Accelerated Gradient Tracking over Time-varying Graphs for Decentralized Optimization

Huan Li [1] Zhouchen Lin [2]

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 the $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^2\sqrt{\frac{L}{\epsilon}})$ and $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ complexities for the practical single loop accelerated gradient tracking 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, ϵ is the desired precision, and L and μ are the smoothness and strong convexity constants, respectively. Our complexities 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 only for static graphs, where $\frac{1}{1-\sigma}$ equals $\frac{\gamma}{1-\sigma_\gamma}$ 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 computation and communication 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.

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 at each time the network may be even unconnected. A typical example is the federated learning (Li et al., 2020c; Kairouz et al., 2019), 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

[1] Institute of Robotics and Automatic Information Systems, College of Artificial Intelligence, Nankai University, Tianjin, China (lihuanss@nankai.edu.cn).
[2] Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University, Beijing, China (zlin@pku.edu.cn)..

$\mathcal{V} = \{1, ..., m\}$ 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 communication rounds and gradient or stochastic gradient computations for strongly convex and smooth problems are well-known (Scaman et al., 2017; 2019; Hendrikx et al., 2020), and optimal accelerated algorithms with upper bounds exactly matching the lower bounds are developed (Koralev et al., 2020; Li et al., 2020b), for the time-varying graphs, the literature is much scarcer. It remains an open problem on 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. Literature Review

In this section, we briefly review the decentralized algorithms over static graphs and time-varying graphs, mainly focusing on the accelerated methods. Tables 1 and 2 sum up the complexity comparisons of the state-of-the-art methods.

1.1.1. 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 & 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 & 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 & Ozdaglar, 2013; Iutzeler et al., 2016; Makhdoumi & Ozdaglar, 2017), and the primal-dual method (Lan et al., 2020; Scaman et al., 2018; Hong et al., 2017; Jakovetić, 2019). 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 & Li, 2020), the robust distributed accelerated stochastic gradient method (Fallah et al., 2019), the multi-step dual accelerated method (Scaman et al., 2017; 2019), accelerated penalty method (APM) (Li et al., 2020a; Dvinskikh & Gasnikov, 2019), the multi-consensus decentralized accelerated gradient descent (Mudag) (Ye et al., 2020a;b), accelerated EXTRA (Li & Lin, 2020; Li et al., 2020b), the decentralized accelerated augmented Lagrangian method (Arjevani et al., 2020), and the accelerated proximal alternating predictor-corrector method (APAPC) (Koralev et al., 2020). Scaman et al. (2017; 2019) proved the $\Omega\left(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon}\right)$ (see the notations in Section 1.3) communication complexity lower bound and the $\Omega\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$ gradient computation complexity lower bound. To the best of our knowledge, only APAPC combined with the Chebyshev acceleration (CA) (Arioli & Scott, 2014) exactly achieves 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 $\mathcal{O}\left(\left(\frac{L}{\mu}\right)^{5/7}\frac{1}{(1-\sigma)^{1.5}}\log\frac{1}{\epsilon}\right)$ complexity, originally proved in (Qu & 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 & Li, 2020), APM (Li et al., 2020a; Dvinskikh & Gasnikov, 2019), accelerated EXTRA (Li & 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 $\Omega\left(\sqrt{\frac{L}{\epsilon(1-\sigma)}}\right)$ communication complexity lower bound and the $\Omega\left(\sqrt{\frac{L}{\epsilon}}\right)$ gradient computation 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 $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$ factor in the communication complexity. Acc-DNGD only has the $\mathcal{O}\left(\frac{1}{\epsilon^{5/7}}\right)$ complexity[1], originally proved in (Qu & Li, 2020). Xu et al. (2020) proposed an accelerated primal dual method, however, their complexity remains $\mathcal{O}\left(\frac{1}{\epsilon}\right)$.

---

[1]The dependence on $1 - \sigma$, a small constant charactering the network connectivity, was not explicitly given in (Qu & Li, 2020).

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.

Methods	gradient computation complexity	communication complexity	single or double loop
Nonstrongly convex and smooth functions			
Gradient tracking (Qu & 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 & 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 & Gasnikov, 2019)	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon}}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\epsilon(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	double
Acc-EXTRA (Li & 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., 2020)	$\mathcal{O}\left(\left(\frac{L}{\mu}+\frac{1}{(1-\sigma)^2}\right)\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\left(\frac{L}{\mu}+\frac{1}{(1-\sigma)^2}\right)\log\frac{1}{\epsilon}\right)$	single
Acc-DNGD (Qu & 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 (Koralev 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)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu(1-\sigma)}}\log\frac{1}{\epsilon}\right)$	\

1.1.2. 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, which means that the joint graph $\{\mathcal{V}, \cup_{r=k}^{k+\gamma-1}\mathcal{E}^r\}$ is connected for all $k = 0, 1, ...$ In the second scenario, the network is assumed to be connected at every time. Some researchers formulate the time-varying communication graphs as random graphs, as in (Hong & Chang, 2017; Jakovetić et al., 2014b; Ananduta et al., 2020). It is beyond the discussion of this paper.

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 & 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 & 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 & Scutari, 2016) and SONATA (Scutari & 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., 2020a; 2021) is the only accelerated method for strongly convex and smooth objectives with explicit complexities in this general time-varying setting. However, the communication 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

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 computation complexity	communication complexity	single or double loop
Nonstrongly convex and smooth functions			
DGD[2] (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., 2020a)	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\left(\log\frac{1}{\epsilon}\right)^2\right)$	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

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 5.

Connected at Every Time. In this scenario, the literatures are rich and many distributed methods originally designed over static graphs, such as Acc-DNGD, can be directly used. Kovalev et al. (2021) proposed a dual based method named ADOM with the state-of-the-art communication complexity of $\mathcal{O}(\frac{1}{1-\sigma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ for strongly convex and smooth problems. However, ADOM needs to compute the gradient of Fenchel conjugate of the local objectives, which is expensive, and ADOM cannot be used for nonstrongly convex problems. Rogozin et al. (2020b) 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 a tiny percentage of the number of iterations.

1.2. Contributions

In this paper, we revisit the accelerated gradient tracking originally proposed in (Qu & Li, 2020) and extend it to the time-varying graphs with sharper complexities. We give our analysis over static graphs and time-varying graphs in a unified framework. The former scenario provides the basis and insights for the latter one. 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 $\mathcal{O}(\frac{\gamma^2}{(1-\sigma_\gamma)^2}\sqrt{\frac{L}{\epsilon}})$ complexity 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 communications and $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ for gradient

---

[2]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, $\bar{\zeta} = \frac{1}{m}\sum_{i=1}^{m}\|\nabla f_{(i)}(x^*)\|^2$.

computations. Our communication cost is lower than that of the state-of-the-art APM (Li et al., 2020a) by a $\mathcal{O}(\log \frac{1}{\epsilon})$ factor, while our gradient computation cost is the same as that of APM.

(b) When the local objective functions are strongly convex and smooth, we prove the $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^{1.5} \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon})$ complexity for the practical single loop accelerated gradient tracking. When combing with the multiple consensus subroutine, we can improve the complexities to $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma} \sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon})$ for communications and $\mathcal{O}(\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon})$ for gradient computations. Our communication cost is lower than that of the state-of-the-art DAGD-C (Rogozin et al., 2020a) by a $\mathcal{O}(\log \frac{1}{\epsilon})$ factor, while our gradient computation cost remains the same as that of DAGD-C.

(c) To the best of our knowledge, this is the first time that the upper bounds with the optimal dependence on the precision ϵ and condition number L/μ are given for both nonstrongly convex and strongly convex problems. Making things more important, 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 $\mathcal{O}(\frac{1}{(1-\sigma)^2} \sqrt{\frac{L}{\epsilon}})$ complexity for the practical single loop accelerated gradient tracking, which significantly improves over the existing $\mathcal{O}(\frac{1}{\epsilon^{5/7}})$ one originally proved in (Qu & Li, 2020). When combing with the Chebyshev acceleration, we can improve the complexities to $\mathcal{O}(\sqrt{\frac{L}{\epsilon(1-\sigma)}})$ for communications and $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ for gradient computations, which exactly match the communication and gradient computation complexity lower bounds, respectively. 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 $\mathcal{O}(\sqrt{\frac{L}{\mu(1-\sigma)^3}} \log \frac{1}{\epsilon})$ complexity for the practical single loop accelerated gradient tracking, which improves over the existing $\mathcal{O}((\frac{L}{\mu})^{5/7} \frac{1}{(1-\sigma)^{1.5}} \log \frac{1}{\epsilon})$ one originally given in (Qu & 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.

1.3. 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

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

Denote $\mathbf{x}^k$ to be the value at iteration k. For scalers, for example, θ, we use θ_k instead of θ^k to denote the value at iteration k, while the latter one represents its kth power. Specially, x^T means the transpose of x. To avoid confusion, this article does not use the value at iteration T. 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. Denote x^* as the optimal solution of problem (1). Define the average variable across all the local variables at iteration k as

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

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

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

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):

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

When the network is static, we make the following standard assumptions for the weight matrix $W \in \mathbb{R}^{m \times m}$ associated to the network:

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 $\mathbf{1}^T W = \mathbf{1}^T$.*

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, $\sigma = \|W - \frac{1}{m}\mathbf{1}\mathbf{1}^T\|_2 < 1$. Moreover, we have the following classical consensus contraction:

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

When the network is time-varying, each graph instance $\mathcal{G}^k$ associates with a weight matrix W^k. We follow (Nedić et al., 2017) to 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 make the following standard assumptions for the sequence of weight 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 $\mathbf{1}^T W^k = \mathbf{1}^T$.*

3. *(Joint spectrum property) There exists a constant integer γ such that*

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

Assumption 3 is weaker than the assumption that every graph $\mathcal{G}^k$ is connected. A typical example of the weight 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 undirected graph $\{\mathcal{V}, \cup_{r=k}^{k+\gamma-1}\mathcal{E}^r\}$ is connected for all $k = 0, 1, \dots$.*

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

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

2. Accelerated Gradient Tracking over Time-varying Graphs

We first review the gradient tracking and its accelerated variant, where the latter one 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 & 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:

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

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

To further accelerate gradient tracking, Qu & 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 & 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

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

$$\mathbf{y}^k = \theta_k \mathbf{z}^k + (1 - \theta_k)\mathbf{x}^k, \tag{13a}$$

$$\mathbf{s}^k = W_1^k \mathbf{s}^{k-1} + \nabla f(\mathbf{y}^k) - \nabla f(\mathbf{y}^{k-1}), \tag{13b}$$

$$\mathbf{z}^{k+1} = \frac{1}{1 + \frac{\mu\alpha}{\theta_k}} \left(W_2^k \left(\frac{\mu\alpha}{\theta_k} \mathbf{y}^k + \mathbf{z}^k \right) - \frac{\alpha}{\theta_k} \mathbf{s}^k \right), \tag{13c}$$

$$\mathbf{x}^{k+1} = \theta_k \mathbf{z}^{k+1} + (1 - \theta_k)W_3^k \mathbf{x}^k, \tag{13d}$$

where we initialize $\mathbf{x}^0 = \mathbf{y}^0 = \mathbf{z}^0$ with $\Pi\mathbf{x}^0 = 0$, $\mathbf{s}^0 = \nabla f(\mathbf{y}^0)$, $\mathbf{z}^1 = W_2^0 \mathbf{z}^0 - \frac{\alpha}{\theta_0 + \mu\alpha}\mathbf{s}^0$, and $\mathbf{x}^1 = \theta_0 \mathbf{z}^1 + (1 - \theta_0)W_3^0 \mathbf{x}^0$. Variables $s_{(i)}^{k-1}$, $x_{(i)}^k$, $y_{(i)}^k$, and $z_{(i)}^k$ can be transmitted together at each iteration. We allow W_1^k, W_2^k, and W_3^k to be different when the variables are transmitted separately. 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 weight matrix. We see that algorithm (13a)-(13d) is equivalent to (12a)-(12d) when the weight matrices are fixed and $\mu = 0$. However, when $\mu > 0$, it is not equivalent to the method proposed in (Qu & 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 & Li, 2020) to rewrite the distributed algorithm in the form of inexact accelerated gradient descent. However, we use a different proof framework from (Qu & Li, 2020) with much simpler proofs, and give sharper complexities. See Remark 7 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 7 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 1 *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 $\alpha \leq \frac{(1-\sigma_\gamma)^4}{21675L\gamma^4}$. Then for algorithm (13a)-(13d), we have*

$$F(\overline{x}^{T\gamma+1}) - F(x^*) \leq \frac{1}{(T\gamma+1)^2} \left(\frac{2}{\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{1-\sigma_\gamma}{5mL\gamma} \max_{r=0,\ldots,\gamma} \sum_{i=1}^m \|s_{(i)}^r - \overline{s}^r\|^2 \right),$$

and

$$\frac{1}{m}\|\Pi\mathbf{x}^{T\gamma}\|^2 \leq \frac{9}{2L(T\gamma+1)^2} \left(\frac{2}{\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{1-\sigma_\gamma}{5mL\gamma} \max_{r=0,\ldots,\gamma} \sum_{i=1}^m \|s_{(i)}^r - \overline{s}^r\|^2 \right).$$

Theorem 2 *Suppose that Assumptions 1 and 3 hold with $\mu > 0$. 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*

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

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 1 that algorithm (13a)-(13d) needs $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^2 \sqrt{\frac{LR}{\epsilon}})$ rounds of communications and gradient computations to find an ϵ-optimal averaged solution, where $R = \mathcal{O}(\|\bar{z}^0 - x^*\|^2 + \frac{(1-\sigma_\gamma)^5}{mL^2\gamma^5} \max_{r=0,...,\gamma} \sum_{i=1}^m \|s_{(i)}^r - \bar{s}^r\|^2)$. When strong convexity is assumed, we see from Theorem 2 that both the communication and gradient computation 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 compared in Table 2, our complexities improve over the state-of-the-art APM (Li et al., 2020a) and DAGD-C (Rogozin et al., 2020a) on the dependence of ϵ since they have an additional $\mathcal{O}(\log \frac{1}{\epsilon})$ factor in their communication complexities. However, our dependence on $\frac{\gamma}{1-\sigma_\gamma}$ is not state-of-the-art. We will improve it in Section 2.4.

Remark 1 *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* $\mathcal{O}((\frac{\gamma}{1-\sigma_\gamma})^2 \sqrt{\frac{LR}{\epsilon}}) + \mathcal{O}(\frac{\gamma}{1-\sigma_\gamma} \log \frac{1}{\epsilon})$ *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 2 *For nonstrongly convex problems, we can also prove the convergence rate measured at the point* $x_{(i)}^{T\gamma+1}$ *for any* i:

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

However, the complexity increases 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 complexity keeps the same no matter measured at* $\bar{x}^{T\gamma+1}$ *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 3 *In Theorems 1 and 2, we measure the convergence rates at the* $(T\gamma+1)$th *iteration for simplicity. For any* $K = T\gamma + r$, *a straightforward modification is to regard the* $(r-1)$th *iteration as the virtual initialization. However, we should modify the proof of Theorem 1 accordingly with* $\theta_0 < 1$. *We omit the details since it complicates the proofs and notations.*

Remark 4 *Due to the physical constraints such as the battery dies, the device shuts down, or the WiFi network is unavailable, the agents may drop out for a period of time. We can formulate this case by the local updates. 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 *and iteration* $k = t+1, t+2, ..., t'$:

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

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

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

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

which are a serious of local updates without communications. Motivated by the local SGD (Stich, 2019; Koloskova et al., 2020), which is widely used in federated learning, we require agent i *to make up the delayed computations by local updates before joining the network again, that is, performing the local updates (14a)-(14d) for* $t' - t$ *iterations. Since (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), it will not take long to keep pace with the other agents.*

2.3. Special Cases over Static Graphs

When we fix $W_1^k = W_2^k = W_3^k = W$, algorithm (13a)-(13d) can be applied to static graphs. As a special case of Theorems 1 and 2, we have the following theorems over static graphs.

Theorem 3 *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 $\alpha \leq \frac{(1-\sigma)^4}{537L}$. Then for algorithm (13a)-(13d) with fixed weight matrix W, we have*

$$F(\overline{x}^{K+1}) - F(x^*) \leq \frac{1}{(K+1)^2}\left(\frac{2}{\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{1-\sigma}{2L}\frac{1}{m}\sum_{i=1}^m \|s_{(i)}^0 - \overline{s}^0\|^2\right),$$

and

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

Theorem 4 *Suppose that Assumptions 1 and 2 hold with $\mu > 0$. Let $\alpha \leq \frac{(1-\sigma)^3}{119L}$ and $\theta_k \equiv \theta = \frac{\sqrt{\mu\alpha}}{2}$. Then for algorithm (13a)-(13d) with fixed weight matrix W, we have*

$$F(\overline{x}^{K+1}) - F(x^*) + \left(\frac{\theta2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^{K+1} - x^*\|^2 \leq (1-\theta)^{K+1}C,$$

and

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

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

The above theorems give the $\mathcal{O}(\frac{1}{(1-\sigma)^2}\sqrt{\frac{L}{\epsilon}})$ and $\mathcal{O}(\sqrt{\frac{L}{\mu(1-\sigma)^3}}\log\frac{1}{\epsilon})$ complexities for nonstrongly convex and strongly convex problems, respectively. As compared in Table 1, our complexities 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 & Li, 2020).

2.4. Improve the 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 becomes 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 $k \geq 1$. Define $L = I - W$ with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{m-1} > \lambda_m = 0$ being its eigenvalues (assume that W is symmetric). 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, $P_t(c_3L)$ is a symmetric gossip matrix satisfying $P_t(c_3L)\mathbf{1} = 0$ 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 & 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 weight matrix W in algorithm (13a)-(13d) by $I - P_t(c_3L)$ because its second largest singular value σ' satisfies $\sigma' \leq 0.65$, which is independent of $1 - \sigma$. From Theorems 3 and 4 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 problems to find an ϵ-optimal solution, respectively, which corresponds to the gradient computation complexity. On the other hand, we can compute the operation $(I - P_t(c_3L))\mathbf{x}$ by the following procedure (Scaman et al., 2017):

Input: $\mathbf{x}$,

Initialize: $a_0 = 1$, $a_1 = c_2$, $\mathbf{z}^0 = \mathbf{x}$, $\mathbf{z}^1 = c_2(I - c_3 L)\mathbf{x}$,

for $s = 1, 2, ..., t - 1$ **do**

    $a_{s+1} = 2c_2 a_s - a_{s-1}$,

    $\mathbf{z}^{s+1} = 2c_2(I - c_3 L)\mathbf{z}^s - \mathbf{z}^{s-1}$.

end for

Output: $(I - P_t(c_3 L))\mathbf{x} = \frac{\mathbf{z}^t}{a_t}$.

Thus, the communication 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 1 *Under the settings of Theorem 3 with symmetric and fixed weight matrix W, algorithm (13a)-(13d) with Chebyshev acceleration requires the time of $\mathcal{O}(\sqrt{\frac{L}{\epsilon(1-\sigma)}})$ communication rounds and $\mathcal{O}(\sqrt{\frac{L}{\epsilon}})$ gradient computations to find an ϵ-optimal averaged solution such that $F(\overline{x}) - F(x^*) \leq \epsilon$.*

Corollary 2 *Under the settings of Theorem 4 with symmetric and fixed weight matrix W, algorithm (13a)-(13d) with Chebyshev acceleration requires the 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 computations 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_1^k, W_2^k, and W_3^k in (13a)-(13d) by three virtual weight matrices $W_1^{k,\zeta}$, $W_2^{k,\zeta}$, and $W_3^{k,\zeta}$ with carefully designed ζ such that

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

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 weight matrices in Assumptions 2 and 3, thus our theorems apply to the virtual weight matrices $W_r^{k,\zeta}$. From Theorems 1 and 2 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 computation complexity. Next, we consider the communication complexity. Letting $\zeta = \lceil \frac{\gamma}{1-\sigma_\gamma} \rceil$, it follows from (9) that

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

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

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

with ζ rounds of communications initialized at $\mathbf{u}^0 = \mathbf{x}$, the communication 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 problems, respectively.

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

Corollary 4 *Under the settings of Theorem 2, algorithm (13a)-(13d) combined with the multiple consensus subroutine requires the 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 computations to find an ϵ-optimal averaged solution such that $F(\overline{x}) - F(x^*) \leq \epsilon$.*

Remark 5 *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 complexities. The similar issue also happens in APM (Li et al., 2020a) and DAGD-C (Rogozin et al., 2020a), 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., 2020). 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 6 *Unlike the scenario over static graphs, the communication complexity lower bounds over 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 $\mathcal{O}(\frac{\gamma}{1-\sigma_\gamma}\sqrt{\frac{L}{\mu}}\log\frac{1}{\epsilon})$ complexities can be further improved. We leave it as an open problem.*

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 one.

3.1. Convergence Rates of the Inexact Accelerated Gradient Descent

Following the proof framework in (Jakovetić et al., 2014a; Qu & Li, 2020), we multiply both sides of (13a)-(13d) by $\frac{1}{m}\mathbf{1}^T$ 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 $\mathbf{1}^T W^k = \mathbf{1}^T$. 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 & Li, 2020). For the completeness and the readers' convenience, we give the proof in the appendix.

Lemma 1 *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 distance 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 distance 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 & Li, 2020).

Lemma 2 *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*

$$\frac{F(\overline{x}^{K+1}) - F(x^*)}{\theta_K^2} + \frac{1}{2\alpha} \|\overline{z}^{K+1} - x^*\|^2 \leq \frac{1}{2\alpha} \|\overline{z}^0 - x^*\|^2$$

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

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

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

$$\leq F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2} \right) \|\overline{z}^0 - x^*\|^2 + \underbrace{\sum_{k=0}^K \frac{L}{2m(1-\theta)^{k+1}} \|\Pi \mathbf{y}^k\|^2}_{\text{term (c)}}$$

$$- \underbrace{\sum_{k=0}^K \left(\frac{1}{(1-\theta)^{k+1}} \left(\frac{\theta^2}{2\alpha} - \frac{L\theta^2}{2} \right) \|\overline{z}^{k+1} - \overline{z}^k\|^2 + \frac{1}{(1-\theta)^k} D_f(\overline{x}^k, \mathbf{y}^k) \right)}_{\text{term (d)}}. \tag{22}$$

Proof 1 *From the inexact smoothness (19), we have*

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

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

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

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

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

where we use (15c) in $\stackrel{d}{=}$. 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 3 *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 weight 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 2 *Multiplying both sides of (13a)-(13d) by Π, using (6) and $\|\Pi\mathbf{x}\| \leq \|\mathbf{x}\|$, we have*

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

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

$$\|\Pi\mathbf{z}^{k+1}\| \leq \sigma\left(\frac{\mu\alpha}{\theta_k + \mu\alpha}\|\Pi\mathbf{y}^k\| + \frac{\theta_k}{\theta_k + \mu\alpha}\|\Pi\mathbf{z}^k\|\right) + \frac{\alpha}{\theta_k + \mu\alpha}\|\Pi\mathbf{s}^k\|$$

$$\overset{a}{\leq} \frac{\sigma(\mu\alpha + 1)\theta_k}{\theta_k + \mu\alpha}\|\Pi\mathbf{z}^k\| + \frac{\sigma\mu\alpha}{\theta_k + \mu\alpha}\|\Pi\mathbf{x}^k\| + \frac{\alpha}{\theta_k + \mu\alpha}\|\Pi\mathbf{s}^k\| \tag{29}$$

$$\overset{b}{\leq} \sigma\|\Pi\mathbf{z}^k\| + \frac{\mu\alpha}{\theta_k}\|\Pi\mathbf{x}^k\| + \frac{\alpha}{\theta_k}\|\Pi\mathbf{s}^k\|,$$

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

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

$$\|\nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k)\|^2 = \sum_{i=1}^{m}\|\nabla f_{(i)}(y_{(i)}^{k+1}) - \nabla f_{(i)}(y_{(i)}^k)\|^2$$

$$\overset{c}{\leq} \sum_{i=1}^{m}(1+\tau)\|\nabla f_{(i)}(y_{(i)}^{k+1}) - \nabla f_{(i)}(\overline{x}^{k+1})\|^2$$

$$+ \sum_{i=1}^{m}\left(1+\frac{1}{\tau}\right)2\left(\|\nabla f_{(i)}(\overline{x}^{k+1}) - \nabla f_{(i)}(\overline{y}^k)\|^2 + \|\nabla f_{(i)}(\overline{y}^k) - \nabla f_{(i)}(y_{(i)}^k)\|^2\right)$$

$$\overset{d}{\leq} 2mL(1+\tau)D_f(\overline{x}^{k+1}, \mathbf{y}^{k+1}) + 2L^2\left(1+\frac{1}{\tau}\right)\sum_{i=1}^{m}\left(\|\overline{x}^{k+1} - \overline{y}^k\|^2 + \|\overline{y}^k - y_{(i)}^k\|^2\right) \tag{31}$$

$$\overset{e}{=} 2mL(1+\tau)D_f(\overline{x}^{k+1}, \mathbf{y}^{k+1}) + 2L^2\left(1+\frac{1}{\tau}\right)\left(m\theta_k^2\|\overline{z}^{k+1} - \overline{z}^k\|^2 + \|\Pi\mathbf{y}^k\|^2\right)$$

$$= \theta_k^2\Phi^k + 2L^2\left(1+\frac{1}{\tau}\right)\|\Pi\mathbf{y}^k\|^2$$

$$\overset{f}{\leq} \theta_k^2\Phi^k + 4L^2\left(1+\frac{1}{\tau}\right)\left(\theta_k^2\|\Pi\mathbf{z}^k\|^2 + \|\Pi\mathbf{x}^k\|^2\right),$$

where $\overset{c}{\leq}$ uses Young's inequality of $\|a-b\|^2 \leq (1+\tau)\|a\|^2 + (1+\frac{1}{\tau})\|b\|^2$ for any $\tau > 0$, $\overset{d}{\leq}$ 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), $\overset{f}{\leq}$ 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

$$\|\Pi\mathbf{s}^{k+1}\|^2 \leq \left(1 + \frac{1-\sigma}{2\sigma}\right)\sigma^2\|\Pi\mathbf{s}^k\|^2 + \left(1 + \frac{2\sigma}{1-\sigma}\right)\|\nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k)\|^2$$

$$= \frac{\sigma + \sigma^2}{2}\|\Pi\mathbf{s}^k\|^2 + \frac{1+\sigma}{1-\sigma}\|\nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k)\|^2 \tag{32}$$

$$\overset{g}{\leq} \frac{1+\sigma}{2}\|\Pi\mathbf{s}^k\|^2 + \frac{2}{1-\sigma}\left(\theta_k^2\Phi^k + c_0\theta_k^2\|\Pi\mathbf{z}^k\|^2 + c_0\|\Pi\mathbf{x}^k\|^2\right),$$

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

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

$$\|\Pi \mathbf{x}^{k+1}\|^2 \le \frac{1+\sigma}{2}\|\Pi \mathbf{x}^k\|^2 + \frac{2\theta_k^2}{1-\sigma}\|\Pi \mathbf{z}^{k+1}\|^2. \tag{34}$$

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

$$c_1\|\Pi \mathbf{s}^{k+1}\|^2 + c_2\theta_{k+1}^2\|\Pi \mathbf{z}^{k+1}\|^2 + c_3\|\Pi \mathbf{x}^{k+1}\|^2$$

$$\overset{h}{\le} c_1\|\Pi \mathbf{s}^{k+1}\|^2 + \left(c_2\theta_k^2 + \frac{2c_3\theta_k^2}{1-\sigma}\right)\|\Pi \mathbf{z}^{k+1}\|^2 + \frac{c_3(1+\sigma)}{2}\|\Pi \mathbf{x}^k\|^2$$

$$\le \left(\frac{c_1(1+\sigma)}{2} + \left(c_2 + \frac{2c_3}{1-\sigma}\right)\frac{4\alpha^2}{1-\sigma}\right)\|\Pi \mathbf{s}^k\|^2$$

$$+ \left(\frac{2c_0 c_1}{1-\sigma} + \left(c_2 + \frac{2c_3}{1-\sigma}\right)\frac{1+\sigma}{2}\right)\theta_k^2\|\Pi \mathbf{z}^k\|^2$$

$$+ \left(\frac{2c_0 c_1}{1-\sigma} + \left(c_2 + \frac{2c_3}{1-\sigma}\right)\frac{4\mu^2\alpha^2}{1-\sigma} + \frac{c_3(1+\sigma)}{2}\right)\|\Pi \mathbf{x}^k\|^2 + \frac{2c_1}{1-\sigma}\theta_k^2\Phi^k,$$

where we use $\theta_{k+1} \le \theta_k$ and (34) in $\overset{h}{\le}$. Letting $c_3 = \frac{9c_0 c_1}{(1-\sigma)^2}$, $c_2 = \frac{80c_0 c_1}{(1-\sigma)^4} \ge \frac{8c_0 c_1}{(1-\sigma)^2} + \frac{8c_3}{(1-\sigma)^2}$, and $\alpha^2 \le \min\left\{\frac{(1-\sigma)^6}{1600c_0}, \frac{(1-\sigma)^4}{1600\mu^2}\right\}$ such that

$$\frac{c_1(1+\sigma)}{2} + \left(c_2 + \frac{2c_3}{1-\sigma}\right)\frac{4\alpha^2}{1-\sigma} \le \frac{c_1(1+\sigma)}{2} + \frac{400c_0 c_1\alpha^2}{(1-\sigma)^5} \le \frac{c_1(3+\sigma)}{4},$$

$$\frac{2c_0 c_1}{1-\sigma} + \left(c_2 + \frac{2c_3}{1-\sigma}\right)\frac{1+\sigma}{2} \le \frac{2c_0 c_1}{1-\sigma} + \frac{c_2(1+\sigma)}{2} + \frac{2c_3}{1-\sigma} \le \frac{c_2(3+\sigma)}{4},$$

$$\frac{2c_0 c_1}{1-\sigma} + \left(c_2 + \frac{2c_3}{1-\sigma}\right)\frac{4\mu^2\alpha^2}{1-\sigma} + \frac{c_3(1+\sigma)}{2} \le \frac{2c_0 c_1}{1-\sigma} + \frac{400c_0 c_1\mu^2\alpha^2}{(1-\sigma)^5} + \frac{c_3(1+\sigma)}{2} \le \frac{c_3(3+\sigma)}{4},$$

we have

$$c_1\|\Pi \mathbf{s}^{k+1}\|^2 + c_2\theta_{k+1}^2\|\Pi \mathbf{z}^{k+1}\|^2 + c_3\|\Pi \mathbf{x}^{k+1}\|^2$$

$$\le \frac{3+\sigma}{4}\left(c_1\|\Pi \mathbf{s}^k\|^2 + c_2\theta_k^2\|\Pi \mathbf{z}^k\|^2 + c_3\|\Pi \mathbf{x}^k\|^2\right) + \frac{2c_1}{1-\sigma}\theta_k^2\Phi^k$$

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

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

$$\|\Pi \mathbf{y}^{k+1}\|^2 \le \frac{2}{c_3}\left(c_1\|\Pi \mathbf{s}^{k+1}\|^2 + c_2\theta_{k+1}^2\|\Pi \mathbf{z}^{k+1}\|^2 + c_3\|\Pi \mathbf{x}^{k+1}\|^2\right)$$

$$\le \left(\frac{3+\sigma}{4}\right)^{k+1}\frac{2c_1}{c_3}\|\Pi \mathbf{s}^0\|^2 + \frac{4c_1}{c_3(1-\sigma)}\sum_{r=0}^{k}\left(\frac{3+\sigma}{4}\right)^{k-r}\theta_r^2\Phi^r$$

$$= \left(\frac{3+\sigma}{4}\right)^{k+1}\frac{(1-\sigma)^2}{18(1+\frac{1}{\tau})L^2}\|\Pi \mathbf{s}^0\|^2 + \frac{1-\sigma}{9(1+\frac{1}{\tau})L^2}\sum_{r=0}^{k}\left(\frac{3+\sigma}{4}\right)^{k-r}\theta_r^2\Phi^r,$$

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 4 *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 weight matrix W, we have*

$$\max\left\{\sum_{k=0}^K \frac{L}{2m\theta_k^2}\|\Pi \mathbf{y}^k\|^2, \sum_{k=0}^K \frac{L}{2m\theta_k^2}\|\Pi \mathbf{x}^k\|^2\right\}$$

$$\leq \frac{16}{3mL(1+\frac{1}{\tau})(1-\sigma)}\|\Pi \mathbf{s}^0\|^2 + \frac{11}{mL(1+\frac{1}{\tau})(1-\sigma)^2}\sum_{r=0}^{K-1}\Phi^r, \tag{35}$$

where τ and Φ^r are defined in Lemma 3.

Proof 3 *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

$$\sum_{k=1}^K \frac{L}{2m\theta_k^2}\|\Pi \mathbf{y}^k\|^2 \leq \sum_{k=1}^K \frac{C_1 L\rho^k}{2m\theta_k^2} + \sum_{k=1}^K \frac{C_2 L}{2m\theta_k^2}\sum_{r=0}^{k-1}\rho^{k-1-r}\theta_r^2\Phi^r$$

$$= \frac{C_1 L}{2m}\sum_{k=1}^K \frac{\rho^k}{\theta_k^2} + \frac{C_2 L}{2m\rho}\sum_{k=1}^K \frac{\rho^k}{\theta_k^2}\sum_{r=0}^{k-1}\frac{\theta_r^2}{\rho^r}\Phi^r \tag{37}$$

$$= \frac{C_1 L}{2m}\sum_{k=1}^K \frac{\rho^k}{\theta_k^2} + \frac{C_2 L}{2m\rho}\sum_{r=0}^{K-1}\frac{\theta_r^2}{\rho^r}\Phi^r \sum_{k=r+1}^K \frac{\rho^k}{\theta_k^2}.$$

Recall that for scalers, 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

$$(1-\rho)^2 S = (1-\rho)S - \rho(1-\rho)S$$

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

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

$$\overset{b}{\leq} \sum_{k=r+1}^K \rho^k + \frac{(1-\rho)\rho^{r+1}}{\theta_r^2} + \frac{\rho^{r+1}}{\theta_r} \leq \frac{\rho^{r+1}}{1-\rho} + \frac{2\rho^{r+1}}{\theta_r^2},$$

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 $\Pi\mathbf{y}^0 = 0$ that

$$\sum_{k=0}^{K} \frac{L}{2m\theta_k^2} \|\Pi\mathbf{y}^k\|^2 \leq \frac{3C_1L\rho}{2m(1-\rho)^3} + \frac{3C_2L}{2m(1-\rho)^3} \sum_{r=0}^{K-1} \Phi^r$$

$$\leq \frac{16}{3mL(1+\frac{1}{\tau})(1-\sigma)} \|\Pi\mathbf{s}^0\|^2 + \frac{11}{mL(1+\frac{1}{\tau})(1-\sigma)^2} \sum_{r=0}^{K-1} \Phi^r,$$

where the last inequality uses the definitions of C_1, C_2, and ρ given in Lemma 3. 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 4.

Lemma 5 *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 weight matrix W, we have*

$$\max\left\{ \sum_{k=0}^{K} \frac{L}{2m(1-\theta)^{k+1}} \|\Pi\mathbf{y}^k\|^2, \sum_{k=0}^{K} \frac{L}{2m(1-\theta)^{k+1}} \|\Pi\mathbf{x}^k\|^2 \right\}$$

$$\leq \frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)} \|\Pi\mathbf{s}^0\|^2 + \frac{8\theta^2}{27mL(1+\frac{1}{\tau})} \sum_{r=0}^{K-1} \frac{\Phi^r}{(1-\theta)^{r+1}}, \tag{39}$$

where τ and Φ^r are defined in Lemma 3.

Proof 4 *From (25), we get*

$$\sum_{k=1}^{K} \frac{L}{2m(1-\theta)^{k+1}} \|\Pi\mathbf{y}^k\|^2$$

$$\leq \sum_{k=1}^{K} \frac{C_1L\rho^k}{2m(1-\theta)^{k+1}} + \sum_{k=1}^{K} \frac{C_2L}{2m(1-\theta)^{k+1}} \sum_{r=0}^{k-1} \rho^{k-1-r}\theta^2\Phi^r$$

$$= \frac{C_1L}{2m(1-\theta)} \sum_{k=1}^{K} \left(\frac{\rho}{1-\theta} \right)^k + \frac{\theta^2C_2L}{2m\rho(1-\theta)} \sum_{k=1}^{K} \left(\frac{\rho}{1-\theta} \right)^k \sum_{r=0}^{k-1} \frac{\Phi^r}{\rho^r}$$

$$= \frac{C_1L}{2m(1-\theta)} \sum_{k=1}^{K} \left(\frac{\rho}{1-\theta} \right)^k + \frac{\theta^2C_2L}{2m\rho(1-\theta)} \sum_{r=0}^{K-1} \frac{\Phi^r}{\rho^r} \sum_{k=r+1}^{K} \left(\frac{\rho}{1-\theta} \right)^k.$$

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 $\sum_{k=r+1}^{K} \left(\frac{\rho}{1-\theta} \right)^k \leq \left(\frac{\rho}{1-\theta} \right)^r \frac{\rho}{1-\theta-\rho}$. It follows from $\Pi\mathbf{y}^0 = 0$ that

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

$$\leq \frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)} \|\Pi\mathbf{s}^0\|^2 + \frac{8\theta^2}{27mL(1+\frac{1}{\tau})} \sum_{r=0}^{K-1} \frac{\Phi^r}{(1-\theta)^{r+1}},$$

where the last inequality uses the definitions of C_1 and C_2 given in Lemma 3 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 3 and 4. We first prove Theorem 3. 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 5 *Plugging (35) into (21) and using the definition of Φ^r in (26), we obtain*

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

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

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

It follows from (35) that

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

From (36), we have the conclusions.

Next, we prove Theorem 4.

Proof 6 *Plugging (39) into (22) and using the definition of Φ^r in (26), we have*

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

$$\leq F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)}\|\Pi\mathbf{s}^0\|^2$$

$$- \sum_{k=0}^{K}\left(\frac{1}{(1-\theta)^k}\left(1 - \frac{16(1+\tau)}{27(1+\frac{1}{\tau})}\right)D_f(\overline{x}^k, \mathbf{y}^k)\right.$$

$$\left. + \frac{1}{(1-\theta)^{k+1}}\left(\frac{\theta^2}{2\alpha} - \frac{L\theta^2}{2} - \frac{16L\theta^2}{27}\right)\|\overline{z}^{k+1} - \overline{z}^k\|^2\right)$$

$$\overset{a}{\leq} F(\overline{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\overline{z}^0 - x^*\|^2 + \frac{4(1-\sigma)}{27m(1+\frac{1}{\tau})L(1-\theta)}\|\Pi\mathbf{s}^0\|^2$$

$$- \sum_{k=0}^{K}\left(\frac{1}{2(1-\theta)^k}D_f(\overline{x}^k, \mathbf{y}^k) + \frac{\theta^2}{4\alpha(1-\theta)^{k+1}}\|\overline{z}^{k+1} - \overline{z}^k\|^2\right)$$

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

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

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

It follows from (39) that

$$\sum_{k=0}^{K}\frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{x}^k\|^2$$

$$\leq \frac{4(1-\sigma)}{27mL(1+\frac{1}{\tau})(1-\theta)}\|\Pi\mathbf{s}^0\|^2 + \frac{8\theta^2}{27mL(1+\frac{1}{\tau})}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}}$$

$$= \frac{4(1-\sigma)}{59mL(1-\theta)}\|\Pi\mathbf{s}^0\|^2 + \frac{8\theta^2}{59mL}\sum_{r=0}^{K-1}\frac{\Phi^r}{(1-\theta)^{r+1}}$$

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

Thus, we have the second conclusion.

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

Remark 7 *As shown in Lemmas 2 and 3, we keep the Bregman distance term $D_f(\overline{x}^k, \mathbf{y}^k)$, and use a constant τ to balance this distance term and $\|\overline{z}^{k+1} - \overline{z}^k\|^2$. As shown in the proofs of Theorems 3 and 4, we make the constant before $D_f(\overline{x}^k, \mathbf{y}^k)$ positive by setting τ small. As a comparison, Qu & Li (2020) did not use this Bregman distance term, and they bounded the term $\|\overline{x}^k - \overline{y}^k\|^2$, which is generated by the consensus errors, 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 & Li, 2020) for the details. To make the constant A_4 in (32) positive, Qu & 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 & Li (2020) only got the slower convergence rate of $\mathcal{O}((1 - (\frac{\mu}{L})^{5/7})^k)$. For nonstrongly convex problems, Qu & Li (2020) set the step size of the order $\mathcal{O}(\frac{1}{k^{0.6+\epsilon}})$ to bound the corresponding term in (53), which gives the slower convergence rate of $\mathcal{O}(\frac{1}{k^{1.4-\epsilon}})$.*

As shown in Lemma 3, 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 & 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 & Li, 2020). Our proof is much simpler than those in (Qu & 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 6 *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 $k \geq \gamma - 1$,*

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

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

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

where we denote $c_0 = 4L^2\left(1 + \frac{1}{\tau}\right)$, and τ and Φ^r are defined in Lemma 3.

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

$$\begin{aligned}
\mathbf{s}^{k+1} &= W_1^{k+1} \mathbf{s}^k + \nabla f(\mathbf{y}^{k+1}) - \nabla f(\mathbf{y}^k) \\
&= \left(\prod_{t=k-\gamma+2}^{k+1} W_1^t\right) \mathbf{s}^{k-\gamma+1} + \sum_{r=k-\gamma+1}^{k} \left(\prod_{t=r+1}^{k} W_1^{t+1}\right) (\nabla f(\mathbf{y}^{r+1}) - \nabla f(\mathbf{y}^r)) \\
&= W_1^{k+1,\gamma} \mathbf{s}^{k-\gamma+1} + \sum_{r=k-\gamma+1}^{k} W_1^{k+1,k-r} (\nabla f(\mathbf{y}^{r+1}) - \nabla f(\mathbf{y}^r)),
\end{aligned}$$

where we denote $\prod_{t=k+1}^{k} W_1^{t+1} = I$. 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

$$\mathbf{x}^{k+1} = (1 - \theta_k)W_3^k\mathbf{x}^k + \theta_k\mathbf{z}^{k+1}$$

$$= \left(\prod_{t=k-\gamma+1}^{k}(1-\theta_t)W_3^t\right)\mathbf{x}^{k-\gamma+1} + \sum_{r=k-\gamma+1}^{k}\left(\prod_{t=r+1}^{k}(1-\theta_t)W_3^t\right)\theta_r\mathbf{z}^{r+1}$$

$$= W_3^{k,\gamma}\mathbf{x}^{k-\gamma+1}\prod_{t=k-\gamma+1}^{k}(1-\theta_t) + \sum_{r=k-\gamma+1}^{k}W_3^{k,k-r}\theta_r\mathbf{z}^{r+1}\prod_{t=r+1}^{k}(1-\theta_t).$$

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

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

and

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

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

$$\mathbf{z}^{k+1} = \frac{\theta_k}{\theta_k+\mu\alpha}W_2^k\mathbf{z}^k + \frac{\mu\alpha}{\theta_k+\mu\alpha}W_2^k\mathbf{y}^k - \frac{\alpha}{\theta_k+\mu\alpha}\mathbf{s}^k$$

$$\overset{a}{=} \frac{\theta_k(1+\mu\alpha)}{\theta_k+\mu\alpha}W_2^k\mathbf{z}^k + \frac{\mu\alpha(1-\theta_k)}{\theta_k+\mu\alpha}W_2^k\mathbf{x}^k - \frac{\alpha}{\theta_k+\mu\alpha}\mathbf{s}^k$$

$$= \left(\prod_{t=k-\gamma+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha}W_2^t\right)\mathbf{z}^{k-\gamma+1}$$

$$+ \sum_{r=k-\gamma+1}^{k}\left(\prod_{t=r+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha}W_2^t\right)\left(\frac{\mu\alpha(1-\theta_r)}{\theta_r+\mu\alpha}W_2^r\mathbf{x}^r - \frac{\alpha}{\theta_r+\mu\alpha}\mathbf{s}^r\right)$$

$$= W_2^{k,\gamma}\mathbf{z}^{k-\gamma+1}\prod_{t=k-\gamma+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha}$$

$$+ \sum_{r=k-\gamma+1}^{k}W_2^{k,k-r}\left(\frac{\mu\alpha(1-\theta_r)}{\theta_r+\mu\alpha}W_2^r\mathbf{x}^r - \frac{\alpha}{\theta_r+\mu\alpha}\mathbf{s}^r\right)\prod_{t=r+1}^{k}\frac{\theta_t(1+\mu\alpha)}{\theta_t+\mu\alpha},$$

and

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

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

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

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

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 3, 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 3. 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 2.

Lemma 7 *Under the settings of Lemma 6, 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 8 *For any t satisfying $k \leq t \leq k+\gamma-1$ with $k \geq \gamma$, we can relax (40) to*

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

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

$$\leq \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{t-\gamma+1}\|^2 + \frac{2\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right)$$

$$+ \frac{2\gamma}{1-\sigma_\gamma}\sum_{r=k-\gamma+1}^{k+\gamma}\left(c_0\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\right)$$

$$= \frac{1+\sigma_\gamma}{2}\|\Pi\mathbf{s}^{t-\gamma+1}\|^2 + \frac{2\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right)$$

$$+ \frac{4\gamma^2}{1-\sigma_\gamma}\left(c_0\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\right).$$

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

$$\mathcal{M}_{\mathbf{s}}^{k+\gamma,\gamma} \leq \frac{1+\sigma_\gamma}{2}\mathcal{M}_{\mathbf{s}}^{k,\gamma} + \frac{2\gamma}{1-\sigma_\gamma}\left(\mathcal{S}_\phi^{k-1,\gamma} + \mathcal{S}_\phi^{k+\gamma-1,\gamma}\right)$$

$$+ \frac{4\gamma^2}{1-\sigma_\gamma}\left(c_0\mathcal{M}_{\mathbf{z}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k,\gamma} + c_0\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma}\right).$$

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

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

$$\mathcal{M}_{\mathbf{x}}^{k+\gamma,\gamma} \leq \frac{1+\sigma_\gamma}{2}\mathcal{M}_{\mathbf{x}}^{k,\gamma} + \frac{11\gamma^2}{1-\sigma_\gamma}\left(\mathcal{M}_{\mathbf{z}}^{k,\gamma} + \mathcal{M}_{\mathbf{z}}^{k+\gamma,\gamma}\right),$$

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

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

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

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

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

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

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

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

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

and

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

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

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

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

$$
\sum_{r=1}^{t} \rho^{\gamma(t-r)} \left(\mathcal{S}_\phi^{r\gamma-1,\gamma} + \mathcal{S}_\phi^{(r+1)\gamma-1,\gamma} \right) = \rho^{\gamma t} \sum_{r=1}^{t} \left(\frac{1}{\rho^\gamma} \right)^r \left(\sum_{s=(r-1)\gamma}^{r\gamma-1} \theta_s^2 \Phi^s + \sum_{s=r\gamma}^{(r+1)\gamma-1} \theta_s^2 \Phi^s \right)
$$

$$
= \rho^{\gamma t} \sum_{s=0}^{t\gamma-1} \left(\frac{1}{\rho^\gamma} \right)^{\lfloor \frac{s}{\gamma} \rfloor + 1} \theta_s^2 \Phi^s + \rho^{\gamma t} \sum_{s=\gamma}^{(t+1)\gamma-1} \left(\frac{1}{\rho^\gamma} \right)^{\lfloor \frac{s}{\gamma} \rfloor} \theta_s^2 \Phi^s
$$

$$
\leq 2\rho^{\gamma t} \sum_{s=0}^{(t+1)\gamma-1} \left(\frac{1}{\rho^\gamma} \right)^{\frac{s}{\gamma}+1} \theta_s^2 \Phi^s = 2 \sum_{s=0}^{(t+1)\gamma-1} \rho^{(t-1)\gamma-s} \theta_s^2 \Phi^s.
$$

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

Remark 8 *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 7, 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 4 provides some insights.

Lemma 8 *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 $\alpha \leq \frac{(1-\sigma_\gamma)^3}{3385 L \gamma^3 \sqrt{1+\frac{1}{\tau}}}$. Then for algorithm (13a)-(13d), we have*

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

where τ and Φ^r are defined in Lemma 3, and C_3 is defined in Lemma 7.

Proof 9 *We first verify $\theta_k \leq 1.62\theta_{k+1}$ for all $k \geq 0$, which is required in Lemmas 6 and 7. 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 $k \geq 0$. Next, we upper and lower bound ρ. From the definition of $\rho = \sqrt[\gamma]{1 - \frac{1-\sigma_\tau}{5}}$ and the fact that $(1 - \frac{x}{\gamma})^\gamma \geq 1 - x$ for any $x \in (0,1)$ and $\gamma \geq 1$, we know*

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

The remaining proof is similar to that of Lemma 4. From the definition of $\mathcal{M}_{\mathbf{y}}^{t\gamma+\gamma,\gamma}$ and (45), we have

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

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

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

Plugging (51) into (50), it follows from (38) and $\Pi\mathbf{y}^0 = 0$ that

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

$$\le \frac{3C_3L}{2m\rho^{\gamma-1}(1-\rho)^3} + \frac{3C_4L}{2m\rho^{3\gamma-1}(1-\rho)^3} \sum_{s=0}^{T\gamma-1} \Phi^s$$

$$\overset{c}{\le} \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} + \frac{10\gamma^2}{mL(1+\frac{1}{\tau})(1-\sigma_\gamma)^2} \sum_{s=0}^{T\gamma-1} \Phi^s,$$

where $\overset{b}{\le}$ uses $\theta_s \le 1$ for $s \le \gamma - 1$ and $\theta_s \le \theta_{s-\gamma}$ for $s \ge \gamma$, $\overset{c}{\le}$ uses (49) and the definition of C_4 given in Lemma 7. 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 5, and the proof is similar to that of the above Lemma 8.

Lemma 9 *Suppose that Assumptions 1 and 3 hold with $\mu > 0$. Let $\alpha \le \frac{(1-\sigma_\gamma)^3}{3385L\gamma^3\sqrt{1+\frac{1}{\tau}}}$ and $\theta_k \equiv \theta = \frac{\sqrt{\mu\alpha}}{2}$. Then for algorithm (13a)-(13d), we have*

$$\max\left\{\sum_{k=0}^{T\gamma} \frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2, \sum_{k=0}^{T\gamma} \frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{x}^k\|^2\right\} \tag{52}$$

$$\le \frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)} + \frac{\theta^2}{7mL(1+\frac{1}{\tau})} \sum_{s=0}^{T\gamma-1} \frac{\Phi^s}{(1-\theta)^{s+1}},$$

where τ and Φ^r are defined in Lemma 3, and C_3 is defined in Lemma 7.

Proof 10 *From the definition of $\mathcal{M}_\mathbf{y}^{t\gamma,\gamma}$ and (45), we have*

$$\sum_{k=1}^{T\gamma} \frac{L}{2m(1-\theta)^{k+1}}\|\Pi\mathbf{y}^k\|^2$$

$$= \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma} \frac{L}{2m(1-\theta)^{t\gamma+r+1}}\|\Pi\mathbf{y}^{t\gamma+r}\|^2 \le \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma} \frac{L}{2m(1-\theta)^{t\gamma+r+1}}\mathcal{M}_\mathbf{y}^{(t+1)\gamma,\gamma}$$

$$\le \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma} \frac{C_3 L\rho^{t\gamma}}{2m(1-\theta)^{t\gamma+r+1}} + \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma} \frac{C_4 L}{2m(1-\theta)^{t\gamma+r+1}} \sum_{s=0}^{(t+1)\gamma-1} \rho^{(t-1)\gamma-s}\theta^2\Phi^s$$

$$\le \frac{C_3 L}{2m(1-\theta)} \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma} \frac{\rho^{t\gamma+r}}{\rho^\gamma(1-\theta)^{t\gamma+r}} + \frac{C_4 L\theta^2}{2m(1-\theta)} \sum_{t=0}^{T-1}\sum_{r=1}^{\gamma} \frac{\rho^{t\gamma+r}}{\rho^{2\gamma}(1-\theta)^{t\gamma+r}} \sum_{s=0}^{(t+1)\gamma-1} \frac{\Phi^s}{\rho^s}$$

$$= \frac{C_3 L}{2m\rho^\gamma(1-\theta)} \sum_{k=1}^{T\gamma} \left(\frac{\rho}{1-\theta}\right)^k + \frac{C_4 L\theta^2}{2m\rho^{2\gamma}(1-\theta)} \sum_{k=1}^{T\gamma} \left(\frac{\rho}{1-\theta}\right)^k \sum_{s=0}^{\lceil\frac{k}{\gamma}\rceil\gamma-1} \frac{\Phi^s}{\rho^s}.$$

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 $\sum_{k=r+1}^{K} \left(\frac{\rho}{1-\theta}\right)^k \leq \left(\frac{\rho}{1-\theta}\right)^r \frac{\rho}{1-\theta-\rho}$. It follows from $\|\Pi\mathbf{y}^0\| = 0$ that

$$\sum_{k=0}^{T\gamma} \frac{L}{2m(1-\theta)^{k+1}} \|\Pi\mathbf{y}^k\|^2 \leq \frac{C_3 L}{2m\rho^{\gamma-1}(1-\theta)(1-\theta-\rho)}$$

$$+ \frac{C_4 L\theta^2}{2m\rho^{2\gamma-1}(1-\theta)(1-\theta-\rho)} \left(\sum_{s=0}^{\gamma-1} \frac{\Phi^s}{(1-\theta)^s}\left(\frac{1-\theta}{\rho}\right)^s + \left(\frac{\rho}{1-\theta}\right)^{-\gamma} \sum_{s=\gamma}^{T\gamma-1} \frac{\Phi^s}{(1-\theta)^s}\right)$$

$$\overset{a}{\leq} \frac{C_3 L}{2m\rho^{\gamma-1}(1-\theta)(1-\theta-\rho)} + \frac{C_4 L\theta^2(1-\theta)^\gamma}{2m\rho^{3\gamma-1}(1-\theta-\rho)} \sum_{s=0}^{T\gamma-1} \frac{\Phi^s}{(1-\theta)^{s+1}}$$

$$\overset{b}{\leq} \frac{3.3C_3 L\gamma}{m(1-\theta)(1-\sigma_\gamma)} + \frac{\theta^2}{7mL(1+\frac{1}{\tau})} \sum_{s=0}^{T\gamma-1} \frac{\Phi^s}{(1-\theta)^{s+1}}.$$

where $\overset{a}{\leq}$ uses $\frac{1-\theta}{\rho} > 1$ such that $\left(\frac{1-\theta}{\rho}\right)^s \leq \left(\frac{1-\theta}{\rho}\right)^\gamma$ for all $s \leq \gamma - 1$, $\overset{b}{\leq}$ uses (49), $1 - \rho - \theta \geq \frac{0.19(1-\sigma_\gamma)}{\gamma}$, and the definition of C_4 given in Lemma 7. 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 1 and 2. We first prove Theorem 1.

Proof 11 *Plugging (48) into (21) and using the definition of Φ^r in (26), we have*

$$\frac{F(\overline{x}^{T\gamma+1}) - F(x^*)}{\theta_{T\gamma}^2} + \frac{1}{2\alpha}\|\overline{z}^{T\gamma+1} - x^*\|^2$$

$$\leq \frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} - \sum_{k=0}^{T\gamma} \left(\left(\frac{1}{2\alpha} - \frac{L}{2} - \frac{20L\gamma^2}{(1-\sigma_\gamma)^2}\right)\|\overline{z}^{t+1} - \overline{z}^t\|^2\right.$$

$$\left. + \frac{1}{\theta_{k-1}^2}\left(1 - \frac{20(1+\tau)\gamma^2}{(1+\frac{1}{\tau})(1-\sigma_\gamma)^2}\right) D_f(\overline{x}^k, \mathbf{y}^k)\right)$$

$$\overset{a}{\leq} \frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} - \sum_{k=0}^{T\gamma} \left(\frac{1}{4\alpha}\|\overline{z}^{t+1} - \overline{z}^t\|^2 + \frac{1}{2\theta_{k-1}^2} D_f(\overline{x}^k, \mathbf{y}^k)\right)$$

$$\leq \frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} - \frac{1}{5mL}\sum_{r=0}^{T\gamma-1} \Phi^r,$$

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

$$F(\overline{x}^{T\gamma+1}) - F(x^*) \leq \theta_{T\gamma}^2 \left(\frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3}\right), \tag{53}$$

$$\frac{1}{5mL}\sum_{r=0}^{T\gamma-1} \Phi^r \leq \frac{1}{2\alpha}\|\overline{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3}. \tag{54}$$

It follows from (48) that

$$\max\left\{\sum_{k=0}^{T\gamma}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{y}^k\|^2, \sum_{k=0}^{T\gamma}\frac{L}{2m\theta_k^2}\|\Pi\mathbf{x}^k\|^2\right\}$$

$$\leq \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} + \frac{10\gamma^2}{mL(1+\frac{1}{\tau})(1-\sigma_\gamma)^2}\sum_{s=0}^{T\gamma-1}\Phi^s \tag{55}$$

$$\leq \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3} + \frac{1}{4mL}\sum_{s=0}^{T\gamma-1}\Phi^s$$

$$\leq \frac{9}{4}\left(\frac{1}{2\alpha}\|\bar{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3}\right).$$

From the definition of C_3 given in Lemma 7, we have

$$\frac{1}{2\alpha}\|\bar{z}^0 - x^*\|^2 + \frac{235\gamma^3 C_3 L}{m(1-\sigma_\gamma)^3}$$

$$\leq \frac{1}{2\alpha}\|\bar{z}^0 - x^*\|^2 + \frac{1-\sigma_\gamma}{27mL\gamma}\mathcal{M}_\mathbf{s}^{\gamma,\gamma} + \frac{103870L\gamma^5}{m(1-\sigma_\gamma)^5}\mathcal{M}_\mathbf{z}^{\gamma,\gamma} + \frac{470L\gamma^3}{m(1-\sigma_\gamma)^3}\mathcal{M}_\mathbf{x}^{\gamma,\gamma} \equiv C_5. \tag{56}$$

The conclusion follows from Lemma 10.

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

Lemma 10 *Under the settings of Theorem 1, we can further bound C_5 by*

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

Proof 12 *From step (13c) with $\mu = 0$, we have for any $k \leq \gamma - 1$,*

$$\theta_{k+1}\|\Pi\mathbf{z}^{k+1}\| \leq \theta_k\|\Pi\mathbf{z}^{k+1}\| \leq \theta_k\|\Pi\mathbf{z}^k\| + \alpha\|\Pi\mathbf{s}^k\|$$

$$\leq \theta_0\|\Pi\mathbf{z}^0\| + \alpha\sum_{t=0}^{k}\|\Pi\mathbf{s}^t\| \leq \alpha\sum_{t=0}^{\gamma-1}\|\Pi\mathbf{s}^t\|$$

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

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

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

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

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

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

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

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

So we have the conclusion.

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

Lemma 11 *Under the settings of Theorem 1, we have for any $t \leq T - 1$,*

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

Proof 13 *We first bound $F(x_{(i)}^k) - F(\bar{x}^k)$ for any i. From Lemma 1, we have*

$$F(x_{(i)}^k) \leq f(\bar{y}^{k-1}, \mathbf{y}^{k-1}) + \left\langle \bar{s}^{k-1}, x_{(i)}^k - \bar{y}^{k-1}\right\rangle + \frac{L}{2}\|x_{(i)}^k - \bar{y}^{k-1}\|^2 + \frac{L}{2m}\|\Pi\mathbf{y}^{k-1}\|^2$$

$$\leq F(\bar{x}^k) + \left\langle \bar{s}^{k-1}, x_{(i)}^k - \bar{x}^k\right\rangle + L\|x_{(i)}^k - \bar{x}^k\|^2 + L\|\bar{x}^k - \bar{y}^{k-1}\|^2 + \frac{L}{2m}\|\Pi\mathbf{y}^{k-1}\|^2$$

$$\overset{a}{\leq} F(\bar{x}^k) + \frac{\theta_{k-1}}{\alpha}\|\bar{z}^k - \bar{z}^{k-1}\|\|\Pi\mathbf{x}^k\| + L\|\Pi\mathbf{x}^k\|^2 + L\theta_{k-1}^2\|\bar{z}^k - \bar{z}^{k-1}\|^2 + \frac{L}{2m}\|\Pi\mathbf{y}^{k-1}\|^2,$$

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 $k \leq T\gamma$,

$$\|\bar{z}^k - \bar{z}^{k-1}\|^2 \leq \frac{\Phi^{k-1}}{2mL^2(1+\frac{1}{\tau})} \overset{b}{\leq} \frac{5(1-\sigma_\gamma)^2}{80L\gamma^2}\left(\frac{1}{2\alpha}\|\bar{z}^0 - x^*\|^2 + \frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}\right),$$

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

$$F(x_{(i)}^{t\gamma+1}) - F(x^*) \leq \theta_{t\gamma}^2\max\left\{\frac{\sqrt{m}(1-\sigma_\gamma)}{L\alpha\gamma}, 8m\right\}\left(\frac{1}{2\alpha}\|\bar{z}^0 - x^*\|^2 + \frac{235\gamma^3C_3L}{m(1-\sigma_\gamma)^3}\right).$$

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

Next, we prove Theorem 2.

Proof 14 *Plugging (52) into (22) and using the definition of Φ^r in (26), we have*

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

$$\leq F(\bar{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\bar{z}^0 - x^*\|^2 + \frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)}$$

$$- \sum_{k=0}^{T\gamma}\left(\frac{1}{(1-\theta)^k}\left(1 - \frac{2(1+\tau)}{7(1+\frac{1}{\tau})}\right)D_f(\bar{x}^k, \mathbf{y}^k) + \frac{1}{(1-\theta)^{k+1}}\left(\frac{\theta^2}{2\alpha} - \frac{L\theta^2}{2} - \frac{2L\theta^2}{7}\right)\|\bar{z}^{k+1} - \bar{z}^k\|^2\right)$$

$$\overset{a}{\leq} F(\bar{x}^0) - F(x^*) + \left(\frac{\theta^2}{2\alpha} + \frac{\mu\theta}{2}\right)\|\bar{z}^0 - x^*\|^2 + \frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)}$$

$$- \sum_{k=0}^{T\gamma}\left(\frac{1}{2(1-\theta)^k}D_f(\bar{x}^k, \mathbf{y}^k) + \frac{\theta^2}{4\alpha(1-\theta)^{k+1}}\|\bar{z}^{k+1} - \bar{z}^k\|^2\right)$$

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

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

$$\sum_{k=0}^{T\gamma} \frac{L}{2m(1-\theta)^{k+1}} \|\Pi\mathbf{x}^k\|^2$$

$$\leq \frac{3.3C_3L\gamma}{m(1-\theta)(1-\sigma_\gamma)} + \frac{\theta^2}{7mL(1+\frac{1}{\tau})} \sum_{s=0}^{T\gamma-1} \frac{\Phi^s}{(1-\theta)^{s+1}}$$

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

Thus, we have the second conclusion by plugging the definition of C_3 in Lemma 7.

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 & Li, 2020). Moreover, Qu & 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:

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

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 weight matrix given in (8).

For strongly convex problems, we compare Acc-GT and Acc-GT-C (Acc-GT with multiple consensus) with DIGing (Nedić et al., 2017), DAGD-C (Rogozin et al., 2020a), 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 & 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 problems, 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 problems 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 problems, and 200000 iterations for nonstrongly convex problems. 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 problems, 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 communication 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 1 and 2, 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 1 and 2, 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 & 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 & 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.

A. Proof of Lemma 1

Proof 15 *From the μ-strong convexity and L-smoothness of $f_{(i)}$, we have*

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

---

[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.
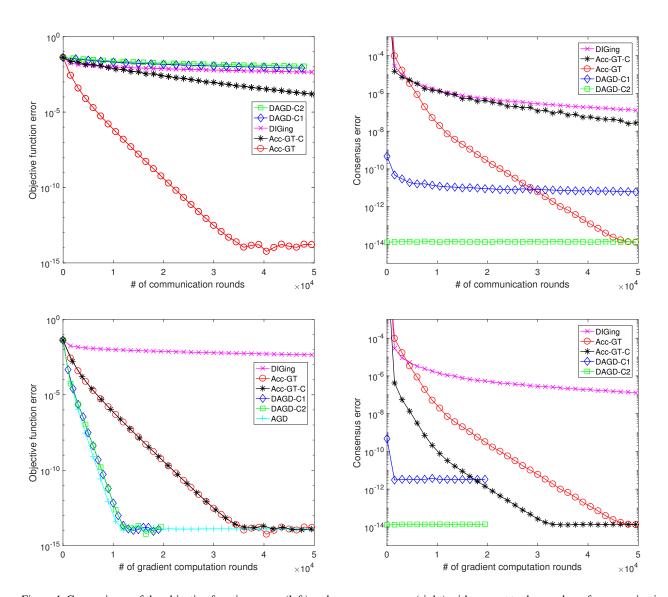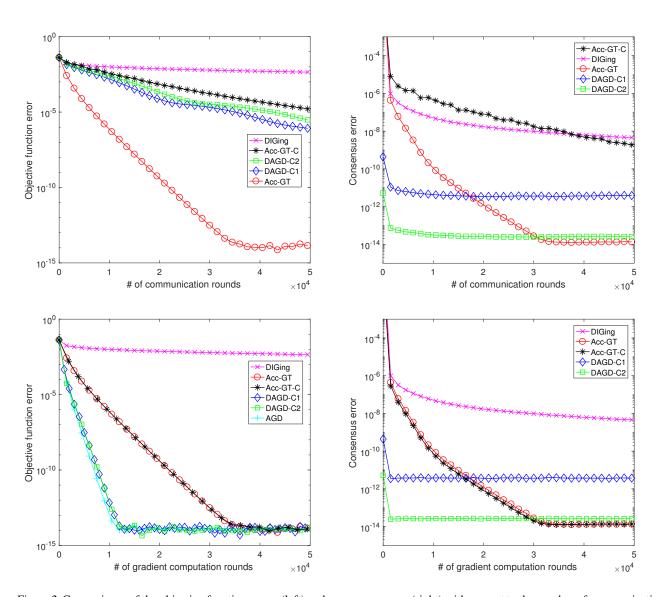
Figure 1. Comparisons of the objective function errors (left) and consensus errors (right) with respect to the number of communication (top) and computation (bottom) rounds for strongly convex problems with $d = 2$.
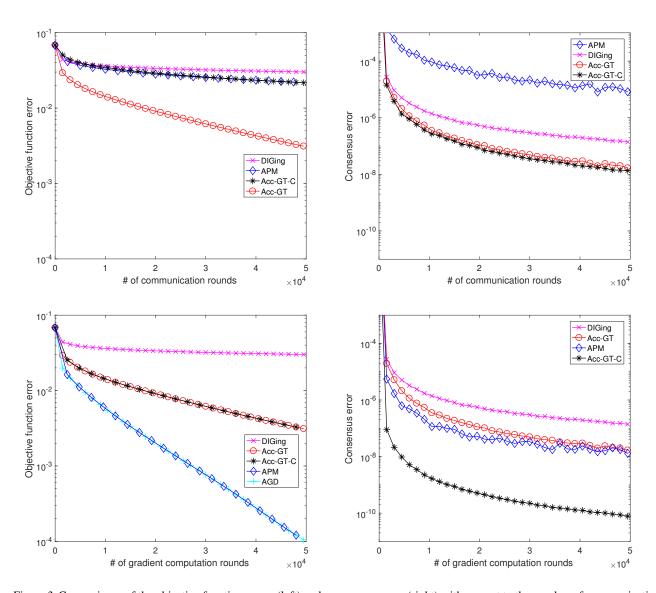
Figure 2. Comparisons of the objective function errors (left) and consensus errors (right) with respect to the number of communication (top) and computation (bottom) rounds for strongly convex problems with $d = 20$.

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 problems with $d = 2$.
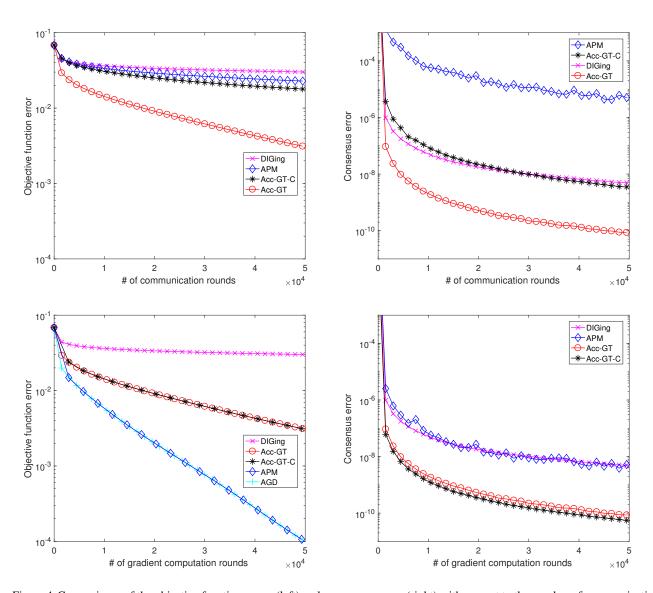
Figure 4. Comparisons of the objective function errors (left) and consensus errors (right) with respect to the number of communication (top) and computation (bottom) rounds for nonstrongly convex problems with $d = 20$.

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

and

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

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).

References

Alghunaim, S. A., Ryu, E. K., Yuan, K., and H.Sayed, A. Decentralized proximal gradient algorithms with linear covnergence rates. *preprint arXiv:1909.06479*, 2020.

Ananduta, W., Ocampo-Martinez, C., and Nedić, A. Accelerated multi-agent optimization method over stochastic networks. In *IEEE Conference on Decision and Control (CDC)*, pp. 14–18, 2020.

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

Arjevani, Y., Bruna, J., Can, B., Gürbüzbalaban, M., Jegelka, S., and Lin, H. Ideal: Inexact decentralized accelerated augmented lagrangian method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Auzinger, W. and Melenk, J. M. Iterative solution of large linear systems. *Lecture notes, TU Wien*, 2017.

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

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

Chen, Y., Hashemi, A., and Vikalo, H. Communication-efficient decentralized optimization over time-varying directed graphs. *preprint arXiv:2005.13189*, 2020.

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

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

Dvinskikh, D. and Gasnikov, A. Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems. *preprint arXiv:1904.09015*, 2019.

Fallah, A., Gurbuzbalaban, M., Ozdaglar, A., Simsekli, U., and Zhu, L. Robust distributed accelerated stochastic gradient methods for multi-agent networks. *preprint arXiv:1910.08701*, 2019.

Hendrikx, H., Bach, F., and Massoulié, L. An optimal algorithm for decentralized finite sum optimization. *preprint arXiv:2005.10675*, 2020.

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

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

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

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

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

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

Kairouz, P., McMahan, H. B., and et al. Advances and open problems in federated learning. *preprint arXiv:1912.04977*, 2019.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning (ICML)*, pp. 5381–5393, 2020.

Koralev, D., Salim, A., and Richtárik, P. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Kovalev, D., Shulgin, E., Richtarik, P., Rogozin, A., and Gasnikov, A. ADOM: accelerated decentralized optimization method for time-varying networks. *preprint arXiv:2102.09234*, 2021.

Lan, G., Lee, S., and Zhou, Y. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180:237–284, 2020.

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

Li, H., Fang, C., Yin, W., and Lin, Z. Decentralized accelerated gradient methods with increasing penalty parameters. *IEEE transactions on Signal Processing*, 68:4855–4870, 2020a.

Li, H., Lin, Z., and Fang, Y. Variance reduced EXTRA and DIGing and their optimal acceleration for strongly convex decentralized optimization. *preprint arXiv:2009.04373*, 2020b.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020c.

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

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 5330–5340, 2017.

Lin, Z., Li, H., and Fang, C. *Accelerated Optimization in Machine Learning: First-Order Algorithms*. Springer, 2020.

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

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

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

Maros, M. and Jalden, J. 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)*, pp. 5257–5261, 2019.

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

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

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

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

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

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

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

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

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

Rogozin, A., Lukoshkin, V., Gasnikov, A., Kovalev, D., and Shulgin, E. Towards accelerated rates for distributed optimization over time-varying networks. *preprint arXiv:2009.11069*, 2020a.

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

Rogozin, A., Bochko, M., Dvurechensky, P., Gasnikov, A., and Lukoshkin, V. An accelerated method for decentralized distributed stochastic optimization over time-varying graphs. *preprint arXiv:2103.15598*, 2021.

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

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *International Conference on Machine Learning (ICML)*, pp. 3027–3036, 2017.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2740–2749, 2018.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

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

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

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

Spiridonoff, A., Olshevsky, A., and Paschalidis, I. C. Robust asynchronous stochastic gradient push: Asymptotically optimal and network independent performance for strongly convex functions. *IEEE Transactions on Control of Network Systems*, 21:1–47, 2020.

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

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

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

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

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

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

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

Xu, J., Zhu, S., Soh, Y. C., and Xie, L. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060, 2015.

Xu, J., Tian, Y., Sun, Y., and Scutari, G. Accelerated primal-dual algorithms for distributed smooth convex optimization over networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2381–2391, 2020.

Ye, H., Luo, L., Zhou, Z., and Zhang, T. Multi-consensus decentralized accelerated gradient descent. *preprint arXiv:2005.00797*, 2020a.

Ye, H., Zhou, Z., Luo, L., and Zhang, T. Decentralized accelerated proximal gradient descent. In *Advances in Neural Information Processing Systems (NIPS)*, 2020b.

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