Double Stochasticity Gazes Faster: Snap-Shot Decentralized Stochastic Gradient Tracking Methods

Hao Di [1 2] **Haishan Ye** [1 3] **Xiangyu Chang** [1] **Guang Dai** [3] **Ivor W. Tsang** [4 5]

Abstract

In decentralized optimization, m agents form a network and only communicate with their neighbors, which gives advantages in data ownership, privacy, and scalability. At the same time, decentralized stochastic gradient descent (SGD) methods, as popular decentralized algorithms for training large-scale machine learning models, have shown their superiority over centralized counterparts. Distributed stochastic gradient tracking (DSGT) (Pu & Nedić, 2021) has been recognized as the popular and state-of-the-art decentralized SGD method due to its proper theoretical guarantees. However, the theoretical analysis of DSGT (Koloskova et al., 2021) shows that its iteration complexity is $\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{m\mu\varepsilon}+\frac{\sqrt{L}\bar{\sigma}}{\mu(1-\lambda_2(W))^{1/2}C_W\sqrt{\varepsilon}}\right)$, where the doubly stochastic matrix W represents the network topology and C_W is a parameter that depends on W. Thus, it indicates that the convergence property of DSGT is heavily affected by the topology of the communication network. To overcome the weakness of DSGT, we resort to the snap-shot gradient tracking skill and propose two novel algorithms, snap-shot DSGT (SS_DSGT) and accelerated snap-shot DSGT (ASS_DSGT). We further justify that SS_DSGT exhibits a lower iteration complexity compared to DSGT in the general communication network topology. Additionally, ASS_DSGT matches DSGT's iteration complexity $\mathcal{O}\left(\frac{\bar{\sigma}^2}{m\mu\varepsilon}+\frac{\sqrt{L}\bar{\sigma}}{\mu(1-\lambda_2(W))^{1/2}\sqrt{\varepsilon}}\right)$ under the same conditions as DSGT. Numerical experiments validate SS_DSGT's superior performance in the general communication network topology and exhibit better practical performance of ASS_DSGT on the specified W compared to DSGT.

1. Introduction

In this paper, we consider the decentralized optimization problem, where there are m agents to cooperatively minimize a common objective function $f(x)$ with the following formulation:

$$f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x) \quad \text{with} \quad f_i(x) := \mathbb{E}_{\xi^{(i)}\sim\mathcal{D}^{(i)}} f_i(x,\xi^{(i)}). \tag{1}$$

The formulation assumes that the objective function $f(x)$ is composed of m-local functions $f_i(x), i \in [m] = \{1,\ldots,m\}$. The i-th agent maintains the private data set $\mathcal{D}^{(i)}$ and its objective function $f_i(x)$. The m agents form a connected network and can only communicate with their neighbors. Shi et al. (2015); Scaman et al. (2019); Ye et al. (2020) indicate that the decentralized optimization has advantages over traditional centralized optimization in data ownership, privacy, and scalability (Nedic, 2020; Kairouz et al., 2021; Even et al., 2021; Shi et al., 2015; Qu & Li, 2017; Alghunaim et al., 2020; Zeng & Yin, 2018).

Due to the imminent need to train large-scale machine models, decentralized SGD methods are attracting significant attention recently because they are easy to implement, and the computation cost of each iteration is cheap (Xin et al., 2021b; Lu & De Sa, 2021; Alghunaim & Yuan, 2022; Xin et al., 2021a). Especially, Lian et al. (2017) provides the first theoretical analysis that indicates decentralized algorithms might outperform centralized algorithms of distributed stochastic gradient descent (SGD). However, the performance of decentralized SGD suffers from the data heterogeneity (Lian et al., 2017; Koloskova et al., 2020), that is, training data is in a non-IID fashion distributed over agents.

Recently, the gradient tracking method developed by Di Lorenzo & Scutari (2016) and Nedic et al. (2017) has been widely used to overcome the data heterogeneity chal-

[1]Center for Intelligent Decision-Making and Machine Learning, School of Management, Xi'an Jiaotong University, China. [2]This work was completed during the internship at SGIT AI Lab, State Grid Corporation of China. [3]SGIT AI Lab, State Grid Corporation of China. [4]College of Computing and Data Science, NTU, Singapore. [5]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore. Correspondence to: Haishan Ye <yehaishan@xjtu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

lenge. Many decentralized algorithms based on the gradient tracking method have been proposed (Qu & Li, 2017; Ye et al., 2020; Song et al., 2023). For instance, Pu & Nedić (2021) applied the gradient tracking to the decentralized SGD and proposed the distributed stochastic gradient tracking method (DSGT). DSGT effectively conquers the dilemma of data heterogeneity, and its dominant computation complexity is the same as its centralized counterpart. Furthermore, DSGT also has an advantage over centralized SGD in communication complexity.

However, the performance of DSGT is heavily affected by the topology of the communication network through which the agents exchange information. For the L-smooth and μ-strongly convex functions, DSGT has the following iteration complexity (Pu & Nedić, 2021) to achieve ε-suboptimality

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{m\mu\varepsilon}+\frac{\sqrt{L}\bar{\sigma}}{\mu(1-\lambda_2(W))^{3/2}\sqrt{\varepsilon}}\right), \tag{2}$$

where $\bar{\sigma}^2$ is the upper bound on the variance of the stochastic noise (see Assumption 2.1) and $\lambda_2(W)$ is the second largest eigenvalue of the doubly stochastic mixing matrix W. The above equation shows that when $\lambda_2(W)$ is close to one, DSGT still suffers from poor performance. Recently, Koloskova et al. (2021) improved the convergence analysis of DSGT, and obtain the following complexity

$$\tilde{\mathcal{O}}\left(\frac{\bar{\sigma}^2}{m\mu\varepsilon}+\frac{\sqrt{L}\bar{\sigma}}{\mu(1-\lambda_2(W))^{1/2}C_W\sqrt{\varepsilon}}\right), \tag{3}$$

where C_W is a parameter no smaller than $1-\lambda_2(W)$. Koloskova et al. (2021) showed that for a large number of communication networks, C_W is a constant independent of $\lambda_2(W)$. In these cases, Eq. (3) provides a better complexity than Eq. (2). Unfortunately, in the general case, C_W is no longer a constant. Eq. (3) may even reduce to Eq. (2) in the worst case. Thus, the result in Eq. (2) is the best iteration complexity of DSGT for general cases. It is still an open question: *can DSGT achieve lower communication and computation complexities than Eq. (2) for all communication networks?* Koloskova et al. (2021) also proposed an open problem: *is the parameter C_W in Eq. (3) tight in general for DSGT?*

Instead of answering the above open questions, we design two novel decentralized stochastic gradient descent tracking algorithms in this paper. We will justify that the proposed algorithm without extra inner communication loops can achieve lower complexities than Eq. (2), which take the same communication strategy as DSGT. We first extend the "snap-shot" gradient tracking method proposed by Song et al. (2023) to the SGD. Then we propose a snap-shot decentralized stochastic gradient tracking (SS_DSGT) algorithm accordingly. SS_DSGT is shown that has the following iteration complexity

$$\mathcal{O}\left(\frac{\bar{\sigma}^2}{m\mu\varepsilon}+\frac{\sqrt{L}\bar{\sigma}}{\mu(1-\lambda_2(W))\sqrt{\varepsilon}}\right),$$

which is better than the one shown in Eq. (2). In addition, we leverage the loopless Chebyshev acceleration technique (Arioli & Scott, 2014; Scaman et al., 2019; Song et al., 2023) to improve the performance of SS_DSGT on the specified W (doubly stochastic, positive semi-definite) and further propose ASS_DSGT with the iteration complexity

$$\mathcal{O}\left(\frac{\bar{\sigma}^2}{m\mu\varepsilon}+\frac{\sqrt{L}\bar{\sigma}}{\mu(1-\lambda_2(W))^{1/2}\sqrt{\varepsilon}}\right), \tag{4}$$

which aligns with the result in Alghunaim & Yuan (2024). In contrast to the method in Alghunaim & Yuan (2024), ASS_DSGT employs a decaying learning rate, which allows a larger initial value of the learning rate and potentially reduces the number of communication rounds required for convergence in practical.

To the best of our knowledge, SS_DSGT achieves the best iteration complexity for the decentralized SGD without inner communication loops in the general communication network, whose iteration complexity is independent of the parameter C_W. Additionally, under the same typology of communication network (i.e., the matrix W is doubly stochastic and positive semi-definite), ASS_DSGT demonstrates an iteration complexity consistent with prior research.

2. Notation and Assumptions

Let $\mathbf{x}$ and $\mathbf{s}$ be two $m\times d$ matrices whose i-th rows $\mathbf{x}^{(i)}$ and $\mathbf{s}^{(i)}$ are the local copy of the decision and gradient-tracking variables for the i-th agent, respectively. Accordingly, we define the averaging variables

$$\overline{x}:=\frac{1}{m}\sum_{i=1}^{m}\mathbf{x}^{(i)}=\frac{1}{m}\mathbf{1}^\top\mathbf{x}\in\mathbb{R}^{1\times d},\quad \overline{s}:=\frac{1}{m}\mathbf{1}^\top\mathbf{s}\in\mathbb{R}^{1\times d}, \tag{5}$$

where $\mathbf{1}$ denotes the vector with all entries equal to 1. Now we introduce the projection matrix

$$\mathbf{\Pi}=\mathbf{I}_m-\frac{\mathbf{1}\mathbf{1}^\top}{m}. \tag{6}$$

Using the projection matrix $\mathbf{\Pi}$, we can represent that

$$\|\mathbf{x}-\mathbf{1}\overline{x}\|=\left\|\mathbf{x}-\frac{\mathbf{1}\mathbf{1}^\top}{m}\mathbf{x}\right\|=\|\mathbf{\Pi}\mathbf{x}\|,$$
$$\|\mathbf{s}-\mathbf{1}\overline{s}\|=\|\mathbf{\Pi}\mathbf{s}\|.$$

We denote an aggregate objective function:

$$F(\mathbf{x}):=\sum_{i=1}^{m}f_i(\mathbf{x}^{(i)}) \tag{7}$$

and its aggregate gradient

$$\nabla F(\mathbf{x}) := [\nabla f_1(\mathbf{x}^{(1)}), \ldots, \nabla f_m(\mathbf{x}^{(m)})]^\top \in \mathbb{R}^{m\times d}.$$

In addition, let $\xi := [\xi^{(1)}, \ldots, \xi^{(m)}] \in \mathbb{R}^m$ and

$$\nabla F(\mathbf{x}, \xi) := \left[\nabla f_1(\mathbf{x}^{(1)}, \xi^{(1)}), \ldots, \nabla f_m(\mathbf{x}^{(m)}, \xi^{(m)})\right].$$

Throughout this paper, we use $\|\cdot\|$ to denote the "Frobenius" norm. That is, for a matrix $\mathbf{x} \in \mathbb{R}^{m\times d}$, it holds that

$$\|\mathbf{x}\|^2 = \sum_{i=1,j=1}^{m,d} \left(\mathbf{x}^{(i,j)}\right)^2.$$

Furthermore, we use $\|\mathbf{x}\|_2$ to denote the spectral norm which is the largest singular value of $\mathbf{x}$. For vectors $x, y \in \mathbb{R}^d$, we use $\langle x, y\rangle$ to denote the standard inner product of x and y.

Now we introduce several assumptions that will be used throughout this paper. First, we state an assumption that the stochastic gradients have bounded noise.

Assumption 2.1 (Bounded Noise)**.** We assume that there exists constant $\bar{\sigma}$ s.t. for any $\mathbf{x}^{(i)} \in \mathbb{R}^d$ with $i \in [m]$,

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi^{(i)}}\left[\left\|\nabla f_i(\mathbf{x}^{(i)}, \xi^{(i)}) - \nabla f_i(\mathbf{x}^{(i)})\right\|^2\right] \le \bar{\sigma}^2. \quad (8)$$

In this paper, we focus on the smooth and strongly-convex functions. That is, the function $f_i(x)$ satisfies the following assumption.

Assumption 2.2. Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth, i.e., for any $x, y \in \mathbb{R}^d$,

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

The agents are connected through a graph $\mathcal{G} = \{V, E\}$ with V and E being the sets of nodes and edges. We assume that the graph is undirected and connected. W is an $m \times m$ mixing matrix with $W_{i,j}$ being positive if and only if there is an edge between i-th and j-th agents. We also assume that W satisfies the following properties.

Definition 2.3 (Mixing matrix)**.** Matrix $W \in [0, 1]^{m\times m}$ is doubly stochastic, that is $W\mathbf{1} = \mathbf{1}$, and $\mathbf{1}^\top W = \mathbf{1}^\top$.

We further suppose that the mixing matrix has the following property to achieve the information average. Specifically, we can represent the information exchange through matrix multiplication.

Assumption 2.4. Letting $W \in \mathbb{R}^{m\times m}$ be a (random) mixing matrix and parameter $\theta \in (0, 1]$, it satisfies that

$$\mathbb{E}_W\left[\|W\mathbf{x} - \mathbf{1}\bar{x}\|^2\right] \le (1-\theta)^2\|\mathbf{x} - \mathbf{1}\bar{x}\|^2,$$
$$\text{with} \quad \theta = 1 - \sqrt{\lambda_2\left(\mathbb{E}\left[W^\top W\right]\right)}. \quad (9)$$

Algorithm 1 Snap-Shot Decentralized Stochastic Gradient Tracking

Input: x_0, mixing matrix W, initial step size η.
Initialization: Set $\mathbf{x}_0 = \mathbf{1}x_0$, $\mathbf{q}_0 = \mathbf{1}x_0$, $\mathbf{s}_0^{(i)} = \nabla f_i(\mathbf{x}_0^{(i)}, \xi_0)$, in parallel for $i \in [m]$, $\tau = 0$.
for $t = 1, \ldots, T$ **do**
Generate ζ_t with probability p.
Sample $\xi_t^{(i)}$ in parallel for all m agents and update

$$\mathbf{x}_{t+1} = W\left(\mathbf{x}_t - \eta_t\left(\mathbf{s}_t + \nabla F(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{q}_t, \xi_\tau)\right)\right). \quad (10)$$

Update

$$\mathbf{q}_{t+1} = \zeta_t\mathbf{x}_t + (1-\zeta_t)\mathbf{q}_t. \quad (11)$$

Update

$$\mathbf{s}_{t+1} = W\mathbf{s}_t + \zeta_t\left(\nabla F(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{q}_t, \xi_\tau)\right). \quad (12)$$

Set

$$\tau = \begin{cases} t, & \text{if } \zeta_t = 1, \\ \tau, & \text{otherwise.}\end{cases}$$

end for

Assumption 2.4 says that the mixing matrix W can achieve averaging in expectation but without any other constraint. As a concrete example, Boyd et al. (2006) showed that randomized gossip matrices with time-varying topologies satisfy Assumption 2.4.

3. Snap-Shot Decentralized Stochastic Gradient Tracking

In this section, we propose the SS_DSGT algorithm. We first give the algorithm description and the intuition behind our algorithm. Then, we provide a detailed convergence analysis of SS_DSGT.

3.1. Algorithm Description

Our work extends the idea of snap-shot gradient tracking (SS_GT) proposed by Song et al. (2023) to the decentralized SGD. The detailed algorithm description is in Algorithm 1.

Following the idea of SS_GT, our algorithm introduces a variable $\mathbf{q}_t$ to record some history position of $\mathbf{x}_t$ and updates it with probability p. Furthermore, instead of updating the gradient tracking variable $\mathbf{s}_t$ with the aggregated stochastic gradient $\nabla F(\mathbf{x}_{t+1}, \xi_{t+1})$ for each iteration in DSGT (refer to Eq. (67)), SS_DSGT updates $\mathbf{s}_t$ with gradient information also with probability p. The value of $\bar{s}_t$ is updated only when $\zeta_t = 1$. If $\zeta_t = 1$, we need to update the τ which records the time update $\mathbf{q}_t$.

Unlike DSGT whose $\mathbf{s}_t$ tracks the average of $\nabla F(\mathbf{x}_t, \xi_t)$ (Pu & Nedić, 2021), $\mathbf{s}_t$ of our algorithm tracks the average of $\nabla F(\mathbf{q}_t, \xi_\tau)$ which is shown by the following lemma.

Lemma 3.1. *Let sequence $\{\mathbf{s}_t\}$ be updated as Eq. (12). Then, for any $0 \le t \le T$, it holds that*

$$\overline{s}_t = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_t^{(i)}, \xi_\tau^{(i)}). \tag{13}$$

Since $\mathbf{s}_t$ tracks the value of $\nabla F(\mathbf{q}_t, \xi_\tau)$ instead of $\nabla F(\mathbf{x}_t, \xi_t)$, our algorithm proposes the update rule (10) in contrast to using $\mathbf{s}_t$ to update $\mathbf{x}_t$ directly which is used in DSGT (See Algorithm 3). Such modification follows from the fact

$$\mathbf{1}^\top \Big(\mathbf{s}_t + \nabla F(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{q}_t, \xi_\tau)\Big) \overset{(13)}{=} \mathbf{1}^\top \nabla F(\mathbf{x}_t, \xi_t).$$

Note that, our algorithm is inspired by the SS_GT proposed by Song et al. (2023), and its idea originates from SVRG (Johnson & Zhang, 2013), L-SVRG (Kovalev et al., 2020) and ANITA (Anita, 2021). However, SS_DSGT is not an easy extension of SS_GT. The original SS_GT strongly correlates to the loopless Katyusha (Kovalev et al., 2020). Extra variables such as $\mathbf{U}_t$ compared to our algorithm and "negative momentum" are important in the building of SS_GT in (Song et al., 2023). Moreover, a large part of the proof of SS_GT follows the framework of loopless Katyusha. Thus, it is unknown whether the idea of SS_GT has a broader application. Our work tries to explore the application range of the idea SS_GT and try to extend it to decentralized stochastic gradient descent.

3.2. Convergence Analysis

We will first study the evolution of $\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{x}_t\|^2\right]$, $\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{s}_t\|^2\right]$ and $\mathbb{E}\left[\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|^2\right]$. We introduce a Lyapunov function to describe the dynamics of consensus errors and $\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|$. Let us denote by $\mathcal{F}_t$ the σ-algebra generated by $\{(\xi_0, \zeta_0), (\xi_1, \zeta_1), \ldots, (\xi_{t-1}, \zeta_{t-1})\}$ and define $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ as the conditional expectation given $\mathcal{F}_t$.

Lemma 3.2. *Suppose Assumptions 2.1-2.4 hold. Let $\{\eta_t\}$ be a non-increasing sequence and satisfy $\eta_t \le \frac{\theta}{16L}$. Setting $C_{1,t} = 4\eta_t^2/\theta^2$, $C_{2,t} = 2\eta_t/(L\theta)$, and $p = \theta$, it holds that*

$$\mathbb{E}\left[\Psi_{t+1} \mid \mathcal{F}_t\right] \le \left(\frac{88mL\eta_t^2}{\theta} + 8m\eta_t\right)\left(f(\overline{x}_t) - f(x^*)\right) + \left(1 - \frac{\theta}{4}\right)\cdot \Psi_t + \frac{18m\eta_t^2}{\theta}\bar{\sigma}^2, \tag{14}$$

where we denote

$$\Psi_t \triangleq \|\mathbf{\Pi}\mathbf{x}_t\|^2 + C_{1,t}\|\mathbf{\Pi}\mathbf{s}_t\|^2 + C_{2,t}\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|^2.$$

Lemma 3.2 shows that Ψ_t will converge to zero under the condition that the step size η_t will decrease to zero and $f(\overline{x}_t) - f(x^*)$ is non-increasing. This implies that $\|\mathbf{\Pi}\mathbf{x}_t\|$ will converge to zero, that is, the distance $\left\|\mathbf{x}_t^{(i)} - \overline{x}_t\right\|^2$ will vanish as t goes. Next, we are going to upper bound the distance $\|\overline{x}_t - x^*\|$.

Lemma 3.3. *Suppose Assumptions 2.1-2.4 hold. Then we have the following inequality:*

$$\begin{aligned}&\mathbb{E}\left[\|\overline{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t\right]\\ &\le \left(1 - \frac{\mu\eta_t}{2}\right)\|\overline{x}_t - x^*\|^2 + \frac{2L\eta_t\left(1 + 2\eta_t L\right)}{m}\|\mathbf{\Pi}\mathbf{x}_t\|^2 + \\ &\quad \eta_t^2 \cdot \frac{\bar{\sigma}^2}{m} - 2\eta_t\left(1 - 2\eta_t L\right)\left(f(\overline{x}_t) - f(x^*)\right).\end{aligned} \tag{15}$$

Lemma 3.2 and 3.3 show that the dynamics of Ψ_t and $\|\overline{x}_t - x^*\|^2$ correlate to each other. Based on above two lemmas, we obtain the following convergence properties.

Lemma 3.4. *Suppose Assumptions 2.1-2.4 hold. Let $\{\eta_t\}$ be a non-increasing sequence and satisfy $\eta_t \le \frac{\theta}{2^6\cdot 3\cdot L}$. It holds that*

$$\begin{aligned}&\mathbb{E}\left[\|\overline{x}_{t+1} - x^*\|^2 + \frac{24L\eta_{t+1}}{m\theta}\Psi_{t+1} \mid \mathcal{F}_t\right]\\ &\le \exp\left(-\frac{\mu\eta_t}{2}\right)\left(\|\overline{x}_t - x^*\|^2 + \frac{24L\eta_t}{m\theta}\Psi_t\right) - \\ &\quad \frac{7\eta_t}{8}\left(f(\overline{x}_t) - f(x^*)\right) + \frac{2^4\cdot 3^3\cdot L\eta_t^3}{\theta^2}\bar{\sigma}^2 + \eta_t^2\cdot\frac{\bar{\sigma}^2}{m}.\end{aligned} \tag{16}$$

Based on Lemma 3.4, we can derive the desired convergence properties shown in the following theorem. The proof is deferred in Appendix C.

Theorem 3.5. *Suppose Assumptions 2.1-2.4 hold. Sequences $\{\mathbf{x}_t\}$, $\{\mathbf{q}_t\}$, and $\{\mathbf{s}_t\}$ are generated by Algorithm 1. Then Algorithm 1 has the following convergence properties:*

- *If $\bar{\sigma}^2 = 0$ and step size $\eta_t = \frac{\theta}{2^6\cdot 3\cdot L}$, it holds that*

$$\begin{aligned}&\mathbb{E}\left[\|\overline{x}_T - x^*\|^2 + \frac{1}{8m}\Psi_T\right]\\ &\le \exp\left(-\frac{\theta\mu}{2^7\cdot 3L}\cdot T\right)\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{1}{8m}\Psi_0\right).\end{aligned} \tag{17}$$

- *If $\bar{\sigma}^2 > 0$, the set the step size sequence $\{\eta_t\}$ and weight sequence $\{\omega_t\}$ as follows:*

$$\eta_t = \frac{6\beta}{L + \beta\mu t}, \quad \text{and} \quad \omega_t = \frac{\eta_t}{\eta_0}\exp\left(\frac{\mu}{2}\sum_{i=0}^{t}\eta_i\right),$$

where $\beta = \frac{\theta}{2^7\cdot 3^2}$. Letting $S_T = \sum_{t=0}^{T}\omega_t$, then it holds

that

$$\frac{1}{S_T}\sum_{t=0}^{T}\omega_t\big(\mathbb{E}[f(\overline{x}_t)]-f(x^*)\big) \le \frac{24\cdot(L+\beta\mu(T+1))^2}{\beta^2\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m}+\frac{2^8\cdot3^5\cdot L}{\mu^2\theta^2}\cdot\frac{\bar{\sigma}^2}{T^2}+ \frac{2^{24}\cdot3^6\cdot L^3}{\theta^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0-x^*\|^2+\frac{\Psi_0}{8m}\right). \tag{18}$$

Based on Theorem 3.5, we can directly obtain the iteration complexities for both $\bar{\sigma}=0$ and $\bar{\sigma}>0$. The proof is deferred to Appendix C.6.

Corollary 3.6. *Suppose Assumptions 2.1-2.4 hold and the mixing matrix W is fixed. Parameters of Algorithm 1 are set as Theorem 3.5, then Algorithm 1 has the following iteration complexity:*

- *If $\bar{\sigma}^2=0$, to achieve ε-suboptimality, the iteration complexity of Algorithm 1 is*

$$T=\mathcal{O}\left(\frac{L}{\mu(1-\lambda_2(W))}\log\frac{1}{\varepsilon}\right). \tag{19}$$

- *If $\bar{\sigma}^2>0$, to achieve ε-suboptimality, the iteration complexity of Algorithm 1 is*

$$T=\mathcal{O}\Big(\frac{\bar{\sigma}^2}{\mu m\varepsilon}+\frac{\sqrt{L}\bar{\sigma}}{(1-\lambda_2(W))\mu\sqrt{\varepsilon}}+ \frac{L}{(1-\lambda_2(W))\mu\varepsilon^{1/3}}\Big). \tag{20}$$

Remark 3.7. According to Corollary 3.6, when either $\bar{\sigma}^2=0$ or $\bar{\sigma}^2>0$, SS_DSGT achieves lower complexity than DSGT. Specifically,

- When $\bar{\sigma}^2=0$, our algorithm achieves a linear convergence rate and its iteration complexity is $\mathcal{O}\left(\frac{L}{\mu(1-\lambda_2(W))}\log\frac{1}{\varepsilon}\right)$. In contrast, the iteration complexity of DSGT is $\mathcal{O}\left(\frac{L}{\mu(1-\lambda_2(W))^2}\log\frac{1}{\varepsilon}\right)$ (Pu & Nedić, 2021; Qu & Li, 2017). Thus, our SS_DSGT has better performance than DSGT theoretically.
- When $\bar{\sigma}^2>0$, we can observe that the iteration complexity of SS_DSGT depends on $(1-\lambda_2(W))^{-1}$ while DSGT depends on $(1-\lambda_2(W))^{-3/2}$ (see Eq. (2)). Thus, SS_DSGT also outperforms DSGT when $\bar{\sigma}^2>0$.

4. Acceleration with Loopless Chebyshev Acceleration

In this section, we try to further improve SS_DSGT and combine it with the loopless Chebyshev acceleration proposed by Song et al. (2023). Because the loopless Chebyshev acceleration only works for the static networks, we assume that for all iterations, it shares the same mixing matrix W in this section. For the static networks, Corollary 3.6 shows that the iteration complexity of SS_DSGT depends on $\theta^{-1}=(1-\lambda_2(W))^{-1}$. In this section, we propose ASS_DSGT which achieves an iteration complexity depending on $\left(1-\lambda_2(W)\right)^{-1/2}$ instead of $\left(1-\lambda_2(W)\right)^{-1}$.

4.1. Algorithm Description

Before introducing ASS_DSGT, we make an additional assumption on the mixing matrix and define some new necessary notations.

Assumption 4.1. The mixing matrix $W\in\mathbb{R}^{m\times m}$ is symmetric positive semi-definite.

The above assumption can be easily satisfied since we can choose $\frac{I+W}{2}$ as the mixing matrix that is positive semi-definite for any mixing matrix W.

Now, we introduce $2m\times2m$ augmented matrices $\widetilde{W}$ and $\widetilde{\mathbf{\Pi}}$ for the mixing matrix W and projection matrix $\mathbf{\Pi}$ defined as follows:

$$\widetilde{W}=\begin{bmatrix}(1+\gamma)W & -\gamma W\\ I_m & \mathbf{0}\end{bmatrix},\quad\text{and}\quad\widetilde{\mathbf{\Pi}}=\begin{bmatrix}\mathbf{\Pi} & \mathbf{0}\\ \mathbf{0} & \mathbf{\Pi}\end{bmatrix}. \tag{21}$$

Accordingly, we define the augmented decision variable $\widetilde{\mathbf{x}}\in\mathbb{R}^{2m\times d}$ and gradient-tracking variable $\widetilde{\mathbf{s}}\in\mathbb{R}^{2m\times d}$. Furthermore, we denote that $\mathbf{x}_t:=\widetilde{\mathbf{x}}_t^{(1:m)}$, that is, $\mathbf{x}_t$ takes the value of the first m rows of $\widetilde{\mathbf{x}}_t$. Given these notations, we describe ASS_DSGT in Algorithm 2.

We can observe that Algorithm 2 shares almost the same algorithmic structure to the one of Algorithm 1. The advantage of ASS_DSGT mainly depends on the following property.

Lemma 4.2 (Lemma 11 of Song et al. (2023))**.** *Under Assumption 4.1, for any $\mathbf{x}\in\mathbb{R}^{m\times d}$ and $t>0$, it holds that*

$$\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^t\widetilde{\mathbf{\Pi}}[\mathbf{x};\ \mathbf{x}]\right\|^2\le\alpha\left(1-\tilde{\theta}\right)^{2t}\|\mathbf{\Pi}\mathbf{x}\|^2, \tag{25}$$

where $\alpha\le14$ and $\tilde{\theta}=\mathcal{O}\left(\sqrt{1-\lambda_2(W)}\right)$.

The above property is also used in the analysis of the heavy ball method and shows that the heavy ball method can achieve a faster convergence rate than the gradient descent (Recht, 2010).

4.2. Convergence Analysis

First, we will show that the first and last m rows of $\widetilde{\mathbf{x}}_t$ share the same mean. This property also holds for $\widetilde{\mathbf{s}}_t$.

Lemma 4.3. *Letting sequences $\{\widetilde{\mathbf{x}}_t\}$ and $\{\widetilde{\mathbf{s}}_t\}$ are gener-*

Algorithm 2 Snap-Shot Decentralized Stochastic Gradient Tracking with Loopless Chebyshev Acceleration (ASS_DSGT)

Input: x_0, mixing matrix W, initial step size η.
Initialization: Set $\widetilde{\mathbf{x}}_0 = [\mathbf{1}x_0; \mathbf{1}x_0]$, $\mathbf{q}_0 = \mathbf{1}x_0$, $\mathbf{s}_0^{(i)} = \nabla f_i(\mathbf{x}_0^{(i)}, \xi_0)$, in parallel for $i \in [m]$, $\widetilde{\mathbf{s}}_0 = [\mathbf{s}_0; \mathbf{s}_0]$, and $\tau = 0$.
for $t = 1, \ldots, T$ **do**
Generate ζ_t with probability p.
Sample $\xi_t^{(i)}$ in parallel for all m agents and update

$$\widetilde{\mathbf{x}}_{t+1} = \widetilde{W}\Big(\widetilde{\mathbf{x}}_t - \eta_t\Big(\widetilde{\mathbf{s}}_t + [\nabla F(\mathbf{x}_t, \xi_t);\ \nabla F(\mathbf{x}_t, \xi_t)] - [\nabla F(\mathbf{q}_t, \xi_\tau);\ \nabla F(\mathbf{q}_t, \xi_\tau)]\Big)\Big). \tag{22}$$

Update

$$\mathbf{q}_{t+1} = \zeta_t \mathbf{x}_t + (1 - \zeta_t)\mathbf{q}_t. \tag{23}$$

Update

$$\widetilde{\mathbf{s}}_{t+1} = \widetilde{W}\widetilde{\mathbf{s}}_t + \zeta_t\big([\nabla F(\mathbf{x}_t, \xi_t);\ \nabla F(\mathbf{x}_t, \xi_t)] - [\nabla F(\mathbf{q}_t, \xi_\tau);\ \nabla F(\mathbf{q}_t, \xi_\tau)]\big). \tag{24}$$

Set

$$\tau = \begin{cases} t, & \text{if } \zeta_t = 1, \\ \tau, & \text{otherwise.} \end{cases}$$

end for

ated by Algorithm 2, then it holds that

$$\sum_{i=1}^{m} \widetilde{\mathbf{x}}_t^{(i)} = \sum_{i=m+1}^{2m} \widetilde{\mathbf{x}}_t^{(i)}, \quad \sum_{i=1}^{m} \widetilde{\mathbf{s}}_t^{(i)} = \sum_{i=m+1}^{2m} \widetilde{\mathbf{s}}_t^{(i)}, \tag{26}$$

and

$$\frac{1}{2m}\sum_{i=1}^{2m} \widetilde{\mathbf{s}}_t^{i} = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_t^{(i)}, \xi_\tau^{(i)}). \tag{27}$$

Above lemma shows that the means of $\widetilde{\mathbf{x}}_t$ and $\widetilde{\mathbf{s}}_t$ equal to $\overline{x}_t$ and $\overline{s}_t$, respectively. Thus, Lemma 3.3 still holds for ASS_DSGT. Next, we will focus on analyzing the convergence properties of consensus errors which are different from the ones of SS_DSGT.

Lemma 4.4. *Letting sequences $\{\widetilde{\mathbf{x}}_t\}$ and $\{\widetilde{\mathbf{s}}_t\}$ are generated by Algorithm 2, it holds that*

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{x}}_t\right\|^2\right] \le \mathcal{E}_{x,t} \quad \textit{and} \quad \mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t\right\|^2\right] \le \mathcal{E}_{s,t} \tag{28}$$

with

$$\begin{aligned}&\mathcal{E}_{s,t+1}\\ &\le (1 - \frac{\tilde{\theta}}{2})\mathcal{E}_{s,t} + 2m\bar{\sigma}^2 + 4\alpha p(\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{q}_t)\|^2)\end{aligned} \tag{29}$$

$$\begin{aligned}&\mathcal{E}_{x,t+1}\\ &\le (1 - \frac{\tilde{\theta}}{2})\mathcal{E}_{x,t} + \frac{3\alpha}{\tilde{\theta}}\eta_t^2(\mathcal{E}_{s,t} + 2\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{q}_t)\|^2) + 2\alpha m\bar{\sigma}^2\eta_t^2\end{aligned} \tag{30}$$

and

$$\mathcal{E}_{s,0} = 2\alpha\|\Pi\mathbf{s}_0\|^2, \quad \mathcal{E}_{x,0} = \alpha\left\|\widetilde{\mathbf{\Pi}}\mathbf{x}_0\right\|^2. \tag{31}$$

Based on the above lemma about the consensus error terms, we can obtain the following lemma similar to Lemma 3.2.

Lemma 4.5. *Suppose Assumptions 2.1-2.2 and Assumption 4.1 hold. Let $\{\eta_t\}$ be a non-increasing sequence and satisfy $\eta_t \le \frac{\tilde{\theta}}{2^4\cdot 3^3\cdot L}$. Setting $C_{1,t} = \frac{12\alpha\eta_t^2}{\tilde{\theta}^2}$, $C_{2,t} = \frac{16(1+8\alpha)\eta_t^2}{\tilde{\theta}^2}$, $p = \tilde{\theta}$, then we can obtain that*

$$\begin{aligned}\widetilde{\Psi}_{t+1} \le \Big(1 - \tilde{\theta}\Big)\widetilde{\Psi}_t + \frac{2^{12}\cdot 3^2\cdot m\eta_t^2}{\tilde{\theta}}\cdot\bar{\sigma}^2 + \\ \frac{2^{12}\cdot 3^2\cdot mL\eta_t^2}{\tilde{\theta}}\cdot\Big(f(\overline{x}_t) - f(x^*)\Big),\end{aligned} \tag{32}$$

where we define

$$\widetilde{\Psi}_t := \mathcal{E}_{x,t} + C_{1,t}\cdot\mathcal{E}_{s,t} + C_{2,t}\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|^2. \tag{33}$$

Combining Lemma 4.5 and Lemma 3.3, we can obtain a lemma similar to Lemma 3.4.

Lemma 4.6. *Suppose Assumptions 2.1-2.2 and Assumption 4.1 hold. Let $\{\eta_t\}$ be a non-increasing sequence and satisfy $\eta_t \le \frac{\tilde{\theta}}{2^8\cdot 3\cdot L}$. Then, it holds that*

$$\begin{aligned}&\mathbb{E}\left[\|\overline{x}_{t+1} - x^*\|^2 + \frac{48L\eta_{t+1}}{m\tilde{\theta}}\widetilde{\Psi}_{t+1}\right]\\ &\le \exp\left(-\frac{\mu\eta_t}{2}\right)\left(\|\overline{x}_t - x^*\|^2 + \frac{48L\eta_t}{m\tilde{\theta}}\cdot\widetilde{\Psi}_t\right) + \frac{\eta_t^2\bar{\sigma}^2}{m} + \\ &\quad\frac{2^{16}\cdot 3^2\cdot L\eta_t^3}{\tilde{\theta}^2}\cdot\bar{\sigma}^2 - \frac{7\eta_t}{8}\Big(f(\overline{x}_t) - f(x^*)\Big)\end{aligned} \tag{34}$$

Based on Lemma 4.6, we can derive the desired convergence properties shown in the following theorem. The proof is deferred in Appendix D.

Theorem 4.7. *Suppose Assumptions 2.1-2.2 and Assumption 4.1 hold. Sequences $\{\mathbf{x}_t\}$, $\{\mathbf{q}_t\}$, and $\{\widetilde{\mathbf{s}}_t\}$ are generated by Algorithm 2. Then Algorithm 2 has the following convergence properties:*

- *If $\bar{\sigma}^2 = 0$ and step size $\eta_t = \frac{\theta}{2^8\cdot 3\cdot L}$, it holds that*

$$\begin{aligned}&\mathbb{E}\left[\|\overline{x}_T - x^*\|^2 + \frac{1}{8m}\Psi_T\right]\\ &\le \exp(-\frac{1}{2^7\cdot 3}\cdot\frac{\theta\mu}{L}\cdot T)\left(\|\overline{x}_0 - x^*\|^2 + \frac{1}{8m}\Psi_0\right).\end{aligned} \tag{35}$$

- *If $\bar{\sigma}^2 > 0$, set the step size sequence $\{\eta_t\}$ and weight sequence $\{\omega_t\}$ as follows:*

$$\eta_t = \frac{6\tilde{\beta}}{L + \tilde{\beta}\mu t}, \quad \text{and} \quad \omega_t = \frac{\eta_t}{\eta_0} \exp\left(\frac{\mu}{2}\sum_{i=0}^{t}\eta_i\right),$$

where $\tilde{\beta} = \frac{\tilde{\theta}}{2^9 \cdot 3^2}$. Letting $S_T = \sum_{t=0}^{T}\omega_t$, then it holds that

$$\begin{aligned}&\frac{1}{S_T}\sum_{t=0}^{T}\omega_t\big(\mathbb{E}[f(\overline{x}_t)] - f(x^*)\big)\\ \leq&\frac{24\cdot(L+\tilde{\beta}\mu(T+1))^2}{\tilde{\beta}^2\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^{20}\cdot 3^4\cdot L}{\tilde{\theta}^2\mu^2}\cdot\frac{\bar{\sigma}^2}{T^2} + \\ &\frac{2^{28}\cdot 3^5\cdot L^3}{\tilde{\theta}^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{\widetilde{\Psi}_0}{16m}\right).\end{aligned} \tag{36}$$

Corollary 4.8. *Suppose Assumptions 2.1-2.2 and Assumption 4.1 hold. Parameters of Algorithm 2 are set as Theorem 4.7, then Algorithm 2 has the following iteration complexity:*

- *If $\bar{\sigma}^2 = 0$, to achieve ε-suboptimality, the iteration complexity of Algorithm 1 is*

$$T = \mathcal{O}\left(\frac{L}{\mu\sqrt{1-\lambda_2(W)}}\log\frac{1}{\varepsilon}\right) \tag{37}$$

- *If $\bar{\sigma}^2 > 0$, to achieve ε-suboptimality, the iteration complexity of Algorithm 1 is*

$$\begin{aligned}T = \mathcal{O}\Big(&\frac{\bar{\sigma}^2}{\mu m\varepsilon} + \frac{\sqrt{L}\bar{\sigma}}{\mu\sqrt{1-\lambda_2(W)}\sqrt{\varepsilon}} + \\ &\frac{L}{\sqrt{1-\lambda_2(W)}\mu\varepsilon^{1/3}}\Big)\end{aligned} \tag{38}$$

Remark 4.9. Eq. (38) shows that the loopless Chebyshev acceleration can effectively reduce the iteration complexity. Comparing Eq. (38) with Eq. (3), we can conclude that ASS_DSGT can achieve good performance comparable to DSGT for all cases (no C_W). In contrast, DSGT can only achieve good performance on the limited kinds of communication networks.

5. Experiment

In this section, we carry out numerical experiments to validate the convergence property of SS_DSGT and ASS_DSGT compared to DSGT on the following l_2-penalized logistic regression problem:

$$f_i(x) = \mathbb{E}_{z^{(i)},y^{(i)}}\log\left(1+\exp(-y^{(i)}x^\top z^{(i)})\right) + \frac{v}{2}\|x\|^2,$$

Table 1. Summary of data sets, the number of agent, the regularization coefficient, and the batch size used in our experiments

Data Set	n	d	m	v	Batch Size
banknote	1360	4	20	10^{-2}	30
a9a	32560	123	20	10^{-2}	200
ijcnn1	49980	22	20	10^{-2}	200

where $z^{(i)} \in \mathbb{R}^d$ is the feature vector, $y^{(i)} \in \{-1, 1\}$ is the label, and v is the regularization coefficient.

In Table 1, we present three datasets used in our experiments along with their respective settings. The 'banknote' dataset is sourced from the UCI Machine Learning Repository website[1], while 'a9a' and 'ijcnn1' are obtained from the LIBSVM website[2]. We utilize $m = 20$ agents, distributing the data randomly and equally among them. These methods are executed in batch mode, and hyperparameters are fine-tuned for optimal performance.

Here, we construct an asymmetric mixing matrix W^{asy} following the approach of Gharesifard & Cortés (2012). This asymmetric topology is common in decentralized settings (Nedić & Olshevsky, 2014; Jiang et al., 2021; Freund et al., 2023), and the W^{asy} is a representation of the diverse general communication network typologies. Moreover, W^{asy} challenges the previous assumption about the communication network topology (Scaman et al., 2019), and the inferior performance of DSGTis expected based on Eq. (4). To elaborate, we initially generate a symmetric mixing matrix W^{cyc} with its elements set as follows:

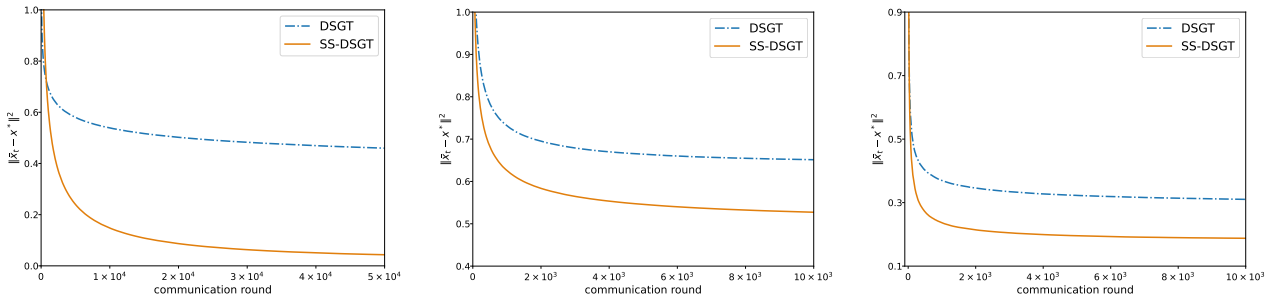
$$W^{\text{cyc}}_{(i,j)} = \begin{cases}\frac{1}{4}, & (i,j)\in E\\ \frac{1}{2}, & i=j\\ 0, & \text{otherwise}\end{cases}$$

where the edge set is given by $E = \{(i, i+1) : 1 \leq i \leq m-1\} \cup \{(m, 1)\}$. Subsequently, the asymmetric mixing matrix w^{asy} is generated by randomly adding additional 20 edges to the cycle W^{cyc}. Simultaneously, we set the symmetric matrix W^{sy} to be identical to W^{cyc}. The values of θ for W^{asy} and W^{sy} are 0.0761 and 0.024, respectively. Since ASS_DSGT is not applicable in this asymmetric setting, we compare the performance of SS_DSGT and DSGT on W^{asy}. For all experiments, we run the centralized gradient descent method to find the optimal point x^* and $f(x^*)$.

In Figure 1, we compare the distance $\|\bar{x}_t - x^*\|$ between SS_DSGT and DSGT on the asymmetric mixing matrix W^{asy}. SS_DSGT demonstrates superior performance across the three datasets, particularly depicted in Figure 1(b). This

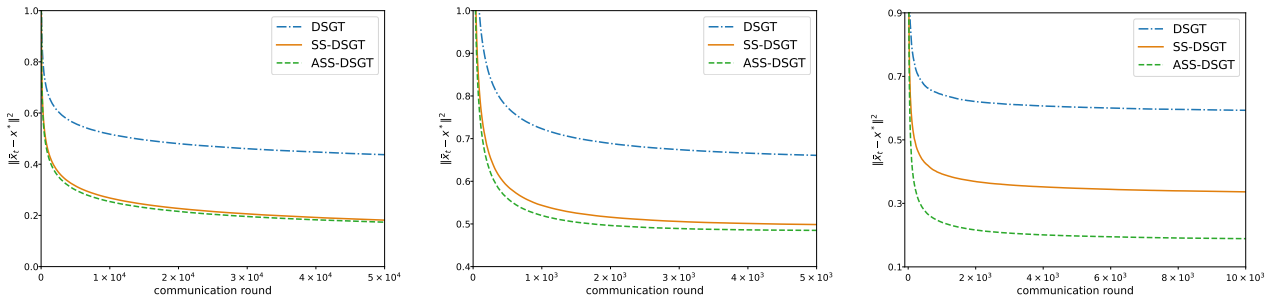
[1] https://archive.ics.uci.edu/dataset/267/banknote+authentication

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html



(a) banknote (b) a9a (c) ijcnn1

Figure 1. Comparison of SS_DSGT and DSGT for the term $\|\bar{x}_t - x^*\|^2$ versus the communication round on the asymmetric mixing matrix W^{asy}.



(a) banknote (b) a9a (c) ijcnn1

Figure 2. Comparison of different methods for the term $\|\bar{x}_t - x^*\|^2$ versus the communication round on the symmetric mixing matrix W^{sy}.

performance superiority validates the lower iteration complexity of SS_DSGT compared to DSGT for the general topology of communication network, as indicated in Eq. (3). Besides, we also conduct a comparison of the training loss for the proposed methods against DSGT, with additional figures presented in Appendix F.

As shown in Figure 2, we compare the distance $\|\bar{x}_t - x^*\|^2$ among the three methods on the symmetric matrix W^{sy}. Notably, although ASS_DSGT and DSGT have the same iteration complexity as DSGT, ASS_DSGT exhibits better practical performance across these three datasets, likely attributed to the incorporation of the acceleration technique. Besides, ASS_DSGT obtains better performance than DSGT, indicating potential improvements in iteration complexity in this symmetric setting.

6. Conclusion

In this paper, we explore the application range of the idea of SS_GT and extend it to design novel decentralized SGD methods. We propose two novel algorithms named SS_DSGT and ASS_DSGT based on the idea of SS_GT. These two algorithms have similar algorithmic structure to DSGT and they both take single loop communication strategy, which is the same as DSGT. SS_DSGT can achieve better convergence rate than DSGT for the general topology of communication network, and the iterative complexity of ASS_DSGT aligns with the result of DSGT (Kairouz et al., 2021) on the mixing matrix W. The numerical experiment validates the lower iteration complexity of SS_DSGT in the general cases compared to DSGT and demonstrates better practical performance of ASS_DSGT.

Acknowledge

This work was supported by the National Natural Science Foundation of China under Grant 12101491, the National Natural Science Foundation for Outstanding Young Scholars of China under Grant 72122018, the MOE Project of Key Research Institute of Humanities and Social Sciences No. 22JJD110001, and A*star Centre for Frontier AI Research.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alghunaim, S. A. and Yuan, K. A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing*, 70:3264–3279, 2022.

Alghunaim, S. A. and Yuan, K. An enhanced gradient-tracking bound for distributed online stochastic convex optimization. *Signal Processing*, 217:109345, 2024.

Alghunaim, S. A., Ryu, E. K., Yuan, K., and Sayed, A. H. Decentralized proximal gradient algorithms with linear convergence rates. *IEEE Transactions on Automatic Control*, 66(6):2787–2794, 2020.

Anita, Z. L. An optimal loopless accelerated variance-reduced gradient method. *arXiv preprint arXiv:2103.11333*, 2021.

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

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

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

Even, M., Hendrikx, H., and Massoulie, L. Decentralized optimization with heterogeneous delays: a continuous-time approach. *arXiv e-prints*, pp. arXiv–2106, 2021.

Freund, D., Lykouris, T., and Weng, W. Efficient decentralized multi-agent learning in asymmetric bipartite queueing systems. *Operations Research*, 2023.

Gharesifard, B. and Cortés, J. Distributed strategies for generating weight-balanced and doubly stochastic digraphs. *European Journal of Control*, 18(6):539–557, 2012.

Jiang, J., Zhang, W., Gu, J., and Zhu, W. Asynchronous decentralized online learning. *Advances in Neural Information Processing Systems*, 34:20185–20196, 2021.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

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

Koloskova, A., Lin, T., and Stich, S. U. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34:11422–11435, 2021.

Kovalev, D., Horváth, S., and Richtárik, P. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pp. 451–467. PMLR, 2020.

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. *Advances in Neural Information Processing Systems*, 30, 2017.

Lu, Y. and De Sa, C. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pp. 7111–7123. PMLR, 2021.

Nedic, A. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.

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

Nedic, 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.

Pu, S. and Nedić, A. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.

Qu, G. and Li, N. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Recht, B. Cs726-lyapunov analysis and the heavy ball method. *Department of Computer Sciences, University of Wisconsin–Madison*, 2010.

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*, 20:1–31, 2019.

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, 2015.

Song, Z., Shi, L., Pu, S., and Yan, M. Optimal gradient tracking for decentralized optimization. *Mathematical Programming*, pp. 1–53, 2023.

Xin, R., Khan, U. A., and Kar, S. A fast randomized incremental gradient method for decentralized nonconvex optimization. *IEEE Transactions on Automatic Control*, 67(10):5150–5165, 2021a.

Xin, R., Khan, U. A., and Kar, S. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69:1842–1858, 2021b.

Ye, H., Zhou, Z., Luo, L., and Zhang, T. Decentralized accelerated proximal gradient descent. *Advances in Neural Information Processing Systems*, 33:18308–18317, 2020.

Zeng, J. and Yin, W. On nonconvex decentralized gradient descent. *IEEE Transactions on Signal Processing*, 66 (11):2834–2848, 2018.

A. Useful Lemmas

In this section, we will introduce several useful lemmas that will be used in our proofs. These lemmas are easy to check or prove. Thus, we omit the detailed proofs of these lemmas.

Lemma A.1. *Let $g(x)$ be a monotonically increasing function in the range $[t_0, T]$, then it holds that*

$$\int_{t_0}^{T} g(x)\, dx \leq \sum_{k=t_0}^{T} g(k) \leq \int_{t_0}^{T+1} g(x)\, dx. \tag{39}$$

If f(x) is monotonically decreasing in the range $[t_0, T]$, then it holds that

$$\int_{t_0}^{T+1} g(x)\, dx \leq \sum_{k=t_0}^{T} g(k) \leq \int_{t_0-1}^{T} g(x)\, dx. \tag{40}$$

Lemma A.2. *If a_i's are independent random variables with expectation $\mathbb{E}[a_i] = 0$, then it holds that*

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m} a_i\right\|^2\right] = \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\left[\|a_i\|^2\right], \tag{41}$$

and for any consistent random variable b being independent of a_i, it holds

$$\mathbb{E}\left[\|a_i + b\|^2\right] = \mathbb{E}\left[\|a_i\|^2 + \|b\|^2\right]. \tag{42}$$

Lemma A.3. *For any matrix $X \in \mathbb{R}^{m\times d}$, it holds that for the projection matrix $\mathbf{\Pi}$ defined in Eq. (6),*

$$\|\mathbf{\Pi} X\| \leq \|X\|. \tag{43}$$

Lemma A.4 (Lemma 6 of (Qu & Li, 2017))**.** *Let Assumption 2.2 hold, then*

$$\left\|\nabla f(\overline{x}_t) - \frac{1}{m}\mathbf{1}^\top \nabla F(\mathbf{x}_t)\right\| \leq \frac{L}{\sqrt{m}}\|\mathbf{\Pi}\mathbf{x}_t\|. \tag{44}$$

Lemma A.5 (Lemma 3 of (Song et al., 2023))**.** *Let $f_i : \mathbb{R}^d \to \mathbb{R}$ satisfy Assumption 2.2. Denoting that $\overline{g}_t = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_t^{(i)})$, it holds that*

$$f(\overline{x}_t) \leq f(x^*) + \langle \overline{g}_t, \overline{x}_t - x^*\rangle - \frac{\mu}{4}\|\overline{x}_t - x^*\|^2 + \frac{L}{m}\|\mathbf{\Pi}\mathbf{x}_t\|^2. \tag{45}$$

B. Important Lemmas Related to Our Algorithms

Lemma B.1. *Letting Assumption 2.2 hold, then we have the following inequalities:*

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_t^{(i)}) - \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{q}_t^{(i)})\right\|^2 \leq \frac{1}{m}\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{q}_t)\|^2, \tag{46}$$

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \leq \frac{2L^2}{m}\|\mathbf{\Pi}\mathbf{x}_t\|^2 + 4L\big(f(\overline{x}_t) - f(x^*)\big). \tag{47}$$

Proof. For the first inequality, we have

$$\left\|\frac{1}{m}\sum_{i=1}^{m}\Big(\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_i(\mathbf{q}_t^{(i)})\Big)\right\|^2 \leq \frac{1}{m}\sum_{i=1}^{m}\left\|\nabla f_i(\mathbf{x}_t^{(i)}) - \nabla f_i(\mathbf{q}_t^{(i)})\right\|^2 = \frac{1}{m}\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{q}_t)\|^2.$$

For the second inequality, we have

$$\begin{aligned}\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 =&2\left\|\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}_t^{(i)})-\nabla f(\overline{x}_t)\right\|^2+2\left\|\nabla f(\overline{x}_t)\right\|^2 \overset{(44)}{\leq} \frac{2L^2}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+2\left\|\nabla f(\overline{x}_t)\right\|^2\\ \leq&\frac{2L^2}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+4L\big(f(\overline{x}_t)-f(x^*)\big),\end{aligned}$$

where the last inequality is because $f(\cdot)$ is L-smooth implied by Assumption 2.2. □

Lemma B.2. *Letting Assumption 2.2 hold, then we have the following inequalities:*

$$\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{1}x^*)\right\|^2 \leq 2L^2\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+4mL\big(f(\overline{x}_t)-f(x^*)\big), \tag{48}$$

$$\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2 \leq 4L^2\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+2\left\|\nabla F(\mathbf{q}_t)-\nabla F(\mathbf{1}x^*)\right\|^2+8mL\big(f(\overline{x}_t)-f(x^*)\big). \tag{49}$$

Proof. First, we have

$$\begin{aligned}\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{1}x^*)\right\|^2 \leq&2\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{1}\overline{x}_t)\right\|^2+2\left\|\nabla F(\mathbf{1}\overline{x}_t)-\nabla F(\mathbf{1}x^*)\right\|^2\\ \leq&2L^2\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+2\left\|\nabla F(\mathbf{1}\overline{x}_t)-\nabla F(\mathbf{1}x^*)\right\|^2,\end{aligned}$$

where the last inequality is because of $\nabla F(\mathbf{x})$ is L-smooth implied by Assumption 2.2. In addition, applying the L-smoothness of f_i again, we can obtain that

$$\begin{aligned}\left\|\nabla F(\mathbf{1}\overline{x}_t)-\nabla F(\mathbf{1}x^*)\right\|^2 =&\sum_{i=1}^{m}\left\|\nabla f_i(\overline{x}_t)-\nabla f_i(x^*)\right\|^2\\ \leq&\sum_{i=1}^{m}2L\left(f_i(\overline{x}_t)-f_i(x^*)-\langle\nabla f_i(x^*),\overline{x}_t-x^*\rangle\right)\\ =&2mL\left(f(\overline{x}_t)-f(x^*)-\left\langle\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(x^*),\overline{x}_t-x^*\right\rangle\right)\\ =&2mL\big(f(\overline{x}_t)-f(x^*)\big),\end{aligned}$$

where the last equality is because of $\frac{1}{m}\sum_{i=1}^{m}\nabla f_i(x^*)=\nabla f(x^*)=0$. Combining above two inequality, we can obtain Eq. (48).

Furthermore,

$$\begin{aligned}\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2 \leq&2\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{1}x^*)\right\|^2+2\left\|\nabla F(\mathbf{q}_t)-\nabla F(\mathbf{1}x^*)\right\|^2\\ \overset{(48)}{\leq}&4L^2\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+2\left\|\nabla F(\mathbf{q}_t)-\nabla F(\mathbf{1}x^*)\right\|^2+8mL\big(f(\overline{x}_t)-f(x^*)\big),\end{aligned}$$

which concludes the proof. □

Lemma B.3. *Letting Assumption 2.1 hold, then it holds that*

$$\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\right\|^2\right]\leq 2m\bar{\sigma}^2+\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2. \tag{50}$$

Proof. First, using the fact that $\mathbb{E}\left[\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)-(\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t))\right]=0$, we can obtain that

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\right\|^2\right]\\ \overset{(42)}{=}&\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)-(\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t))\right\|^2\right]+\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2.\end{aligned}$$

Similarly, it holds that $\mathbb{E}\left[\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{x}_t)\right]=0$ and $\mathbb{E}\left[\nabla F(\mathbf{q}_t,\xi_\tau)-\nabla F(\mathbf{q}_t)\right]=0$. Consequently,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)-(\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t))\right\|^2\right]\\
=&\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{x}_t)\right\|^2+\left\|\nabla F(\mathbf{q}_t,\xi_\tau)-\nabla F(\mathbf{q}_t)\right\|^2\right]\\
=&\mathbb{E}\left[\sum_{i=1}^m\left(\left\|\nabla f_i(\mathbf{x}_t^{(i)},\xi_t^{(i)})-\nabla f_i(\mathbf{x}_t^{(i)})\right\|^2+\left\|\nabla f_i(\mathbf{q}_t^{(i)},\xi_\tau^{(i)})-\nabla f_i(\mathbf{q}_t^{(i)})\right\|^2\right)\right]\\
\le&2m\bar{\sigma}^2.
\end{aligned}$$

Combining above two equations, we can obtain that

$$\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\right\|^2\right]\le 2m\bar{\sigma}^2+\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2.$$

□

Lemma B.4. *Let A, B, and C be three positive constants. Non-negative sequences $\{r_t\}$, $\{e_t\}$, and $\{\eta_t\}$ satisfies the follow property*

$$r_{t+1}\le\exp\left(-\frac{\mu\eta_t}{2}\right)r_t-\eta_t e_t A+\eta_t^2 B+\eta_t^3 C. \tag{51}$$

Set step size sequence and weight sequence as follows

$$\eta_t=6\beta\left(L+\beta\mu t\right)^{-1},\quad \textit{and}\quad \omega_t=\frac{\eta_t}{\eta_0}\exp\left(\frac{\mu}{2}\sum_{i=0}^t\eta_i\right).$$

Letting $S_T=\sum_{t=0}^T\omega_t$, then it holds that

$$\frac{A}{S_T}\sum_{t=0}^T e_t\omega_t\le\frac{L^3}{2\beta^3\mu^2}\cdot\frac{1}{T^3}\cdot r_0+\frac{18BL^3(L+\beta\mu(T+1))^2}{\beta^2(L-\beta\mu)^3\mu^3T^3}+\frac{108L^3}{\mu^2(L-\beta\mu)^3}\cdot\frac{C}{T^2} \tag{52}$$

Proof. Diving η_t both sides of Eq. (51) and rearranging it, we can obtain that

$$\begin{aligned}
A\sum_{t=0}^T\omega_t e_t\le&\sum_{t=0}^T\left(\frac{\exp\left(-\frac{\mu\eta_t}{2}\right)\omega_t}{\eta_t}\cdot r_t-\frac{\omega_t}{\eta_t}r_{t+1}+B\omega_t\eta_t+C\omega_t\eta_t^2\right)\\
=&\frac{\exp\left(-\frac{\mu\eta_0}{2}\right)\omega_0}{\eta_0}\cdot r_0-\frac{\omega_T}{\eta_T}r_{T+1}+B\sum_{t=0}^T\omega_t\eta_t+C\sum_{t=0}^T\omega_t\eta_t^2,
\end{aligned} \tag{53}$$

where the equality is because of $\omega_t=\frac{\eta_t}{\eta_{t-1}}\exp\left(\frac{\mu}{2}\eta_t\right)\omega_{t-1}$.

Now, we are going to upper bound ω_t as follows

$$\begin{aligned}
\omega_t=&\frac{L}{L+\beta\mu t}\exp\left(3\beta\mu\sum_{i=0}^t(L+\beta\mu i)^{-1}\right)\overset{(40)}{\le}\frac{L}{L+\beta\mu t}\exp\left(3\beta\mu\int_{-1}^t(L+\beta\mu i)^{-1}\,di\right)\\
=&\frac{L}{L+\beta\mu t}\cdot\left(\frac{L+\beta\mu t}{L-\beta\mu}\right)^3=\frac{L(L+\beta\mu t)^2}{(L-\beta\mu)^3}.
\end{aligned} \tag{54}$$

We lower bound S_T as follows:

$$\begin{aligned} S_T =& \sum_{t=0}^{T} \omega_t = \sum_{t=0}^{T} \frac{L}{L+\beta\mu t} \exp\left(3\mu\beta \sum_{i=0}^{t} (L+\beta\mu i)^{-1}\right) \\ \overset{(40)}{\geq}& \sum_{t=0}^{T} \frac{L}{L+\beta\mu t} \exp\left(3\mu\beta \int_{i=0}^{t} (L+\beta\mu i)^{-1}\, di\right) \\ =& \sum_{t=0}^{T} \frac{L}{L+\beta\mu t} \left(\frac{L+\beta\mu t}{L}\right)^3 \overset{(39)}{\geq} \int_0^T \left(\frac{L+\beta\mu t}{L}\right)^2 dt \\ =& \frac{L}{3\beta\mu}\left(\left(1+\frac{\beta\mu}{L}T\right)^3 - 1\right) \geq \frac{\beta^2\mu^2}{3L^2}T^3. \end{aligned} \tag{55}$$

We also have

$$\begin{aligned} \sum_{t=0}^{T} \omega_t\eta_t \overset{(54)}{\leq}& \sum_{t=0}^{T} \frac{L(L+\beta\mu t)^2}{(L-\beta\mu)^3} \cdot \frac{6\beta}{L+\beta\mu t} \overset{(39)}{\leq} \frac{6L\beta}{(L-\beta\mu)^3} \int_0^{T+1} (L+\beta\mu t)\, dt \\ \leq& \frac{6L(L+\beta\mu(T+1))^2}{(L-\beta\mu)^3\mu}, \end{aligned} \tag{56}$$

and

$$\sum_{t=0}^{T} \omega_t\eta_t^2 \overset{(54)}{\leq} \sum_{t=0}^{T} \frac{L(L+\beta\mu t)^2}{(L-\beta\mu)^3} \cdot \left(\frac{6\beta}{L+\beta\mu t}\right)^2 = \frac{36L\beta^2}{(L-\beta\mu)^3}T. \tag{57}$$

Dividing S_T both sides of Eq. (53), we can obtain that

$$\begin{aligned} \frac{A}{S_T}\sum_{i=0}^{T} \omega_t e_t \leq& \frac{\exp\left(-\frac{\mu\eta_0}{2}\right)\omega_0}{S_T\eta_0} \cdot r_0 + \frac{B}{S_T}\sum_{t=0}^{T} \omega_t\eta_t + \frac{C}{S_T}\sum_{t=0}^{T} \omega_t\eta_t^2 \\ \overset{(55)(56)(57)}{\leq}& \frac{L^3}{2\beta^3\mu^2} \cdot \frac{1}{T^3} \cdot r_0 + \frac{18BL^3(L+\beta\mu(T+1))^2}{\beta^2(L-\beta\mu)^3\mu^3T^3} + \frac{108L^3}{\mu^2(L-\beta\mu)^3} \cdot \frac{C}{T^2}. \end{aligned}$$

□

C. Proofs for Section 3

C.1. Proof of Lemma 3.1

Proof of Lemma 3.1. We prove the result by the induction. For $t = 0$, by the initialization, we have

$$\overline{s}_0 = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_t^{(0)}, \xi_0^{(i)}).$$

If $\zeta_t = 0$, then $\mathbf{q}_{t+1} = \mathbf{q}_t$. By the induction hypothesis and the fact $\mathbf{1}^\top W = \mathbf{1}^\top$, we have

$$\overline{s}_{t+1} = \overline{s}_t = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_t^{(i)}, \xi_\tau^{(i)}) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_{t+1}^{(i)}, \xi_\tau^{(i)}).$$

If $\zeta_t = 1$, then $\mathbf{q}_{t+1} = \mathbf{x}_t$. By the induction hypothesis,

$$\begin{aligned} \overline{s}_{t+1} =& \overline{s}_t + \frac{1}{m}\sum_{i=1}^{m} \left(\nabla f_i(\mathbf{x}_t^{(i)}, \xi_t^{(i)}) - \nabla f_i(\mathbf{q}_t^{(i)}, \xi_\tau^{(i)})\right) \\ =& \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_t^{(i)}, \xi_\tau^{(i)}) + \frac{1}{m}\sum_{i=1}^{m} \left(\nabla f_i(\mathbf{x}_t^{(i)}, \xi_t^{(i)}) - \nabla f_i(\mathbf{q}_t^{(i)}, \xi_\tau^{(i)})\right) \\ =& \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{x}_t^{(i)}, \xi_t^{(i)}) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_{t+1}^{(i)}, \xi_t^{(i)}) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(\mathbf{q}_{t+1}^{(i)}, \xi_\tau^{(i)}), \end{aligned}$$

where the last equality is because of $\tau = t$ if $\zeta_t = 1$. □

C.2. Proof of Lemma 3.2

Before the detailed proof, we first introduce the following lemma which describes the evolution of consensus error terms.

Lemma C.1. *Suppose Assumptions 2.1-2.4 hold. Sequences $\{\mathbf{x}_t\}$, $\{\mathbf{s}_t\}$, and $\{\mathbf{q}_t\}$ are generated by Algorithm 1. We have the following inequalities:*

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{x}_{t+1}\|^2 \mid \mathcal{F}_t\right]\\ \le&(1-\theta)\cdot\|\mathbf{\Pi}\mathbf{x}_t\|^2+\frac{3\eta_t^2}{\theta}\|\mathbf{\Pi}\mathbf{s}_t\|^2+\left(\eta_t^2+\frac{2\eta_t^2}{\theta}\right)\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2+2m\eta_t^2\bar{\sigma}^2\end{aligned} \tag{58}$$

and

$$\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{s}_{t+1}\|^2 \mid \mathcal{F}_t\right]\le(1-\theta)\|\mathbf{\Pi}\mathbf{s}_t\|^2+2p\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2+4mp\bar{\sigma}^2 \tag{59}$$

and

$$\begin{aligned}&\mathbb{E}\left[\|\nabla F(\mathbf{q}_{t+1})-\nabla F(\mathbf{1}x^*)\|^2 \mid \mathcal{F}_t\right]\\ \le&(1-p)\|\nabla F(\mathbf{q}_t)-\nabla F(\mathbf{1}x^*)\|^2+2pL^2\|\mathbf{\Pi}\mathbf{x}_t\|^2+4mLp\big(f(\overline{x}_t)-f(x^*)\big).\end{aligned} \tag{60}$$

Proof. First, we have

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{\Pi}\left(\mathbf{x}_t-\eta_t\left(\mathbf{s}_t+\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\right)\right)\|^2 \mid \mathcal{F}_t\right]\\ =&\mathbb{E}\left[\|\mathbf{\Pi}\left(\mathbf{x}_t-\eta_t\mathbf{s}_t\right)-\eta_t\cdot\mathbf{\Pi}\left(\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\right)\|^2 \mid \mathcal{F}_t\right]\\ =&\|\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t)\|^2+\eta_t^2\cdot\mathbb{E}\left[\left\|\mathbf{\Pi}\big(\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\big)\right\|^2 \mid \mathcal{F}_t\right]\\ &-2\eta_t\cdot\mathbb{E}\left[\left\langle\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t),\mathbf{\Pi}\big(\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\big)\right\rangle \mid \mathcal{F}_t\right]\\ =&\|\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t)\|^2+\eta_t^2\cdot\mathbb{E}\left[\left\|\mathbf{\Pi}\big(\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\big)\right\|^2 \mid \mathcal{F}_t\right]\\ &-2\eta_t\left\langle\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t),\mathbf{\Pi}\big(\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\big)\right\rangle\\ \le&\|\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t)\|^2+\eta_t^2\mathbb{E}\left[\left\|\mathbf{\Pi}\big(\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\big)\right\|^2 \mid \mathcal{F}_t\right]\\ &+\frac{\theta}{2}\|\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t)\|^2+\frac{2\eta_t^2}{\theta}\left\|\mathbf{\Pi}\big(\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\big)\right\|^2\\ \overset{(43)}{\le}&\left(1+\frac{\theta}{2}\right)\|\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t)\|^2+\eta_t^2\mathbb{E}\left[\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\|^2\right]\\ &+\frac{2\eta_t^2}{\theta}\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2\\ \overset{(50)}{\le}&\left(1+\frac{\theta}{2}\right)\|\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t)\|^2+\left(\eta_t^2+\frac{2\eta_t^2}{\theta}\right)\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2+2m\eta_t^2\bar{\sigma}^2,\end{aligned} \tag{61}$$

where the first inequality is because of the Cauchy's inequality.

Using above equation, we can obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{x}_{t+1}\|^2 \mid \mathcal{F}_t\right] \\
=&\mathbb{E}\left[\|\mathbf{\Pi}W\left(\mathbf{x}_t-\eta_t\left(\mathbf{s}_t+\zeta_t\left(\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\right)\right)\right)\|^2 \mid \mathcal{F}_t\right] \\
\overset{(9)}{\leq}&(1-\theta)^2\cdot\mathbb{E}\left[\|\mathbf{\Pi}\left(\mathbf{x}_t-\eta_t\left(\mathbf{s}_t+\zeta_t\left(\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\right)\right)\right)\|^2 \mid \mathcal{F}_t\right] \\
\overset{(61)}{\leq}&(1-\theta)^2\left(1+\frac{\theta}{2}\right)\|\mathbf{\Pi}(\mathbf{x}_t-\eta_t\mathbf{s}_t)\|^2+\left(\eta_t^2+\frac{2\eta_t^2}{\theta}\right)\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2+2m\eta_t^2\bar{\sigma}^2 \\
\leq&(1-\theta)^2\left(1+\frac{\theta}{2}\right)\left(\left(1+\frac{\theta}{2}\right)\|\mathbf{\Pi}\mathbf{x}_t\|+\left(1+\frac{2}{\theta}\right)\eta_t^2\|\mathbf{\Pi}\mathbf{s}_t\|^2\right) \\
&+\left(\eta_t^2+\frac{2\eta_t^2}{\theta}\right)\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2+2m\eta_t^2\bar{\sigma}^2 \\
\leq&(1-\theta)\|\mathbf{\Pi}\mathbf{x}_t\|^2+\frac{3\eta_t^2}{\theta}\|\mathbf{\Pi}\mathbf{s}_t\|^2+\frac{3\eta_t^2}{\theta}\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2+2m\eta_t^2\bar{\sigma}^2,
\end{aligned}
$$

which proves the result of Eq. (58).

Now, we are going to prove Eq. (59) and we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{s}_{t+1}\|^2 \mid \mathcal{F}_t\right] \\
=&(1-p)\mathbb{E}\left[\|\mathbf{\Pi}W\mathbf{s}_t\|^2 \mid \mathcal{F}_t\right]+p\mathbb{E}\left[\|\mathbf{\Pi}(W\mathbf{s}_t+\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau))\|^2 \mid \mathcal{F}_t\right] \\
\leq&(1-p)\mathbb{E}\left[\|\mathbf{\Pi}W\mathbf{s}_t\|^2 \mid \mathcal{F}_t\right]+2p\mathbb{E}\left[\|\mathbf{\Pi}W\mathbf{s}_t\|^2+\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\|^2 \mid \mathcal{F}_t\right] \\
=&(1+p)\mathbb{E}\left[\|\mathbf{\Pi}W\mathbf{s}_t\|^2 \mid \mathcal{F}_t\right]+2p\mathbb{E}\left[\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\|^2 \mid \mathcal{F}_t\right] \\
\overset{(9)}{\leq}&(1+p)(1-\theta)^2\cdot\|\mathbf{\Pi}\mathbf{s}_t\|^2+2p\mathbb{E}\left[\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\|^2 \mid \mathcal{F}_t\right] \\
\leq&(1-\theta)\,\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{s}_t\|^2\right]+2p\mathbb{E}\left[\|\nabla F(\mathbf{x}_t,\xi_t)-\nabla F(\mathbf{q}_t,\xi_\tau)\|^2 \mid \mathcal{F}_t\right],
\end{aligned}
$$

where the last inequality is because of $p\leq\theta$. Using Eq. (50), we can obtain that

$$
\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{s}_{t+1}\|^2 \mid \mathcal{F}_t\right]\overset{(50)}{\leq}(1-\theta)\|\mathbf{\Pi}\mathbf{s}_t\|^2+2p\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\|^2+4mp\bar{\sigma}^2.
$$

By the update rule of $\mathbf{q}_t$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\mathbf{q}_{t+1})-\nabla F(\mathbf{1}x^*)\|^2 \mid \mathcal{F}_t\right] \\
\overset{(11)}{=}&p\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{1}x^*)\|^2+(1-p)\|\nabla F(\mathbf{q}_t)-\nabla F(\mathbf{1}x^*)\|^2 \\
\overset{(48)}{\leq}&2pL^2\|\mathbf{\Pi}\mathbf{x}_t\|^2+4mLp\big(f(\overline{x}_t)-f(x^*)\big)+(1-p)\|\nabla F(\mathbf{q}_t)-\nabla F(\mathbf{1}x^*)\|^2.
\end{aligned}
$$

□

By the above lemma and the setting of parameters, we can prove Lemma 3.2 as follows.

Proof of Lemma 3.2. First, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{\Pi}\mathbf{x}_{t+1}\|^2 + C_{1,t+1}\|\mathbf{\Pi}\mathbf{s}_{t+1}\|^2 + C_{2,t+1}\|\nabla F(\mathbf{q}_{t+1}) - \nabla F(\mathbf{1}x^*)\|^2\right] \\
\overset{(58)(59)(60)}{\leq}& \left(1-\theta+2pL^2C_{2,t}\right)\cdot\|\mathbf{\Pi}\mathbf{x}_t\|^2 + \left(1-\theta+\frac{3\eta_t^2}{\theta C_{1,t}}\right)C_{1,t}\|\mathbf{\Pi}\mathbf{s}_t\|^2 \\
&+ (1-p)\,C_{2,t}\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|^2 + \left(\eta_t^2 + \frac{2\eta_t^2}{\theta} + 2pC_{1,t}\right)\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{q}_t)\|^2 \\
&+ \left(2mn\eta_t^2 + 4mpC_{1,t}\right)\bar{\sigma}^2 + 4mLpC_{2,t}\big(f(\overline{x}_t) - f(x^*)\big) \\
\overset{(49)}{\leq}& \left(1-\theta+2pL^2C_{2,t} + 4L^2\eta_t^2 + \frac{8L^2\eta_t^2}{\theta} + 8pL^2C_{1,t}\right)\cdot\|\mathbf{\Pi}\mathbf{x}_t\|^2 + \left(1-\theta+\frac{3\eta_t^2}{\theta C_{1,t}}\right)C_{1,t}\|\mathbf{\Pi}\mathbf{s}_t\|^2 \\
&+ \left(1-p+\frac{2\eta_t^2}{C_{2,t}} + \frac{4\eta_t^2}{\theta C_{2,t}} + \frac{4pC_{1,t}}{C_{2,t}}\right)C_{2,t}\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|^2 \\
&+ \left(2mn\eta_t^2 + 4mpC_{1,t}\right)\bar{\sigma}^2 + 4mL\left(pC_{2,t} + 2\eta_t^2 + \frac{4\eta_t^2}{\theta} + 4pC_{1,t}\right)\big(f(\overline{x}_t) - f(x^*)\big) \\
=& \left(1-\theta+4L\eta_t + \frac{(4\theta+8+32)L^2\eta_t^2}{\theta}\right)\cdot\|\mathbf{\Pi}\mathbf{x}_t\|^2 + \left(1-\theta+\frac{3\theta}{4}\right)C_{1,t}\|\mathbf{\Pi}\mathbf{s}_t\|^2 \\
&+ (1-\theta+\eta_tL\theta+2L\eta_t+8\eta_tL)\cdot C_{2,t}\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|^2 \\
&+ \left(2mn\eta_t^2 + \frac{16mn\eta_t^2}{\theta}\right)\bar{\sigma}^2 + 4mL\left(\frac{2\eta_t}{L} + 2\eta_t^2 + \frac{4\eta_t^2}{\theta} + \frac{16\eta_t^2}{\theta}\right)\big(f(\overline{x}_t) - f(x^*)\big) \\
\leq& \left(1-\frac{\theta}{4}\right)\cdot\left(\|\mathbf{\Pi}\mathbf{x}_t\|^2 + C_{1,t}\|\mathbf{\Pi}\mathbf{s}_t\|^2 + C_{2,t}\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\|^2\right) \\
&+ \frac{18mn\eta_t^2}{\theta}\bar{\sigma}^2 + \left(\frac{88mL\eta_t^2}{\theta} + 8mn\eta_t\right)\big(f(\overline{x}_t) - f(x^*)\big),
\end{aligned}
$$

where the first equality is because of $C_{1,t} = 4\eta_t^2/\theta^2$, $C_{2,t} = 2\eta_t/(L\theta)$, $p = \theta$ and the last inequality is because of $\eta_t \leq \frac{\theta}{16L}$. □

C.3. Proof of Lemma 3.3

Proof of Lemma 3.3. By the update rule of $\mathbf{x}_t$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\|\overline{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t\right] \\
=& \|\overline{x}_t - x^*\|^2 - 2\eta_t\mathbb{E}\left[\left\langle \overline{s}_t + \frac{1}{m}\sum_{i=1}^m\left(\nabla f_i(\mathbf{x}_t^{(i)},\xi_t^{(i)}) - \nabla f_i(\mathbf{q}_t^{(i)}),\xi_\tau^{(i)}\right), \overline{x}_t - x^*\right\rangle \mid \mathcal{F}_t\right] \\
&+ \eta_t^2\mathbb{E}\left[\left\|\overline{s}_t + \frac{1}{m}\sum_{i=1}^m\left(\nabla f_i(\mathbf{x}_t^{(i)},\xi_t^{(i)}) - \nabla f_i(\mathbf{q}_t^{(i)}),\xi_\tau^{(i)}\right)\right\|^2 \mid \mathcal{F}_t\right] \\
\overset{(13)}{=}& \|\overline{x}_t - x^*\|^2 - 2\eta_t\mathbb{E}\left[\left\langle \frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}_t^{(i)},\xi_t^{(i)}), \overline{x}_t - x^*\right\rangle \mid \mathcal{F}_t\right] + \eta_t^2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}_t^{(i)},\xi_t^{(i)})\right\|^2 \mid \mathcal{F}_t\right] \qquad (62) \\
=& \|\overline{x}_t - x^*\|^2 - 2\eta_t\left\langle \frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}_t^{(i)}), \overline{x}_t - x^*\right\rangle + \eta_t^2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}_t^{(i)},\xi_t^{(i)})\right\|^2 \mid \mathcal{F}_t\right] \\
\overset{(45)}{\leq}& \left(1-\frac{\mu\eta_t}{2}\right)\|\overline{x}_t - x^*\|^2 - 2\eta_t\big(f(\overline{x}_t) - f(x^*)\big) + \frac{2L\eta_t}{m}\|\mathbf{\Pi}\mathbf{x}_t\|^2 \\
&+ \eta_t^2\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}_t^{(i)},\xi_t^{(i)})\right\|^2 \mid \mathcal{F}_t\right].
\end{aligned}
$$

Furthermore, using the fact that $\mathbb{E}\left[\nabla f_i(\mathbf{x}_t^{(i)}, \xi_t^{(i)}) - \nabla f_i(\mathbf{x}_t^{(i)}) \mid \mathcal{F}_t\right] = 0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_t^{(i)}, \xi_t^{(i)})\right\|^2 \mid \mathcal{F}_t\right] \\
\overset{(42)}{\leq} &\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_t^{(i)}, \xi_t^{(i)}) - \frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \mid \mathcal{F}_t\right] + \left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \\
\overset{(41)}{=} &\frac{1}{m^2}\sum_{i=1}^m \mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_t^{(i)}, \xi_t^{(i)}) - \nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \mid \mathcal{F}_t\right] + \left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \\
\overset{(8)}{\leq} &\frac{\bar{\sigma}^2}{m} + \left\|\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}_t^{(i)})\right\|^2 \\
\overset{(47)}{\leq} &\frac{2L^2}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2 + 4L\big(f(\overline{x}_t) - f(x^*)\big) + \frac{\bar{\sigma}^2}{m}.
\end{aligned}
$$

Combining above results, we can obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\overline{x}_{t+1} - x^*\right\|^2 \mid \mathcal{F}_t\right] \\
\leq &\left(1 - \frac{\mu\eta_t}{2}\right)\left\|\overline{x}_t - x^*\right\|^2 - 2\eta_t\big(f(\overline{x}_t) - f(x^*)\big) + \frac{2L\eta_t}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2 + \eta_t^2 \cdot \frac{\bar{\sigma}^2}{m} \\
&+ \frac{2\eta_t^2 L^2}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2 + 4\eta_t^2 L\big(f(\overline{x}_t) - f(x^*)\big) \\
= &\left(1 - \frac{\mu\eta_t}{2}\right)\left\|\overline{x}_t - x^*\right\|^2 - 2\eta_t\left(1 - 2\eta_t L\right)\big(f(\overline{x}_t) - f(x^*)\big) + \eta_t^2 \cdot \frac{\bar{\sigma}^2}{m} \\
&+ \frac{2L\eta_t\left(1 + 2\eta_t L\right)}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2.
\end{aligned}
$$

□

C.4. Proof of Lemma 3.4

Proof of Lemma 3.4. We have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\overline{x}_{t+1}-x^*\right\|^2+\frac{24L\eta_{t+1}}{m\theta}\Psi_{t+1}\mid\mathcal{F}_t\right]\\
&\overset{(15)(14)}{\leq}\left(1-\frac{\mu\eta_t}{2}\right)\left\|\overline{x}_t-x^*\right\|^2-2\eta_t\left(1-2\eta_tL\right)\left(f(\overline{x}_t)-f(x^*)\right)+\eta_t^2\cdot\frac{\bar{\sigma}^2}{m}\\
&\quad+\frac{2L\eta_t\left(1+2\eta_tL\right)}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+\left(1-\frac{\theta}{4}\right)\cdot\frac{24L\eta_t}{m\theta}\Psi_t\\
&\quad+\frac{2^4\cdot3^3\cdot L\eta_t^3}{\theta^2}\bar{\sigma}^2+\left(\frac{2^8\cdot3^2\cdot L^2\eta_t^3}{\theta^2}+\frac{2^6\cdot3\cdot L\eta_t^2}{\theta}\right)\cdot\left(f(\overline{x}_t)-f(x^*)\right)\\
&=\left(1-\frac{\mu\eta_t}{2}\right)\left\|\overline{x}_t-x^*\right\|^2-2\eta_t\left(1-2\eta_tL-\frac{2^7\cdot3^2\cdot L^2\eta_t^2}{\theta^2}-\frac{2^5\cdot3\cdot L\eta_t}{\theta}\right)\left(f(\overline{x}_t)-f(x^*)\right)\\
&\quad+\eta_t^2\cdot\frac{\bar{\sigma}^2}{m}+\frac{2^4\cdot3^3\cdot L\eta_t^3}{\theta^2}\bar{\sigma}^2+\frac{3L\eta_t}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2+\left(1-\frac{\theta}{4}\right)\cdot\frac{24L\eta_t}{m\theta}\Psi_t\\
&\overset{\eta_t\leq\frac{\theta}{2^6\cdot3\cdot L}}{\leq}\left(1-\frac{\mu\eta_t}{2}\right)\left\|\overline{x}_t-x^*\right\|^2-\frac{7\eta_t}{8}\left(f(\overline{x}_t)-f(x^*)\right)+\left(1-\frac{\theta}{4}+\frac{3L\eta_t}{m}\cdot\frac{m\theta}{24L\eta_t}\right)\cdot\frac{24L\eta_t}{m\theta}\Psi_t\\
&\quad+\frac{2^4\cdot3^3\cdot L\eta_t^3}{\theta^2}\bar{\sigma}^2+\eta_t^2\cdot\frac{\bar{\sigma}^2}{m}\\
&=\left(1-\frac{\mu\eta_t}{2}\right)\left\|\overline{x}_t-x^*\right\|^2+\left(1-\frac{\theta}{8}\right)\cdot\frac{24L\eta_t}{m\theta}\Psi_t-\frac{7\eta_t}{8}\left(f(\overline{x}_t)-f(x^*)\right)\\
&\quad+\frac{2^4\cdot3^3\cdot L\eta_t^3}{\theta^2}\bar{\sigma}^2+\eta_t^2\cdot\frac{\bar{\sigma}^2}{m}\\
&\leq\left(1-\frac{\mu\eta_t}{2}\right)\left(\left\|\overline{x}_t-x^*\right\|^2+\frac{24L\eta_t}{m\theta}\Psi_t\right)-\frac{7\eta_t}{8}\left(f(\overline{x}_t)-f(x^*)\right)+\frac{2^4\cdot3^3\cdot L\eta_t^3}{\theta^2}\bar{\sigma}^2+\eta_t^2\cdot\frac{\bar{\sigma}^2}{m}\\
&\leq\exp\left(-\frac{\mu\eta_t}{2}\right)\left(\left\|\overline{x}_t-x^*\right\|^2+\frac{24L\eta_t}{m\theta}\Psi_t\right)-\frac{7\eta_t}{8}\left(f(\overline{x}_t)-f(x^*)\right)+\frac{2^4\cdot3^3\cdot L\eta_t^3}{\theta^2}\bar{\sigma}^2+\eta_t^2\cdot\frac{\bar{\sigma}^2}{m},
\end{aligned}
$$

where the last inequality is because of $1-x\leq\exp(-x)$ when $0<x<1$. □

C.5. Proof of Theorem 3.5

Proof of Theorem 3.5. For the case $\bar{\sigma}^2=0$, Eq. (16) reduces to

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\overline{x}_{t+1}-x^*\right\|^2+\frac{24L\eta_{t+1}}{m\theta}\Psi_{t+1}\mid\mathcal{F}_t\right]\\
&\leq\exp\left(-\frac{\mu\eta_t}{2}\right)\left(\left\|\overline{x}_t-x^*\right\|^2+\frac{24L\eta_t}{m\theta}\Psi_t\right)-\frac{7\eta_t}{8}\left(f(\overline{x}_t)-f(x^*)\right)\\
&\leq\exp\left(-\frac{\mu\eta_t}{2}\right)\left(\left\|\overline{x}_t-x^*\right\|^2+\frac{24L\eta_t}{m\theta}\Psi_t\right).
\end{aligned}
$$

Using above equation recursively and replacing $\eta_t=\frac{\theta}{2^6\cdot3\cdot L}$, we can obtain the first result.

The result for the case $\bar{\sigma}^2>0$ follows from Lemma B.4 with $e_t=f(\overline{x}_t)-f(x^*)$, $r_t=\left\|\overline{x}_t-x^*\right\|^2+\frac{24L\eta_t}{m\theta}\Psi_t$, $A=\frac{7}{8}$,

$B = \frac{\bar{\sigma}^2}{m}$, and $C = \frac{2^4 \cdot 3^3 \cdot L}{\theta^2}\bar{\sigma}^2$. Specifically, we have

$$
\begin{aligned}
&\frac{1}{S_T}\sum_{t=0}^{T}\omega_t\big(\mathbb{E}[f(\overline{x}_t)] - f(x^*)\big)\\
\overset{(52)}{\leq}&\frac{2^{24}\cdot 3^6\cdot L^3}{\theta^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{24L\eta_0}{m\theta}\Psi_0\right) + \frac{18\cdot 8\cdot L^3(L+\beta\mu(T+1))^2}{7\cdot\beta^2(L-\beta\mu)^3\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m}\\
&+\frac{2^9\cdot 3^6\cdot L^4}{7\mu^2\theta^2(L-\beta\mu)^3}\cdot\frac{\bar{\sigma}^2}{T^2}\\
\leq&\frac{2^{24}\cdot 3^6\cdot L^3}{\theta^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{\Psi_0}{8m}\right) + \frac{18\cdot 8\cdot L^3(L+\beta\mu(T+1))^2}{7\cdot(6L^3/7)\cdot\beta^2\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^9\cdot 3^6\cdot L^4}{7\mu^2\theta^2\cdot(6L^3/7)}\cdot\frac{\bar{\sigma}^2}{T^2}\\
=&\frac{2^{24}\cdot 3^6\cdot L^3}{\theta^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{\Psi_0}{8m}\right) + \frac{24\cdot(L+\beta\mu(T+1))^2}{\beta^2\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^8\cdot 3^5\cdot L}{\mu^2\theta^2}\cdot\frac{\bar{\sigma}^2}{T^2},
\end{aligned}
$$

where the second inequality is because of $0 < \theta < 1$ and $\mu \leq L$ and

$$
L - \beta\mu = L - \frac{\theta}{2^7\cdot 3^2}\mu \geq L - \frac{L}{2^7\cdot 3^2} \geq L\left(\frac{6}{7}\right)^{1/3}.
$$

□

C.6. Proof of Corollary 3.6

Proof of Corollary 3.6. If $\bar{\sigma}^2 = 0$, to achieve ε-suboptimality, by Eq. (17), it requires that

$$
\exp\left(-\frac{1}{2^7\cdot 3}\cdot\frac{\theta\mu}{L}\cdot T\right)\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{1}{8m}\Psi_0\right) \leq \varepsilon,
$$

which leads to

$$
\begin{aligned}
T &= \frac{2^7\cdot 3\cdot L}{\mu\theta}\log\frac{\|\overline{x}_0 - x^*\|^2 + \frac{1}{8m}\Psi_0}{\varepsilon}\\
&= \mathcal{O}\left(\frac{L}{\mu(1-\lambda_2(W))}\log\frac{1}{\varepsilon}\right),
\end{aligned}
$$

where the last equality is because of the fact that $\theta = 1 - \lambda_2(W)$ when W is a fixed mixing matrix.

If $\bar{\sigma}^2 > 0$, supposing T is sufficient large that $\beta\mu(T+1)$ dominates L, in this case, Eq. (18) reduces to

$$
\begin{aligned}
&\frac{1}{S_T}\sum_{t=0}^{T}\omega_t\big(\mathbb{E}[f(\overline{x}_t)] - f(x^*)\big)\\
=&\mathcal{O}\Big(\frac{1}{T}\cdot\frac{\bar{\sigma}^2}{m\mu} + \frac{L}{\mu^2\theta^2}\cdot\frac{\bar{\sigma}^2}{T^2} +\\
&\frac{L^3}{\theta^3\mu^3}\cdot\frac{1}{T^3}\cdot(\|\overline{x}_0 - x^*\|^2 + \frac{\Psi_0}{8m})\Big).
\end{aligned}
$$

Thus, to achieve ε-suboptimality, the iteration complexity is

$$
T = \mathcal{O}\left(\frac{\bar{\sigma}^2}{\mu m\varepsilon} + \frac{\sqrt{L}\bar{\sigma}}{\mu\theta\sqrt{\varepsilon}} + \frac{L}{\theta\mu\varepsilon^{1/3}}\right).
$$

Replacing $\theta = 1 - \lambda_2(W)$ to above equation concludes the proof. □

D. Proofs of Section 4

D.1. Proof of Lemma 4.3

Proof of Lemma 4.3. We prove the result by the induction. For the case $t = 0$, the it holds that $\sum_{i=1}^{m} \widetilde{\mathbf{x}}_0^{(i)} = \sum_{i=m+1}^{2m} \widetilde{\mathbf{x}}_0^{(i)}$ trivially by the initialization of Algorithm 2. Assuming that Eq. (26) holds for t, then we have

$$\begin{aligned}\sum_{i=1}^{m} \widetilde{\mathbf{x}}_{t+1}^{(i)} =&(1+\gamma)\mathbf{1}^\top W \widetilde{\mathbf{x}}_t^{(1:m)} - \gamma \mathbf{1}^\top W \widetilde{\mathbf{x}}_t^{(m+1:2m)} + \mathbf{1}^\top \left(\nabla F(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{q}_t, \xi_\tau)\right)\\ =&\mathbf{1}^\top W \widetilde{\mathbf{x}}_t^{(1:m)} + \mathbf{1}^\top \left(\nabla F(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{q}_t, \xi_\tau)\right)\\ =&\sum_{i=m+1}^{2m} \widetilde{\mathbf{x}}_{t+1}^{(i)}\end{aligned}$$

where the second equality is because of the induction assumption, the first and third equality are because of definition of $\widetilde{W}$ and update rule of $\widetilde{\mathbf{x}}_t$. The result $\sum_{i=1}^{m} \widetilde{\mathbf{s}}_t^{(i)} = \sum_{i=m+1}^{2m} \widetilde{\mathbf{s}}_t^{(i)}$ can be proved similarly.

Since $\sum_{i=1}^{m} \widetilde{\mathbf{s}}_t^{(i)} = \sum_{i=m+1}^{2m} \widetilde{\mathbf{s}}_t^{(i)}$, then Eq. (27) can be proved as the same to the one of Eq. (13).

□

D.2. Proof of Lemma 4.4

Proof of Lemma 4.4. First, for notation convenience, we denote that

$$A_{\#} := \begin{bmatrix} A \\ A \end{bmatrix}, \quad \forall\, A \in \mathbb{R}^{m\times d}.$$

By the update rule of $\widetilde{\mathbf{s}}_t$, we can obtain that

$$\begin{aligned}&\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_{t+1}\right\|^2\right]\\ =&(1-p)\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t\right\|^2 + p\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t + \widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t, \xi_t)_{\#} - \nabla F(\mathbf{q}_t, \xi_\tau)_{\#}\big)\right\|^2\right]\\ \le&(1+p)\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t\right\|^2 + 2p\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t, \xi_t)_{\#} - \nabla F(\mathbf{q}_t, \xi_\tau)_{\#}\right\|^2\right].\end{aligned}$$

Similar to above equation, we have

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t\right\|^2\right] \le (1+p)\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^2\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_{t-1}\right\|^2 + 2p\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_{t-1}, \xi_{t-1})_{\#} - \nabla F(\mathbf{q}_{t-1}, \xi_\tau)_{\#}\big)\right\|^2.$$

Using above equation recursively, we can obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_{t+1}\right\|^2\right]\\
\leq&(1+p)\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t\right\|^2+2p\mathbb{E}\left[\left\|\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\right\|^2\right]\\
\leq&\mathbb{E}\left[\sum_{i=0}^{t}2p(1+p)^{t-i}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^{t-i}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_i,\xi_i)_{\#}-\nabla F(\mathbf{q}_i,\xi_{\tau_i})_{\#}\big)\right\|^2\right]\\
&+(1+p)^{t+1}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^{t+1}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_{\#,0}\right\|^2\\
\overset{(25)}{\leq}&\mathbb{E}\left[\sum_{i=0}^{t}2\alpha p(1+p)^{t-i}\left(1-\tilde{\theta}\right)^{2(t-i)}\left\|\nabla F(\mathbf{x}_i,\xi_i)_{\#}-\nabla F(\mathbf{q}_i,\xi_{\tau_i})_{\#}\right\|^2\right]\\
&+\alpha(1+p)^{t+1}(1-\tilde{\theta})^{2(t+1)}\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_0\right\|^2\\
\leq&\mathbb{E}\left[\sum_{i=0}^{t}4\alpha p\left(1-\tilde{\theta}\right)^{t-i}\left\|\nabla F(\mathbf{x}_i,\xi_i)-\nabla F(\mathbf{q}_i,\xi_{\tau_i})\right\|^2\right]+2\alpha\left(1-\tilde{\theta}\right)^{t+1}\left\|\mathbf{\Pi}\mathbf{s}_0\right\|^2\\
\overset{(50)}{\leq}&\mathbb{E}\left[\sum_{i=0}^{t}4\alpha p\left(1-\tilde{\theta}\right)^{t-i}\left(\left\|\nabla F(\mathbf{x}_i)-\nabla F(\mathbf{q}_i)\right\|^2+2m\bar{\sigma}^2\right)\right]+2\alpha\left(1-\tilde{\theta}\right)^{t+1}\left\|\mathbf{\Pi}\mathbf{s}_0\right\|^2\\
=&\mathcal{E}_{s,t+1}
\end{aligned}
$$

where the forth inequality is because of $p\leq\theta$. By the definition of $\mathcal{E}_{s,t}$, we have

$$
\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t\right\|^2\right]\leq\mathcal{E}_{s,t} \tag{63}
$$

and

$$
\mathcal{E}_{s,t+1}\leq\left(1-\tilde{\theta}\right)\mathcal{E}_{s,t}+4\alpha p\left(\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2+2m\bar{\sigma}^2\right). \tag{64}
$$

Now, we are going to prove the consensus error related to $\widetilde{\mathbf{x}}_t$. First, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\left(\widetilde{\mathbf{x}}_t-\eta_t\Big(\widetilde{\mathbf{s}}_t+\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\Big)\right)\right\|^2\right]\\
\leq&\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}(\widetilde{\mathbf{x}}_t-\eta_t\widetilde{\mathbf{s}}_t)\right\|^2\right]+\eta_t^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\|^2\right]\\
&-2\eta_t\mathbb{E}\left[\left\langle\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{x}}_t,\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\widetilde{\mathbf{s}}_t+\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\rangle\right]\\
=&\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}(\widetilde{\mathbf{x}}_t-\eta_t\widetilde{\mathbf{s}}_t)\right\|^2\right]+\eta_t^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\|^2\right]\\
&-2\eta_t\left\langle\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}(\widetilde{\mathbf{x}}_t-\eta_t\widetilde{\mathbf{s}}_t),\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t)_{\#}-\nabla F(\mathbf{q}_t)_{\#}\big)\right\rangle\\
\leq&\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}(\widetilde{\mathbf{x}}_t-\eta_t\widetilde{\mathbf{s}}_t)\right\|^2\right]+\eta_t^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\|^2\right]\\
&+\frac{\tilde{\theta}}{2}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}(\widetilde{\mathbf{x}}_t-\eta_t\widetilde{\mathbf{s}}_t)\right\|+\frac{2\eta_t^2}{\tilde{\theta}}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t)_{\#}-\nabla F(\mathbf{q}_t)_{\#}\big)\right\|^2\\
=&\left(1+\frac{\tilde{\theta}}{2}\right)\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}(\widetilde{\mathbf{x}}_t-\eta_t\widetilde{\mathbf{s}}_t)\right\|^2+\eta_t^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\|^2\right]\\
&+\frac{2\eta_t^2}{\tilde{\theta}}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t)_{\#}-\nabla F(\mathbf{q}_t)_{\#}\big)\right\|^2\\
\leq&\left(1+\frac{\tilde{\theta}}{2}\right)\left(\left(1+\frac{\tilde{\theta}}{2}\right)\left\|\mathbf{\Pi}\widetilde{W}\mathbf{\Pi}\widetilde{\mathbf{x}}_t\right\|^2+\left(1+\frac{2}{\tilde{\theta}}\right)\eta_t^2\left\|\mathbf{\Pi}\widetilde{W}\mathbf{\Pi}\widetilde{\mathbf{s}}_t\right\|^2\right)\\
&+\eta_t^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\|^2\right]+\frac{2\eta_t^2}{\tilde{\theta}}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t)_{\#}-\nabla F(\mathbf{q}_t)_{\#}\big)\right\|^2\\
\leq&\left(1+\frac{\tilde{\theta}}{2}\right)^2\left\|\mathbf{\Pi}\widetilde{W}\mathbf{\Pi}\widetilde{\mathbf{x}}_t\right\|^2+\frac{6\eta_t^2}{\tilde{\theta}}\left\|\mathbf{\Pi}\widetilde{W}\mathbf{\Pi}\widetilde{\mathbf{s}}_t\right\|^2\\
&+\eta_t^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\|^2\right]+\frac{2\eta_t^2}{\tilde{\theta}}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t)_{\#}-\nabla F(\mathbf{q}_t)_{\#}\big)\right\|^2\\
=&\left(1+\frac{\tilde{\theta}}{2}\right)^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^2\widetilde{\mathbf{\Pi}}\left(\widetilde{\mathbf{x}}_{t-1}-\eta_{t-1}\Big(\widetilde{\mathbf{s}}_{t-1}+\nabla F(\mathbf{x}_{t-1},\xi_{t-1})_{\#}-\nabla F(\mathbf{q}_{t-1},\xi_\tau)_{\#}\Big)\right)\right\|^2\right]\\
&+\frac{6\eta_t^2}{\tilde{\theta}}\left\|\mathbf{\Pi}\widetilde{W}\mathbf{\Pi}\widetilde{\mathbf{s}}_t\right\|^2+\eta_t^2\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\big)\right\|^2\right]\\
&+\frac{2\eta_t^2}{\tilde{\theta}}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\big(\nabla F(\mathbf{x}_t)_{\#}-\nabla F(\mathbf{q}_t)_{\#}\big)\right\|^2,
\end{aligned}
$$

where the second and third inequality are because of Cauchy's inequality.

Using above equation recursively, we can obtain that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\left(\widetilde{\mathbf{x}}_t-\eta_t\Big(\widetilde{\mathbf{s}}_t+\nabla F(\mathbf{x}_t,\xi_t)_{\#}-\nabla F(\mathbf{q}_t,\xi_\tau)_{\#}\Big)\right)\right\|^2\right]\\
\leq&\left(1+\frac{\tilde{\theta}}{2}\right)^{2t}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^{t+1}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{x}}_0\right\|^2+\frac{6}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1+\frac{\tilde{\theta}}{2}\right)^{2(t-i)}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^{t+1-i}\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_i\right\|^2\\
&+\sum_{i=0}^{t}\eta_i^2\left(1+\frac{\tilde{\theta}}{2}\right)^{2(t-i)}\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^{t+1-i}\widetilde{\mathbf{\Pi}}\Big(\nabla F(\mathbf{x}_i,\xi_i)_{\#}-\nabla F(\mathbf{q}_i,\xi_{\tau_i})_{\#}\Big)\right\|^2\right]\\
&+\frac{2}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1+\frac{\tilde{\theta}}{2}\right)^{2(t-i)}\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}^{t+1-i}\widetilde{\mathbf{\Pi}}\Big(\nabla F(\mathbf{x}_i)_{\#}-\nabla F(\mathbf{q}_i)_{\#}\Big)\right\|^2\\
\leq&\alpha\left(1+\frac{\tilde{\theta}}{2}\right)^{2t}(1-\tilde{\theta})^{2(t+1)}\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{x}}_0\right\|^2+\frac{6\alpha}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1+\frac{\tilde{\theta}}{2}\right)^{2(t-i)}(1-\tilde{\theta})^{2(t+1-i)}\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_i\right\|^2\\
&+\sum_{i=0}^{t}\eta_i^2\left(1+\frac{\tilde{\theta}}{2}\right)^{2(t-i)}\Big(1-\tilde{\theta}\Big)^{2(t+1-i)}\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\left(\nabla F(\mathbf{x}_i,\xi_i)_{\#}-\nabla F(\mathbf{q}_i,\xi_{\tau_i})_{\#}\right)\right\|^2\right]\\
&+\frac{2}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1+\frac{\tilde{\theta}}{2}\right)^{2(t-i)}\Big(1-\tilde{\theta}\Big)^{2(t+1-i)}\left\|\widetilde{\mathbf{\Pi}}\Big(\nabla F(\mathbf{x}_i)_{\#}-\nabla F(\mathbf{q}_i)_{\#}\Big)\right\|^2\\
\leq&\alpha\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1}\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{x}}_0\right\|^2+\frac{6\alpha}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1-i}\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_i\right\|^2\\
&+\alpha\sum_{i=0}^{t}\eta_i^2\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1-i}\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\left(\nabla F(\mathbf{x}_i,\xi_i)_{\#}-\nabla F(\mathbf{q}_i,\xi_{\tau_i})_{\#}\right)\right\|^2\right]\\
&+\frac{2\alpha}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1-i}\left\|\widetilde{\mathbf{\Pi}}\Big(\nabla F(\mathbf{x}_i)_{\#}-\nabla F(\mathbf{q}_i)_{\#}\Big)\right\|^2\\
\overset{(50)}{\leq}&\alpha\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1}\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{x}}_0\right\|^2+\frac{6\alpha}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1-i}\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_i\right\|^2\\
&+\frac{3\alpha}{\tilde{\theta}}\sum_{i=0}^{t}\eta_i^2\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1-i}\left\|\widetilde{\mathbf{\Pi}}\Big(\nabla F(\mathbf{x}_i)_{\#}-\nabla F(\mathbf{q}_i)_{\#}\Big)\right\|^2\\
&+2\alpha m\bar{\sigma}^2\sum_{i=0}^{t}\eta_i^2\left(1-\frac{\tilde{\theta}}{2}\right)^{t+1-i}\\
=&\mathcal{E}_{x,t+1}.
\end{aligned}
$$

By the definition of $\mathcal{E}_{x,t+1}$, we can obtain the following inequality

$$
\begin{aligned}
&\mathcal{E}_{x,t+1}\\
\leq&\left(1-\frac{\tilde{\theta}}{2}\right)\mathcal{E}_{x,t}+\frac{3\alpha}{\tilde{\theta}}\left(1-\frac{\tilde{\theta}}{2}\right)\eta_t^2\left(\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{s}}_t\right\|^2+2\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2\right)+2\alpha m\bar{\sigma}^2\left(1-\frac{\tilde{\theta}}{2}\right)\eta_t^2\\
\leq&\left(1-\frac{\tilde{\theta}}{2}\right)\mathcal{E}_{x,t}+\frac{3\alpha}{\tilde{\theta}}\eta_t^2\left(\mathcal{E}_{s,t}+2\left\|\nabla F(\mathbf{x}_t)-\nabla F(\mathbf{q}_t)\right\|^2\right)+2\alpha m\bar{\sigma}^2\eta_t^2.
\end{aligned}
$$

By the update rule of $\widetilde{\mathbf{x}}_t$, it holds that

$$
\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{\mathbf{x}}_{t+1}\right\|^2\right]=\mathbb{E}\left[\left\|\widetilde{\mathbf{\Pi}}\widetilde{W}\widetilde{\mathbf{\Pi}}\left(\widetilde{\mathbf{x}}_t-\eta_t\big(\widetilde{\mathbf{s}}_t+\zeta_t\left(\nabla F(\mathbf{x}_t)_{\#}-\nabla F(\mathbf{q}_t)_{\#}\right)\big)\right)\right\|^2\right]\leq\mathcal{E}_{x,t+1}. \tag{65}
$$

□

D.3. Proof of Lemma 4.5

Proof of Lemma 4.5. By the definition of $\widetilde{\Psi}_{t+1}$, we have

$$
\begin{aligned}
&\mathcal{E}_{x,t+1} + C_{1,t+1}\cdot\mathcal{E}_{s,t+1} + C_{2,t+1}\cdot\left\|\nabla F(\mathbf{q}_{t+1}) - \nabla F(\mathbf{1}x^*)\right\|^2 \\
\overset{(29)(30)(60)}{\leq}& \left(1-\frac{\tilde{\theta}}{2}\right)\mathcal{E}_{x,t} + \left(1-\frac{\tilde{\theta}}{2}+\frac{3\alpha\eta_t^2}{\tilde{\theta}C_{1,t}}\right)C_{1,t}\cdot\mathcal{E}_{s,t} + (1-p)C_{2,t}\left\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\right\|^2 \\
&+ \left(\frac{6\alpha\eta_t^2}{\tilde{\theta}} + 4\alpha pC_{1,t}\right)\left\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{q}_t)\right\|^2 + \left(2\alpha m\eta_t^2 + 8\alpha mpC_{1,t}\right)\bar{\sigma}^2 \\
&+ 4mLpC_{2,t}\Big(f(\overline{x}_t) - f(x^*)\Big) + 2pL^2C_{2,t}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2 \\
\overset{(49)(28)}{\leq}& \left(1-\frac{\tilde{\theta}}{2}+\frac{24\alpha L^2\eta_t^2}{\tilde{\theta}} + 16\alpha pL^2C_{1,t} + 2pL^2C_{2,t}\right)\cdot\mathcal{E}_{x,t} \\
&+ \left(1-\frac{\tilde{\theta}}{2}+\frac{3\alpha\eta_t^2}{\tilde{\theta}C_{1,t}}\right)C_{1,t}\cdot\mathcal{E}_{s,t} \\
&+ \left(1-p+\frac{12\alpha\eta_t^2}{\tilde{\theta}C_{2,t}} + 8\alpha p\cdot\frac{C_{1,t}}{C_{2,t}}\right)C_{2,t}\left\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\right\|^2 \\
&+ \left(2\alpha m\eta_t^2 + 8\alpha mpC_{1,t}\right)\bar{\sigma}^2 + \left(4mLpC_{2,t} + \frac{48\alpha mL\eta_t^2}{\tilde{\theta}} + 32\alpha pmLC_{1,t}\right)\Big(f(\overline{x}_t) - f(x^*)\Big) \\
\leq& \left(1-\frac{\tilde{\theta}}{4}\right)\Big(\mathcal{E}_{x,t} + C_{1,t}\cdot\mathcal{E}_{s,t} + C_{2,t}\left\|\nabla F(\mathbf{q}_t) - \nabla F(\mathbf{1}x^*)\right\|^2\Big) \\
&\frac{2^{12}\cdot 3^2\cdot m\eta_t^2}{\tilde{\theta}}\cdot\bar{\sigma}^2 + \frac{2^{12}\cdot 3^2\cdot mL\eta_t^2}{\tilde{\theta}}\cdot\Big(f(\overline{x}_t) - f(x^*)\Big),
\end{aligned}
$$

where the last inequality is due to the setting of parameters. □

D.4. Proof of Lemma 4.6

Proof of Lemma 4.6. By Lemma 4.3, we can conclude that Lemma 3.3 still holds.

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\overline{x}_{t+1} - x^*\right\|^2 + \frac{48L\eta_{t+1}}{m\tilde{\theta}}\widetilde{\Psi}_{t+1}\right] \\
\overset{(15)(32)}{\leq}& \Big(1-\frac{\mu\eta_t}{2}\Big)\left\|\overline{x}_t - x^*\right\|^2 - 2\eta_t\left(1-2\eta_tL\right)\left(f(\overline{x}_t) - f(x^*)\right) \\
&+ \eta_t^2\cdot\frac{\bar{\sigma}^2}{m} + \frac{2L\eta_t\left(1+2\eta_tL\right)}{m}\left\|\mathbf{\Pi}\mathbf{x}_t\right\|^2 + \left(1-\frac{\tilde{\theta}}{4}\right)\frac{48L\eta_t}{m\tilde{\theta}}\widetilde{\Psi}_t \\
&+ \frac{2^{16}\cdot 3^2\cdot L\eta_t^3}{\tilde{\theta}^2}\bar{\sigma}^2 + \frac{2^{16}\cdot 3^2\cdot L^2\eta_t^3}{\tilde{\theta}^2}\cdot\left(f(\overline{x}_t) - f(x^*)\right) \\
\leq& \Big(1-\frac{\mu\eta_t}{2}\Big)\left\|\overline{x}_t - x^*\right\|^2 - 2\eta_t\left(1-2\eta_tL - \frac{2^{15}\cdot 3^2\cdot L^2\eta_t^2}{\tilde{\theta}^2}\right)\left(f(\overline{x}_t) - f(x^*)\right) \\
&+ \left(1-\frac{\tilde{\theta}}{4}\right)\frac{48L\eta_t}{m\tilde{\theta}}\widetilde{\Psi}_t + \frac{6L\eta_t\cdot\mathcal{E}_{x,t}}{m} + \eta_t^2\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^{16}\cdot 3^2\cdot L\eta_t^3}{\tilde{\theta}^2}\cdot\bar{\sigma}^2 \\
\overset{\eta_t\leq\frac{\tilde{\theta}}{2^8\cdot 3\cdot L}}{\leq}& \Big(1-\frac{\mu\eta_t}{2}\Big)\left(\left\|\overline{x}_t - x^*\right\|^2 + \frac{48L\eta_t}{m\tilde{\theta}}\cdot\widetilde{\Psi}_t\right) - \frac{7\eta_t}{8}\Big(f(\overline{x}_t) - f(x^*)\Big) + \eta_t^2\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^{16}\cdot 3^2\cdot L\eta_t^3}{\tilde{\theta}^2}\cdot\bar{\sigma}^2 \\
\leq& \exp\Big(-\frac{\mu\eta_t}{2}\Big)\left(\left\|\overline{x}_t - x^*\right\|^2 + \frac{48L\eta_t}{m\tilde{\theta}}\cdot\widetilde{\Psi}_t\right) - \frac{7\eta_t}{8}\Big(f(\overline{x}_t) - f(x^*)\Big) + \eta_t^2\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^{16}\cdot 3^2\cdot L\eta_t^3}{\tilde{\theta}^2}\cdot\bar{\sigma}^2,
\end{aligned}
$$

where the last inequality is because of $1-x \leq \exp(-x)$ for $0 \leq x < 1$. □

D.5. Proof of Theorem 4.7

Proof of Theorem 4.7. For the case $\bar{\sigma}^2 = 0$, Eq. (34) reduces to

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{x}_{t+1} - x^*\|^2 + \frac{48L\eta_{t+1}}{m\tilde{\theta}}\widetilde{\Psi}_{t+1}\right] \\
\leq &\exp\left(-\frac{\mu\eta_t}{2}\right)\left(\|\overline{x}_t - x^*\|^2 + \frac{48L\eta_t}{m\tilde{\theta}}\widetilde{\Psi}_t\right) - \frac{7\eta_t}{8}\big(f(\overline{x}_t) - f(x^*)\big) \\
\leq &\exp\left(-\frac{\mu\eta_t}{2}\right)\left(\|\overline{x}_t - x^*\|^2 + \frac{24L\eta_t}{m\tilde{\theta}}\widetilde{\Psi}_t\right).
\end{aligned}$$

Using above equation recursively and replacing $\eta_t = \frac{\theta}{2^8\cdot 3\cdot L}$, we can obtain the first result.

The result for the case $\bar{\sigma}^2 > 0$ follows from Lemma B.4 with $e_t = f(\overline{x}_t) - f(x^*)$, $r_t = \|\overline{x}_t - x^*\|^2 + \frac{48L\eta_t}{m\tilde{\theta}}\Psi_t$, $A = \frac{7}{8}$, $B = \frac{\bar{\sigma}^2}{m}$, and $C = \frac{2^{16}\cdot 3^3\cdot L}{\tilde{\theta}^2}\bar{\sigma}^2$. Specifically, we have

$$\begin{aligned}
&\frac{1}{S_T}\sum_{t=0}^{T}\omega_t\big(\mathbb{E}[f(\overline{x}_t)] - f(x^*)\big) \\
\overset{(52)}{\leq} &\frac{2^{28}\cdot 3^5\cdot L^3}{\tilde{\theta}^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{48L\eta_0}{m\tilde{\theta}}\widetilde{\Psi}_0\right) + \frac{18\cdot 8\cdot L^3(L+\tilde{\beta}\mu(T+1))^2}{7\cdot\tilde{\beta}^2(L-\tilde{\beta}\mu)^3\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m} \\
&+ \frac{2^{21}\cdot 3^5\cdot L^4}{7\mu^2\theta^2(L-\tilde{\beta}\mu)^3}\cdot\frac{\bar{\sigma}^2}{T^2} \\
\leq &\frac{2^{28}\cdot 3^5\cdot L^3}{\tilde{\theta}^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{\widetilde{\Psi}_0}{16m}\right) + \frac{18\cdot 8\cdot L^3(L+\tilde{\beta}\mu(T+1))^2}{7\cdot(6L^3/7)\cdot\tilde{\beta}^2\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^{21}\cdot 3^5\cdot L^4}{7\mu^2\theta^2\cdot(6L^3/7)}\cdot\frac{\bar{\sigma}^2}{T^2} \\
= &\frac{2^{28}\cdot 3^5\cdot L^3}{\tilde{\theta}^3\mu^2}\cdot\frac{1}{T^3}\cdot\left(\|\overline{x}_0 - x^*\|^2 + \frac{\widetilde{\Psi}_0}{16m}\right) + \frac{24\cdot(L+\tilde{\beta}\mu(T+1))^2}{\tilde{\beta}^2\mu^3T^3}\cdot\frac{\bar{\sigma}^2}{m} + \frac{2^{20}\cdot 3^4\cdot L}{\mu^2\tilde{\theta}^2}\cdot\frac{\bar{\sigma}^2}{T^2},
\end{aligned}$$

where the second inequality is because of $0 < \theta < 1$ and $\mu \leq L$ and

$$L - \tilde{\beta}\mu = L - \frac{\theta}{2^7\cdot 3^2}\mu \geq L - \frac{L}{2^7\cdot 3^2} \geq L\left(\frac{6}{7}\right)^{1/3}.$$

□

E. Decentralized Stochastic Gradient Tracking

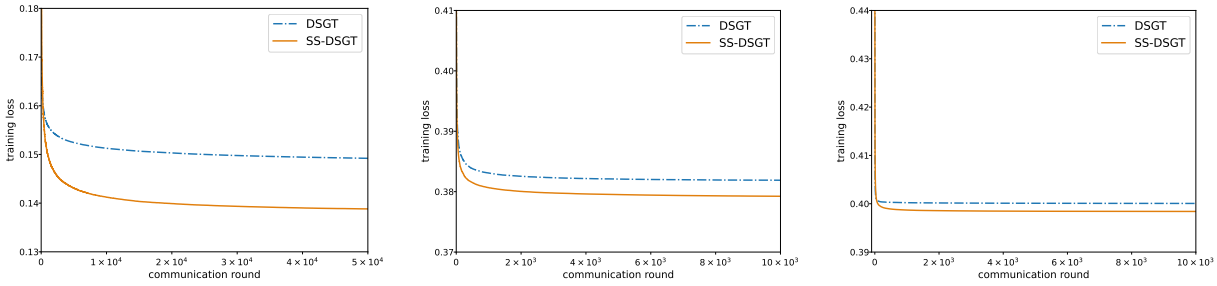
In Algorithm 3, we present the algorithm of DSGT.

Algorithm 3 Decentralized Stochastic Gradient Tracking

Input: x_0, mixing matrix W, initial step size η.
Initialization: Set $\mathbf{x}_0 = \mathbf{1}x_0$, $\mathbf{q}_0 = \mathbf{1}x_0$, $\mathbf{s}_0^{(i)} = \nabla f_i(\mathbf{x}_0^{(i)}, \xi_0)$, in parallel for $i \in [m]$, $\tau = 0$.
for $t = 1, \ldots, T$ **do**
Update

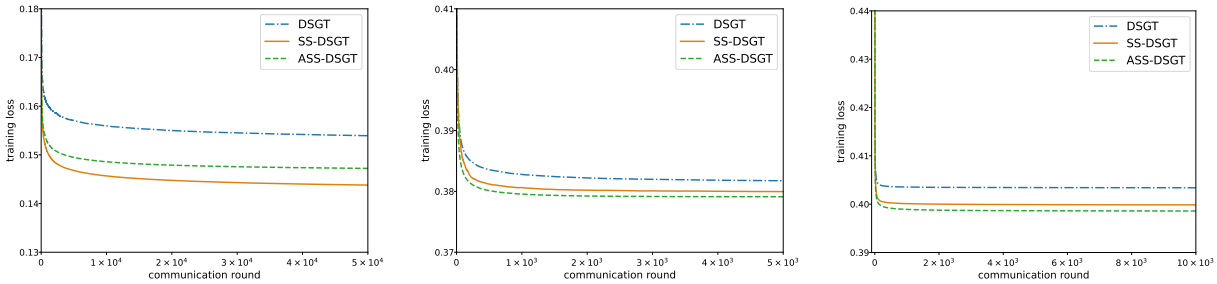
$$\mathbf{x}_{t+1} = W\left(\mathbf{x}_t - \eta_t\mathbf{s}_t\right), \tag{66}$$
$$\mathbf{s}_{t+1} = W\mathbf{s}_t + \nabla F(\mathbf{x}_{t+1}, \xi_{t+1}) - \nabla F(\mathbf{x}_t, \xi_t). \tag{67}$$

end for



(a) banknote (b) a9a (c) ijcnn1

Figure 3. Comparison of SS_DSGT and DSGT for the training loss versus the communication round on the asymmetric mixing matrix W^{asy}. The optimal values $f(x^*)$ on these three data sets are 0.1336, 0.3727, and 0.3965, respectively.



(a) banknote (b) a9a (c) ijcnn1

Figure 4. Comparison of different methods for the training loss versus the communication round on the symmetric mixing matrix W^{sy}. The optimal values $f(x^*)$ on these three data sets are 0.1335, 0.3727, and 0.3965, respectively.

F. Additional Experiment Results

In Figure 3 and Figure 4, we compare the training loss of the methods across these three data sets on the asymmetric mixing matrix W^{asy} and symmetric mixing matrix W^{sy}, respetively.