Phi^r\\
&\le\frac{5}{8\alpha}\|\overline{z}^0-x^*\|^2+\frac{9(1-\sigma)}{32mL}\|\Pi\mathbf{s}^0\|^2.
\end{aligned}$$

From (36), we have the conclusions. ■

Next, we prove Theorem 10.

Proof Plugging (39) into (22) and using the definition of Φ^r in (26), we have

$$\begin{aligned}
&\frac{1}{(1-\theta)^{K+1}}\left(F(\overline{x}^{K+1})-F(x^*)+\left(\frac{\theta^2}{2\alpha}+\frac{\mu\theta}{2}\right)\|\overline{z}^{K+1}-x^*\|^2\right)\\
&\le F(\overline{x}^0)-F(x^*)+\left(\frac{\theta^2}{2\alpha}+\frac{\mu\theta}{2}\right)\|\overline{z}^0-x^*\|^2+\frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)}\|\Pi\mathbf{s}^0\|^2\\
&\quad-\sum_{k=0}^{K}\left(\frac{1}{(1-\theta)^k}\left(1-\frac{16(1+\tau)}{27(1+\frac{1}{\tau})}\right)D_f(\overline{x}^k,\mathbf{y}^k)\right.\\
&\qquad\left.+\frac{1}{(1-\theta)^{k+1}}\left(\frac{\theta^2}{2\alpha}-\frac{L\theta^2}{2}-\frac{16L\theta^2}{27}\right)\|\overline{z}^{k+1}-\overline{z}^k\|^2\right)\\
&\overset{a}{\le} F(\overline{x}^0)-F(x^*)+\left(\frac{\theta^2}{2\alpha}+\frac{\mu\theta}{2}\right)\|\overline{z}^0-x^*\|^2+\frac{4(1-\sigma)}{27m(1+\frac{1}{\tau})L(1-\theta)}\|\Pi\mathbf{s}^0\|^2\\
&\quad-\sum_{k=0}^{K}\left(\frac{1}{2(1-\theta)^k}D_f(\overline{x}^k,\mathbf{y}^k)+\frac{\theta^2}{4\alpha(1-\theta)^{k+1}}\|\overline{z}^{k+1}-\overline{z}^k\|^2\right)\\
&\le F(\overline{x}^0)-F(x^*)+\left(\frac{\theta^2}{2\alpha}+\frac{\mu\theta}{2}\right)\|\overline{z}^0-x^*\|^2+\frac{4(1-\sigma)}{59mL(1-\theta)}\|\Pi\mathbf{s}^0\|^2-\frac{8\theta^2}{59mL}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}},
\end{aligned}$$

where in $\overset{a}{\le}$ we let $\tau=\frac{27}{32}$ so to have $\frac{16(1+\tau)}{27(1+\frac{1}{\tau})}=\frac{1}{2}$, $\alpha\le\frac{(1-\sigma)^3}{119L}\le\frac{(1-\sigma)^3}{80L\sqrt{1+\frac{1}{\tau}}}$, and $\frac{1}{4\alpha}\ge\frac{L}{2}+\frac{16L}{27}$. Thus, we have the first conclusion and

$$\frac{8\theta^2}{59mL}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}}\le F(\overline{x}^0)-F(x^*)+\left(\frac{\theta^2}{2\alpha}+\frac{\mu\theta}{2}\right)\|\overline{z}^0-x^*\|^2+\frac{4(1-\sigma)}{59mL(1-\theta)}\|\Pi\mathbf{s}^0\|^2.$$

It follows from (39) that

$$\begin{aligned}
&\sum_{k=0}^{K}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{x}^k\|^2\\
&\quad\le\frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)}\|\Pi\mathbf{s}^0\|^2+\frac{8\theta^2}{27mL(1+\frac{1}{\tau})}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}}\\
&\quad=\frac{4(1-\sigma)}{59mL(1-\theta)}\|\Pi\mathbf{s}^0\|^2+\frac{8\theta^2}{59mL}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}}\\
&\quad\le 2\left(F(\overline{x}^0)-F(x^*)+\left(\frac{\theta^2}{2\alpha}+\frac{\mu\theta}{2}\right)\|\overline{z}^0-x^*\|^2+\frac{4(1-\sigma)}{59mL(1-\theta)}\|\Pi\mathbf{s}^0\|^2\right).
\end{aligned}$$

Thus, we have the second conclusion. ■

We end this section by summarizing the differences from (Qu and Li, 2020) and the reasons of the convergence rates improvement.

Remark 23 *As shown in Lemmas 19 and 20, we keep the Bregman divergence term $D_f(\overline{x}^k, \mathbf{y}^k)$, and use a constant τ to balance this divergence term and $\|\overline{z}^{k+1} - \overline{z}^k\|^2$. As shown in the proofs of Theorems 9 and 10, we make the constant before $D_f(\overline{x}^k, \mathbf{y}^k)$ positive by setting τ small. As a comparison, Qu and Li (2020) did not use this Bregman divergence term, and they bounded the term $\|\overline{x}^k - \overline{y}^k\|^2$, which is generated by the consensus errors and is an analogy to our term $D_f(\overline{x}^k, \mathbf{y}^k)$ generated in (31), by setting much smaller step sizes than ours. See (32) and (53) in (Qu and Li, 2020) for the details. To make the constant A_4 in their (32) positive, Qu and Li (2020) set the step size of the order $\alpha = \mathcal{O}(\frac{1}{L}(\frac{\mu}{L})^{3/7})$. Since $\sqrt{\mu\alpha}$ dominates the convergence rate for strongly convex problems, Qu and Li (2020) only got the slower convergence rate of $\mathcal{O}((1-(\frac{\mu}{L})^{5/7})^k)$. For nonstrongly convex problems, Qu and Li (2020) set the step size of the order $\mathcal{O}(\frac{1}{k^{0.6+\epsilon}})$ to bound the corresponding term in their (53), which gives the slower convergence rate of $\mathcal{O}(\frac{1}{k^{1.4-\epsilon}})$.*

As shown in Lemma 20, to bound the consensus errors, we construct a linear combination of the consensus errors such that it shrinks geometrically with an additional error term. As a comparison, Qu and Li (2020) used the linear system inequality, which needs to upper bound the spectral radius of a system matrix and thus it is quite involved. See the proofs of Lemmas 7 and 13-15 in (Qu and Li, 2020). Our proof is much simpler than those in (Qu and Li, 2020), and it can be extended to the time-varying graphs in a unified framework.

3.3 Bounding the Consensus Errors over Time-varying Graphs

In this section, we consider algorithm (13a)-(13d) over time-varying graphs. Our analysis follows the same proof framework in the previous section for static graphs, but with more involved details. In the next lemma, we first give the analogy counterparts of (32)-(34).

Lemma 24 *Suppose that Assumptions 1 and 3 hold with $\mu \geq 0$. Let the sequence $\{\theta_k\}_{k=0}^K$ satisfy $\frac{\theta_k}{1.62} \leq \theta_{k+1} \leq \theta_k \leq 1$. Then, we have for any $k \geq \gamma - 1$,*

$$\|\Pi\mathbf{s}^{k+1}\|^2 \leq \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{k-\gamma+1}\|^2 + \frac{2\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+1}^{k}\left(\theta_r^2\Phi^r + c_0\theta_r^2\|\Pi\mathbf{z}^r\|^2 + c_0\|\Pi\mathbf{x}^r\|^2\right), \quad (40)$$

$$\|\Pi\mathbf{x}^{k+1}\|^2 \leq \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{x}^{k-\gamma+1}\|^2 + \frac{5.5\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+2}^{k+1}\theta_r^2\|\Pi\mathbf{z}^r\|^2, \quad (41)$$

$$\theta_{k+1}^2\|\Pi\mathbf{z}^{k+1}\|^2 \leq \frac{1+\sigma_\gamma}{2}\theta_{k-\gamma+1}^2\|\Pi\mathbf{z}^{k-\gamma+1}\|^2 + \frac{4\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+1}^{k}\left(\mu^2\alpha^2\|\Pi\mathbf{x}^r\|^2 + \alpha^2\|\Pi\mathbf{s}^r\|^2\right), \quad (42)$$

where we denote $c_0 = 4L^2\left(1+\frac{1}{\tau}\right)$, and τ and Φ^r are defined in Lemma 20.

Proof From (13b) and the definition of $W^{k,\gamma}$ in (7), we have for any $k \geq \gamma - 1$,

$$\begin{aligned}\mathbf{s}^{k+1} =& W^{k+1}\mathbf{s}^k + \nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k)\\ =& \left(\prod_{t=k-\gamma+2}^{k+1} W^t\right)\mathbf{s}^{k-\gamma+1} + \sum_{r=k-\gamma+1}^{k}\left(\prod_{t=r+1}^{k} W^{t+1}\right)\left(\nabla f(\mathbf{y}^{r+1}) - \nabla f(\mathbf{y}^r)\right)\end{aligned}$$

$$=W^{k+1,\gamma}\mathbf{s}^{k-\gamma+1}+\sum_{r=k-\gamma+1}^{k}W^{k+1,k-r}(\nabla f(\mathbf{y}^{r+1})-\nabla f(\mathbf{y}^r)),$$

where we denote $\prod_{t=k+1}^{k}W^{t+1}=I$. Multiplying both sides by Π, using (9) and (10), it gives

$$\|\Pi\mathbf{s}^{k+1}\|\leq\sigma_\gamma\|\Pi\mathbf{s}^{k-\gamma+1}\|+\sum_{r=k-\gamma+1}^{k}\|\nabla f(\mathbf{y}^{r+1})-\nabla f(\mathbf{y}^r)\|. \tag{43}$$

Similar to (32), squaring both sides of (43) yields

$$\begin{aligned}\|\Pi\mathbf{s}^{k+1}\|^2\leq&\frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{k-\gamma+1}\|^2+\frac{2}{1-\sigma_\gamma}\left(\sum_{r=k-\gamma+1}^{k}\|\nabla f(\mathbf{y}^{r+1})-\nabla f(\mathbf{y}^r)\|\right)^2\\ \leq&\frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{k-\gamma+1}\|^2+\frac{2\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+1}^{k}\|\nabla f(\mathbf{y}^{r+1})-\nabla f(\mathbf{y}^r)\|^2.\end{aligned} \tag{44}$$

From (31), we have (40). It follows from (13d) that

$$\begin{aligned}\mathbf{x}^{k+1}=&(1-\theta_k)W^k\mathbf{x}^k+\theta_k\mathbf{z}^{k+1}\\ =&\left(\prod_{t=k-\gamma+1}^{k}(1-\theta_t)W^t\right)\mathbf{x}^{k-\gamma+1}+\sum_{r=k-\gamma+1}^{k}\left(\prod_{t=r+1}^{k}(1-\theta_t)W^t\right)\theta_r\mathbf{z}^{r+1}\\ =&W^{k,\gamma}\mathbf{x}^{k-\gamma+1}\prod_{t=k-\gamma+1}^{k}(1-\theta_t)+\sum_{r=k-\gamma+1}^{k}W^{k,k-r}\theta_r\mathbf{z}^{r+1}\prod_{t=r+1}^{k}(1-\theta_t).\end{aligned}$$

Similar to (43) and (44), we also have

$$\|\Pi\mathbf{x}^{k+1}\|\leq\sigma_\gamma\|\Pi\mathbf{x}^{k-\gamma+1}\|+\sum_{r=k-\gamma+1}^{k}\theta_r\|\Pi\mathbf{z}^{r+1}\|=\sigma_\gamma\|\Pi\mathbf{x}^{k-\gamma+1}\|+\sum_{r=k-\gamma+2}^{k+1}\theta_{r-1}\|\Pi\mathbf{z}^{r}\|,$$

and

$$\|\Pi\mathbf{x}^{k+1}\|^2\leq\frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{x}^{k-\gamma+1}\|^2+\frac{2\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+2}^{k+1}\theta_{r-1}^2\|\Pi\mathbf{z}^{r}\|^2.$$

Using $\theta_{r-1}\leq 1.62\theta_r$, we obtain (41). Similarly, for (13c), we have

$$\begin{aligned}\mathbf{z}^{k+1}=&\frac{\theta_k}{\theta_k+\mu\alpha}W^k\mathbf{z}^k+\frac{\mu\alpha}{\theta_k+\mu\alpha}W^k\mathbf{y}^k-\frac{\alpha}{\theta_k+\mu\alpha}\mathbf{s}^k\\ \overset{a}{=}&\frac{\theta_k(1+\mu\alpha)}{\theta_k+\mu\alpha}W^k\mathbf{z}^k+\frac{\mu\alpha(1-\theta_k)}{\theta_k+\mu\alpha}W^k\mathbf{x}^k-\frac{\alpha}{\theta_k+\mu\alpha}\mathbf{s}^k\\ =&\left(\prod_{t=k-\gamma+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha}W^t\right)\mathbf{z}^{k-\gamma+1}\\ &+\sum_{r=k-\gamma+1}^{k}\left(\prod_{t=r+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha}W^t\right)\left(\frac{\mu\alpha(1-\theta_r)}{\theta_r+\mu\alpha}W^r\mathbf{x}^r-\frac{\alpha}{\theta_r+\mu\alpha}\mathbf{s}^r\right)\end{aligned}$$

$$=W^{k,\gamma}\mathbf{z}^{k-\gamma+1}\prod_{t=k-\gamma+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha}$$
$$+\sum_{r=k-\gamma+1}^{k}W^{k,k-r}\left(\frac{\mu\alpha(1-\theta_r)}{\theta_r+\mu\alpha}W^r\mathbf{x}^r-\frac{\alpha}{\theta_r+\mu\alpha}\mathbf{s}^r\right)\prod_{t=r+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha},$$

and

$$\|\Pi\mathbf{z}^{k+1}\|\overset{b}{\leq}\sigma_\gamma\|\Pi\mathbf{z}^{k-\gamma+1}\|+\sum_{r=k-\gamma+1}^{k}\left(\frac{\mu\alpha(1-\theta_r)}{\theta_r+\mu\alpha}\|\Pi\mathbf{x}^r\|+\frac{\alpha}{\theta_r+\mu\alpha}\|\Pi\mathbf{s}^r\|\right)$$
$$\leq\sigma_\gamma\|\Pi\mathbf{z}^{k-\gamma+1}\|+\sum_{r=k-\gamma+1}^{k}\left(\frac{\mu\alpha}{\theta_r}\|\Pi\mathbf{x}^r\|+\frac{\alpha}{\theta_r}\|\Pi\mathbf{s}^r\|\right),$$

where we use (13a) in $\overset{a}{=}$, $\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha}\leq 1$ with $\theta_t\leq 1$ in $\overset{b}{\leq}$. Similar to (44), squaring both sides yields

$$\|\Pi\mathbf{z}^{k+1}\|^2\leq\frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{z}^{k-\gamma+1}\|^2+\frac{4\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+1}^{k}\left(\frac{\mu^2\alpha^2}{\theta_r^2}\|\Pi\mathbf{x}^r\|^2+\frac{\alpha^2}{\theta_r^2}\|\Pi\mathbf{s}^r\|^2\right).$$

Multiplying both sides by θ_{k+1}^2 and using the non-increasing of $\{\theta_k\}$, it further gives (42) ■

Motivated by the proof of Lemma 20, we want to construct a linear combination of the consensus errors. However, due to the time-varying graphs and the γ-step joint spectrum property in Assumption 3, we see from (40)-(42) that they shrink every γ iterations, rather than every iteration. By exploiting the special structures in (40)-(42), we define the following quantities:

$$\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma}=\max_{r=k+1,\ldots,k+\gamma}\|\Pi\mathbf{s}^r\|^2,\quad\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}=\max_{r=k+1,\ldots,k+\gamma}\|\Pi\mathbf{x}^r\|^2,$$
$$\mathcal{M}_{\mathbf{y}}^{k+\gamma,\gamma}=\max_{r=k+1,\ldots,k+\gamma}\|\Pi\mathbf{y}^r\|^2,\quad\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma}=\max_{r=k+1,\ldots,k+\gamma}\theta_r^2\|\Pi\mathbf{z}^r\|^2.$$

Motivated by (25), we define the following quantity in the form of summation, instead of the maximum, and we sum up to $k+\gamma-1$, rather than $k+\gamma$,

$$\mathcal{S}_{\phi}^{k+\gamma-1,\gamma}=\sum_{r=k}^{k+\gamma-1}\theta_r^2\Phi^r.$$

The next lemma is an analogy counterpart of Lemma 20. Unlike the classical analysis relying on the small gain theorem (Nedić et al., 2017), which is unclear how to be used to the accelerated methods, and especially for nonstrongly convex problems, our main idea is to construct a linear combination of $\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma}$, $\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}$, and $\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma}$ with carefully designed weights such that it shrinks geometrically with the additional error term $\mathcal{S}_{\phi}^{k+\gamma-1,\gamma}$, which is crucial to extend our analysis over static graphs to time-varying graphs in a unified

framework, both for nonstrongly convex and strongly convex problems. Moreover, our proof technique to bound the consensus errors can be embedded into many algorithm frameworks, because it is separated from the analysis of the inexact accelerated gradient descent in Lemma 19.

Lemma 25 *Under the settings of Lemma 24, letting* $\alpha \leq \frac{(1-\sigma_\gamma)^3}{3385L\gamma^3\sqrt{1+\frac{1}{\tau}}}$, *we have for any* $t \geq 0$,

$$\max\left\{\mathcal{M}_{\mathbf{y}}^{(t+1)\gamma,\gamma}, \mathcal{M}_{\mathbf{x}}^{(t+1)\gamma,\gamma}\right\} \leq C_3\rho^{t\gamma} + C_4\sum_{s=0}^{(t+1)\gamma-1}\rho^{(t-1)\gamma-s}\theta_s^2\Phi^s. \tag{45}$$

where $\rho = \sqrt[\gamma]{1-\frac{1-\sigma_\gamma}{5}}$, $C_3 = \left(\frac{(1-\sigma_\gamma)^2}{162L^2\gamma^2(1+\frac{1}{\tau})}\mathcal{M}_{\mathbf{s}}^{\gamma,\gamma} + \frac{442\gamma^2}{(1-\sigma_\gamma)^2}\mathcal{M}_{\mathbf{z}}^{\gamma,\gamma} + 2\mathcal{M}_{\mathbf{x}}^{\gamma,\gamma}\right)$, *and* $C_4 = \frac{1-\sigma_\gamma}{38L^2\gamma(1+\frac{1}{\tau})}$.

Proof For any t satisfying $k \leq t \leq k+\gamma-1$ with $k \geq \gamma$, we can relax (40) to

$$\begin{aligned}
\|\Pi\mathbf{s}^{t+1}\|^2 \leq& \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{t-\gamma+1}\|^2 + \frac{2\gamma}{1-\sigma_\gamma}\sum_{r=t-\gamma+1}^{t}\left(\theta_r^2\Phi^r + c_0\theta_r^2\|\Pi\mathbf{z}^r\|^2 + c_0\|\Pi\mathbf{x}^r\|^2\right)\\
\leq& \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{t-\gamma+1}\|^2 + \frac{2\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma}^{k+\gamma-1}\theta_r^2\Phi^r + \frac{2\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+1}^{k+\gamma}\left(c_0\theta_r^2\|\Pi\mathbf{z}^r\|^2 + c_0\|\Pi\mathbf{x}^r\|^2\right)\\
\leq& \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{t-\gamma+1}\|^2 + \frac{2\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right)\\
&+ \frac{2\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+1}^{k+\gamma}\left(c_0\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\right)\\
=& \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{t-\gamma+1}\|^2 + \frac{2\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right)\\
&+ \frac{4\gamma^2}{1-\sigma_\gamma}\left(c_0\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\right).
\end{aligned}$$

Taking the maximum over $t = k, k+1, ..., k+\gamma-1$ on both sides, we have

$$\begin{aligned}
\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma} \leq& \frac{1+\sigma_\gamma}{2}\mathcal{M}_{\mathbf{s}}^{k,\gamma} + \frac{2\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right)\\
&+ \frac{4\gamma^2}{1-\sigma_\gamma}\left(c_0\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\right).
\end{aligned}$$

Similarly, for (42) and (41), we also have

$$\begin{aligned}
\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} \leq& \frac{1+\sigma_\gamma}{2}\mathcal{M}_{\mathbf{z}}^{k,\gamma} + \frac{8\gamma^2}{1-\sigma_\gamma}\left(\mu^2\alpha^2\mathcal{M}_{\mathbf{x}}^{k,\gamma} + \mu^2\alpha^2\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma} + \alpha^2\mathcal{M}_{\mathbf{s}}^{k,\gamma} + \alpha^2\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma}\right),\\
\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma} \leq& \frac{1+\sigma_\gamma}{2}\mathcal{M}_{\mathbf{x}}^{k,\gamma} + \frac{11\gamma^2}{1-\sigma_\gamma}\left(\mathcal{M}_{\mathbf{z}}^{k,\gamma} + \mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma}\right),
\end{aligned}$$

where for the second one, the relaxation of $\sum_{r=t-\gamma+2}^{t+1}\theta_r^2\|\Pi\mathbf{z}^r\|^2 \le \sum_{r=k-\gamma+1}^{k+\gamma}\theta_r^2\|\Pi\mathbf{z}^r\|^2$ also holds for any t satisfying $k \le t \le k+\gamma-1$.

Adding the above three inequalities together with weights c_1, c_2, and c_3, respectively, we have

$$\begin{aligned}
&c_1\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma} + c_2\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_3\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\\
&\le \frac{8c_2\gamma^2\alpha^2}{1-\sigma_\gamma}\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma} + \left(\frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{11c_3\gamma^2}{1-\sigma_\gamma}\right)\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + \left(\frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{8c_2\gamma^2\mu^2\alpha^2}{1-\sigma_\gamma}\right)\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\\
&\quad + \left(\frac{c_1(1+\sigma_\gamma)}{2} + \frac{8c_2\gamma^2\alpha^2}{1-\sigma_\gamma}\right)\mathcal{M}_{\mathbf{s}}^{k,\gamma} + \left(\frac{c_2(1+\sigma_\gamma)}{2} + \frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{11c_3\gamma^2}{1-\sigma_\gamma}\right)\mathcal{M}_{\mathbf{z}}^{k,\gamma}\\
&\quad + \left(\frac{c_3(1+\sigma_\gamma)}{2} + \frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{8c_2\gamma^2\mu^2\alpha^2}{1-\sigma_\gamma}\right)\mathcal{M}_{\mathbf{x}}^{k,\gamma} + \frac{2c_1\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right).
\end{aligned}$$

We want to choose c_1, c_2, c_3, and α such that the following inequalities hold,

$$\frac{8c_2\gamma^2\alpha^2}{1-\sigma_\gamma} \le \frac{c_1(1-\sigma_\gamma)}{20}, \qquad \frac{c_1(1+\sigma_\gamma)}{2} + \frac{8c_2\gamma^2\alpha^2}{1-\sigma_\gamma} \le \frac{c_1(3+\sigma_\gamma)}{4},$$

$$\frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{11c_3\gamma^2}{1-\sigma_\gamma} \le \frac{c_2(1-\sigma_\gamma)}{20}, \qquad \frac{c_2(1+\sigma_\gamma)}{2} + \frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{11c_3\gamma^2}{1-\sigma_\gamma} \le \frac{c_2(3+\sigma_\gamma)}{4},$$

$$\frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{8c_2\gamma^2\mu^2\alpha^2}{1-\sigma_\gamma} \le \frac{c_3(1-\sigma_\gamma)}{20}, \qquad \frac{c_3(1+\sigma_\gamma)}{2} + \frac{4c_1c_0\gamma^2}{1-\sigma_\gamma} + \frac{8c_2\gamma^2\mu^2\alpha^2}{1-\sigma_\gamma} \le \frac{c_3(3+\sigma_\gamma)}{4},$$

which are satisfied if the three inequalities in the left column hold. Accordingly, we can choose $c_3 = \frac{81c_1c_0\gamma^2}{(1-\sigma_\gamma)^2}$, $c_2 = \frac{17900c_1c_0\gamma^4}{(1-\sigma_\gamma)^4} \ge \frac{80c_1c_0\gamma^2}{(1-\sigma_\gamma)^2} + \frac{220c_3\gamma^2}{(1-\sigma_\gamma)^2}$, and $\alpha^2 \le \min\left\{\frac{(1-\sigma_\gamma)^6}{2864000c_0\gamma^6}, \frac{(1-\sigma_\gamma)^4}{2864000\mu^2\gamma^4}\right\}$. Thus, we have for any $k \ge \gamma$,

$$\begin{aligned}
&\frac{19+\sigma_\gamma}{20}\left(c_1\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma} + c_2\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_3\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\right)\\
&\le \frac{3+\sigma_\gamma}{4}\left(c_1\mathcal{M}_{\mathbf{s}}^{k,\gamma} + c_2\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_3\mathcal{M}_{\mathbf{x}}^{k,\gamma}\right) + \frac{2c_1\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right)\\
&\le \frac{19+\sigma_\gamma}{20}\left(1-\frac{1-\sigma_\gamma}{5}\right)\left(c_1\mathcal{M}_{\mathbf{s}}^{k,\gamma} + c_2\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_3\mathcal{M}_{\mathbf{x}}^{k,\gamma}\right) + \frac{2c_1\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right),
\end{aligned}$$

and

$$\begin{aligned}
&c_1\mathcal{M}_{\mathbf{s}}^{(t+1)\gamma,\gamma} + c_2\mathcal{M}_{\mathbf{z}}^{(t+1)\gamma,\gamma} + c_3\mathcal{M}_{\mathbf{x}}^{(t+1)\gamma,\gamma}\\
&\le \left(1-\frac{1-\sigma_\gamma}{5}\right)^t\left(c_1\mathcal{M}_{\mathbf{s}}^{\gamma,\gamma} + c_2\mathcal{M}_{\mathbf{z}}^{\gamma,\gamma} + c_3\mathcal{M}_{\mathbf{x}}^{\gamma,\gamma}\right)\\
&\quad + \frac{40c_1\gamma}{19(1-\sigma_\gamma)}\sum_{r=1}^{t}\left(1-\frac{1-\sigma_\gamma}{5}\right)^{t-r}\left(\mathcal{S}_\phi^{r\gamma-1,\gamma} + \mathcal{S}_\phi^{(r+1)\gamma-1,\gamma}\right).
\end{aligned} \tag{46}$$

It follows from (27) and $c_2 > c_3$ that

$$\frac{c_3}{2}\max\left\{\mathcal{M}_{\mathbf{y}}^{(t+1)\gamma,\gamma}, \mathcal{M}_{\mathbf{x}}^{(t+1)\gamma,\gamma}\right\} \le c_1\mathcal{M}_{\mathbf{s}}^{(t+1)\gamma,\gamma} + c_2\mathcal{M}_{\mathbf{z}}^{(t+1)\gamma,\gamma} + c_3\mathcal{M}_{\mathbf{x}}^{(t+1)\gamma,\gamma}.$$

On the other hand, denoting $\rho = \sqrt[\gamma]{1 - \frac{1-\sigma_\gamma}{5}}$, we have

$$\begin{aligned}
&\sum_{r=1}^{t} \rho^{\gamma(t-r)} \left(\mathcal{S}_\phi^{r\gamma-1,\gamma} + \mathcal{S}_\phi^{(r+1)\gamma-1,\gamma} \right) \\
&= \rho^{\gamma t} \sum_{r=1}^{t} \left(\frac{1}{\rho^\gamma} \right)^r \left(\sum_{s=(r-1)\gamma}^{r\gamma-1} \theta_s^2 \Phi^s + \sum_{s=r\gamma}^{(r+1)\gamma-1} \theta_s^2 \Phi^s \right) \\
&= \rho^{\gamma t} \sum_{s=0}^{t\gamma-1} \left(\frac{1}{\rho^\gamma} \right)^{\lfloor \frac{s}{\gamma} \rfloor + 1} \theta_s^2 \Phi^s + \rho^{\gamma t} \sum_{s=\gamma}^{(t+1)\gamma-1} \left(\frac{1}{\rho^\gamma} \right)^{\lfloor \frac{s}{\gamma} \rfloor} \theta_s^2 \Phi^s \\
&\leq 2\rho^{\gamma t} \sum_{s=0}^{(t+1)\gamma-1} \left(\frac{1}{\rho^\gamma} \right)^{\frac{s}{\gamma}+1} \theta_s^2 \Phi^s = 2 \sum_{s=0}^{(t+1)\gamma-1} \rho^{(t-1)\gamma-s} \theta_s^2 \Phi^s.
\end{aligned}$$

Plugging the above two inequalities and the settings of c_3 and c_0 into (46), we have the conclusion. ■

Remark 26 *We briefly demonstrate the advantage of introducing the quantities of $\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma}$, $\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}$, $\mathcal{M}_{\mathbf{y}}^{k+\gamma,\gamma}$, and $\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma}$. As discussed in Remark 23, researchers in the control community often use linear system inequality to prove the convergence, which is quite challenging to use over time-varying graphs. For example, Saadatniaki et al. (2020) constructed a γth order linear system inequality in the form of*

$$\begin{pmatrix} \alpha^{k+\gamma} \\ \alpha^{k+\gamma-1} \\ \alpha^{k+\gamma-2} \\ \vdots \\ \alpha^{k+1} \end{pmatrix} \leq \begin{pmatrix} M_1 & M_2 & \cdots & M_{\gamma-1} & M_\gamma \\ I & & & & \\ & I & & & \\ & & \ddots & & \\ & & & I & \end{pmatrix} \begin{pmatrix} \alpha^{k+\gamma-1} \\ \alpha^{k+\gamma-2} \\ \alpha^{k+\gamma-3} \\ \vdots \\ \alpha^k \end{pmatrix} \tag{47}$$

for the $\mathcal{AB}$/push-pull method, which is an extension of gradient tracking to time-varying directed graphs. They only proved that the spectral radius of the system matrix is strictly less than 1 without any explicit upper bound. Thus, no explicit convergence rate was given in (Saadatniaki et al., 2020).

On the other hand, the system (47) can be simplified by defining similar quantities of $\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma}$, $\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}$, $\mathcal{M}_{\mathbf{y}}^{k+\gamma,\gamma}$, and $\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma}$. Moreover, the proof can be further simplified by avoiding analyzing the spectral radius if our technical trick of constructing the linear combination is used.

Following the same proof framework over static graphs, our next step is to bound the weighted cumulative consensus errors. However, the details are much more complex. The proof of Lemma 21 provides some insights.

Lemma 27 *Suppose that Assumptions 1 and 3 hold with $\mu = 0$. Let the sequence $\{\theta_k\}_{k=0}^{T\gamma}$ satisfy $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ with $\theta_0 = 1$, let $\alpha \le \frac{(1-\sigma_\gamma)^3}{3385L\gamma^3\sqrt{1+\frac{1}{\tau}}}$. Then for algorithm (13a)-(13d), we have*

$$\begin{aligned}&\max\left\{\sum_{k=0}^{T\gamma}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{y}^k\|^2, \sum_{k=0}^{T\gamma}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{x}^k\|^2\right\}\\ &\le \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} + \frac{10\gamma^2}{mL(1+\frac{1}{\tau})(1-\sigma_\gamma)^2}\sum_{s=0}^{T\gamma-1}\Phi^s,\end{aligned} \tag{48}$$

where τ and Φ^r are defined in Lemma 20, and C_3 is defined in Lemma 25.

Proof We first verify $\theta_k \le 1.62\theta_{k+1}$ for all $k \ge 0$, which is required in Lemmas 24 and 25. In fact, from $\frac{1-\theta_{k+1}}{\theta_{k+1}^2} = \frac{1}{\theta_k^2}$ and $\theta_0 = 1$, we have $\frac{\theta_k}{\theta_{k+1}} = \frac{1}{\sqrt{1-\theta_{k+1}}} \in (1, \frac{1}{\sqrt{1-\theta_1}}] \in (1, 1.62]$ for any $k \ge 0$. Next, we upper and lower bound ρ. From the definition of $\rho = \sqrt[\gamma]{1-\frac{1-\sigma_\tau}{5}}$ and the fact that $(1-\frac{x}{\gamma})^\gamma \ge 1-x$ for any $x \in (0,1)$ and $\gamma \ge 1$, we know

$$\rho \le 1 - \frac{1-\sigma_\gamma}{5\gamma}, \qquad \rho^\gamma \ge \frac{4}{5}. \tag{49}$$

The remaining proof is similar to that of Lemma 21. From the definition of $\mathcal{M}_{\mathbf{y}}^{t\gamma+\gamma,\gamma}$ and (45), we have

$$\begin{aligned}&\sum_{k=1}^{T\gamma}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{y}^k\|^2 = \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{L}{2m\theta_{t\gamma+r}^2}\|\Pi\mathbf{y}^{t\gamma+r}\|^2 \le \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{L}{2m\theta_{t\gamma+r}^2}\mathcal{M}_{\mathbf{y}}^{(t+1)\gamma,\gamma}\\ &\le \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{C_3 L\rho^{t\gamma}}{2m\theta_{t\gamma+r}^2} + \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{C_4 L}{2m\theta_{t\gamma+r}^2}\sum_{s=0}^{(t+1)\gamma-1}\rho^{(t-1)\gamma-s}\theta_s^2\Phi^s\\ &\le \frac{C_3 L}{2m}\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{\rho^{t\gamma+r}}{\rho^\gamma\theta_{t\gamma+r}^2} + \frac{C_4 L}{2m}\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{\rho^{t\gamma+r}}{\rho^{2\gamma}\theta_{t\gamma+r}^2}\sum_{s=0}^{(t+1)\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s\\ &\overset{a}{=} \frac{C_3 L}{2m\rho^\gamma}\sum_{k=1}^{T\gamma}\frac{\rho^k}{\theta_k^2} + \frac{C_4 L}{2m\rho^{2\gamma}}\sum_{k=1}^{T\gamma}\frac{\rho^k}{\theta_k^2}\sum_{s=0}^{\lceil\frac{k}{\gamma}\rceil\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s,\end{aligned} \tag{50}$$

where $(t+1)\gamma - 1 = \lceil\frac{k}{\gamma}\rceil\gamma - 1$ in $\overset{a}{=}$ comes from the variable substitution $k = t\gamma + r$ with $r = 1, 2, ..., \gamma$. Next, we compute the second part in $\overset{a}{=}$. It gives

$$\begin{aligned}&\sum_{k=1}^{T\gamma}\frac{\rho^k}{\theta_k^2}\sum_{s=0}^{\lceil\frac{k}{\gamma}\rceil\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s\\ &\le \sum_{k=1}^{T\gamma-\gamma}\frac{\rho^k}{\theta_k^2}\sum_{s=0}^{k+\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s + \sum_{k=T\gamma-\gamma+1}^{T\gamma}\frac{\rho^k}{\theta_k^2}\sum_{s=0}^{T\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s\\ &= \left(\sum_{k=1}^{T\gamma-\gamma}\frac{\rho^k}{\theta_k^2}\sum_{s=0}^{\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s + \sum_{k=1}^{T\gamma-\gamma}\frac{\rho^k}{\theta_k^2}\sum_{s=\gamma}^{k+\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s\right) + \sum_{s=0}^{T\gamma-1}\frac{\theta_s^2}{\rho^s}\Phi^s\sum_{k=T\gamma-\gamma+1}^{T\gamma}\frac{\rho^k}{\theta_k^2}\end{aligned} \tag{51}$$

$$
\begin{aligned}
&= \sum_{s=0}^{\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=1}^{T\gamma-\gamma} \frac{\rho^k}{\theta_k^2} + \sum_{s=\gamma}^{T\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=s-\gamma+1}^{T\gamma-\gamma} \frac{\rho^k}{\theta_k^2} \\
&\quad + \left(\sum_{s=0}^{\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=T\gamma-\gamma+1}^{T\gamma} \frac{\rho^k}{\theta_k^2} + \sum_{s=\gamma}^{T\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=T\gamma-\gamma+1}^{T\gamma} \frac{\rho^k}{\theta_k^2} \right) \\
&= \sum_{s=0}^{\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=1}^{T\gamma} \frac{\rho^k}{\theta_k^2} + \sum_{s=\gamma}^{T\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=s-\gamma+1}^{T\gamma} \frac{\rho^k}{\theta_k^2}.
\end{aligned}
$$

Plugging (51) into (50), it follows from $\Pi \mathbf{y}^0 = 0$ that

$$
\begin{aligned}
&\sum_{k=0}^{T\gamma} \frac{L}{2m\theta_k^2} \|\Pi \mathbf{y}^k\|^2 \\
&\le \frac{C_3 L}{2m\rho^\gamma} \sum_{k=1}^{T\gamma} \frac{\rho^k}{\theta_k^2} + \frac{C_4 L}{2m\rho^{2\gamma}} \left(\sum_{s=0}^{\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=1}^{T\gamma} \frac{\rho^k}{\theta_k^2} + \sum_{s=\gamma}^{T\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \sum_{k=s-\gamma+1}^{T\gamma} \frac{\rho^k}{\theta_k^2} \right) \\
&\overset{b}{\le} \frac{C_3 L}{2m\rho^\gamma} \frac{3\rho}{(1-\rho)^3} + \frac{C_4 L}{2m\rho^{2\gamma}} \left(\frac{3\rho}{(1-\rho)^3} \sum_{s=0}^{\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s + \sum_{s=\gamma}^{T\gamma-1} \frac{\theta_s^2}{\rho^s}\Phi^s \frac{3\rho^{s-\gamma+1}}{(1-\rho)^3 \theta_{s-\gamma}^2} \right) \\
&\overset{c}{\le} \frac{3C_3 L}{2m\rho^{\gamma-1}(1-\rho)^3} + \frac{C_4 L}{2m\rho^{2\gamma}} \left(\frac{3\rho}{(1-\rho)^3 \rho^{\gamma-1}} \sum_{s=0}^{\gamma-1} \Phi^s + \frac{3\rho^{-\gamma+1}}{(1-\rho)^3} \sum_{s=\gamma}^{T\gamma-1} \Phi^s \right) \\
&\le \frac{3C_3 L}{2m\rho^{\gamma-1}(1-\rho)^3} + \frac{3C_4 L}{2m\rho^{3\gamma-1}(1-\rho)^3} \sum_{s=0}^{T\gamma-1} \Phi^s \\
&\overset{d}{\le} \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} + \frac{10\gamma^2}{mL(1+\frac{1}{\tau})(1-\sigma_\gamma)^2} \sum_{s=0}^{T\gamma-1} \Phi^s,
\end{aligned}
$$

where $\overset{b}{\le}$ uses (38), $\overset{c}{\le}$ uses $\theta_s \le 1$ for $s \le \gamma - 1$ and $\theta_s \le \theta_{s-\gamma}$ for $s \ge \gamma$, $\overset{d}{\le}$ uses (49) and the definition of C_4 given in Lemma 25. Replacing $\|\Pi \mathbf{y}^k\|$ by $\|\Pi \mathbf{x}^k\|$ in the above analysis, we have the same bound for $\|\Pi \mathbf{x}^k\|^2$. ■

The next lemma is an analogy counterpart of Lemma 22, and the proof is similar to that of the above Lemma 27.

Lemma 28 *Suppose that Assumptions 1 and 3 hold with $\mu > 0$. Let $\alpha \le \frac{(1-\sigma_\gamma)^3}{3385 L\gamma^3 \sqrt{1+\frac{1}{\tau}}}$ and $\theta_k \equiv \theta = \frac{\sqrt{\mu\alpha}}{2}$. Then for algorithm (13a)-(13d), we have*

$$
\begin{aligned}
&\max \left\{ \sum_{k=0}^{T\gamma} \frac{L}{2m(1-\theta)^{k+1}} \|\Pi \mathbf{y}^k\|^2, \sum_{k=0}^{T\gamma} \frac{L}{2m(1-\theta)^{k+1}} \|\Pi \mathbf{x}^k\|^2 \right\} \\
&\le \frac{3.3 C_3 L\gamma}{m(1-\theta)(1-\sigma_\gamma)} + \frac{\theta^2}{7mL(1+\frac{1}{\tau})} \sum_{s=0}^{T\gamma-1} \frac{\Phi^s}{(1-\theta)^{s+1}},
\end{aligned} \tag{52}
$$

where τ and Φ^r are defined in Lemma 20, and C_3 is defined in Lemma 25.

Proof From the definition of $\mathcal{M}_{\mathbf{y}}^{t\gamma,\gamma}$ and (45), we have

$$\begin{aligned}
&\sum_{k=1}^{T\gamma}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2\\
&=\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{L}{2m(1-\theta)^{t\gamma+r+1}}\|\Pi\mathbf{y}^{t\gamma+r}\|^2\le\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{L}{2m(1-\theta)^{t\gamma+r+1}}\mathcal{M}_{\mathbf{y}}^{(t+1)\gamma,\gamma}\\
&\le\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{C_3L\rho^{t\gamma}}{2m(1-\theta)^{t\gamma+r+1}}+\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{C_4L}{2m(1-\theta)^{t\gamma+r+1}}\sum_{s=0}^{(t+1)\gamma-1}\rho^{(t-1)\gamma-s}\theta^2\Phi^s\\
&\le\frac{C_3L}{2m(1-\theta)}\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{\rho^{t\gamma+r}}{\rho^{\gamma}(1-\theta)^{t\gamma+r}}+\frac{C_4L\theta^2}{2m(1-\theta)}\sum_{t=0}^{T-1}\sum_{r=1}^{\gamma}\frac{\rho^{t\gamma+r}}{\rho^{2\gamma}(1-\theta)^{t\gamma+r}}\sum_{s=0}^{(t+1)\gamma-1}\frac{\Phi^s}{\rho^s}\\
&=\frac{C_3L}{2m\rho^{\gamma}(1-\theta)}\sum_{k=1}^{T\gamma}\left(\frac{\rho}{1-\theta}\right)^k+\frac{C_4L\theta^2}{2m\rho^{2\gamma}(1-\theta)}\sum_{k=1}^{T\gamma}\left(\frac{\rho}{1-\theta}\right)^k\sum_{s=0}^{\lceil\frac{k}{\gamma}\rceil\gamma-1}\frac{\Phi^s}{\rho^s}.
\end{aligned}$$

Similar to (51), we have

$$\sum_{k=1}^{T\gamma}\left(\frac{\rho}{1-\theta}\right)^k\sum_{s=0}^{\lceil\frac{k}{\gamma}\rceil\gamma-1}\frac{\Phi^s}{\rho^s}\le\sum_{s=0}^{\gamma-1}\frac{\Phi^s}{\rho^s}\sum_{k=1}^{T\gamma}\left(\frac{\rho}{1-\theta}\right)^k+\sum_{s=\gamma}^{T\gamma-1}\frac{\Phi^s}{\rho^s}\sum_{k=s-\gamma+1}^{T\gamma}\left(\frac{\rho}{1-\theta}\right)^k.$$

From the settings of θ and α, we know $\theta\le\frac{1-\sigma_\gamma}{116\gamma}$. From (49), we further have $\frac{\rho}{1-\theta}<1$ and $1-\rho-\theta\ge\frac{0.19(1-\sigma_\gamma)}{\gamma}$. So we have $\sum_{k=r+1}^{K}\left(\frac{\rho}{1-\theta}\right)^k\le\left(\frac{\rho}{1-\theta}\right)^r\frac{\rho}{1-\theta-\rho}$. It follows from $\|\Pi\mathbf{y}^0\|=0$ that

$$\begin{aligned}
&\sum_{k=0}^{T\gamma}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2\le\frac{C_3L}{2m\rho^{\gamma-1}(1-\theta)(1-\theta-\rho)}\\
&\quad+\frac{C_4L\theta^2}{2m\rho^{2\gamma-1}(1-\theta)(1-\theta-\rho)}\left(\sum_{s=0}^{\gamma-1}\frac{\Phi^s}{(1-\theta)^s}\left(\frac{1-\theta}{\rho}\right)^s+\left(\frac{\rho}{1-\theta}\right)^{-\gamma}\sum_{s=\gamma}^{T\gamma-1}\frac{\Phi^s}{(1-\theta)^s}\right)\\
&\overset{a}{\le}\frac{C_3L}{2m\rho^{\gamma-1}(1-\theta)(1-\theta-\rho)}+\frac{C_4L\theta^2(1-\theta)^{\gamma}}{2m\rho^{3\gamma-1}(1-\theta-\rho)}\sum_{s=0}^{T\gamma-1}\frac{\Phi^s}{(1-\theta)^{s+1}}\\
&\overset{b}{\le}\frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)}+\frac{\theta^2}{7mL(1+\frac{1}{\tau})}\sum_{s=0}^{T\gamma-1}\frac{\Phi^s}{(1-\theta)^{s+1}}.
\end{aligned}$$

where $\overset{a}{\le}$ uses $\frac{1-\theta}{\rho}>1$ such that $\left(\frac{1-\theta}{\rho}\right)^s\le\left(\frac{1-\theta}{\rho}\right)^{\gamma}$ for all $s\le\gamma-1$, $\overset{b}{\le}$ uses (49), $1-\rho-\theta\ge\frac{0.19(1-\sigma_\gamma)}{\gamma}$, and the definition of C_4 given in Lemma 25. Replacing $\|\Pi\mathbf{y}^k\|$ by $\|\Pi\mathbf{x}^k\|$ in the above analysis, we have the same bound for $\|\Pi\mathbf{x}^k\|^2$. ■

Now, we are ready to prove Theorems 2 and 3. We first prove Theorem 2.

Proof Plugging (48) into (21) and using the definition of Φ^r in (26), we have

$$\begin{aligned}
&\frac{F(\overline{x}^{T\gamma+1})-F(x^*)}{\theta_{T\gamma}^2}+\frac{1}{2\alpha}\|\overline{z}^{T\gamma+1}-x^*\|^2\\
&\le \frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}-\sum_{k=0}^{T\gamma}\left(\left(\frac{1}{2\alpha}-\frac{L}{2}-\frac{20L\gamma^2}{(1-\sigma_\gamma)^2}\right)\|\overline{z}^{t+1}-\overline{z}^t\|^2\right.\\
&\qquad\left.+\frac{1}{\theta_{k-1}^2}\left(1-\frac{20(1+\tau)\gamma^2}{(1+\frac{1}{\tau})(1-\sigma_\gamma)^2}\right)D_f(\overline{x}^k,\mathbf{y}^k)\right)\\
&\overset{a}{\le} \frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}-\sum_{k=0}^{T\gamma}\left(\frac{1}{4\alpha}\|\overline{z}^{t+1}-\overline{z}^t\|^2+\frac{1}{2\theta_{k-1}^2}D_f(\overline{x}^k,\mathbf{y}^k)\right)\\
&\le \frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}-\frac{1}{5mL}\sum_{r=0}^{T\gamma-1}\Phi^r,
\end{aligned}$$

where in $\overset{a}{\le}$ we let $\tau=\frac{(1-\sigma_\gamma)^2}{40\gamma^2}$ so to have $\frac{20(1+\tau)\gamma^2}{(1+\frac{1}{\tau})(1-\sigma_\gamma)^2}=\frac{1}{2}$, $\alpha=\frac{(1-\sigma_\gamma)^4}{21675L\gamma^4}\le\frac{(1-\sigma_\gamma)^3}{3385L\gamma^3\sqrt{1+\frac{1}{\tau}}}$, and $\frac{1}{4\alpha}\ge\frac{L}{2}+\frac{20L\gamma^2}{(1-\sigma_\gamma)^2}$. So we have

$$F(\overline{x}^{T\gamma+1})-F(x^*)\le\theta_{T\gamma}^2\left(\frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}\right),\tag{53}$$

$$\frac{1}{5mL}\sum_{r=0}^{T\gamma-1}\Phi^r\le\frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}.\tag{54}$$

It follows from (48) that

$$\begin{aligned}
&\max\left\{\sum_{k=0}^{T\gamma}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{y}^k\|^2,\sum_{k=0}^{T\gamma}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{x}^k\|^2\right\}\\
&\le\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}+\frac{10\gamma^2}{mL(1+\frac{1}{\tau})(1-\sigma_\gamma)^2}\sum_{s=0}^{T\gamma-1}\Phi^s\\
&\le\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}+\frac{1}{4mL}\sum_{s=0}^{T\gamma-1}\Phi^s\\
&\le\frac{9}{4}\left(\frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}\right).
\end{aligned}\tag{55}$$

From the definition of C_3 given in Lemma 25, we have

$$\begin{aligned}
&\frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}\\
&\le\frac{1}{2\alpha}\|\overline{z}^0-x^*\|^2+\frac{1-\sigma_\gamma}{27mL\gamma}\mathcal{M}_{\mathbf{s}}^{\gamma,\gamma}+\frac{103870L\gamma^5}{m(1-\sigma_\gamma)^5}\mathcal{M}_{\mathbf{z}}^{\gamma,\gamma}+\frac{470L\gamma^3}{m(1-\sigma_\gamma)^3}\mathcal{M}_{\mathbf{x}}^{\gamma,\gamma}\equiv C_5.
\end{aligned}\tag{56}$$

The conclusion follows from Lemma 29. ■

The next lemma gives a sharper bound of the constant C_5 appeared in the above proof.

Lemma 29 *Under the settings of Theorem 2, we can further bound C_5 by*

$$C_5 \leq \frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{1-\sigma_\gamma}{20mL\gamma}\max_{r=0,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2. \tag{57}$$

Proof From step (13c) with $\mu = 0$, we have for any $k \leq \gamma - 1$,

$$\begin{aligned}\theta_{k+1}\|\Pi \mathbf{z}^{k+1}\| \leq & \theta_k\|\Pi \mathbf{z}^{k+1}\| \leq \theta_k\|\Pi \mathbf{z}^k\| + \alpha\|\Pi \mathbf{s}^k\| \\ \leq & \theta_0\|\Pi \mathbf{z}^0\| + \alpha\sum_{t=0}^{k}\|\Pi \mathbf{s}^t\| \leq \alpha\sum_{t=0}^{\gamma-1}\|\Pi \mathbf{s}^t\|\end{aligned}$$

where we use $\Pi \mathbf{z}^0 = 0$. Squaring both sides gives

$$\theta_{k+1}^2\|\Pi \mathbf{z}^{k+1}\|^2 \leq \alpha^2\gamma\sum_{t=0}^{\gamma-1}\|\Pi \mathbf{s}^t\|^2 \leq \alpha^2\gamma^2\max_{r=0,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2.$$

From the setting of α and the definition of $\mathcal{M}_{\mathbf{z}}^{\gamma,\gamma}$, we have

$$\frac{103870L\gamma^5}{m(1-\sigma_\gamma)^5}\mathcal{M}_{\mathbf{z}}^{\gamma,\gamma} \leq \frac{1-\sigma_\gamma}{4523mL\gamma}\max_{r=0,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2.$$

On the other hand, it follows from step (13d) that

$$\|\Pi \mathbf{x}^{k+1}\| \leq \theta_k\|\Pi \mathbf{z}^{k+1}\| + \|\Pi \mathbf{x}^k\| \leq \sum_{t=0}^{k}\theta_t\|\Pi \mathbf{z}^{t+1}\| \leq 1.62\sum_{t=0}^{\gamma-1}\theta_{t+1}\|\Pi \mathbf{z}^{t+1}\|.$$

Squaring both sides gives

$$\|\Pi \mathbf{x}^{k+1}\|^2 \leq 2.63\gamma\sum_{t=0}^{\gamma-1}\theta_{t+1}^2\|\Pi \mathbf{z}^{t+1}\|^2 \leq 2.63\gamma^4\alpha^2\max_{r=0,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2.$$

From the setting of α and the definition of $\mathcal{M}_{\mathbf{x}}^{\gamma,\gamma}$, we have

$$\frac{470L\gamma^3}{m(1-\sigma_\gamma)^3}\mathcal{M}_{\mathbf{x}}^{\gamma,\gamma} \leq \frac{1-\sigma_\gamma}{380070mL\gamma}\max_{r=0,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2.$$

So we have the conclusion. ■

In the next lemma, we measure the convergence rate at $x_{(i)}^{t\gamma+1}$ for any $i = 1, \ldots, m$.

Lemma 30 *Under the settings of Theorem 2, we have for any $t \leq T - 1$,*

$$\begin{aligned}&F(x_{(i)}^{t\gamma+1}) - F(x^*) \\ &\leq \frac{1}{(t\gamma+1)^2}\max\left\{\frac{\sqrt{m}(1-\sigma_\gamma)}{L\alpha\gamma}, 8m\right\}\left(\frac{2}{\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{1-\sigma_\gamma}{5mL\gamma}\max_{r=0,\ldots,\gamma}\|\Pi \mathbf{s}^r\|^2\right).\end{aligned}$$

Proof We first bound $F(x_{(i)}^k) - F(\overline{x}^k)$ for any i. From Lemma 18, we have

$$\begin{aligned}
F(x_{(i)}^k) \leq & f(\overline{y}^{k-1}, \mathbf{y}^{k-1}) + \left\langle \overline{s}^{k-1}, x_{(i)}^k - \overline{y}^{k-1} \right\rangle + \frac{L}{2}\|x_{(i)}^k - \overline{y}^{k-1}\|^2 + \frac{L}{2m}\|\Pi \mathbf{y}^{k-1}\|^2 \\
\leq & F(\overline{x}^k) + \left\langle \overline{s}^{k-1}, x_{(i)}^k - \overline{x}^k \right\rangle + L\|x_{(i)}^k - \overline{x}^k\|^2 + L\|\overline{x}^k - \overline{y}^{k-1}\|^2 + \frac{L}{2m}\|\Pi \mathbf{y}^{k-1}\|^2 \\
\overset{a}{\leq} & F(\overline{x}^k) + \frac{\theta_{k-1}}{\alpha}\|\overline{z}^k - \overline{z}^{k-1}\|\|\Pi \mathbf{x}^k\| + L\|\Pi \mathbf{x}^k\|^2 + L\theta_{k-1}^2\|\overline{z}^k - \overline{z}^{k-1}\|^2 + \frac{L}{2m}\|\Pi \mathbf{y}^{k-1}\|^2,
\end{aligned}$$

where we use (15c) with $\mu = 0$, (15a), and (15d) in $\overset{a}{\leq}$. From the definition of Φ^r in (26), it follows from (54) that for any $k \leq T\gamma$,

$$\|\overline{z}^k - \overline{z}^{k-1}\|^2 \leq \frac{\Phi^{k-1}}{2mL^2(1+\frac{1}{\tau})} \overset{b}{\leq} \frac{5(1-\sigma_\gamma)^2}{80L\gamma^2}\left(\frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3}\right),$$

where $\overset{b}{\leq}$ uses the setting of $\tau = \frac{(1-\sigma_\gamma)^2}{40\gamma^2}$ given in the proof of Theorem 2. From (53) and (55), we have for any $t\gamma + 1$ with $t \leq T - 1$

$$F(x_{(i)}^{t\gamma+1}) - F(x^*) \leq \theta_{t\gamma}^2 \max\left\{\frac{\sqrt{m}(1-\sigma_\gamma)}{L\alpha\gamma}, 8m\right\}\left(\frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3}\right).$$

From (56), (57), and (36), we have the conclusion. ■

Next, we prove Theorem 3.

Proof Plugging (52) into (22) and using the definition of Φ^r in (26), we have

$$\begin{aligned}
& \frac{1}{(1-\theta)^{T\gamma+1}}\left(F(\overline{x}^{T\gamma+1}) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^{T\gamma+1} - x^*\|^2\right) \\
& \leq F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{3.3C_3 L\gamma}{m(1-\theta)(1-\sigma_\gamma)} \\
& \quad - \sum_{k=0}^{T\gamma}\left(\frac{1}{(1-\theta)^k}\left(1 - \frac{2(1+\tau)}{7(1+\frac{1}{\tau})}\right)D_f(\overline{x}^k, \mathbf{y}^k)\right. \\
& \qquad \left. + \frac{1}{(1-\theta)^{k+1}}\left(\frac{\theta^2}{2\alpha} - \frac{L\theta^2}{2} - \frac{2L\theta^2}{7}\right)\|\overline{z}^{k+1} - \overline{z}^k\|^2\right) \\
& \overset{a}{\leq} F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{3.3C_3 L\gamma}{m(1-\theta)(1-\sigma_\gamma)} \\
& \quad - \sum_{k=0}^{T\gamma}\left(\frac{1}{2(1-\theta)^k}D_f(\overline{x}^k, \mathbf{y}^k) + \frac{\theta^2}{4\alpha(1-\theta)^{k+1}}\|\overline{z}^{k+1} - \overline{z}^k\|^2\right) \\
& \leq F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{3.3C_3 L\gamma}{m(1-\theta)(1-\sigma_\gamma)} - \frac{\theta^2}{11mL}\sum_{r=0}^{T\gamma-1}\frac{\Phi^r}{(1-\theta)^{r+1}},
\end{aligned}$$

where in $\overset{a}{\leq}$ we let $\tau = \frac{7}{4}$ so to have $\frac{2(1+\tau)}{7(1+\frac{1}{\tau})} = \frac{1}{2}$, $\alpha = \frac{(1-\sigma_\gamma)^3}{4244L\gamma^3} \leq \frac{(1-\sigma_\gamma)^3}{3385L\gamma^3\sqrt{1+\frac{1}{\tau}}}$, and $\frac{1}{4\alpha} \geq \frac{L}{2} + \frac{2L}{7}$. Thus, we have the first conclusion and

$$\frac{\theta^2}{11mL}\sum_{r=0}^{T\gamma-1}\frac{\Phi^r}{(1-\theta)^{r+1}} \leq F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)}.$$

It follows from (52) that

$$\begin{aligned}
&\sum_{k=0}^{T\gamma}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{x}^k\|^2 \\
&\quad\leq \frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)} + \frac{\theta^2}{7mL(1+\frac{1}{\tau})}\sum_{s=0}^{T\gamma-1}\frac{\Phi^s}{(1-\theta)^{s+1}} \\
&\quad\leq 2\left(F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)}\right).
\end{aligned}$$

Thus, we have the second conclusion by plugging the definition of C_3 in Lemma 25. ■

Remark 31 *We rewrite the convergence rates in Theorems 2 and 3 in the form of complexities. For the nonstrongly convex case, letting* $F(\overline{x}^{T\gamma+1}) - F(x^*) \leq \frac{2C}{\alpha(T\gamma+1)^2} = \epsilon$, *we have* $T\gamma + 1 = \sqrt{\frac{2C}{\alpha\epsilon}} = \mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^2\sqrt{\frac{LC}{\epsilon}})$. *Each iteration only requires* $\mathcal{O}(1)$ *communication round and gradient oracle call. For the strongly convex case, letting* $F(\overline{x}^{T\gamma+1}) - F(x^*) \leq (1-\theta)^{T\gamma+1}C = \epsilon$, *we have* $T\gamma + 1 = \mathcal{O}(\frac{1}{\theta}\log\frac{C}{\epsilon}) = \mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$.

4. Numerical Experiments

In this section, we test the performance of the accelerated gradient tracking (Acc-GT) over time-varying graphs. The performance of Acc-GT over static graphs has already been verified in (Qu and Li, 2020). Moreover, Qu and Li (2020) reported in their experiment that algorithm (12a)-(12d) with fixed step size (our theoretical setting) performs faster than the one with vanishing step sizes (their theoretical setting). Thus, we omit the comparisons over static graphs.

We consider the following decentralized regularized logistic regression problem:

$$\min_{x\in\mathbb{R}^p}\sum_{i=1}^m f_{(i)}(x), \quad \text{where} \quad f_{(i)}(x) = \frac{\mu}{2}\|x\|^2 + \frac{1}{n}\sum_{j=1}^n \log\left(1 + \exp(-y_{(i),j}A_{(i),j}^\top x)\right),$$

where $(A_{(i),j}, y_{(i),j}) \in \mathbb{R}^p \times \{1,-1\}$ is the data point with $A_{(i),j}$ being the feature vector, and $y_{(i),j}$ the label. We use the cifar10 dataset with $p = 3072$, $n = 50$, and $m = 1000$. Each feature vector is normalized to have unit norm, and the data are divided into two classes to fit the logistic regression model. We observe that $L = \max_i \frac{\|A_{(i)}\|_2^2}{4n} \approx 0.215$. We consider both strongly convex ($\mu = 10^{-6}$) and nonstrongly convex ($\mu = 0$) problems. We

test the performance on the 2D grid graphs, where at each iteration, m nodes are uniformly placed in a $\lceil 5\sqrt{m} \rceil \times \lceil 5\sqrt{m} \rceil$ region in random, and each node is connected with the nodes around it within the distance of d. We test on $d = 20$ and $d = 2$, which correspond to $(\gamma, \sigma_\gamma) \approx (1, 0.9858)$ and $(\gamma, \sigma_\gamma) \approx (32, 0.9471)$, respectively. When $d = 20$, the network is connected almost every time. When $d = 2$, we observe that at each iteration, almost 61 percent of the nodes drop out from the communication network in average, which means that they have no connection with the other nodes. We use the Metropolis gossip matrix given in (8).

For strongly convex problem, we compare Acc-GT and Acc-GT-C (Acc-GT with multiple consensus) with DIGing (Nedić et al., 2017), DAGD-C (Rogozin et al., 2021b), as well as the classical non-distributed accelerated gradient descent (AGD), where AGD runs on a single machine, and it gives the upper limit of the practical performance of the distributed algorithms. We do not compare with the time-varying $\mathcal{AB}$/push-pull method (Saadatniaki et al., 2020) and the push-sum based methods (Nedić and Olshevsky, 2016, 2015; Nedić et al., 2017) because they are designed for directed graphs. We tune the step sizes $\alpha = \frac{0.1}{L}$ for Acc-GT and Acc-GT-C, $\alpha = \frac{0.5}{L}$ for DIGing, and $\alpha = \frac{1}{L}$ for AGD. For DAGD-C, when $d = 2$, we test on the number of inner iterations sa $T = \frac{\gamma}{3(1-\sigma_\gamma)} \approx 201$ and $T = \frac{\gamma}{2(1-\sigma_\gamma)} \approx 302$, and name the methods DAGD-C1 and DAGD-C2, respectively. When $d = 20$, we test on $T = \frac{\gamma}{5(1-\sigma_\gamma)} \approx 14$ and $T = \frac{\gamma}{4(1-\sigma_\gamma)} \approx 17$, respectively. For Acc-GT-C, we set the number of inner iterations as $T = \frac{\gamma}{50(1-\sigma_\gamma)} \approx 12$ and $T = \frac{\gamma}{10(1-\sigma_\gamma)} \approx 7$ for $d = 2$ and $d = 20$, respectively. The other parameter settings follow the corresponding theorems of each method. For nonstrongly convex problem, we compare Acc-GT and Acc-GT-C with DIGing (Nedić et al., 2017), APM (Li et al., 2020a), and AGD, and set the same step sizes as above. We tune the step size $\alpha = \frac{1}{L}$ for APM, and set the number of inner iterations as $T_k = \frac{\gamma \log(k+1)}{100(1-\sigma_\gamma)}$ and $T_k = \frac{\gamma \log(k+1)}{10(1-\sigma_\gamma)}$ at each outer loop iteration for $d = 2$ and $d = 20$, respectively. Although the convergence of DIGing was only proved for strongly convex problem in (Nedić et al., 2017), it also converges for nonstrongly convex ones by using our proof techniques.

Figures 1-4 plot the results, where the objective function error is measured by $F(\overline{x}^k) - F(x^*)$, and the consensus error is measured by $\sqrt{\frac{\sum_{i=1}^m \|x_{(i)}^k - \overline{x}^k\|^2}{m\|\overline{x}^k\|^2}}$. Since $F(x^*)$ is unknown, we approximate it by the output of the classical non-distributed AGD with 50000 iterations for strongly convex problem, and 200000 iterations for nonstrongly convex one. One round of communications means that all the nodes, if they are active, receive information from their neighbors once, and one round of gradient computations means that all the nodes compute their gradient $\nabla f_{(i)}(x)$ once in parallel. Especially, for AGD, one round of gradient computations means computing the full gradient $\sum_{i=1}^m \nabla f_{(i)}(x)$ once. We have the following observations:

1. Acc-GT converges faster than DIGing, both on the decrease of the objective function errors and consensus errors. This verifies the efficiency of the acceleration technique. Moreover, for strongly convex problem, Acc-CT is only three times slower than the classical non-distributed AGD.

2. Acc-GT-C needs more communication rounds than Acc-GT to reach the same precision of the objective function error, although Acc-GT-C has lower theoretically communi-

cation round complexity. Thus, Acc-GT-C is only for the theoretical interest, and it is not suggested in practice.

3. DAGD-C and APM need less gradient computation rounds than Acc-GT to reach the same precision of the objective function error, but they require more communication rounds. This supports that the multiple consensus subroutine places more communication burdens in practice. But on the other hand, DAGD-C and APM have almost the same computation cost as the classical non-distributed AGD. Comparing DAGD-C1 with DAGD-C2, we see that less inner iterations give larger consensus errors, and our settings of the inner iteration numbers are fair to DAGD-C.

4. The network connectivity, that is, the different settings of d in our experiment, has little influence on the decrease of the objective function errors for both DIGing and Acc-GT[3]. We think this is because we set the same step sizes for $d = 2$ and $d = 20$. From Theorems 2 and 3, we see that the network connectivity constants impact on the step sizes, and the step sizes impact on the decrease speed of the objective function errors. On the other hand, from the proofs of Theorems 2 and 3, we see that the decrease speed of the consensus errors given in the two theorems is not tight, and we observe in the experiment that the consensus errors decrease faster when $d = 20$ for both DIGing and Acc-GT.

5. Conclusion

This paper extends the widely used accelerated gradient tracking to time-varying network, which was originally proposed in (Qu and Li, 2020) only for static network. We prove the state-of-the-art complexities for both nonstrongly convex and strongly convex problems with the optimal dependence on the precision ϵ and the condition number L/μ, matching that of the classical centralized accelerated gradient descent. When the network is static, our complexities improve significantly over the previous ones proved in (Qu and Li, 2020). When combing with the Chebyshev acceleration, Our complexities exactly match the lower bounds for both nonstrongly convex and strongly convex problems over static graphs.

This paper only considers the γ-connectivity of time-varying graphs. Some researchers formulate the time-varying graphs as random graphs (Hong and Chang, 2017; Jakovetić et al., 2014b; Ananduta et al., 2020) and use the mean connectivity in expectation. It is an interesting future work to extend our proof techniques to random graphs. Another interesting direction is to study the acceleration over time-varying unbalanced directed graphs (Nedić et al., 2017). Besides gradient tracking, EXTRA is another important family of decentralized optimization algorithms. However, it remains an open problem to extend EXTRA to time-varying graphs.

3. This phenomenon depends on the data. We also test on the simulated data with $p = 100$, $n = 50$, and $m = 1000$, where each element of the feature vectors is generated randomly in $[0, 1]$ from the uniform distribution, we observe that Acc-GT with $d = 20$ performs about 1.1 times as fast as that with $d = 2$. The difference is not significant.

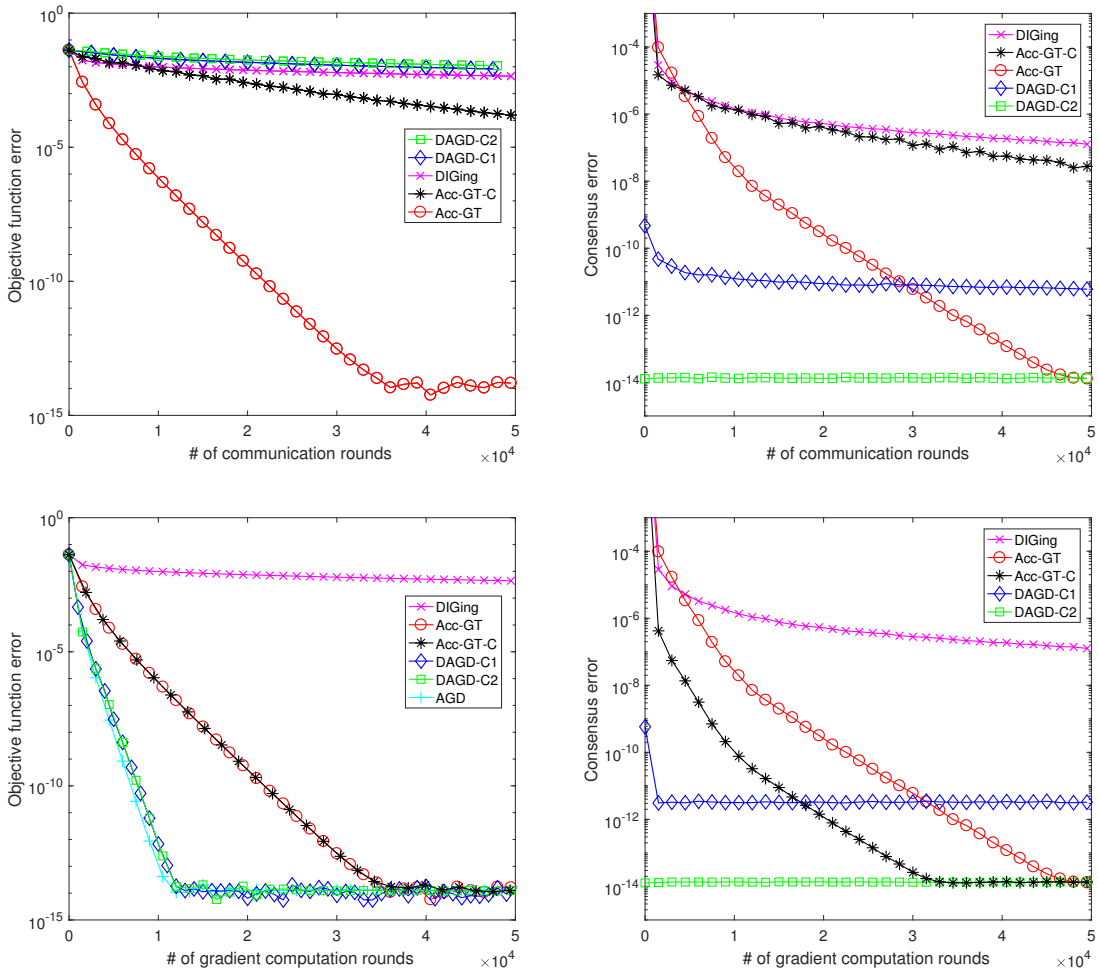

Figure 1: Comparisons of the objective function errors (left) and consensus errors (right) with respect to the number of communication (top) and computation (bottom) rounds for strongly convex problem with $d = 2$.

Appendix A. Proof of Lemma 18

Proof From the μ-strong convexity and L-smoothness of $f_{(i)}$, we have

$$\begin{aligned} F(w) =& \frac{1}{m}\sum_{i=1}^{m} f_{(i)}(w) \\ \geq& \frac{1}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k) + \left\langle \nabla f_{(i)}(y_{(i)}^k), w - y_{(i)}^k \right\rangle + \frac{\mu}{2}\|w - y_{(i)}^k\|^2\right) \end{aligned}$$

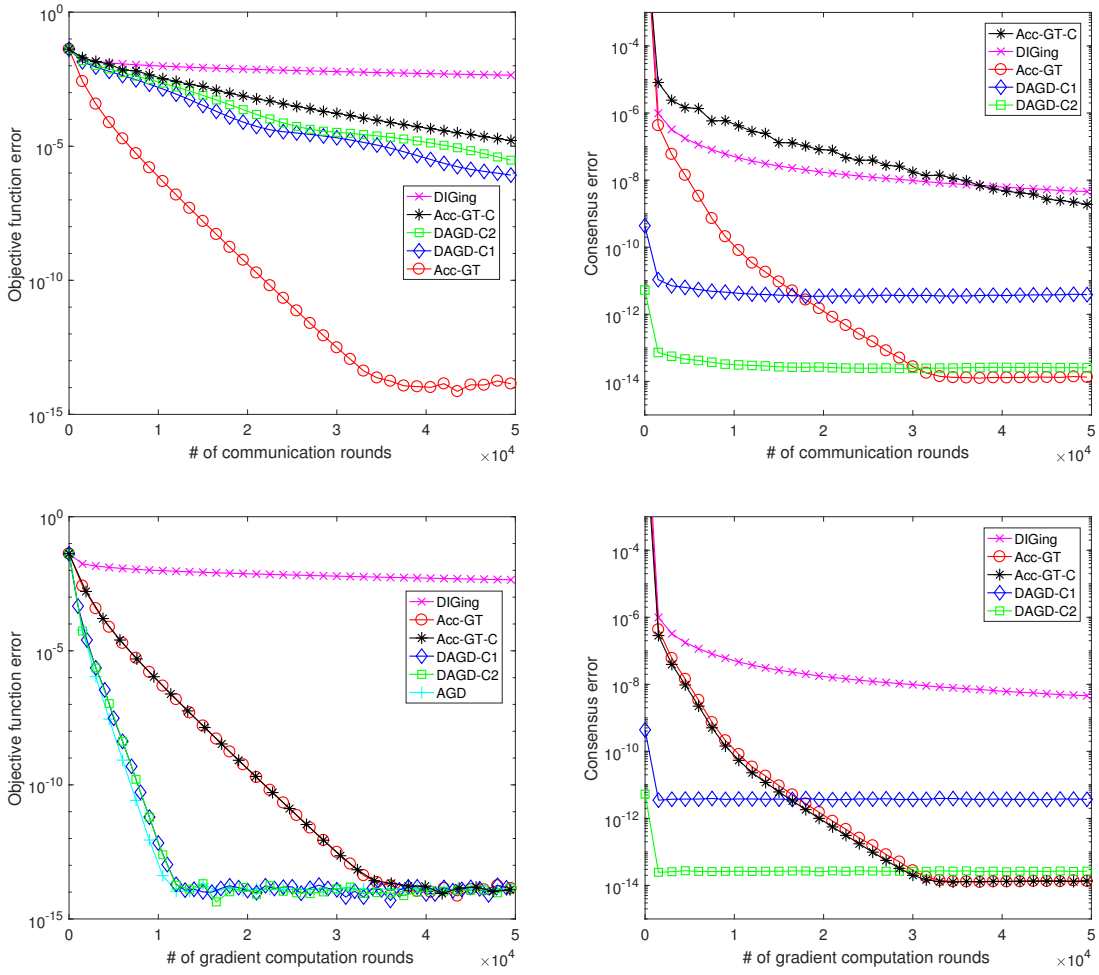

Figure 2: Comparisons of the objective function errors (left) and consensus errors (right) with respect to the number of communication (top) and computation (bottom) rounds for strongly convex problem with $d = 20$.

$$
\begin{aligned}
=&\frac{1}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k)+\left\langle\nabla f_{(i)}(y_{(i)}^k),w-y_{(i)}^k\right\rangle+\frac{\mu}{2}\|w-\overline{y}^k\|^2\right.\\
&\left.+\frac{\mu}{2}\|\overline{y}^k-y_{(i)}^k\|^2+\mu\left\langle w-\overline{y}^k,\overline{y}^k-y_{(i)}^k\right\rangle\right)\\
\overset{a}{=}&\frac{1}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k)+\left\langle\nabla f_{(i)}(y_{(i)}^k),w-y_{(i)}^k\right\rangle+\frac{\mu}{2}\|w-\overline{y}^k\|^2+\frac{\mu}{2}\|\overline{y}^k-y_{(i)}^k\|^2\right)\\
\geq&\frac{1}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k)+\left\langle\nabla f_{(i)}(y_{(i)}^k),w-y_{(i)}^k\right\rangle+\frac{\mu}{2}\|w-\overline{y}^k\|^2\right)\\
\overset{b}{=}&f(\overline{y}^k,\mathbf{y}^k)+\left\langle\overline{s}^k,w-\overline{y}^k\right\rangle+\frac{\mu}{2}\|w-\overline{y}^k\|^2,
\end{aligned}
$$

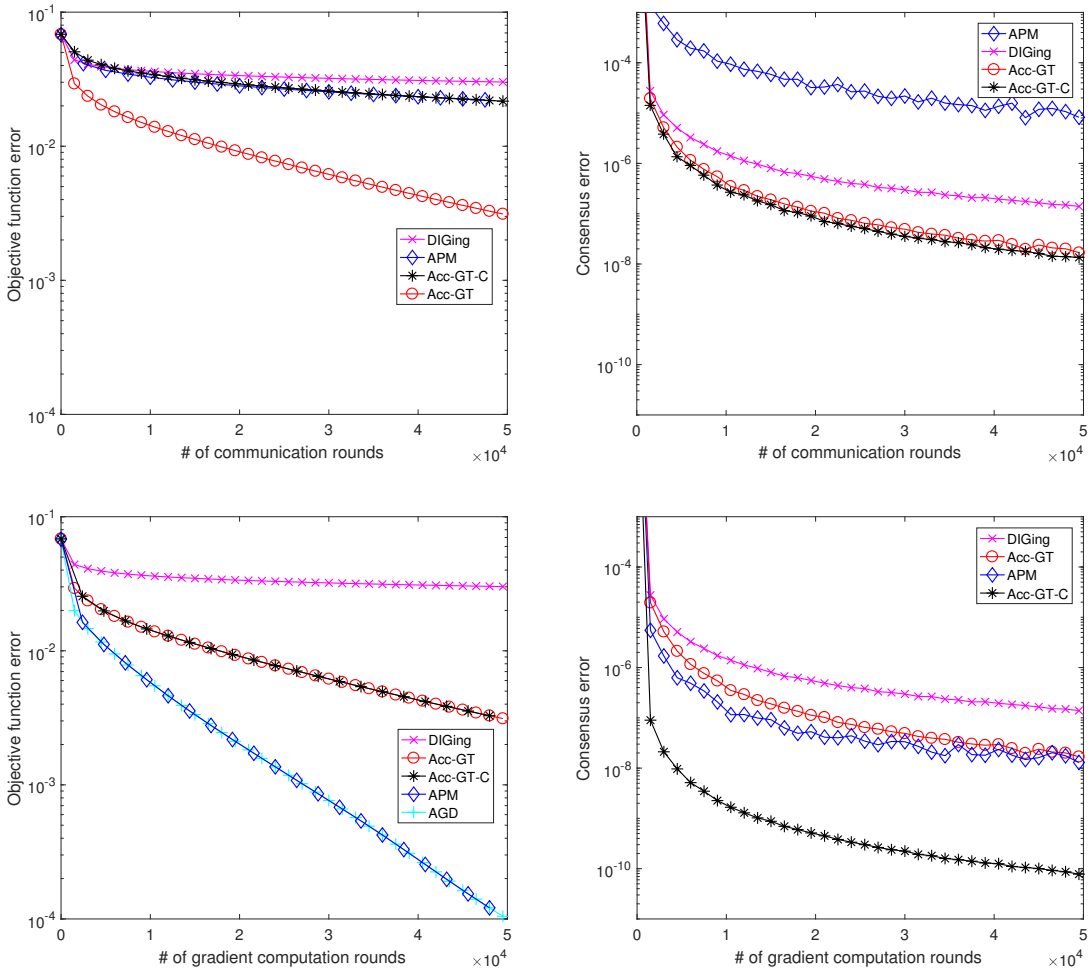

Figure 3: Comparisons of the objective function errors (left) and consensus errors (right) with respect to the number of communication (top) and computation (bottom) rounds for nonstrongly convex problem with $d = 2$.

and

$$\begin{aligned}
F(w) \leq & \frac{1}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k)+\left\langle \nabla f_{(i)}(y_{(i)}^k), w-y_{(i)}^k\right\rangle+\frac{L}{2}\|w-y_{(i)}^k\|^2\right)\\
= & \frac{1}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k)+\left\langle \nabla f_{(i)}(y_{(i)}^k), w-y_{(i)}^k\right\rangle+\frac{L}{2}\|w-\overline{y}^k\|^2\right.\\
& \left.+\frac{L}{2}\|\overline{y}^k-y_{(i)}^k\|^2+L\left\langle w-\overline{y}^k, \overline{y}^k-y_{(i)}^k\right\rangle\right)\\
\overset{c}{=} & \frac{1}{m}\sum_{i=1}^{m}\left(f_{(i)}(y_{(i)}^k)+\left\langle \nabla f_{(i)}(y_{(i)}^k), w-y_{(i)}^k\right\rangle+\frac{L}{2}\|w-\overline{y}^k\|^2+\frac{L}{2}\|\overline{y}^k-y_{(i)}^k\|^2\right)\\
\overset{d}{=} & f(\overline{y}^k,\mathbf{y}^k)+\left\langle \overline{s}^k, w-\overline{y}^k\right\rangle+\frac{L}{2}\|w-\overline{y}^k\|^2+\frac{L}{2m}\|\Pi\mathbf{y}^k\|^2,
\end{aligned}$$

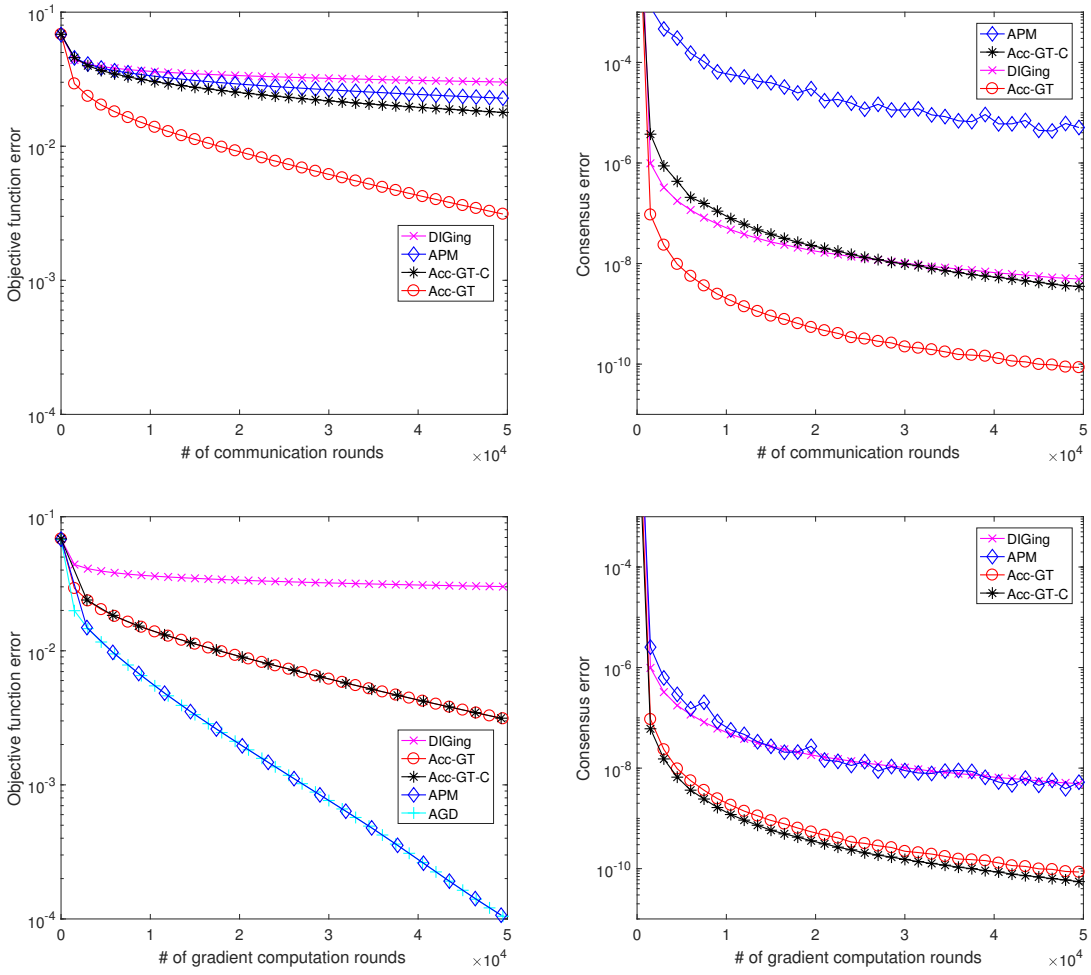

Figure 4: Comparisons of the objective function errors (left) and consensus errors (right) with respect to the number of communication (top) and computation (bottom) rounds for nonstrongly convex problem with $d = 20$.

where $\overset{a}{=}$ and $\overset{c}{=}$ use the definition of $\overline{y}^k$ in (3), $\overset{b}{=}$ and $\overset{d}{=}$ use the definition of $f(\overline{y}^k, \mathbf{y}^k)$ in (17), (16), and the definition of $\Pi\mathbf{y}$ in (4). ■

Acknowledgments

The authors were supported by National Key R&D Program of China (2022ZD0160300), NSF China (grant no.s 62476142, 62006116, 62276004), and Qualcomm.

References

Sulaiman A. Alghunaim, Ernest K. Ryu, Kun Yuan, and Ali H.Sayed. Decentralized proximal gradient algorithms with linear covnergence rates. *IEEE Transactions on Automatic Control*, 66(6):2787–2794, 2021.

Wicak Ananduta, Carlos Ocampo-Martinez, and Angelia Nedić. Accelerated multi-agent optimization method over stochastic networks. In *IEEE Conference on Decision and Control (CDC)*, pages 14–18, 2020.

Mario Arioli and Jennifer Scott. Chebyshev acceleration of iterative refinement. *Numerical Algorithms*, 66(3):591–608, 2014.

Yossi Arjevani, Joan Bruna, Bugra Can, Mert Gürbüzbalaban, Stefanie Jegelka, and Hongzhou Lin. Ideal: Inexact decentralized accelerated augmented lagrangian method. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 20648–20659, 2020.

Winfried Auzinger and Jens Markus Melenk. Iterative solution of large linear systems. *Lecture notes, TU Wien*, 2017.

Dimitri P. Bertsekas. Distributed asynchromous computation of fixed points. *Mathmatical Programming*, 27:107–120, 1983.

Keith Bonawitz, Hubert Eichner, and et al. Towards federated learning at scale: System design. In *Conference on Machine Learning and Systems (MLSys)*, 2019.

Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo. Communication-efficient variance-reduced decentralized stochastic optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 67(12):6583–6594, 2022.

Olivier Devolder, Francois Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146:37–75, 2014.

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

Darina Dvinskikh and Alexander Gasnikov. Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems. *Journal of Inverse and Ill-posed Problems*, 29(3):385–405, 2021.

Alireza Fallah, Mert Gurbuzbalaban, Asuman Ozdaglar, Umut Simsekli, and Lingjiong Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *Journal of Machine Learning Research*, 23(220):1–96, 2022.

Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. An optimal algorithm for decentralized finite sum optimization. *SIAM Journal on Optimization*, 31(4):2753–2783, 2021.

Mingyi Hong and Tsung-Hui Chang. Stochastic proximal gradient consensus over random networks. *IEEE Transactions on Signal Processing*, 65(11):2933–2948, 2017.

Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning (ICML)*, pages 1529–1538, 2017.

Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Explicit convergence rate of a distributed alternating direction method of multipliers. *IEEE Transactions on Automatic Control*, 61(4):892–904, 2016.

Dusan Jakovetić. A unification and generatliztion of exact distributed first order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2019.

Dusan Jakovetić, Joao Xavier, and José M. F. Moura. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146, 2014a.

Dusan Jakovetić, Joao Xavier, and Jose M. F. Moura. Convergence rates of distributed nesterov-like gradient methods on random networks. *IEEE Transactions on Signal Processing*, 62(4):868–882, 2014b.

Peter Kairouz, H. Brendan McMahan, and et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14:1–210, 2021.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, pages 5381–5393, 2020.

Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18342–18352, 2020.

Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 22325–22335, 2021a.

Dmitry Kovalev, Egor Shulgin, Peter Richtarik, Alexander Rogozin, and Alexander Gasnikov. ADOM: accelerated decentralized optimization method for time-varying networks. In *International Conference on Machine Learning (ICML)*, pages 5784–5793, 2021b.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180:237–284, 2020.

Huan Li and Zhouchen Lin. Revisiting EXTRA for smooth distributed optimization. *SIAM Journal Optimization*, 30(3):1795–1821, 2020.

Huan Li, Cong Fang, Wotao Yin, and Zhouchen Lin. Decentralized accelerated gradient methods with increasing penalty parameters. *IEEE transactions on Signal Processing*, 68: 4855–4870, 2020a.

Huan Li, Zhouchen Lin, and Yongchun Fang. Variance reduced EXTRA and DIGing and their optimal acceleration for strongly convex decentralized optimization. *Journal of Machine Learning Research*, 23(222):1–41, 2022.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60, 2020b.

Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17):4494–4506, 2019.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5330–5340, 2017.

Zhouchen Lin, Huan Li, and Cong Fang. *Accelerated Optimization in Machine Learning: First-Order Algorithms*. Springer, 2020.

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

Ali Makhdoumi and Asuman Ozdaglar. Convergence rate of distributed ADMM over networks. *IEEE Transactions on Automatic Control*, 62(10):5082–5095, 2017.

Marie Maros and Joakim Jalden. PANDA: A dual linearly converging method for distributed optimization over time-varying undirected graphs. In *IEEE Conference on Decision and Control (CDC)*, pages 6520–6525, 2018.

Marie Maros and Joakim Jalden. Eco-panda: A computationally economic, geometrically converging dual optimization method on time-varying undirected graphs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5257–5261, 2019.

Agelia Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Angelia Nedić. Asynchronous broadcast-based convex optimization over a network. *IEEE Transactions on Automatic Control*, 56(6):1337–1351, 2011.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.

Angelia Nedić and Alex Olshevsky. Stochastic gradient-push for strongly convexfunctions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, 2016.

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

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic, Boston, 2004.

Duong Thuy Anh Nguyen, Duong Tung Nguyen, and Angelia Nedić. Accelerated ab/push-pull methods for distributed optimization over time-varying directed networks. *IEEE Transactions on Control of Network Systems. DOI: 10.1109/TCNS.2023.3338236*, 2024.

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

Guannan Qu and Na Li. Accelerated distributed Nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2020.

S. Sundhar Ram, Angelia Nedić, and Venugopal V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of Optimization Theory and Applications*, 147:516–545, 2010.

Alexander Rogozin, César A. Uribe, Alexander V. Gasnikov, Nikolay Malkovsky, and Angelia Nedić. Optimal distributed convex optimization on slowly time-varying graphs. *IEEE Transactions on Control of Network Systems*, 7(2):829–841, 2020.

Alexander Rogozin, Mikhail Bochko, Pavel Dvurechensky, Alexander Gasnikov, and Vladislav Lukoshkin. An accelerated method for decentralized distributed stochastic optimization over time-varying graphs. In *IEEE Conference on Decision and Control (CDC)*, pages 3367–3373, 2021a.

Alexander Rogozin, Vladislav Lukoshkin, Alexander Gasnikov, Dmitry Kovalev, and Egor Shulgin. Towards accelerated rates for distributed optimization over time-varying networks. In *OPTIMA 2021: Optimization and Applications*, pages 258–272, 2021b.

Fakhteh Saadatniaki, Ran Xin, and Usman A. Khan. Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices. *IEEE Transactions on Automatic Control*, 65(11):4769–4780, 2020.

Kevin Scaman, Francis Bach, Sebastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning (ICML)*, pages 3027–3036, 2017.

Kevin Scaman, Francis Bach, Sebastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2740–2749, 2018.

Kevin Scaman, Francis Bach, Sebastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

Gesualdo Scutari and Ying Sun. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176:497–544, 2019.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing*, 63(23):6013–6023, 2015a.

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

Artin Spiridonoff, Alex Olshevsky, and Ioannis Ch. Paschalidis. Robust asynchronous stochastic gradient push: Asymptotically optimal and network independent performance for strongly convex functions. *Journal of Machine Learning Research*, 21:1–47, 2020.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

Ying Sun, Gesualdo Scutari, and Amir Daneshmand. Distributed optimization based on gradient-tracking revisited: Enhancing convergence rate via surrogation. *SIAM Journal Optimization*, 32(2):354–385, 2022.

Håkan Terelius, Ufuk Topcu, and Richard M. Murray. Decentralized multi-agent optimization via dual decomposition. *IFAC proceedings volumes*, 44(1):11245–11251, 2011.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, Seattle, 2008.

John N. Tsitsiklis, Dimitri P. Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transaction on Automatic Control*, 31(9):803–812, 1986.

César A. Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. *Optimization Methods and Software*, 36(1):171–210, 2021.

Ermin Wei and Asuman Ozdaglar. On the $o(1/k)$ convergence of asynchronous distributed alternating direction method of multipliers. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 551–554, 2013.

Ran Xin, Usman A. Khan, and Soummya Kar. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.

Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *IEEE Conference on Decision and Control (CDC)*, pages 2055–2060, 2015.

Jinming Xu, Ye Tian, Ying Sun, and Gesualdo Scutari. Accelerated primal-dual algorithms for distributed smooth convex optimization over networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2381–2391, 2020.

Haishan Ye, Ziang Zhou, Luo Luo, and Tong Zhang. Decentralized accelerated proximal gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 18308–18317, 2020.

Haishan Ye, Luo Luo, Ziang Zhou, and Tong Zhang. Multi-consensus decentralized accelerated gradient descent. *Journal of Machine Learning Research*, 24(306):1–50, 2023.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal Optimization*, 26(3):1835–1854, 2016.