Gradient Tracking Methods for Distributed Stochastic Optimization Problems with Decision-dependent Distributions

Licheng Deng and Yongchao Liu*

Abstract. This paper aims to seek the performative stable solution and the optimal solution of the distributed stochastic optimization problem with decision-dependent distributions, which is a finite-sum stochastic optimization problem over a network and the distribution depends on the decision variables. For the performative stable solution, we provide an algorithm, DSGTD-GD, which combines the distributed stochastic gradient tracking descent method with the greedy deployment scheme. Under the constant step size policy, we show that the iterates generated by DSGTD-GD converge linearly, in expectation, to a neighborhood of the performative stable solution. Under the diminishing step size policy, we show that the iterates generated by DSGTD-GD converge to the performative stable solution with rate $\mathcal{O}\left(\frac{1}{k}\right)$. Moreover, we establish that the deviation between the averaged iterates of DSGTD-GD and the performative stable solution converges in distribution to a normal random vector. For the optimal solution, we provide an algorithm, DSGTD-AG, which combines the distributed stochastic gradient tracking descent method with the adaptive gradient scheme. Under the constant step size policy, we show that the iterates generated by DSGTD-AG converge to a stationary solution with rate of $\mathcal{O}(\frac{\ln K}{\sqrt{K}})$, where K is the number of iterations. The effectiveness of DSGTD-GD and DSGTD-AG is further demonstrated numerically with synthetic and real-world data.

Key words. Stochastic optimization with decision-dependent distributions, gradient tracking method, performative stable solution, optimal solution, asymptotic normality.

1 Introduction

Distributed stochastic optimization problems have attracted much attention in recent years due to their many applications such as large-scale machine learning [3, 4], sensor networks [2, 7] and parameter estimation [42]. Traditional stochastic optimization problems crucially rely on the assumption that data follows a static distribution. However, many real-world tasks are dynamical, involving data that could be influenced by the decision [11, 17], which may be characterized by the following distributed stochastic optimization problem with decision-

*School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China, e-mail: denglicheng@mail.dlut.edu.cn (Licheng Deng), lyc@dlut.edu.cn (Yongchao Liu)

dependent distributions (distributed SO-DD)

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \mathbf{E}_{\xi_i\sim\mathcal{D}_i(x)} \left[l_i(x;\xi_i)\right], \tag{1}$$

where $\mathbf{E}_{\xi_i\sim\mathcal{D}_i(x)} \left[l_i(x;\xi_i)\right]$ is the cost of the ith agent with respect to the samples from the ith population of users and $l_i : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is measurable function parameterized by decision vector $x \in \mathbb{R}^d$ and random variable ξ_i. Different from the traditional distributed stochastic optimization problems, there is an interactive mechanism between agents and the populations of users: the ith agent draws samples from the ith population of users, while the samples are characterized by the distribution map $\mathcal{D}_i(x)$ supported on $\Xi \subseteq \mathbb{R}^p$. In other words, the samples are parameterized by the agent's decision vector.

SO-DD can be traced to early stochastic programming [1, 15, 21]. Since the distribution depends on the argument x, SO-DD is usually a non-convex problem and difficult to minimize. More recently, [35] propose the concept of two solutions, namely performative stable solution

$$x_{ps} := \arg\min_x \mathbf{E}_{\xi\sim\mathcal{D}(x_{ps})} \left[l(x;\xi)\right], \tag{2}$$

and optimal solution

$$x_{op} := \arg\min_x \mathbf{E}_{\xi\sim\mathcal{D}(x)} \left[l(x;\xi)\right]. \tag{3}$$

Under some mild conditions, the authors show that the performative stable solution is close to the optimal solution of SO-DD. Moreover, they propose two algorithms, repeated risk minimization and repeated gradient descent, to compute the performative stable solution.

Since the seminal work [35], there is a growing literature [5, 12, 20, 24, 27, 29, 30, 40, 43] in analyzing SO-DD. Most works [5, 12, 24, 29, 40, 43] focus on the performative stable solution. Based on the repeated gradient descent method [35], [29] propose two stochastic approximation-based algorithms, one is the greedy deployment method and another is the lazy deployment method. The algorithm with greedy deployment performs an update at each iteration where data is from the current distribution. Contrary to the one with greedy deployment, the algorithm with lazy deployment starts from a decision x_k and performs $n_k > 1$ updates using samples from the distribution $\mathcal{D}(x_k)$. For the standard stochastic gradient method with decreasing step size, [29] show that the algorithm with greedy deployment converges to performative stable solution at rate $\mathcal{O}(\frac{1}{k})$, while the one with lazy deployment converges at rate $\mathcal{O}(\frac{1}{k^a})$ provided that drawing $\mathcal{O}(k^{1.1a})$ samples between the kth and $(k+1)$th iterate, where a is any positive constant. [12] show that typical stochastic algorithms—originally designed for static problems—can be applied directly for finding such equilibria with little loss in efficiency. As opposed to the setting in [29] with independent and identically distributed samples taken from the shifted distribution, [24] consider the setting that the population-provided samples adapted to the agent's and users' previous states. They propose a stochastic gradient descent algorithm, which can be modeled with biased stochastic gradients driven by a controlled Markov chain, and show that the algorithm converges to the performative stable solution with rate $\mathcal{O}(\frac{1}{k})$. [5] propose a theoretical framework where the population-provided samples are modeled as a function of the agent and the current state (distribution) of the population. The authors analyze the necessary and sufficient conditions that repeated risk minimization method converges to the performative

stable solution. [40] propose two versions of the stochastic gradient descent algorithm for the setting in which the decision-maker has a first-order gradient oracle or simply a loss function oracle, where data distribution is decision-dependent and evolves dynamically with time according to a geometric decay process. [43] consider a time-varying stochastic saddle point problem with decision-dependent distributions and propose an online stochastic primal-dual algorithm for tracking the performative stable point.

Under some standard assumptions, the performative stable point and the optimal point lie in a small neighborhood around each other [35]. In general, the performative stable point may be far from optimal and even fail to converge at all [20]. In the past years, some algorithms are proposed to seek the optimal solution. [20] introduce a performative gradient descent algorithm with a finite difference method to estimate the part of the gradient with distribution shift under a parametric condition that the distribution can be approximately expressed to a mixture of Gaussian distribution. [30] develop a two-stage algorithm, where the distribution map is estimated via the least squares method in the first stage and then the algorithm directly solves the performative risk with the estimated distribution map in the second stage, under the assumption that the distribution map is a global linear structure with respect to the decision variable. [27] consider the SO-DD without the global structure of the distribution map $\mathcal{D}(x)$. They develop a two-step derivative-free method for seeking the optimal point, where the distribution map is predicted by using local linear regression at the current iterate and then a candidate solution is sought by the trust region method. [32] study the decision-dependent game with the condition that $\mathcal{D}(x)$ is a global linear structure. Different from [30], they propose a stochastic gradient descent algorithm with the adaptive gradient method, which alternately runs a adaptive gradient step for learning the distribution map $\mathcal{D}(x)$ and a stochastic gradient descent step for updating the decision variable at each iteration.

More recently, the multi-agent stochastic optimization problems with decision-dependent distributions are studied [25, 32, 36]. [25] consider the case where n agents seek a common decision vector that minimizes the sum of loss functions over an undirected and connected communication network, while the agents acquire data that react to the agent's decisions from different population of users, which is precisely the distributed SO-DD (1). On the other hand, [36] and [32] reformulate the multi-agent problems as the non-cooperative games. Theoretically, [25] provide the necessary and sufficient condition that distributed SO-DD admits a unique performative stable solution. Numerically, they propose a distributed stochastic gradient descent method with greedy deployment (DSGD-GD) and show that the proposed method achieves convergence rate $\mathcal{O}(\frac{1}{k})$.

Compared with distributed gradient method, the gradient tracking descent method aims to track the averaged stochastic gradient via the agent-based auxiliary variables instead of the local gradient in the distributed stochastic gradient descent method, which is more robustness to heterogeneous data and maintains the stability of the performance. In the past few years, the gradient tracking technique has been extensively adopted in distributed optimization [23, 28, 33, 34, 38, 39]. This motivates us to propose two gradient tracking-based methods to seek the performative stable solution and the optimal solution of distributed SO-DD. As far as we are concerned, the contribution of the paper can be summarized as follows.

- We provide a distributed stochastic gradient tracking descent method with the greedy deployment (DSGTD-GD) scheme to seek the performative stable solution. Under the constant step size policy, we show that the iterates of DSGTD-GD achieve a linear convergence rate to a neighborhood of the performative stable solution, where the neighborhood size is proportional to both the step size and the spectral norm related to the communication network.
- Under the diminishing step size policy, we show that the square of the distances between the iterates generated by DSGTD-GD and the performative stable solution converges to zero with rate $\mathcal{O}\left(\frac{1}{k}\right)$. Furthermore, we show that the deviation between the averaged iterates of DSGTD-GD and the performative stable solution converges in distribution to a normal random vector. As far as we know, the result seems to be the first asymptotic normality of stochastic approximation-based method for the distributed SO-DD.
- We provide a distributed stochastic gradient tracking descent method with the adaptive gradient (DSGTD-AG) scheme to seek the optimal solution of non-convex distributed SO-DD. Under the constant step size policy, we show that the iterates of DSGTD-AG converge to the optimal solution with rate of $\mathcal{O}(\frac{\ln K}{\sqrt{K}})$, where K is the number of iterations. The effectiveness of DSGTD-GD and DSGTD-AG is further demonstrated numerically with synthetic and real-world data.

The rest of this paper is organized as follows. Section 2 introduces DSGTD-GD and DSGTD-AG for the distributed SO-DD and presents some standard assumptions. Section 3 studies the convergence of DSGTD-GD, where Subsection 3.1 focuses on DSGTD-GD with constant step size and Subsection 3.2 focuses on DSGTD-GD with diminishing step size. The convergence analysis of DSGTD-AG is presented in Section 4. Numerical experiments are provided in Section 5.

Throughout this paper, vectors default to columns if not otherwise specified. $\mathbb{R}^d$ denotes the d-dimension Euclidean space endowed with norm $\|x\| = \sqrt{\langle x, x\rangle}$. Denote $\mathbf{1} := (1\ 1 \dots 1)^\intercal \in \mathbb{R}^d$ and $\mathbf{0} := (0\ 0 \dots 0)^\intercal \in \mathbb{R}^d$. $\mathbf{I} \in \mathbb{R}^{n\times n}$ stands for the identity matrix. The inner product of two matrices A, B is denoted by $\langle A, B\rangle$. For matrices, $\|\cdot\|$ and $\|\cdot\|_2$ represent the Frobenius norm and 2-norm. Denote $f(x; x') := \sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i(x')}\left[l_i(x;\xi_i)\right]$ and $f(x) := \sum_{i=1}^n f_i(x)$, where $f_i(x) := \mathbf{E}_{\xi_i\sim\mathcal{D}_i(x)}\left[l_i(x;\xi_i)\right]$. To clarify the expression, we denote $\nabla l_i(x;\xi)$, $\nabla f(x;x')$ as the gradients taken with respect to the first argument x, $\nabla_\xi l_i(x;\xi)$ as the gradients taken with respect to the second argument ξ and $\nabla f(x)$, $\nabla f_i(x)$ as the gradient taken with respect to x. For the given σ-algebra $\mathcal{F}$, $\mathbf{E}[\cdot|\mathcal{F}]$ is conditional expectation on $\mathcal{F}$. For a sequence of random vectors $\{\xi_k\}$ and a random vector ξ, $\xi_k \xrightarrow{d} \xi$ denotes the convergence in distribution and $\mathrm{Cov}(\xi)$ denotes the covariance matrix of random vector ξ. $\mathcal{N}(z, \Sigma)$ is normal distribution with mean z and covariance matrix Σ. For kth iteration of ith agent, we use the notation $x_{i,k}$, $y_{i,k}$ $\xi_{i,k}$ to denote the iterates, the auxiliary variables and the samples, respectively. Finally, we use bold letter to represent the matrix with rows $z_{i,k}$, i.e.

$$\mathbf{z}_k := \begin{bmatrix} z_{1,k}^\intercal \\ \vdots \\ z_{n,k}^\intercal \end{bmatrix} \in \mathbb{R}^{n\times d}, \tag{4}$$

where $z_{i,k}$ could be $x_{i,k}$, $y_{i,k}$, $\xi_{i,k}$, $\bar{x}_k$, $\bar{y}_k$ and $\bar{x}_k := \frac{1}{n}\sum_{i=1}^n x_{i,k}$, $\bar{y}_k := \frac{1}{n}\sum_{i=1}^n y_{i,k}$.

2 Preliminaries and algorithms

In this section, we introduce DSGTD-GD and DSGTD-AG in Subsection 2.1 and Subsection 2.2, respectively. In Subsection 2.3, we give assumptions to support the algorithms. Moreover, throughout this paper, we consider a set of agents $\mathcal{V} = \{1, ..., n\}$ connected on a communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents links or edges among the agents. The corresponding weight matrix $\mathbf{W} = [w_{ij}]_{n\times n}$ satisfies that $w_{ij} > 0$ if $(i, j) \in \mathcal{E}$, otherwise $w_{ij} = 0$. We assume the communication network and the weight matrix satisfy the following condition.

Assumption 1. [38] [**Networks and weight matrices**] The network is undirected and strongly connected. The matrix $\mathbf{W}$ is symmetric and doubly stochastic, i.e., $\mathbf{W1} = \mathbf{W}^\intercal\mathbf{1} = \mathbf{1}$, and the diagonal entries of $\mathbf{W}$ are positive.

Assumption 1 is a standard condition on the network and the matrix $\mathbf{W}$ [19, 38, 44] and implies that $\mathbf{11}^\intercal\mathbf{W} = \mathbf{W11}^\intercal = \mathbf{11}^\intercal$, the spectral norm ρ of the matrix $\mathbf{W} - \frac{1}{n}\mathbf{11}^\intercal$ satisfies $\rho \in (0, 1)$ [38, Lemma 1].

2.1 DSGTD-GD

DSGTD-GD is to seek the performative stable solution

$$x^\star := \arg\min_x \sum_{i=1}^n \mathbf{E}_{\xi_i\sim\mathcal{D}_i(x^\star)} \left[l_i(x; \xi_i)\right]. \tag{5}$$

In what follows, we present DSGTD-GD in Algorithm 1.

For each iteration k, the scheme includes three phases: (i) the ith agent updates its vector $x_{i,k}$ by mixing its intermediate variable and ones from its neighbors, the current decisions adjusted along the direction of average gradient estimates $y_{j,k}$, $j \in \{i\} \cup \mathcal{N}_i$, with specific weights, where $\mathcal{N}_i$ denotes the set of all neighbors of the ith agent; (ii) the ith population of users reveals a sample $\xi_{i,k+2}$ to the ith agent; (iii) each agent updates its gradient tracker $y_{i,k+1}$ by mixing its own and its neighbors' gradient trackers $y_{j,k}$, $j \in \{i\} \cup \mathcal{N}_i$, with specific weights and adding stochastic gradient-difference term $\nabla l_i(x_{i,k+1}; \xi_{i,k+2}) - \nabla l_i(x_{i,k}; \xi_{i,k+1})$, where $\nabla l_i(x_{i,k+1}; \xi_{i,k+2})$ takes into account only the new information.

For ease of presentation, we write Algorithm 1 in a compact form as

$$\begin{aligned} \mathbf{x}_{k+1} &:= \mathbf{W}\left(\mathbf{x}_k - \gamma\mathbf{y}_k\right), \\ \mathbf{y}_{k+1} &:= \mathbf{W}\mathbf{y}_k + \mathbf{L}(\mathbf{x}_{k+1}; \xi_{k+2}) - \mathbf{L}(\mathbf{x}_k; \xi_{k+1}), \end{aligned} \tag{8}$$

where $\mathbf{W} = [w_{ij}]_{n\times n}$ denotes the coupling matrix of agents and

$$\mathbf{L}(\mathbf{x}_k; \xi_{k+1}) := \begin{bmatrix} \nabla l_1(x_{1,k}; \xi_{1,k+1})^\intercal \\ \vdots \\ \nabla l_n(x_{n,k}; \xi_{n,k+1})^\intercal \end{bmatrix}. \tag{9}$$

Algorithm 1 DSGTD with Greedy Deployment (DSGTD-GD):

Require: initial values $x_{i,0} \in \mathbb{R}^d$, $\xi_{i,1} \sim \mathcal{D}_i(x_{i,0})$, $y_{i,0} = \nabla l_i(x_{i,0}; \xi_{i,1})$ for any $i \in \mathcal{V}$; step size $\gamma > 0$ can be either constant or diminishing; non-negative weight matrices $\mathbf{W} = [w_{ij}]_{n\times n}$.

1: **For** $k = 0, 1, 2, \cdots$ **do**

2: **The ith agent's decision vector update:** for any $i \in \mathcal{V}$,

$$x_{i,k+1} = \sum_{j=1}^{n} w_{ij} \left(x_{j,k} - \gamma y_{j,k}\right). \tag{6}$$

3: **The ith agent's sample update:** for any $i \in \mathcal{V}$, draw $\xi_{i,k+2} \sim \mathcal{D}_i(x_{i,k+1})$.

4: **Gradient tracking update:** for any $i \in \mathcal{V}$,

$$y_{i,k+1} = \sum_{j=1}^{n} w_{ij} y_{j,k} + \nabla l_i(x_{i,k+1}; \xi_{i,k+2}) - \nabla l_i(x_{i,k}; \xi_{i,k+1}). \tag{7}$$

5: **end for**

Obviously, by the double stochasticity of $\mathbf{W}$, we have

$$\bar{x}_{k+1} = \bar{x}_k - \gamma \bar{y}_k, \qquad \bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} \nabla l_i(x_{i,k}; \xi_{i,k+1}), \tag{10}$$

where $\bar{x}_k := \frac{1}{n}\sum_{i=1}^{n} x_{i,k}$ and $\bar{y}_k := \frac{1}{n}\sum_{i=1}^{n} y_{i,k}$.

2.2 DSGTD-AG

DSGTD-AG is to seek the optimal solution

$$x^{\star\star} := \arg\min_{x} \sum_{i=1}^{n} \mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} \left[l_i(x; \xi_i)\right]. \tag{11}$$

Since the problem (1) is potentially a nonconvex optimization problem, we resort to seeking a stationary solution to the problem (1). In what follows, we present DSGTD-AG in Algorithm 2.

For each iteration k, the scheme includes four phases: (i) the ith agent forms a new estimation of the distribution map $\mathcal{D}_{i,k+1}(\cdot)$ based on the current estimation $\mathcal{D}_{i,k}(\cdot)$ and the decision variables $x_{i,k}$ by making use of the data samples generated from the real distribution map; (ii) the ith agent updates its vector $x_{i,k}$ in the same way as the first phase in Algorithm 1; (iii) the ith population of users reveals sample $\xi_{i,k+2}$ to the ith agent; (iv) the ith agent updates its gradient tracker $y_{i,k+1}$ by mixing its own and its neighbors' gradient trackers $y_{j,k}$, $j \in \{i\} \cup \mathcal{N}_i$, with specific weights and adding stochastic gradient-difference term $v_{i,k+1} - v_{i,k}$, where $v_{i,k}$ takes into account the complete gradient information of the loss function $l_i(x_{i,k}, \xi_{i,k+1})$ based on $\mathcal{D}_{i,k}(x_{i,k})$.

Algorithm 2 DSGTD with Adaptive Gradient (DSGTD-AG):

Require: initial values $x_{i,0} \in \mathbb{R}^d$, $\xi_{i,1} \sim \mathcal{D}_{i,0}(x_{i,0})$ and $y_{i,0} = \nabla l_i(x_{i,0}; \xi_{i,1})$, for any $i \in \mathcal{V}$; step sizes $\gamma > 0$; non-negative weight matrices $\mathbf{W} = [w_{ij}]_{n\times n}$.

1: **For** $k = 0, 1, 2, \cdots$ **do**

2: **The ith agent's distribution map update:** for any $i \in \mathcal{V}$,

$$\mathcal{D}_{i,k+1}(\cdot) = \pi\left(\mathcal{D}_{i,k}(\cdot), x_{i,k}, \xi_{i,k+1}\right). \tag{12}$$

3: **The ith agent's decision vector update:** for any $i \in \mathcal{V}$,

$$x_{i,k+1} = \sum_{j=1}^{n} w_{ij}\left(x_{j,k} - \gamma y_{j,k}\right). \tag{13}$$

4: **The ith agent's sample update:** for any $i \in \mathcal{V}$, draw $\xi_{i,k+2} \sim \mathcal{D}_i(x_{i,k+1})$.

5: **Gradient tracking update:** for any $i \in \mathcal{V}$,

$$y_{i,k+1} = \sum_{j=1}^{n} w_{ij} y_{j,k} + v_{i,k+1} - v_{i,k}. \tag{14}$$

6: **end for**

Similar to (8) and (10), we write (13) (14) in Algorithm 1 in a compact form as

$$\begin{aligned} \mathbf{x}_{k+1} &:= \mathbf{W}\left(\mathbf{x}_k - \gamma \mathbf{y}_k\right), \\ \mathbf{y}_{k+1} &:= \mathbf{W}\mathbf{y}_k + \mathbf{v}_{k+1} - \mathbf{v}_k, \end{aligned} \tag{15}$$

and then

$$\bar{x}_{k+1} = \bar{x}_k - \gamma \bar{y}_k, \qquad \bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} v_{i,k}, \tag{16}$$

where $\mathbf{W} = [w_{ij}]_{n\times n}$ denotes the coupling matrix of agents, $\bar{x}_k := \frac{1}{n}\sum_{i=1}^{n} x_{i,k}$ and $\bar{y}_k := \frac{1}{n}\sum_{i=1}^{n} y_{i,k}$.

2.3 Assumptions

For studying the convergence of Algorithms 1 and 2, we need the following assumptions.

Assumption 2. [25] [**Objective function and stochastic gradient**] Let $x^\star$ be the performative stable solution of problem (1).

(a) Fix any $\bar{x} \in \mathbb{R}^d$, function $f(x; \bar{x})$ is $\mu - strongly\ convex$ ($\mu > 0$), that is,

$$\langle \nabla f(x; \bar{x}) - \nabla f(x', \bar{x}), x - x' \rangle \geq \mu \|x - x'\|^2, \quad \forall x, x' \in \mathbb{R}^d. \tag{17}$$

(b) For any $i \in \mathcal{V}$, the loss function $l_i(x; \xi)$ is $L - smooth$ ($L > 0$), that is,

$$\|\nabla l_i(x; \xi) - \nabla l_i(x'; \xi')\| \leq L\{\|x - x'\| + \|\xi - \xi'\|\}, \quad \forall x, x' \in \mathbb{R}^d, \xi, \xi' \in \Xi. \tag{18}$$

(c) Fix any $x \in \mathbb{R}^d$,

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} \left[\nabla l_i(x;\xi_i)\right] = \nabla \mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} \left[l_i(x;\xi_i)\right], \quad \forall i \in \mathcal{V}. \tag{19}$$

(d) For any $i \in \mathcal{V}$ and $x \in \mathbb{R}^d$, there exists $\sigma > 0$ such that

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} \left[\| \nabla l_i(x;\xi_i) - \nabla \mathbf{E}_{\xi_i' \sim \mathcal{D}_i(x)} \left[l_i(x;\xi_i') \right] \|^2 \right] \leq \sigma^2 (1 + \|x - x^\star\|^2). \tag{20}$$

(e) There exist constants $p \geq 2$ and $c_g > 0$ such that

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x^\star)} \left[\|\nabla l_i(x^\star;\xi_i)\|^p \right] \leq c_g^{\frac{p}{2}}. \tag{21}$$

(f) $\nabla R(x)$ is positive definite, where $R(x) := \sum_{i=1}^n \nabla f_i(x;x)$.

Assumption 3. [25] [**Lipschitz distribution**] For any $i \in \mathcal{V}$, there exists $\epsilon_i > 0$ such that

$$\mathcal{W}_1 \left(\mathcal{D}_i(x), \mathcal{D}_i(x') \right) \leq \ \epsilon_i \ \|x - x'\|, \quad \forall x, x' \in \mathbb{R}^d, \tag{22}$$

where $\mathcal{W}_1(\mathcal{D}, \mathcal{D}')$ denotes the Wasserstein-1 distance between the distributions $\mathcal{D}$, $\mathcal{D}'$.

Assumption 4. [9] [**Joint smoothness**] Let $x^\star$ be the performative stable solution of problem (1). The map $x \mapsto \sum_{i=1}^n \nabla \mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} \left[l_i(x;\xi_i) \right]$ is smooth and has a Lipschitz continuous Jacobian on a neighborhood of $x^\star$.

Assumption 5. [32] [**Objective function and gradient**]

(a) For any $i \in \mathcal{V}$, the gradient $\nabla f_i(x)$ are $L - Lipschitz\ continuous$ $(L > 0)$, that is,

$$\|\nabla f_i(x) - \nabla f_i(x')\| \leq L \|x - x'\|, \quad \forall x, x' \in \mathbb{R}^d. \tag{23}$$

(b) For any $i \in \mathcal{V}$ and $x \in \mathbb{R}^d$, there exists $\delta > 0$ such that

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} \left[\|\nabla l_i(x;\xi_i)\| \right] \leq \delta. \tag{24}$$

(c) For any $i \in \mathcal{V}$ and $x \in \mathbb{R}^d$, there exists $\sigma > 0$ such that

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} \left[\| \nabla_{x,\xi_i} l_i(x;\xi_i) - \mathbf{E}_{\xi_i' \sim \mathcal{D}_i(x)} \left[\nabla_{x,\xi_i'} l_i(x;\xi_i') \right] \|^2 \right] \leq \sigma^2. \tag{25}$$

Assumption 6. [32] [**Distribution map**] For any $i \in \mathcal{V}$, there exist probability measures $\mathcal{P}_i$ such that

$$\xi_i \sim \mathcal{D}_i(x) \Longleftrightarrow \xi_i = A_i x + \zeta_i, \ \zeta_i \sim \mathcal{P}_i, \tag{26}$$

where $A_i \in \mathbb{R}^{m \times d}$, $\mu_i := \mathbf{E}_{\zeta_i \sim \mathcal{P}_i} \zeta_i$ and $\Sigma_i := \mathbf{E}_{\zeta_i \sim \mathcal{P}_i} [(\zeta_i - \mu_i)(\zeta_i - \mu_i)]$.

In Assumption 2, conditions (a) and (b) are widely used for the SO-DD [12, 25, 29] and require the loss functions to be strongly convex and smooth; (c) is an unbiasedness condition and (d) bounds the variance of the stochastic gradient $\nabla l_i(x;\xi_i)$; conditions (e) and (f) are for establishing the asymptotic normality of the averaged iterates generated by DSGTD-GD for each agent [13, 37]. Assumption 3 implies that the amount of distribution shift resulting from

the reaction of the ith population to the agent's decision increases linearly with the difference in decision, where ϵ_i denotes the sensitivity of decision-dependent data distributions for the ith agent [25, 29, 35]. Again, Assumption 4 is for studying the asymptotic normality of DSGTD-GD [9, 37]. We should note that the strong convexity of Assumption 2 is not only necessary for finding the unique performative stable solution, but also essential for the convergence of repeated gradient descent method [35]. In Assumption 5, condition (a) requires the objective functions to be smooth, while conditions (b) and (c) bound the norm of the stochastic gradient and its variance, respectively. Assumption 6 implies that the problem (1) has a standard stochastic optimization formulation as follows

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^{n} \mathbf{E}_{\zeta_i\sim\mathcal{P}_i}\left[l_i(x; A_i x + \zeta_i)\right], \tag{27}$$

which is equivalent to a non-convex stochastic optimization problem, and by using the chain rule, we can derive the gradient of $f_i(x)$,

$$\nabla f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i(x)}\left[\nabla l_i(x;\xi_i) + A_i^{\intercal}\nabla_{\xi_i} l_i(x;\xi_i)\right]. \tag{28}$$

Again, under the linear structure of the distribution map, we estimate the distribution map $\mathcal{D}_i(\cdot)$ by estimating the parameter A with A_k dynamically. More specifically, for kth iteration of ith agent, we form the new estimation making use of the data sample from $\mathcal{D}_i(x)$ with the adaptive gradient method [32], that is

$$A_{i,k+1} = A_{i,k} + \nu_k(q_{i,k+1} - \xi_{i,k+1} - A_{i,k}u_{i,k})u_{i,k}^{\intercal},$$

where $\nu_k = \frac{2}{k+6d}$, $q_{i,k+1}\sim\mathcal{D}_i(x_{i,k}+u_{i,k})$, $\xi_{i,k+1}\sim\mathcal{D}_i(x_{i,k})$ and $u_{i,k}\sim\mathcal{N}(0,1)$. Note that the choice of the distribution for $u_{i,k}$ and step size ν_k, [32, Lemma 21] holds. Furthermore, at each iteration in Algorithm 2, we can have the following true stochastic gradient of $l_i(x)$

$$\nabla l_i(x_{i,k};\xi_{i,k+1}) + A_i^{\intercal}\nabla_{\xi_{i,k+1}} l_i(x_{i,k};\xi_{i,k+1}), \tag{29}$$

and with the estimation $A_{i,k}$, we form its estimator as

$$v_{i,k} = \nabla l_i(x_{i,k};\xi_{i,k+1}) + A_{i,k}^{\intercal}\nabla_{\xi_{i,k+1}} l_i(x_{i,k};\xi_{i,k+1}), \tag{30}$$

which is used as a biased stochastic gradient by the algorithm.

3 Convergence of DSGTD-GD

In this section, we study the convergence of DSGTD-GD, where Subsection 3.1 focuses on DSGTD-GD with constant step size and Subsection 3.2 focuses on DSGTD-GD with diminishing step size. Under the constant step size policy, we show that DSGTD-GD converges linearly, in expectation, to a neighborhood of the performative stable solution. Under the diminishing step size policy, we show that DSGTD-GD achieves convergence rate $\mathcal{O}\left(\frac{1}{k}\right)$ and the deviation between the averaged iterates of DSGTD-GD and the performative stable solution is asymptotically normal.

We first provide a technical result that plays a key role in analyzing the convergence of DSGTD-GD in Subsections 3.1 and 3.2.

Lemma 1. *Let $x^\star$ be the performative stable solution of problem (1). Denote*

$$\begin{aligned}
&\epsilon_{max} := max_{i=1,2,...,n}\epsilon_i, \qquad \epsilon_{avg} := \frac{1}{n}\sum_{i=1}^{n}\epsilon_i, \qquad \bar{\mu} := \mu - (1+\delta)\epsilon_{avg}\, L, \\
&c_1 := 2L^2(1+\epsilon_{max})^2 + \frac{2\delta^2}{n}, \qquad c_2 := \frac{L}{2n\delta\epsilon_{avg}}(1+\epsilon_{max})^2, \qquad c_3 := L(1+\epsilon_{max}), \\
&m_\sigma := \frac{\sigma^2}{1-\rho}\left[c_3^2\gamma^2 + 2n\gamma\, L(1+\epsilon_{avg}) + 2n + 2\sigma^2\gamma^2\right],
\end{aligned} \tag{31}$$

where δ is a positive constant. Suppose that $\epsilon_{avg} < \frac{\mu}{(1+\delta)\,L}$ and $\gamma \le \frac{\bar{\mu}}{c_1}$, then under Assumptions 1, 2, and 3,
(i):

$$\mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2|\mathcal{F}_k\right] \le (1-\bar{\mu}\gamma)\|\bar{x}_k - x^\star\|^2 + \frac{c_2\gamma + c_1\gamma^2}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma^2}{n}. \tag{32}$$

(ii):

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 \le \frac{1+\rho^2}{2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \gamma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2}\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2. \tag{33}$$

(iii):

$$\begin{aligned}
\mathbf{E}\left[\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2|\mathcal{F}_k\right] \le& \frac{2}{1-\rho}\left[c_3^2\gamma^2(\sigma^2 + c_3^2) + 2n\gamma c_3\sigma^2 + n\sigma^2(2-\bar{\mu}\gamma)\right]\|\bar{x}_k - x^\star\|^2 \\
&+ \frac{1}{1-\rho}\left[c_3^2(1+\frac{1}{\beta})\|\mathbf{W}-\mathbf{I}\|_2^2 + \frac{2\gamma^2(nc_1\sigma^2 + c_3^2(\sigma^2+c_3^2))}{n}\right. \\
&\left.+2\gamma\sigma^2(2c_3+c_2) + (3+\rho^2)\sigma^2\right]\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \left[\frac{\rho^2\gamma^2}{1-\rho}((1+\beta)c_3^2\right. \\
&\left.+2\sigma^2\frac{1+\rho^2}{1-\rho^2}) + \rho\right]\mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2|\mathcal{F}_k\right] + m_\sigma,
\end{aligned} \tag{34}$$

where β is any positive constant.

Proof. For easy of analysis, we denote,

$$\begin{aligned}
f_i(x;y) &:= \mathbf{E}_{\xi_i\sim\mathcal{D}_i(y)}\left[l_i(x;\xi_i)\right], \\
h(\mathbf{x}_k) &:= \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k};x_{i,k}),
\end{aligned} \tag{35}$$

where $\nabla f_i(x;x)$ denotes the gradient taken with respect to the first argument x.

Part (i). By the updating recursion (10),

$$\mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2|\mathcal{F}_k\right] = \|\bar{x}_k - x^\star\|^2 - 2\gamma\mathbf{E}\left[\langle\bar{x}_k - x^\star, \bar{y}_k\rangle|\mathcal{F}_k\right] + \gamma^2\mathbf{E}\left[\|\bar{y}_k\|^2|\mathcal{F}_k\right], \tag{36}$$

where $\mathcal{F}_k$ denotes the σ-algebra generated by $\{\xi_1, ..., \xi_k\}$. According to the unbiasedness condition (c) in Assumption 2 and the definition of $h(\mathbf{x}_k)$ in (35),

$$\mathbf{E}[\bar{y}_k|\mathcal{F}_k] = h(\mathbf{x}_k).$$

Note also that $\bar{x}_k$ is $\mathcal{F}_k$ measurable,

$$\begin{aligned}\mathbf{E}\left[\|\bar{x}_{k+1}-x^\star\|^2|\mathcal{F}_k\right] =&\|\bar{x}_k-x^\star\|^2-2\gamma\langle\bar{x}_k-x^\star,h(\mathbf{x}_k)\rangle\\ &+\gamma^2(\mathbf{E}\left[\|\bar{y}_k-h(\mathbf{x}_k)\|^2|\mathcal{F}_k\right]+\|h(\mathbf{x}_k)\|^2).\end{aligned}\tag{37}$$

For the third term on the right-hand side of (37),

$$\begin{aligned}\mathbf{E}\left[\|\bar{y}_k-h(\mathbf{x}_k)\|^2|\mathcal{F}_k\right] =&\frac{1}{n^2}\sum_{i=1}^n\mathbf{E}\left[\|\nabla l_i(x_{i,k};\xi_{i,k+1})-\nabla f_i(x_{i,k};x_{i,k})\|^2|\mathcal{F}_k\right]\\ \leq&\frac{\sigma^2}{n}+\frac{\sigma^2}{n^2}\sum_{i=1}^n\|x_{i,k}-x^\star\|^2,\end{aligned}$$

where the equality follows from the fact that $\bar{y}_k=\frac{1}{n}\sum_{i=1}^n\nabla l_i(x_{i,k};\xi_{i,k+1})$ in (10) and the definition of $h(\mathbf{x}_k)$, the inequality is due to Assumption 2 (d). On the other hand,

$$\begin{aligned}\|h(\mathbf{x}_k)\|^2=\|h(\mathbf{x}_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(x^\star;x^\star)\|^2&\\ \leq\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x_{i,k};x_{i,k})-\nabla f_i(x^\star;x^\star)\|^2&\\ \leq\frac{L^2}{n}\sum_{i=1}^n(1+\epsilon_i)^2\|x_{i,k}-x^\star\|^2,&\end{aligned}$$

where the equality and the last inequality follow from the fact that $\frac{1}{n}\sum_{i=1}^n\nabla f_i(x^\star;x^\star)=0$ and [25, Lemma 2], respectively. Then

$$\begin{aligned}\mathbf{E}\left[\|\bar{y}_k-h(\mathbf{x}_k)\|^2|\mathcal{F}_k\right]+\|h(\mathbf{x}_k)\|^2\leq&\frac{\sigma^2}{n}+\frac{\sigma^2}{n^2}\sum_{i=1}^n\|x_{i,k}-x^\star\|^2+\frac{L^2}{n}\sum_{i=1}^n(1+\epsilon_i)^2\|x_{i,k}-x^\star\|^2\\ \leq&\frac{\sigma^2}{n}+\frac{2(\sigma^2+c_3^2)}{n^2}\left[n\|\bar{x}_k-x^\star\|^2+\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^2\right],\end{aligned}\tag{38}$$

where $c_3:=L(1+\epsilon_{max})$ and the last inequality follows from the fact that $\sum_{i=1}^n\|x_{i,k}-x^\star\|^2\leq 2[n\|\bar{x}_k-x^\star\|^2+\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^2]$.

For the inner product term on the right-hand side of (37),

$$\begin{aligned}\langle\bar{x}_k-x^\star,h(\mathbf{x}_k)\rangle=&\frac{1}{n}\sum_{i=1}^n\langle\bar{x}_k-x^\star,\nabla f_i(x_{i,k};x_{i,k})\rangle\\ =&\frac{1}{n}\sum_{i=1}^n\langle\bar{x}_k-x^\star,\nabla f_i(x_{i,k};x_{i,k})-\nabla f_i(\bar{x}_k;x^\star)\rangle\\ &+\frac{1}{n}\sum_{i=1}^n\langle\bar{x}_k-x^\star,\nabla f_i(\bar{x}_k;x^\star)-\nabla f_i(x^\star;x^\star)\rangle\\ \geq&-\frac{L\|\bar{x}_k-x^\star\|}{n}\sum_{i=1}^n(\|x_{i,k}-\bar{x}_k\|+\epsilon_i\|x_{i,k}-x^\star\|)+\mu\|\bar{x}_k-x^\star\|^2\end{aligned}$$

$$\geq (\mu - L\epsilon_{avg})\|\bar{x}_k - x^\star\|^2 - \frac{L}{n}(1+\epsilon_{max})\sum_{i=1}^{n}\|\bar{x}_k - x^\star\|\|x_{i,k} - \bar{x}_k\|$$
$$\geq \left[\mu - L\epsilon_{avg} - \frac{\alpha L}{2n}(1+\epsilon_{max})\right]\|\bar{x}_k - x^\star\|^2 - \frac{L}{2n\alpha}(1+\epsilon_{max})\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2, \tag{39}$$

where α is any positive constant, the first inequality follows from the Cauchy-Schwarz inequality and [12, Lemma 2.1], the last inequality follows from the Young's inequality.

Substituting (38) and (39) back to the equation (37) gives us the following desired result,

$$\begin{aligned}\mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2|\mathcal{F}_k\right] &\leq \left[1 - 2\gamma(\mu - L\epsilon_{avg} - \frac{\alpha L}{2n}(1+\epsilon_{max})) + c_1\gamma^2\right]\|\bar{x}_k - x^\star\|^2 \\ &+ \left[\frac{\gamma L}{n\alpha}(1+\epsilon_{max}) + \frac{c_1\gamma^2}{n}\right]\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma^2}{n} \\ &\leq (1-\bar{\mu}\gamma)\|\bar{x}_k - x^\star\|^2 + \frac{c_2\gamma + c_1\gamma^2}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma^2}{n},\end{aligned} \tag{40}$$

where $\bar{\mu} := \mu - (1+\delta)\epsilon_{avg}L$, $c_1 := 2L^2(1+\epsilon_{max})^2 + \frac{2\delta^2}{n}$, $c_2 := \frac{L}{2n\delta\epsilon_{avg}}(1+\epsilon_{max})^2$, $c_3 := L(1+\epsilon_{max})$, $\alpha = \frac{2n\delta\epsilon_{avg}}{1+\epsilon_{max}}$ and the last inequality follows from the fact that $\gamma \leq \frac{\bar{\mu}}{c_1}$.

Part (ii). By the definition of $\mathbf{x}_{k+1}$ in (8) and the update recursion (10),

$$\begin{aligned}\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 &= \|\mathbf{W}\mathbf{x}_k - \gamma\mathbf{W}\mathbf{y}_k - \bar{\mathbf{x}}_k + \gamma\bar{\mathbf{y}}_k\|^2 \\ &\leq \rho^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + 2\gamma\rho^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|\|\mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k\| + \gamma^2\rho^2\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 \\ &\leq \rho^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \gamma\rho^2\left[\frac{1-\rho^2}{2\gamma\rho^2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right. \\ &\left.+\frac{2\gamma\rho^2}{1-\rho^2}\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2\right] + \gamma^2\rho^2\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 \\ &= \frac{1+\rho^2}{2}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \gamma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2}\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2,\end{aligned} \tag{41}$$

where $\rho \in (0,1)$, the first inequality and the last inequality follow from [38, Lemma 1] and the Young's inequality, respectively.

Part (iii). Denote

$$\mathbf{L}_k := \mathbf{L}(\mathbf{x}_k; \xi_{k+1}), \tag{42}$$
$$\mathbf{F}_k := [\nabla f_1(x_{1,k}; x_{1,k}), \nabla f_2(x_{2,k}; x_{2,k}), ..., \nabla f_n(x_{n,k}; x_{n,k})]^\intercal, \tag{43}$$

we have by the definition of $\mathbf{y}_{k+1}$ in (8) that

$$\begin{aligned}\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1} &= (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)\mathbf{y}_{k+1} \\ &= (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)(\mathbf{W}\mathbf{y}_k + \mathbf{L}_{k+1} - \mathbf{L}_k) \\ &= \mathbf{W}\mathbf{y}_k - \bar{\mathbf{y}}_k + (\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)(\mathbf{L}_{k+1} - \mathbf{L}_k),\end{aligned}$$

where the last equality follows from the fact that $\mathbf{W}$ is a doubly stochastic matrix. Thus, for any $\eta > 0$,

$$\begin{aligned}\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2 &\leq (1+\eta)\|\mathbf{W}\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 + (1+\frac{1}{\eta})\|(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)(\mathbf{L}_{k+1} - \mathbf{L}_k)\|^2 \\ &\leq (1+\eta)\rho^2\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 + (1+\frac{1}{\eta})\|(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)(\mathbf{L}_{k+1} - \mathbf{L}_k)\|^2 \\ &= \rho\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 + \frac{1}{1-\rho}\|(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)(\mathbf{L}_{k+1} - \mathbf{L}_k)\|^2,\end{aligned}$$

where the first inequality is due to the Young's inequality, the second inequality is due to [38, Lemma 1], the equality follows by setting $\eta = \frac{1-\rho}{\rho}$ and the fact that $\rho \in (0,1)$ by [38, Lemma 1].

By the definition of $\mathbf{L}_k$,

$$\begin{aligned}\mathbf{E}\left[\|(\mathbf{L}_{k+1} - \mathbf{L}_k)\|^2|\mathcal{F}_k\right] &= \mathbf{E}\left[\|\mathbf{F}_{k+1} - \mathbf{F}_k\|^2|\mathcal{F}_k\right] + \mathbf{E}\left[\|\mathbf{L}_{k+1} - \mathbf{L}_k - \mathbf{F}_{k+1} + \mathbf{F}_k\|^2|\mathcal{F}_k\right] \\ &+ 2\mathbf{E}\left[\langle \mathbf{F}_{k+1} - \mathbf{F}_k, \mathbf{L}_{k+1} - \mathbf{L}_k - \mathbf{F}_{k+1} + \mathbf{F}_k\rangle|\mathcal{F}_k\right] \\ &= \mathbf{E}\left[\|\mathbf{F}_{k+1} - \mathbf{F}_k\|^2|\mathcal{F}_k\right] + \mathbf{E}\left[\|\mathbf{L}_{k+1} - \mathbf{F}_{k+1}\|^2|\mathcal{F}_k\right] \\ &+ \mathbf{E}\left[\|\mathbf{L}_k - \mathbf{F}_k\|^2|\mathcal{F}_k\right] + 2\mathbf{E}\left[\langle \mathbf{F}_{k+1}, -\mathbf{L}_k + \mathbf{F}_k\rangle|\mathcal{F}_k\right] \\ &\leq \mathbf{E}\left[\|\mathbf{F}_{k+1} - \mathbf{F}_k\|^2|\mathcal{F}_k\right] + 2\mathbf{E}\left[\langle \mathbf{F}_{k+1}, -\mathbf{L}_k + \mathbf{F}_k\rangle|\mathcal{F}_k\right] \\ &+ \sigma^2(2n + \sum_{i=1}^n \mathbf{E}\left[\|x_{i,k+1} - x^\star\|^2|\mathcal{F}_k\right] + \sum_{i=1}^n \|x_{i,k} - x^\star\|^2),\end{aligned}$$

where the second equality and the inequality follow from Assumption 2 (c) and (d), respectively. Then

$$\begin{aligned}\mathbf{E}\left[\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2|\mathcal{F}_k\right] \leq &\rho\mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2|\mathcal{F}_k\right] + \frac{1}{1-\rho}\left[2\mathbf{E}\left[\langle \mathbf{F}_{k+1}, -\mathbf{L}_k + \mathbf{F}_k\rangle|\mathcal{F}_k\right]\right. \\ &+\mathbf{E}\left[\|\mathbf{F}_{k+1} - \mathbf{F}_k\|^2|\mathcal{F}_k\right] + \sigma^2(2n + \sum_{i=1}^n \mathbf{E}\left[\|x_{i,k+1} - x^\star\|^2|\mathcal{F}_k\right] \\ &\left.+\sum_{i=1}^n \|x_{i,k} - x^\star\|^2)\right]. \qquad (44)\end{aligned}$$

For the second term on the right-hand side of (44),

$$\begin{aligned}\mathbf{E}\left[\langle \mathbf{F}_{k+1}, -\mathbf{L}_k + \mathbf{F}_k\rangle|\mathcal{F}_k\right] &= \sum_{i=1}^n \mathbf{E}\left[\langle \nabla f_i(x_{i,k+1}; x_{i,k+1}), -\nabla l_i(x_{i,k}; \xi_{i,k+1}) + \nabla f_i(x_{i,k}; x_{i,k})\rangle|\mathcal{F}_k\right] \\ &= \sum_{i=1}^n \mathbf{E}\left[\left\langle \nabla f_i(\sum_{j=1}^n w_{ij}x_{j,k} - \gamma\sum_{j=1}^n w_{ij}\sum_{l=1}^n w_{jl}y_{l,k-1}\right.\right. \\ &-\gamma\sum_{j=1}^n w_{ij}l_j(x_{j,k}; \xi_{j,k+1}) + \gamma\sum_{j=1}^n w_{ij}l_j(x_{j,k-1}; \xi_{j,k}); x_{i,k+1}\rangle \\ &\left.-\nabla f_i(\sum_{j=1}^n w_{ij}x_{j,k} - \gamma\sum_{j=1}^n w_{ij}\sum_{l=1}^n w_{jl}y_{l,k-1} - \gamma\sum_{j\neq i}^n w_{ij}l_j(x_{j,k}; \xi_{j,k+1})\right.\end{aligned}$$

$$
\begin{aligned}
&-\gamma w_{ii}\nabla f_i(x_{i,k};x_{i,k}) + \gamma\sum_{j=1}^{n} w_{ij} l_j(x_{j,k-1};\xi_{j,k});x_{i,k+1}),\\
&-\nabla l_i(x_{i,k};\xi_{i,k+1}) + \nabla f_i(x_{i,k};x_{i,k})\rangle\,|\mathcal{F}_k]\\
&\le \sum_{i=1}^{n}\gamma\, L(1+\epsilon_i)\mathbf{E}\left[\|\nabla l_i(x_{i,k};\xi_{i,k+1}) - \nabla f_i(x_{i,k};x_{i,k})\|^2|\mathcal{F}_k\right]\\
&\le \gamma\sigma^2\, L\sum_{i=1}^{n}(1+\epsilon_i)(1+\|x_{i,k}-x^\star\|^2)\\
&\le 2\gamma\sigma^2\, L(1+\epsilon_{max})\left[n\|\bar{x}_k - x^\star\|^2 + \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right]\\
&+ \gamma n\sigma^2\, L(1+\epsilon_{avg}),
\end{aligned}
\tag{45}
$$

where the second equality follows from the formulas (6) and (7) in Algorithm 1, the first inequality is due to [12, Lemma 2.1] and Assumption 2 (b), and the second inequality is due to Assumption 2 (d).

By the definitions of $\mathbf{x}_{k+1}$ in (8) and c_3 in (31), for any $\beta > 0$, we have

$$
\begin{aligned}
\|\mathbf{F}_{k+1} - \mathbf{F}_k\|^2 &= \sum_{i=1}^{n}\|\nabla f_i(x_{i,k+1};x_{i,k+1}) - \nabla f_i(x_{i,k};x_{i,k})\|^2\\
&\le c_3^2\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2\\
&= c_3^2\|(\mathbf{W}-\mathbf{I})(\mathbf{x}_k - \bar{\mathbf{x}}_k) - \gamma\mathbf{W}\mathbf{y}_k\|^2\\
&= c_3^2[\|\mathbf{W}-\mathbf{I}\|_2^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 - 2\gamma\langle(\mathbf{W}-\mathbf{I})(\mathbf{x}_k - \bar{\mathbf{x}}_k), \mathbf{W}\mathbf{y}_k - \bar{\mathbf{y}}_k\rangle\\
&+ \gamma^2(\|\mathbf{W}\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 + n\|\bar{y}_k\|^2)]\\
&\le c_3^2[(1+\frac{1}{\beta})\|\mathbf{W}-\mathbf{I}\|_2^2\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + (1+\beta)\gamma^2\rho^2\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 + \gamma^2 n\|\bar{y}_k\|^2],
\end{aligned}
\tag{46}
$$

where the first inequality follows from [12, Lemma 2.1] and the last inequality follows from [38, Lemma 1] and the Young's inequality.

Combining (38), (40), (41), (45), (46) with the inequality (44) gives us the desired result,

$$
\begin{aligned}
\mathbf{E}\left[\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2|\mathcal{F}_k\right] \le &\frac{2}{1-\rho}\left[c_3^2\gamma^2(\sigma^2 + c_3^2) + 2n\gamma c_3\sigma^2 + n\sigma^2(2-\bar{\mu}\gamma)\right]\|\bar{x}_k - x^\star\|^2\\
&+ \frac{1}{1-\rho}\left[c_3^2(1+\frac{1}{\beta})\|\mathbf{W}-\mathbf{I}\|_2^2 + \frac{2\gamma^2(nc_1\sigma^2 + c_3^2(\sigma^2+c_3^2))}{n}\right.\\
&\left.+2\gamma\sigma^2(2c_3 + c_2) + (3+\rho^2)\sigma^2\right]\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \left[\frac{\rho^2\gamma^2}{1-\rho}((1+\beta)c_3^2\right.\\
&\left.+2\sigma^2\frac{1+\rho^2}{1-\rho^2}) + \rho\right]\mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2|\mathcal{F}_k\right] + m_\sigma,
\end{aligned}
$$

where $\bar{\mu}$, c_1, c_2, c_3, m_σ are defined in (31) and β is any positive constant. □

3.1 DSGTD-GD with constant step size

The stochastic gradient descent (SGD) method with constant step size is often used in practice as it may achieve an exponentially fast convergence rate [10, 41]. In this subsection,

we show that DSGTD-GD with constant step size inherits the properties of converging to a neighborhood of the optimal solution of SGD in linear convergence rate.

Theorem 1. *Let $x^\star$ be the performative stable solution of problem (1). Denote*

$$
\begin{aligned}
A &:= \begin{bmatrix} 1-\gamma\bar{\mu} & \frac{c_2\gamma+c_1\gamma^2}{n} & 0 \\ 0 & \frac{1+\rho^2}{2} & \frac{(1+\rho^2)\rho^2\gamma^2}{1-\rho^2} \\ a_{31} & a_{32} & \rho+((1+\hat{\beta})c_3^2+2\sigma^2\frac{1+\rho^2}{1-\rho^2})\frac{\rho^2\gamma^2}{1-\rho} \end{bmatrix}, \\
c_W &:= 2\sigma^2\frac{1+\rho^2}{1-\rho^2}+c_3^2, \qquad \hat{\beta} := \frac{\frac{(1-\rho)^2}{2}-c_W\gamma^2\rho^2}{\gamma^2c_3^2\rho^2}, \\
a_{31} &:= \frac{2}{1-\rho}\left[\gamma^2c_3^2(\sigma^2+c_3^2)+2n\gamma c_3\sigma^2+n\sigma^2(2-\bar{\mu}\gamma)\right], \\
a_{32} &:= \frac{1}{1-\rho}\left[c_3^2(1+\frac{1}{\hat{\beta}})\|\mathbf{W}-\mathbf{I}\|_2^2+\frac{2\gamma^2(nc_1\sigma^2+(\sigma^2+c_3^2)c_3^2)}{n}+2\gamma(2c_3+c_2)\sigma^2\right. \\
&\quad \left.+(3+\rho^2)\sigma^2\right],
\end{aligned}
\tag{47}
$$

where $\bar{\mu}$, c_1, c_2 and c_3 are defined in (31). Suppose that the constant step size

$$
\begin{aligned}
\gamma \le \min\Bigg\{ & \frac{(1-\rho)^2(1+\rho)}{4\left[\frac{(1-\rho)^2c_3^2(\sigma^2+c_3^2)}{2c_W}+2n\sigma^2+\frac{n\sigma^2(1-\rho)}{\sqrt{2c_W}}(2c_3-\bar{\mu})\right]^{\frac{1}{2}}}\left[\frac{n\bar{\mu}(\kappa-1)}{\kappa(\kappa+1)(c_2+\frac{c_1(1-\rho)}{\sqrt{2c_W}})}\right]^{\frac{1}{2}}, \\
& \frac{(1-\rho^2)(1-\rho)}{2\sqrt{2\kappa}\max\left\{\frac{\sqrt{4\sigma^2+c_3^2}c_3\|\mathbf{W}-\mathbf{I}\|_2}{\sigma},(1-\rho)\sqrt{\frac{(nc_1\sigma^2+c_3^2(\sigma^2+c_3^2))}{nc_W}},\sqrt{\frac{8\sigma^2(1-\rho)(2c_3+c_2)}{\sqrt{2c_W}}},4\sigma\right\}}, \frac{1-\rho}{\sqrt{2c_W}}\Bigg\},
\end{aligned}
\tag{48}
$$

where $\kappa > 1$ is arbitrarily chosen. Then under the conditions of Lemma 1,
(i):

$$
\sup_{t\ge k}\mathbf{E}\left[\|\bar{x}_t-x^\star\|^2\right], \qquad \sup_{t\ge k}\mathbf{E}\left[\|\mathbf{x}_t-\bar{\mathbf{x}}_t\|^2\right] \tag{49}
$$

converge to a neighborhood of 0 at the linear rate $\mathcal{O}(\rho(A)^k)$, where $\rho(A)<1$ is the spectral radius of the matrix A.
(ii):

$$
\limsup_{k\to\infty}\mathbf{E}\left[\|\bar{x}_k-x^\star\|^2\right] \le \frac{\gamma\sigma^2(\kappa+1)}{\kappa n\bar{\mu}}+\frac{4\gamma^2\rho^2m_\sigma(\kappa+1)(c_2+\gamma c_1)(1+\rho^2)}{n\bar{\mu}(\kappa-1)(1-\rho)(1-\rho^2)^2}, \tag{50}
$$

$$
\limsup_{k\to\infty}\mathbf{E}\left[\|\mathbf{x}_k-\bar{\mathbf{x}}_k\|^2\right] \le \frac{4\gamma^2\rho^2(\kappa+1)(1+\rho^2)(a_{31}\gamma\sigma^2+\bar{\mu}nm_\sigma)}{n\bar{\mu}(\kappa-1)(1-\rho)(1-\rho^2)^2}, \tag{51}
$$

where m_σ is defined in (31).

Proof. Part (i). By the definition of c_W in (47),

$$
\hat{\beta}=\frac{\frac{(1-\rho)^2}{2}-\gamma^2\rho^2c_W}{\gamma^2c_3^2\rho^2}\ge\frac{(1-\rho)(1+\rho)c_W}{c_3^2\rho^2}>\frac{4\sigma^2}{c_3^2}>0, \tag{52}
$$

where the first inequality follows from the fact that $\gamma \leq \frac{1-\rho}{\sqrt{2c_W}}$ in (48). Taking expectation on both sides of (32)-(34) with $\beta = \hat{\beta}$ and constant step size γ, we obtain the following linear system of inequalities

$$\begin{bmatrix} \mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2\right] \\ \mathbf{E}\left[\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2\right] \\ \mathbf{E}\left[\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2\right] \end{bmatrix} \leq A \begin{bmatrix} \mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] \\ \mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] \\ \mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2\right] \end{bmatrix} + \begin{bmatrix} \frac{\gamma^2\sigma^2}{n} \\ 0 \\ m_\sigma \end{bmatrix}, \tag{53}$$

where m_σ and the matrix $A = (a_{ij})_{3\times 3}$ are defined in (31) and (47), respectively. Then,

$$\begin{bmatrix} \mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] \\ \mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] \\ \mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2\right] \end{bmatrix} \leq A^k \begin{bmatrix} \mathbf{E}\left[\|\bar{x}_0 - x^\star\|^2\right] \\ \mathbf{E}\left[\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2\right] \\ \mathbf{E}\left[\|\mathbf{y}_0 - \bar{\mathbf{y}}_0\|^2\right] \end{bmatrix} + \sum_{l=0}^{k-1} A^l \begin{bmatrix} \frac{\gamma^2\sigma^2}{n} \\ 0 \\ m_\sigma \end{bmatrix}. \tag{54}$$

By some calculations,

$$\|\sum_{l=0}^{k-1} A^l B\| \leq \sum_{l=0}^{k-1} \rho(A)^l \|B\| \leq \sum_{l=0}^{\infty} \rho(A)^l \|B\| = \lim_{k\to\infty} \frac{(1-\rho(A)^k)\|B\|}{1-\rho(A)}, \tag{55}$$

where $B := [\frac{\gamma^2\sigma^2}{n}, 0, m_\sigma]^\intercal$. Obviously, if $\rho(A) < 1$, we may conclude that

$$\sup_{t\geq k} \mathbf{E}\left[\|\bar{x}_t - x^\star\|^2\right], \quad \sup_{t\geq k} \mathbf{E}\left[\|\mathbf{x}_t - \bar{\mathbf{x}}_t\|^2\right], \quad \sup_{t\geq k} \mathbf{E}\left[\|\mathbf{y}_t - \bar{\mathbf{y}}_t\|^2\right]$$

converge to a $\frac{\|B\|}{1-\rho(A)}$ neighborhood of 0 at the linear rate $\mathcal{O}(\rho(A)^k)$.

Next, we employ Lemma 4 in Appendix to show $\rho(A) < 1$, where we just need to verify the conditions of Lemma 4. By the fact that $\bar{\mu} > 0$ and $\rho \in (0,1)$, $a_{11} < 1$, $a_{22} < 1$ and

$$a_{33} = \rho + \frac{\gamma^2\rho^2}{1-\rho}((1+\hat{\beta})c_3^2 + 2\sigma^2\frac{1+\rho^2}{1-\rho^2}) = \frac{1+\rho}{2} < 1. \tag{56}$$

By (48),

$$\begin{aligned} a_{23}a_{32} &= \gamma^2\frac{(1+\rho^2)\rho^2}{(1-\rho^2)(1-\rho)}[c_3^2(1+\frac{1}{\hat{\beta}})\|\mathbf{W}-\mathbf{I}\|_2^2 + \frac{2\gamma^2(c_1 n\sigma^2 + c_3^2(\sigma^2+c_3^2))}{n} \\ &\quad + 2\gamma\sigma^2(2c_3+c_2) + (3+\rho^2)\sigma^2] \\ &\leq \gamma^2\frac{(1+\rho^2)\rho^2}{(1-\rho^2)(1-\rho)}[\frac{4\sigma^2+c_3^2}{4\sigma^2}c_3^2\|\mathbf{W}-\mathbf{I}\|_2^2 + \frac{(1-\rho)^2(c_1 n\sigma^2 + c_3^2(\sigma^2+c_3^2))}{nc_W} \\ &\quad + 2\sigma^2\frac{(1-\rho)(2c_3+c_2)}{\sqrt{2c_W}} \\ &\leq \frac{(1-a_{22})(1-a_{33})}{\kappa}, \end{aligned} \tag{57}$$

where c_1, c_2, c_3 are defined in (31), $\kappa > 1$ is arbitrarily chosen, and the first inequality follows from the fact that $\hat{\beta} > \frac{4\sigma^2}{c_3^2}$ and $\gamma \leq \frac{1-\rho}{\sqrt{2c_W}}$. Then

$$\begin{aligned} a_{12}a_{23}a_{31} &= 2\gamma^3 \frac{(c_2+\gamma c_1)(1+\rho^2)\rho^2}{n(1+\rho)(1-\rho)^2}\left[\gamma^2 c_3^2(\sigma^2+c_3^2) + 2\gamma c_3\sigma^2 n + n\sigma^2(2-\bar{\mu}\gamma)\right] \\ &\leq \frac{\kappa-1}{4\kappa(\kappa+1)}\bar{\mu}\gamma(1-\rho)^2(1+\rho) \\ &= \frac{\kappa-1}{\kappa(\kappa+1)}(1-a_{11})(1-a_{22})(1-a_{33}) \\ &\leq \frac{(1-a_{11})[(1-a_{22})(1-a_{33})-a_{23}a_{32}]}{\kappa+1}, \end{aligned} \tag{58}$$

where $\bar{\mu}$, c_1, c_2, c_3 are defined in (31), $\rho \in (0,1)$ due to [38, Lemma 1], the first inequality and the last inequality follow from (48) and (57), respectively. Subsequently, by (56), (57) and (58), we have

$$\begin{aligned} \det(\mathbf{I}-A) &= (1-a_{11})(1-a_{22})(1-a_{33}) - (1-a_{11})a_{23}a_{32} - a_{12}a_{23}a_{31} \\ &\geq \frac{\kappa}{\kappa+1}(1-a_{11})[(1-a_{22})(1-a_{33})-a_{23}a_{32}] \\ &\geq \frac{\kappa-1}{\kappa+1}(1-a_{11})(1-a_{22})(1-a_{33}) > 0. \end{aligned} \tag{59}$$

Summarize above, the conditions of Lemma 4 in Appendix hold and then $\rho(A) < 1$.

Part (ii). In what follows, we give a more precise bound on the convergence neighborhoods of $\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right]$ and $\mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right]$. Denoting $[(\mathbf{I}-A)^{-1}B]_j$ as the jth element of the vector $[(\mathbf{I}-A)^{-1}B]$, we have

$$\begin{aligned} \limsup_{k\to\infty} \mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] &\leq [(\mathbf{I}-A)^{-1}B]_1 \\ &= \frac{(1-a_{11})[(1-a_{22})(1-a_{33})-a_{23}a_{32}]\frac{\gamma^2\sigma^2}{n} + a_{12}a_{23}m_\sigma}{\det(\mathbf{I}-A)} \\ &\leq \frac{\kappa+1}{\kappa}\frac{\gamma\sigma^2}{\bar{\mu}n} + \frac{\kappa+1}{\kappa-1}\frac{a_{12}a_{23}m_\sigma}{(1-a_{11})(1-a_{22})(1-a_{33})} \\ &= \frac{\gamma\sigma^2(\kappa+1)}{\kappa n\bar{\mu}} + \frac{4\gamma^2\rho^2 m_\sigma(\kappa+1)(c_2+\gamma c_1)(1+\rho^2)}{n\bar{\mu}(\kappa-1)(1-\rho)(1-\rho^2)^2}, \end{aligned}$$

where $\bar{\mu}$, c_1, c_2, m_σ are defined in (31), γ is defined in (48), the first inequality follows from (53) and the second inequality follows from (57) and (59). Similarly, we have

$$\begin{aligned} \limsup_{k\to\infty} \mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] &\leq [(\mathbf{I}-A)^{-1}B]_2 \\ &= \frac{a_{31}a_{23}\frac{\gamma^2\sigma^2}{n} + a_{23}(1-a_{11})m_\sigma}{\det(\mathbf{I}-A)} \\ &\leq \frac{\kappa+1}{\kappa-1}\frac{a_{23}}{(1-a_{11})(1-a_{22})(1-a_{33})}\left[\frac{\gamma^2\sigma^2 a_{31}}{n} + (1-a_{11})m_\sigma\right] \\ &= \frac{4\gamma\rho^2(\kappa+1)(1+\rho^2)(a_{31}\sigma^2\gamma^2 + \gamma\bar{\mu}nm_\sigma)}{n\bar{\mu}(\kappa-1)(1-\rho^2)^2(1-\rho)}. \end{aligned}$$

The proof is complete. □

As shown in Theorem 1, the iterates generated by DSGTD-GD converge to a neighborhood of the performative stable point at the linear rate when the step size satisfies (48). By the definitions of c_3 and $\bar{\mu}$ in (31),

$$2c_3 - \bar{\mu} = L(2(1+\epsilon_{max}) + (1+\delta)\epsilon_{avg}) - \mu > 0.$$

Combined with the fact that $\rho \in (0,1)$, all the terms in (48) are positive, which means the step size γ is obtainable. On the other hand, in light of (50) and (51),

$$\begin{aligned}\limsup_{k\to\infty} \mathbf{E}\left[\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right] &\le \limsup_{k\to\infty} 2n\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] + \limsup_{k\to\infty} 2\mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] \\ &= \gamma\mathcal{O}(\frac{\sigma^2}{\bar{\mu}}) + \frac{\gamma^2}{(1-\rho)^4}\mathcal{O}(\frac{c_2\sigma^2}{\bar{\mu}}),\end{aligned}$$

which means the distance $\mathbf{E}\left[\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right]$ is proportional to the step size and the spectral norm ρ. In other words, the smaller the step size and the spectral norm lead to better accuracy.

3.2 DSGTD-GD with diminishing step size

Different from the constant step size policy, a diminishing step size policy may guarantee that SGD converges to the optimal solution rather than a neighborhood of an optimum [26, 31]. In what follows, we study DSGTD-GD with a diminishing step size policy.

Theorem 2. *Let $x^\star$ be the performative stable solution of problem (1). Suppose that the time-varying step sizes $\{\gamma_k\}_{k\ge1}$ in Algorithm 1 satisfy $\gamma_k = r(k+b)^{-a}$ with $a \in (\frac{1}{2}, 1)$ and r, a, b satisfy*

$$\begin{cases} 2r^2\sigma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2} + r^2c_3^2\rho^2 < \frac{b^{2a}(1-\rho)^2}{2} \\ r > \frac{b^a-(b-1)^a}{\bar{\mu}} \\ \frac{b^{2a}(1-\rho^2)(1-\rho)}{2(1+\rho^2)\rho^2}(\frac{b^{2a}}{(b+1)^{2a}} - \frac{1+\rho^2}{2}) > \frac{K_1r^3(c_2b^a+rc_1)}{b^an(\bar{\mu}r+(b-1)^a-b^a)} + K_2r^2 \end{cases}, \tag{60}$$

where

$$\begin{aligned}\hat{\beta}_k :=& \frac{\frac{(1-\rho)^2}{2} - 2\gamma_k^2\sigma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2} - \gamma_k^2c_3^2\rho^2}{\gamma_k^2c_3^2\rho^2}, \\ K_1 :=& \frac{2}{(1-\rho)b^{2a}}\left[r^2c_3^2b^a(\sigma^2+c_3^2) + 2rc_3\sigma^2n + 2nb^{2a}\sigma^2\right], \\ K_2 :=& \frac{1}{1-\rho}\left[c_3^2(1+\frac{1}{\hat{\beta}_0})\|\mathbf{W}-\mathbf{I}\|_2^2 + \frac{2r^2(c_1n\sigma^2 + c_3^2(\sigma^2+c_3^2))}{nb^{2a}} + \frac{2r\sigma^2(2c_3+c_2)}{b^a}\right. \\ &\left.+(3+\rho^2)\sigma^2\right]\end{aligned} \tag{61}$$

and $\bar{\mu}$, c_1, c_2, and c_3 are defined in (31). Then under the conditions of Lemma 1,

$$\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right], \qquad \mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] \tag{62}$$

converge to 0 at rate $\mathcal{O}(\gamma_k)$ and $\mathcal{O}(\gamma_k^2)$, respectively. Moreover, $\bar{x}_k$ converges to $x^\star$ almost surely.

Proof. By the definition of $\hat{\beta}_k$,

$$\hat{\beta}_k = \frac{\frac{(1-\rho)^2}{2} - 2\gamma_k^2\sigma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2} - \gamma_k^2 c_3^2\rho^2}{\gamma_k^2 c_3^2\rho^2} \geq \frac{\frac{b^{2a}(1-\rho)^2}{2} - 2r^2\sigma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2} - r^2 c_3^2\rho^2}{r^2 c_3^2\rho^2} > 0, \tag{63}$$

where the first inequality is due to $\gamma_k = r(k+b)^{-a}$ with $a \in (\frac{1}{2}, 1)$ and the second inequality follows from the definitions of r, a and b in (60). Taking expectation on both sides of (32)-(34) with $\beta = \hat{\beta}_k$ and $\gamma = \gamma_k$, we have

$$\begin{bmatrix} \mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2\right] \\ \mathbf{E}\left[\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2\right] \\ \mathbf{E}\left[\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2\right] \end{bmatrix} \leq A_k \begin{bmatrix} \mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] \\ \mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] \\ \mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2\right] \end{bmatrix} + \begin{bmatrix} \frac{\gamma_k^2\sigma^2}{n} \\ 0 \\ m_k \end{bmatrix}, \tag{64}$$

where

$$\begin{aligned}
A_k =& \begin{bmatrix} 1-\gamma_k\bar{\mu} & \frac{c_2\gamma_k + c_1\gamma_k^2}{n} & 0 \\ 0 & \frac{1+\rho^2}{2} & \gamma_k^2\frac{(1+\rho^2)\rho^2}{1-\rho^2} \\ a'_{31} & a'_{32} & \rho + \frac{\rho^2\gamma_k^2}{1-\rho}((1+\hat{\beta}_k)c_3^2 + 2\sigma^2\frac{1+\rho^2}{1-\rho^2}) \end{bmatrix}, \\
a'_{31} =& \frac{2}{1-\rho}\left[c_3^2\gamma_k^2(\sigma^2 + c_3^2) + 2\gamma_k c_3\sigma^2 n + n\sigma^2(2 - \bar{\mu}\gamma_k)\right], \\
a'_{32} =& \frac{1}{1-\rho}\left[c_3^2(1 + \frac{1}{\hat{\beta}_k})\|\mathbf{W} - \mathbf{I}\|_2^2 + \frac{2\gamma_k^2(c_1 n\sigma^2 + c_3^2(\sigma^2 + c_3^2))}{n} + 2\gamma_k\sigma^2(2c_3 + c_2)\right. \\
& \left. +(3+\rho^2)\sigma^2\right], \\
m_k =& \frac{\sigma^2}{1-\rho}\left[\gamma_k^2(c_3^2 + 2\sigma^2) + 2\gamma_k n\, L(1+\epsilon_{avg}) + 2n\right]
\end{aligned}$$

and $\bar{\mu}$, c_1, c_2, c_3 are defined in (31).

Before formally proving the results, we first denote

$$\begin{aligned}
&\triangle_1 := \frac{1-\rho^2}{(1+\rho^2)\rho^2}(\frac{b^{2a}}{(b+1)^{2a}} - \frac{1+\rho^2}{2}), \qquad \triangle_2 := \frac{2}{(1-\rho)b^{2a}}[K_1 b^{2a}\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2 + b^{2a}m_0], \\
&\triangle_3 := \frac{2K_2 r^2}{(1-\rho)b^{2a}}, \qquad \triangle_4 := \frac{2}{1-\rho}[\frac{K_1 r^2\sigma^2}{b^a n(\bar{\mu}r + (b-1)^a - b^a)} + m_0], \\
&\triangle_5 := \frac{2r^2}{(1-\rho)b^{2a}}[\frac{K_1 r(c_2 b^a + r c_1)}{b^a n(\bar{\mu}r + (b-1)^a - b^a)} + K_2], \\
&N_1 := \max\{\frac{r[(rc_2 b^a + r^2 c_1)N_2 + \sigma^2 b^{2a}]}{b^{2a}n(\bar{\mu}r + (b-1)^a - b^a)}, \frac{b^a\|\bar{x}_0 - x^\star\|^2}{r}\}, \\
&N_2 := \max\{\frac{b^{2a}\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2}{r^2}, \frac{\mathbf{E}\left[\|\mathbf{y}_0 - \bar{\mathbf{y}}_0\|^2\right]}{\triangle_1}, \frac{\triangle_2}{\triangle_1 - \triangle_3}, \frac{\triangle_4}{\triangle_1 - \triangle_5}\}, \\
&N_3 := \max\{\mathbf{E}\left[\|\mathbf{y}_0 - \bar{\mathbf{y}}_0\|^2\right], \frac{1-\rho}{2}(\frac{rK_1N_1}{b^a} + \frac{K_2N_2r^2}{b^{2a}} + m_0)\},
\end{aligned} \tag{65}$$

where $\bar{\mu}$, c_1, c_2 are defined in (31), K_1, K_2 are defined in (61), the relations between r, a, b are defined in (60).

In what follows, we show that

$$\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] \le N_1\gamma_k, \quad \mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] \le N_2\gamma_k^2, \quad \mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2\right] \le N_3 \tag{66}$$

by induction. Obviously, when $k = 0$, (66) holds. Next, we show that (66) holds for $k+1$ provided that (66) holds for some $k > 0$.

By (64) and the fact that $\rho + \frac{\gamma^2\rho^2}{1-\rho}((1+\hat{\beta}_k)c_3^2 + 2\sigma^2\frac{1+\rho^2}{1-\rho^2}) = \frac{1+\rho}{2}$, we have

$$\begin{aligned}
\mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2\right] &\le (1-\bar{\mu}\gamma_k)\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] + \gamma_k\frac{(c_2+c_1\gamma_k)}{n}\mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] + \frac{\gamma_k^2\sigma^2}{n},\\
\mathbf{E}\left[\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2\right] &\le \frac{1+\rho^2}{2}\mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] + \gamma_k^2\frac{(1+\rho^2)\rho^2}{1-\rho^2}\mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2\right],\\
\mathbf{E}\left[\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2\right] &\le a'_{31}\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] + a'_{32}\mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] + \frac{1+\rho}{2}\mathbf{E}\left[\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2\right] + m_k.
\end{aligned}$$

Then

$$\begin{aligned}
\mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2\right] &\le \gamma_k(1-\bar{\mu}\gamma_k)N_1 + \gamma_k^3\frac{(c_2+c_1\gamma_k)N_2}{n} + \frac{\gamma_k^2\sigma^2}{n}\\
&= \gamma_k(1-\bar{\mu}\gamma_k)N_1 + \gamma_k^3\frac{(c_2+c_1\gamma_k)N_2}{n} + \frac{\gamma_k^2\sigma^2}{n} - N_1\gamma_{k+1} + N_1\gamma_{k+1}\\
&\le (\frac{r}{(k+b)^a} - \frac{\bar{\mu}r^2}{(k+b)^{2a}} - \frac{r}{(k+b+1)^a})N_1 + \gamma_0^3\frac{(c_2+\gamma_0 c_1)N_2}{n}\\
&+ \frac{\gamma_0^2\sigma^2}{n} + N_1\gamma_{k+1}\\
&\le N_1\gamma_{k+1},
\end{aligned}$$

where r, a and b are defined in (60) and the first inequality, the second inequality and the last inequality follow from the fact that (66) holds for some $k > 0$, the definition of γ_k and the definition of N_1 in (65), respectively. Similarly, we have

$$\begin{aligned}
\mathbf{E}\left[\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2\right] &\le \gamma_k^2\frac{(1+\rho^2)N_2}{2} + \gamma_k^2\frac{(1+\rho^2)\rho^2N_3}{1-\rho^2}\\
&\le (\frac{(1+\rho^2)r^2}{2(k+b)^{2a}} - \frac{r^2}{(k+1+b)^{2a}})N_2 + \gamma_0^2\frac{(1+\rho^2)\rho^2N_3}{1-\rho^2} + N_2\gamma_{k+1}^2\\
&\le N_2\gamma_{k+1}^2,\\
\mathbf{E}\left[\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2\right] &\le \gamma_k K_1N_1 + \gamma_k^2K_2N_2 + \frac{(1+\rho)N_3}{2} + m_0\\
&\le \frac{rK_1N_1}{b^a} + \frac{K_2N_2r^2}{b^{2a}} + \frac{(\rho-1)N_3}{2} + m_0 + N_3\\
&\le N_3,
\end{aligned}$$

where $m_0 = \frac{\sigma^2}{(1-\rho)b^{2a}}\left[r^2(2\sigma^2+c_3^2) + 2rn\,L(1+\epsilon_{avg})b^a + 2nb^{2a}\right]$, K_1, K_2 are defined in (61).

Summarize above, we can conclude that (66) holds for any $k \ge 0$. Obviously, N_1, N_2, and N_3 are bounded constants and then $\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right]$ and $\mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right]$ converge to 0 at rate $\mathcal{O}(\gamma_k)$ and $\mathcal{O}(\gamma_k^2)$, respectively.

Next, we use Lemma 5 in Appendix to show that $\bar{x}_k$ converges to $x^\star$ almost surely.

Choosing $\gamma = \gamma_k$ in (32), we have

$$\mathbf{E}\left[\|\bar{x}_{k+1} - x^\star\|^2 | \mathcal{F}_k\right] \leq (1 - \bar{\mu}\gamma_k)\|\bar{x}_k - x^\star\|^2 + \frac{c_2\gamma_k + c_1\gamma_k^2}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma_k^2}{n}. \tag{67}$$

Denote

$$A_k = \|\bar{x}_k - x^\star\|^2, \quad B_k = 0, \quad C_k = \frac{c_2\gamma_k + c_1\gamma_k^2}{n}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma_k^2}{n}, \quad D_k = \bar{\mu}\gamma_k\|\bar{x}_k - x^\star\|^2,$$

(67) falls into the setting of Lemma 5.

In what follows, we verify the conditions of Lemma 5.

Obviously, $\sum_{k=0}^{\infty} B_k < \infty$ and we have

$$\begin{aligned}\sum_{k=0}^{\infty}\mathbf{E}[C_k] &= \sum_{k=0}^{\infty}\frac{c_2\gamma_k + c_1\gamma_k^2}{n}\mathbf{E}\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \sum_{k=0}^{\infty}\frac{\sigma^2\gamma_k^2}{n} \\ &\leq \sum_{k=0}^{\infty} N_2\frac{c_2\gamma_k + c_1\gamma_k^2}{n}\gamma_k^2 + \sum_{k=0}^{\infty}\frac{\sigma^2\gamma_k^2}{n} < \infty,\end{aligned}$$

where the first inequality and the second inequality follow from (66) and the fact that $\gamma_k = r(k+b)^{-a}$ with $a \in (\frac{1}{2}, 1)$, respectively. Monotone convergence theorem implies $\sum_{k=0}^{\infty} C_k < \infty$. Then the conditions of Lemma 5 hold, which implies that there is a non-negative finite random variable A_∞ such that $A_k \to A_\infty$ and $\sum_k D_k < \infty$ almost surely which implies

$$\|\bar{x}_k - x^\star\|^2 \to A_\infty, \qquad \sum_{k=0}^{\infty}\gamma_k\|\bar{x}_k - x^\star\|^2 < \infty$$

almost surely. Noting that $\sum_{k=0}^{\infty}\gamma_k$ diverges, we conclude that $\lim_{k\to\infty}\|\bar{x}_k - x^\star\|^2 = 0$ almost surely. □

Theorem 2 shows that under diminishing step size, the averaged iterates of DSGTD-GD converge to the performative stable solution at rate $\mathcal{O}(\gamma_k)$ and consensus errors of iterates converge to 0 at rate $\mathcal{O}(\gamma_k^2)$. Moreover, we have by (66) that

$$\mathbf{E}\left[\|\mathbf{x}_k - \mathbf{x}^\star\|^2\right] \leq 2n\mathbf{E}\left[\|\bar{x}_k - x^\star\|^2\right] + 2\mathbf{E}\left[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2\right] = \mathcal{O}(\gamma_k),$$

which means DSGTD-GD with diminishing step size achieves the convergence rate $\mathcal{O}(\gamma_k)$.

Next, we move to study the asymptotic normality of DSGTD-GD.

Theorem 3. *Let $x^\star$ be the performative stable solution of problem (1). Under the conditions of Theorem 2, the average iterates $\frac{1}{k}\sum_{t=0}^{k-1} x_{i,t}$ generated by DSGTD-GD satisfy*

$$\sqrt{k}(\frac{1}{k}\sum_{t=0}^{k-1} x_{i,t} - x^\star) \xrightarrow{d} \mathcal{N}(0, \nabla R(x^\star)^{-1} \cdot \Sigma \cdot \nabla R(x^\star)^{-1}), \quad \forall i \in \mathcal{V}, \tag{68}$$

where $R(x) := \sum_{i=1}^n \nabla f_i(x; x)$ and $\Sigma := \mathrm{Cov}(\sum_{i=1}^n \nabla l_i(x^\star; \xi_{i,k}))$.

Proof. By (66) and the fact that $\sum_{k=0}^{\infty} \frac{\gamma_k}{\sqrt{k}} < \infty$, we have

$$\begin{aligned}\mathbf{E}\left[\|\sqrt{k}(\frac{1}{k}\sum_{t=0}^{k-1} x_{i,t} - x^\star) - \sqrt{k}(\frac{1}{k}\sum_{t=0}^{k-1}\bar{x}_t - x^\star)\|\right] &= \frac{1}{\sqrt{k}}\mathbf{E}\left[\|\sum_{t=0}^{k-1}(x_{i,t} - \bar{x}_t)\|\right] \\ &\leq \frac{1}{\sqrt{k}}\sum_{t=0}^{k-1}\sqrt{\mathbf{E}\|(\mathbf{x}_t - \bar{\mathbf{x}}_t)\|^2} \\ &\leq \frac{\sqrt{N_2}}{\sqrt{k}}\sum_{t=0}^{k-1}\gamma_t,\end{aligned}$$

where the last term converges to 0 due to Kronecker lemma. Then Slutsky's theorem implies (68) if

$$\sqrt{k}(\frac{1}{k}\sum_{t=0}^{k-1}\bar{x}_t - x^\star) \xrightarrow{d} N(0, \nabla R(x^\star)^{-1}\cdot\Sigma\cdot\nabla R(x^\star)^{-1}). \tag{69}$$

Next, we show (69) by Lemma 3 in Appendix.

Firstly, we rewrite the recursion $\bar{x}_k - x^\star$ in the form of (91) in Lemma 3. By the update recursion in (10) and the fact $\bar{y}_k = \frac{1}{n}\sum_{i=1}^n \nabla l_i(x_{i,k};\xi_{i,k+1})$, we have

$$\begin{aligned}\bar{x}_{k+1} - x^\star &= \bar{x}_k - \gamma_k\bar{y}_k - x^\star \\ &= (\mathbf{I} - \frac{\gamma_k}{n}\nabla R(x^\star))(\bar{x}_k - x^\star) - \gamma_k\left[\frac{1}{n}\sum_{i=1}^n \nabla l_i(x_{i,k};\xi_{i,k+1}) - \frac{1}{n}\sum_{i=1}^n \nabla l_i(\bar{x}_k;\xi_{i,k+1})\right] \\ &\quad - \gamma_k\left[\frac{1}{n}\nabla f(\bar{x}_k;\bar{x}_k) - \frac{1}{n}\nabla R(x^\star)(\bar{x}_k - x^\star)\right] - \gamma_k\left[\frac{1}{n}\sum_{i=1}^n \nabla l_i(\bar{x}_k;\xi_{i,k+1})\right. \\ &\quad \left. - \frac{1}{n}\nabla f(\bar{x}_k;\bar{x}_k)\right],\end{aligned} \tag{70}$$

where $R(x) := \sum_{i=1}^n \nabla f_i(x;x)$.

Denote

$$\Theta_k = \bar{x}_k - x^\star, \quad \mathbf{G} = \frac{1}{n}\nabla R(x^\star), \quad \alpha_k = \gamma_k, \quad \mu_k = -\frac{1}{n}\sum_{i=1}^n \nabla l_i(\bar{x}_k;\xi_{i,k+1}) + \frac{1}{n}\nabla f(\bar{x}_k;\bar{x}_k)$$

and

$$\eta_k = -(\frac{1}{n}\nabla f(\bar{x}_k;\bar{x}_k) - \frac{1}{n}\nabla R(x^\star)(\bar{x}_k - x^\star)) - (\frac{1}{n}\sum_{i=1}^n \nabla l_i(x_{i,k};\xi_{i,k+1}) - \frac{1}{n}\sum_{i=1}^n \nabla l_i(\bar{x}_k;\xi_{i,k+1})),$$

the recursion (70) can be rewritten as

$$\Theta_{k+1} = (\mathbf{I} - \alpha_k\mathbf{G})\Theta_k + \alpha_k(\mu_k + \eta_k),$$

which falls into the setting of Lemma 3.

In what follows, we verify the conditions (a)-(d) of Lemma 3.

According to the definition of γ_k and the strong convexity of $f(\cdot\,;x)$, conditions (a), (b) of Lemma 3 hold obviously.

Denote

$$\mu_k^{(1)} := -\frac{1}{n}\sum_{i=1}^{n}\nabla l_i(x^\star;\xi_{i,k+1}),$$
$$\mu_k^{(2)} := -\frac{1}{n}\sum_{i=1}^{n}\nabla l_i(\bar{x}_k;\xi_{i,k+1}) + \frac{1}{n}\nabla f(\bar{x}_k;\bar{x}_k) + \frac{1}{n}\sum_{i=1}^{n}\nabla l_i(x^\star;\xi_{i,k+1}).$$

Obviously, $\{\mu_k^{(1)}\}$ and $\{\mu_k^{(2)}\}$ are martingale difference sequences and μ_k can be decomposed into $\mu_k = \mu_k^{(1)} + \mu_k^{(2)}$. By Assumption 2 (d) and (e), there exists constant c such that

$$\mathbf{E}[\|\mu_k^{(2)}\|^2|\mathcal{F}_k] \le c\|\Theta_k\|^2, \qquad \mathbf{E}[\|\mu_k^{(1)}\|^2|\mathcal{F}_k] \le c.$$

Moreover, we apply [6, Lemma 3.3.1] to verify (92) in condition (c) of Lemma 3. Denote

$$\zeta_{k,t} = \frac{\mu_t^{(1)}}{\sqrt{k}}, \qquad \mathbf{H}_{k,t} = \mathbf{E}[\zeta_{k,t}\zeta_{k,t}^\intercal], \qquad \mathbf{K}_{k,t} = \mathbf{E}[\zeta_{k,t}\zeta_{k,t}^\intercal|\zeta_{k,0},...,\zeta_{k,t-1}], \qquad \mathbf{H}_k = \sum_{t=0}^{k-1}\mathbf{H}_{k,t}.$$

Noting that $\{\mu_t^{(1)}\}$ is a martingale difference sequence, we have

$$\mathbf{E}[\zeta_{k,t}|\zeta_{k,0},...,\zeta_{k,t-1}] = 0,$$

which implies the condition (3.3.1) of [6, Lemma 3.3.1]. Next, we verify the conditions (3.3.2)-(3.3.3) of [6, Lemma 3.3.1]. By the definition of $\zeta_{k,t}$,

$$\mathbf{E}[\|\zeta_{k,t}\|^p] = \mathbf{E}_{\xi_{i,t+1}\sim\mathcal{D}_i(x^\star)}[\|\frac{\frac{1}{n}\sum_{i=1}^{n}\nabla l_i(x^\star;\xi_{i,t+1})}{\sqrt{k}}\|^p] \le \frac{\sum_{i=1}^{n}\mathbf{E}[\|\nabla l_i(x^\star;\xi_{i,k+1})\|^p]}{nk^{\frac{p}{2}}} \le \frac{c_g^{\frac{p}{2}}}{k^{\frac{p}{2}}},$$

where the last inequality follows from Assumption 2 (e). Then

$$\sup_{k\ge1}\sum_{t=0}^{k-1}\mathbf{E}[\|\zeta_{k,t}\|^2] \le \sup_{k\ge1}\sum_{t=0}^{k-1}(\mathbf{E}[\|\zeta_{k,t}\|^p])^{\frac{2}{p}} \le \sup_{k\ge1}\frac{kc_g}{k} = c_g.$$

Note that

$$\mathbf{H}_k = \sum_{t=0}^{k-1}\mathbf{H}_{k,t} = \sum_{t=0}^{k-1}\frac{1}{kn^2}Cov(\sum_{i=1}^{n}\nabla l_i(x^\star;\xi_{i,k})) = \frac{1}{n^2}\Sigma,$$

and then condition (3.3.2) of [6, Lemma 3.3.1] holds. Moreover, the fact that $\mathbf{H}_{k,t} = \mathbf{K}_{k,t}$ almost surely implies the condition (3.3.3) of [6, Lemma 3.3.1] directly. By Assumption 2 (e), we have for any $\tau > 0$,

$$\begin{aligned}
\mathbf{E}[\|\zeta_{k,t}\|^2\mathbf{1}_{\{\|\zeta_{k,t}\|\ge\tau\}}] &\le (\mathbf{E}[\|\zeta_{k,t}\|^p])^{\frac{2}{p}}(\mathbf{E}[\mathbf{1}^q_{\{\|\zeta_{k,t}\|\ge\tau\}}])^{\frac{1}{q}} \\
&= (\mathbf{E}[\|\zeta_{k,t}\|^p])^{\frac{2}{p}}\mathbf{P}^{\frac{1}{q}}(\|\zeta_{k,t}\|\ge\tau) \\
&\le (\mathbf{E}[\|\zeta_{k,t}\|^p])^{\frac{2}{p}}(\frac{\mathbf{E}[\|\zeta_{k,t}\|]}{\tau})^{\frac{1}{q}} \\
&\le \frac{c_g^{1+\frac{1}{2q}}}{k^{1+\frac{1}{2q}}\tau^{\frac{1}{q}}},
\end{aligned}$$

where $p > 2$ is defined in Assumption 2 (e), constant q satisfies $\frac{2}{p} + \frac{1}{q} = 1$, the first inequality and the second inequality follow from *Hölder* inequality and Markov inequality, respectively. Then

$$\lim_{k\to\infty} \sum_{t=0}^{k-1} \mathbf{E}[\|\zeta_{k,t}\|^2 \mathbf{1}_{\{\|\zeta_{k,t}\|\geq\tau\}}] \leq \lim_{k\to\infty} \frac{k c_g^{1+\frac{1}{2q}}}{k^{1+\frac{1}{2q}} \tau^{\frac{1}{q}}} = 0,$$

which means the condition (3.3.4) of [6, Lemma 3.3.1] holds. Summarizing above, all the conditions of [6, Lemma 3.3.1] hold and then

$$\frac{1}{\sqrt{k}} \sum_{t=1}^{k} \mu_t^{(1)} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \frac{1}{n^2}\Sigma),$$

where $\Sigma := \mathrm{Cov}(\sum_{i=1}^n \nabla l_i(x^\star; \xi_{i,k+1}))$. Then condition (c) of Lemma 3 holds.

Next, we verify condition (d) of Lemma 3.

By the definition of the performative stable point, $R(x^\star) = 0$. By Assumption 4, $\nabla R(x)$ is Lipschitz on a neighborhood of $x^\star$ implying

$$R(\bar{x}_k) - \nabla R(x^\star)(\bar{x}_k - x^\star) = O(\|\bar{x}_k - x^\star\|^2), \quad as\ \bar{x}_k \to x^\star.$$

Then, by the definition of η_k, there exists constant $c' > 0$ such that

$$\mathbf{E}[\|\eta_k\|] \leq c' \mathbf{E}[\|\bar{x}_k - x^\star\|^2] + \frac{L(1+\epsilon_{max})}{\sqrt{n}} \mathbf{E}[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|] = \mathcal{O}(\gamma_k),$$

where the second term in the right of the inequality is due to [12, Lemma 2.1]. Subsequently, we have $\sum_{t=0}^{\infty} \frac{\mathbf{E}[\|\eta_k\|]}{\sqrt{t+1}} < \infty$. Monotone convergence theorem further implies $\sum_{t=0}^{\infty} \frac{\|\eta_k\|}{\sqrt{t+1}} < \infty$, thereby condition (d) of Lemma 3 holds by Kronecker lemma.

By the fact that $\bar{x}_t - x^\star \to 0$ almost surely in Theorem 2, monotone convergence theorem implies

$$\sum_{t=0}^{\infty} \frac{\|\bar{x}_t - x^\star\|^2}{\sqrt{t+1}} < \infty$$

almost surely. Then Kronecker lemma induces the condition (e) of Lemma 3.

Summarize above, all the conditions of Lemma 3 in Appendix hold. Then

$$\sqrt{k}(\frac{1}{k} \sum_{t=0}^{k-1} x_{i,t} - x^\star) \xrightarrow{d} \mathcal{N}(0, \nabla R(x^\star)^{-1} \cdot \Sigma \cdot \nabla R(x^\star)^{-1}).$$

The proof is complete. □

Theorem 3 shows that the deviation between the averaged iterates generated by DSGTD-GD and the performative stable solution converges in distribution to a normal random vector with the covariance matrix $\nabla R(x^\star)^{-1} \cdot \Sigma \cdot \nabla R(x^\star)^{-1}$, where

$$\nabla R(x^\star) = \sum_{i=1}^{n} \nabla^2 f_i(x^\star; x^\star) + \sum_{i=1}^{n} \frac{d}{dy} \nabla f_i(x^\star; y)|_{y=x^\star}. \tag{71}$$

Comparing with the asymptotic normality results of stochastic approximation method for standard stochastic optimization problems [6, 8, 16, 37], the second term $\sum_{i=1}^{n} \frac{d}{dy} \nabla f_i(x^\star; y)|_{y=x^\star}$ captures the performative effects of distributional shift. To the best of our knowledge, Theorem 3 seems to be the first result on the asymptotic normality of the stochastic approximation method for distributed SO-DD.

4 Convergence of DSGTD-AG

In this section, we study the convergence of DSGTD-AG. Under the constant step size policy, we show that DSGTD-AG converges sublinearly to a neighborhood of the optimal solution based on the technical lemma.

We first state the descent lemma when the objective function is non-convex and the stochastic gradient is biased, which plays a key role in analyzing the convergence of DSGTD-AG.

Lemma 2. *Suppose that $\gamma \leq \frac{n}{6L}$, Assumptions 1, 5 and 6 hold. Then, for $\forall k$, the averaged iterations $\bar{x}_k$ generated by DSGTD-AG satisfy*

$$\begin{aligned}\mathbf{E}\left[f(\bar{x}_{k+1})|\mathcal{F}_k\right] \leq & f(\bar{x}_k) - \frac{\gamma}{4n}\|\nabla f(\bar{x}_k)\|^2 + \frac{3\gamma L^2}{2}\sum_{i=1}^{n}\|\bar{x}_k - x_{i,k}\|^2 \\ & + (\frac{3L\delta^2\gamma^2}{n} + \gamma\delta^2 + \frac{3L\sigma^2\gamma^2}{n})\|A_k - A\|^2 + \frac{3L\sigma^2\gamma^2}{n}\|A\|^2 + \frac{3L\sigma^2\gamma^2}{2n}, \end{aligned} \tag{72}$$

where $A := [A_1, ..., A_n]$ and $A_k := [A_{1,k}, ..., A_{n,k}]$.

Proof. By the updating recursion (16) and the definition of $\mathbf{v}_k$,

$$\begin{aligned}\mathbf{E}\left[f(\bar{x}_{k+1})|\mathcal{F}_k\right] &= \mathbf{E}\left[f(\bar{x}_k - \frac{\gamma}{n}\sum_{i=1}^{n} v_{i,k})|\mathcal{F}_k\right] \\ &\leq f(\bar{x}_k) - \mathbf{E}\left[\langle\nabla f(\bar{x}_k), \frac{\gamma}{n}\sum_{i=1}^{n} v_{i,k}\rangle|\mathcal{F}_k\right] + \frac{L\gamma^2}{2}\mathbf{E}\left[\|\frac{1}{n}\sum_{i=1}^{n} v_{i,k}\|^2|\mathcal{F}_k\right] \\ &= f(\bar{x}_k) - \frac{\gamma}{n}\|\nabla f(\bar{x}_k)\|^2 + \frac{\gamma}{n}\mathbf{E}\left[\langle\nabla f(\bar{x}_k), \nabla f(\bar{x}_k) - \sum_{i=1}^{n} v_{i,k}\rangle|\mathcal{F}_k\right] \\ &\quad + \frac{L\gamma^2}{2}\mathbf{E}\left[\|\frac{1}{n}\sum_{i=1}^{n} v_{i,k}\|^2|\mathcal{F}_k\right] \\ &\leq f(\bar{x}_k) - \frac{\gamma}{2n}\|\nabla f(\bar{x}_k)\|^2 + \frac{\gamma}{2}\sum_{i=1}^{n}\|\nabla f_i(\bar{x}_k) - \mathbf{E}\left[v_{i,k}|\mathcal{F}_k\right]\|^2 \\ &\quad + \frac{L\gamma^2}{2}\mathbf{E}\left[\|\frac{1}{n}\sum_{i=1}^{n} v_{i,k}\|^2|\mathcal{F}_k\right], \end{aligned} \tag{73}$$

where $\mathcal{F}_k$ denotes the σ-algebra generated by $(x_{i,l}, u_{i,l})_{l=0,...,k-1}, i \in \mathcal{V}$, the first inequality and the last inequality follow from the smoothness of the objective function and the Young's

inequality, respectively. For the last term on the right-hand side of (73),

$$\begin{aligned}\mathbf{E}\left[\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k}\|^2|\mathcal{F}_k\right] &= \mathbf{E}\left[\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k}+\frac{1}{n}\nabla f(\bar{x}_k)-\frac{1}{n}\nabla f(\bar{x}_k)+\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}v_{i,k}\right.\\ &\left.-\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}v_{i,k}\|^2|\mathcal{F}_k\right]\\ &\leq 3\mathbf{E}\left[\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k}-\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}v_{i,k}\|^2|\mathcal{F}_k\right]+\frac{3}{n^2}\|\nabla f(\bar{x}_k)\|^2\\ &+3\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}v_{i,k}-\frac{1}{n}\nabla f(\bar{x}_k)\|^2\\ &\leq \frac{3\sigma^2}{n}\max\{1,\|A_k\|^2\}+\frac{3}{n}\sum_{i=1}^{n}\|\mathbf{E}v_{i,k}-\nabla f_i(\bar{x}_k)\|^2+\frac{3}{n^2}\|\nabla f(\bar{x}_k)\|^2,\end{aligned}$$

where the last inequality is due to the definition of $v_{i,k}$ in (30) and Assumption 5 (c). Then

$$\begin{aligned}\mathbf{E}\left[f(\bar{x}_{k+1})|\mathcal{F}_k\right] \leq & f(\bar{x}_k)+(\frac{3L\gamma^2}{2n^2}-\frac{\gamma}{2n})\|\nabla f(\bar{x}_k)\|^2\\ &+(\frac{3L\gamma^2}{2n}+\frac{\gamma}{2})\sum_{i=1}^{n}\|\nabla f_i(\bar{x}_k)-\mathbf{E}\left[v_{i,k}|\mathcal{F}_k\right]\|^2+\frac{3L\sigma^2\gamma^2}{2n}\max\{1,\|A_k\|^2\}.\end{aligned}$$

Again, by the definition of $v_{i,k}$ in (30),

$$\begin{aligned}\sum_{i=1}^{n}\|\nabla f_i(\bar{x}_k)-\mathbf{E}\left[v_{i,k}|\mathcal{F}_k\right]\|^2 &= \sum_{i=1}^{n}\|\nabla f_i(\bar{x}_k)-\mathbf{E}\left[v_{i,k}|\mathcal{F}_k\right]+\nabla f_i(x_{i,k})-\nabla f_i(x_{i,k})\|^2\\ &\leq 2\sum_{i=1}^{n}\|\nabla f_i(\bar{x}_k)-\nabla f_i(x_{i,k})\|^2+2\sum_{i=1}^{n}\|\mathbf{E}\left[v_{i,k}|\mathcal{F}_k\right]-\nabla f_i(x_{i,k})\|^2\\ &\leq 2L^2\sum_{i=1}^{n}\|\bar{x}_k-x_{i,k}\|^2+2\delta^2\|A_k-A\|^2,\end{aligned}$$

where the last inequality is due to Assumption 5 (a) and (b). Then

$$\begin{aligned}\mathbf{E}\left[f(\bar{x}_{k+1})|\mathcal{F}_k\right] &\leq f(\bar{x}_k)+(\frac{3L\gamma^2}{2n^2}-\frac{\gamma}{2n})\|\nabla f(\bar{x}_k)\|^2+(\frac{3\gamma^2L^3}{n}+\gamma L^2)\sum_{i=1}^{n}\|\bar{x}_k-x_{i,k}\|^2\\ &+(\frac{3L\gamma^2}{n}+\gamma)\delta^2\|A_k-A\|^2+\frac{3L\sigma^2\gamma^2}{2n}\max\{1,\|A_k\|^2\}\\ &\leq f(\bar{x}_k)-\frac{\gamma}{4n}\|\nabla f(\bar{x}_k)\|^2+\frac{3\gamma L^2}{2}\sum_{i=1}^{n}\|\bar{x}_k-x_{i,k}\|^2\\ &+\frac{3L\sigma^2\gamma^2}{n}\|A\|^2+\frac{3L\sigma^2\gamma^2}{2n}+(\frac{3L\delta^2\gamma^2}{n}+\gamma\delta^2+\frac{3L\sigma^2\gamma^2}{n})\|A_k-A\|^2,\end{aligned}$$

where the last inequality follows from the fact $\gamma \leq \frac{n}{6L}$, $\max\{1,\|A_k\|^2\} \leq 1+\|A_k\|^2$ and $\|A_k\|^2 \leq 2\|A\|^2+2\|A_k-A\|^2$. The proof is complete. □

Different from the descent lemma [22, Lemma 27], the last term on the right-hand side of (72) depends on the error between the true distribution parameter A and the estimated distribution parameter A_k. In other words, with A_k gradually tending to A, the term tends towards 0.

Now, we are ready to study the convergence of DSGTD-AG.

Theorem 4. *Let* $C_1 = 2560$, $C_2 = 12000$, $c_l := \max\{7\|A_0 - A\|^2, 8\sum_{i=1}^n tr(\Sigma_i)\}$, $f^\star := \min_{x\in\mathbb{R}^d} f(x)$ *and denote*

$$
\begin{aligned}
&c_4 := 8n,\\
&c_5 := 12nL^2\tau^2\left[\frac{c_lC_2(2\sigma^2+\delta^2)+2C_2\sigma^2\|A\|^2+C_2\sigma^2}{(1-\rho)^2}+C_1\sigma^2+c_lC_1(2\sigma^2+\delta^2)+2C_1\sigma^2\|A\|^2\right],\\
&c_6 := 24c_lL(\delta^2+\sigma^2)+24L\sigma^2\|A\|^2+12L\sigma^2, \qquad \tau := \frac{2}{1-\rho}log(\frac{50}{1-\rho}(1+log\frac{1}{1-\rho}))+1,\\
&F_0 := f(\bar{x}_0)-f^\star, \qquad \phi_0 := 16\|\mathbf{x}_0-\bar{\mathbf{x}}_0\|^2+\frac{24\gamma^2}{(1-\rho)^2}\|\mathbf{y}_0-\bar{\mathbf{y}}_0\|^2.
\end{aligned}
\tag{74}
$$

Suppose that (i) $K > \frac{2}{1-\rho}log(\frac{50}{1-\rho}(1+log\frac{1}{1-\rho}))$, *(ii) Assumptions 1, 5, 6 hold, (iii) the step size satisfies*

$$
\gamma := \min\left\{\left(\frac{c_4F_0}{c_5(K+1)}\right)^{\frac{1}{3}}, \left(\frac{c_4F_0}{c_6(K+1)}\right)^{\frac{1}{2}}, \frac{1-\rho}{64\sqrt{10}L\tau}\right\}. \tag{75}
$$

Then the averaged iterations $\bar{x}_k$ *generated by DSGTD-AG satisfy*

$$
\begin{aligned}
\frac{1}{K+1}\sum_{k=0}^{K}\mathbf{E}\|\nabla f(\bar{x}_k)\|^2 \leq& 2c_5^{\frac{1}{3}}\left(\frac{c_4F_0}{K+1}\right)^{\frac{2}{3}}+2\left(\frac{c_4c_6F_0}{K+1}\right)^{\frac{1}{2}}+\frac{64\sqrt{10}c_4L\tau F_0}{(1-\rho)(K+1)}\\
&+\frac{768n\tau\phi_0L^2}{K+1}+\frac{8nc_l\delta^2ln(K+6d)}{K+1}.
\end{aligned}
\tag{76}
$$

Proof. Define $\psi_k := \left[\mathbf{x}_k-\bar{\mathbf{x}}_k, \gamma(\mathbf{y}_k-\bar{\mathbf{y}}_k)\right]$, we have

$$
\begin{aligned}
\mathbf{E}\left[f(\bar{x}_{k+1})|\mathcal{F}_k\right] \leq& f(\bar{x}_k)-\frac{\gamma}{4n}\|\nabla f(\bar{x}_k)\|^2+\frac{3\gamma L^2}{2}\mathbf{E}\|\psi_k\|^2\\
&+(\frac{3L\delta^2\gamma^2}{n}+\gamma\delta^2+\frac{3L\sigma^2\gamma^2}{n})\|A_k-A\|^2+\frac{3L\sigma^2\gamma^2}{n}\|A\|^2+\frac{3L\sigma^2\gamma^2}{2n},
\end{aligned}
\tag{77}
$$

where the inequality is due to Lemma 2. Rearranging it and summing up, we have

$$
\begin{aligned}
&\frac{1}{K+1}\sum_{k=0}^{K}\mathbf{E}\|\nabla f(\bar{x}_k)\|^2\\
\leq&\frac{4nF_0}{\gamma(K+1)}+12L\gamma\sigma^2\|A\|^2+6L\gamma\sigma^2\\
&+\frac{4(3L\gamma\delta^2+n\delta^2+3L\gamma\sigma^2)}{K+1}\sum_{k=0}^{K}\mathbf{E}\left[\|A_k-A\|^2\right]+\frac{6nL^2}{K+1}\sum_{k=0}^{K}\mathbf{E}\|\psi_k\|^2,
\end{aligned}
\tag{78}
$$

where $F_0 := f(\bar{x}_0) - f^\star$. Note that the first three terms on the right-hand side of (78) do not depend on iteration k, we may provide the bound for them by selecting the step size, that is,

$$\frac{4nF_0}{\gamma(K+1)} + 12L\gamma\sigma^2\|A\|^2 + 6L\gamma\sigma^2 \leq 6L\sigma^2\left(2\|A\|^2+1\right)\left(\frac{c_4F_0}{c_6(K+1)}\right)^{\frac{1}{2}} + \frac{32\sqrt{10}c_4L\tau F_0}{(1-\rho)(K+1)}, \tag{79}$$

where c_4 and c_6 are defined in (74), the inequality is due to the definition of γ in (75). Similarly, as the fourth term on the right-hand side of (78) only depends on the error of $\|A_k - A\|^2$, we may bound it with [32, Lemma 21],

$$\begin{aligned}
&\frac{4(3L\gamma\delta^2 + n\delta^2 + 3L\gamma\sigma^2)}{K+1}\sum_{k=0}^{K}\mathbf{E}\left[\|A_k - A\|^2\right] \\
\leq&\frac{4c_l(3L\gamma\delta^2 + n\delta^2 + 3L\gamma\sigma^2)}{K+1}\sum_{k=0}^{K}\frac{1}{k+6d} \\
\leq&12c_lL(\delta^2+\sigma^2)\left(\frac{c_4F_0}{c_6(K+1)}\right)^{\frac{1}{2}} + \frac{4nc_l\delta^2 ln(K+6d)}{K+1},
\end{aligned} \tag{80}$$

where $c_l := \max\{7\|A_1 - A\|^2, 8\sum_{i=1}^{n} tr(\Sigma_i)\}$ and the last inequality follows from the definition of γ in (75).

Next, we derive the bound of $\sum_{k=0}^{K}\mathbf{E}\|\psi_k\|^2$.

Define

$$J := \begin{bmatrix} \tilde{\mathbf{W}} & 0 \\ -\tilde{\mathbf{W}} & \tilde{\mathbf{W}} \end{bmatrix}, \quad \Xi_k := \begin{bmatrix} 0 \\ (\mathbf{v}_{k+1} - \mathbf{v}_k)(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal) \end{bmatrix}^\intercal, \quad \tilde{\mathbf{W}} := \mathbf{W} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal,$$

we may rewrite Algorithm 2 in the form as

$$\psi_{k+1} = \psi_k J + \gamma\Xi_k, \tag{81}$$

where $\psi_k = \Big[\mathbf{x}_k - \bar{\mathbf{x}}_k, \gamma(\mathbf{y}_k - \bar{\mathbf{y}}_k)\Big]$. By unrolling (81) for $\tau \leq t \leq 2\tau$ steps with $\tau = \frac{2}{1-\rho}log(\frac{50}{1-\rho}(1+log\frac{1}{1-\rho})) + 1$,

$$\psi_{k+t} = \psi_k J^t + \gamma\sum_{m=1}^{t}\Xi_{k+m-1}J^{t-m}.$$

Then,

$$\begin{aligned}
\mathbf{E}\|\psi_{k+t}\|^2 \leq&\frac{3}{4}\mathbf{E}\|\psi_k\|^2 + 5\gamma^2\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m} - \mathbf{v}_{k+m-1})\tilde{\mathbf{W}}^{t-m}\|^2 \\
&+ 5\gamma^2\mathbf{E}\|\sum_{m=1}^{t-1}(\mathbf{v}_{k+m} - \mathbf{v}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m}\|^2,
\end{aligned} \tag{82}$$

where the inequality follows from the Young's inequality, [22, Lemma 4] and the fact that $\|\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal\|_2\leq 1$.

In what follows, we give the upper bounds for the last two items on the right-hand side of (82).

Denote

$$\mathbf{F}_k:=[\nabla f_1(x_{1,k}),...,\nabla f_n(x_{n,k})],\quad \bar{\mathbf{F}}_k:=[\nabla f_1(\bar{x}_k),...,\nabla f_n(\bar{x}_k)].$$

For the second term on the right-hand side of (82),

$$\begin{aligned}
&\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m}-\mathbf{v}_{k+m-1})\tilde{\mathbf{W}}^{t-m}\|^2\\
&\leq 3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m}-\mathbf{F}_{k+m})\|^2+3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m-1}-\mathbf{F}_{k+m-1})\|^2+3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{F}_{k+m}-\mathbf{F}_{k+m-1})\|^2\\
&\leq 3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m}-\mathbf{F}_{k+m})\|^2+3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m-1}-\mathbf{F}_{k+m-1})\|^2+3t\sum_{m=1}^{t}\mathbf{E}\|\mathbf{F}_{k+m}-\mathbf{F}_{k+m-1}\|^2\\
&\leq 3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m}-\mathbf{F}_{k+m})\|^2+3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m-1}-\mathbf{F}_{k+m-1})\|^2+9t\sum_{m=1}^{t}\mathbf{E}\|\mathbf{F}_{k+m}-\bar{\mathbf{F}}_{k+m}\|^2\\
&+9t\sum_{m=1}^{t}\mathbf{E}\|\mathbf{F}_{k+m-1}-\bar{\mathbf{F}}_{k+m-1}\|^2++9t\sum_{m=1}^{t}\mathbf{E}\|\bar{\mathbf{F}}_{k+m}-\bar{\mathbf{F}}_{k+m-1}\|^2\\
&\leq 3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m}-\mathbf{F}_{k+m})\|^2+3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m-1}-\mathbf{F}_{k+m-1})\|^2+9tL^2\sum_{m=1}^{t}\mathbf{E}\|\mathbf{x}_{k+m}-\bar{\mathbf{x}}_{k+m}\|^2\\
&+9tL^2\sum_{m=1}^{t}\mathbf{E}\|\mathbf{x}_{k+m-1}-\bar{\mathbf{x}}_{k+m-1}\|^2+9tnL^2\sum_{m=1}^{t}\mathbf{E}\|\bar{x}_{k+m}-\bar{x}_{k+m-1}\|^2,
\end{aligned}\tag{83}$$

where the first inequality and the last inequality follow from the fact that $\|\tilde{\mathbf{W}}^j\|\leq 1$, for $\forall j\geq 1$, and Assumption 5 (a), respectively. For the third term on the right-hand side of (82),

$$\begin{aligned}
&\mathbf{E}\|\sum_{m=1}^{t-1}(\mathbf{v}_{k+m}-\mathbf{v}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m}\|^2\\
&\leq 3\mathbf{E}\|\sum_{m=1}^{t-1}(\mathbf{v}_{k+m}-\mathbf{F}_{k+m}-\mathbf{v}_{k+m-1}+\mathbf{F}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m}\|^2\\
&+3\mathbf{E}\|\sum_{m=1}^{t-1}(\mathbf{F}_{k+m}-\bar{\mathbf{F}}_{k+m}-\mathbf{F}_{k+m-1}+\bar{\mathbf{F}}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m}\|^2\\
&+3\mathbf{E}\|\sum_{m=1}^{t-1}(\bar{\mathbf{F}}_{k+m}-\bar{\mathbf{F}}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m}\|^2\\
&=3\mathbf{E}\|(\mathbf{v}_{k+t-1}-\mathbf{F}_{k+t-1})\tilde{\mathbf{W}}-(\mathbf{v}_k-\mathbf{F}_k)(t-1)\tilde{\mathbf{W}}^{t-1}\\
&+\sum_{m=1}^{t-2}(\mathbf{v}_{k+m}-\mathbf{F}_{k+m})\left[(t-m)\tilde{\mathbf{W}}^{t-m}-(t-m-1)\tilde{\mathbf{W}}^{t-m-1}\right]\|^2\\
&+3\mathbf{E}\|(\mathbf{F}_{k+t-1}-\bar{\mathbf{F}}_{k+t-1})\tilde{\mathbf{W}}-(\mathbf{F}_k-\bar{\mathbf{F}}_k)(t-1)\tilde{\mathbf{W}}^{t-1}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{m=1}^{t-2}(\mathbf{F}_{k+m}-\bar{\mathbf{F}}_{k+m})\left[(t-m)\tilde{\mathbf{W}}^{t-m}-(t-m-1)\tilde{\mathbf{W}}^{t-m-1}\right]\|^2\\
&+3\mathbf{E}\|\sum_{m=1}^{t-1}(\bar{\mathbf{F}}_{k+m}-\bar{\mathbf{F}}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m}\|^2\\
&\le 3t\mathbf{E}\|(\mathbf{v}_{k+t-1}-\mathbf{F}_{k+t-1})\tilde{\mathbf{W}}\|^2+3t\mathbf{E}\|(\mathbf{v}_k-\mathbf{F}_k)(t-1)\tilde{\mathbf{W}}^{t-1}\|^2\\
&+3t\sum_{m=1}^{t-2}\|(\mathbf{v}_{k+m}-\mathbf{F}_{k+m})\left[(t-m)\tilde{\mathbf{W}}^{t-m}-(t-m-1)\tilde{\mathbf{W}}^{t-m-1}\right]\|^2\\
&+3t\mathbf{E}\|(\mathbf{F}_{k+t-1}-\bar{\mathbf{F}}_{k+t-1})\tilde{\mathbf{W}}\|^2+3t\mathbf{E}\|(\mathbf{F}_k-\bar{\mathbf{F}}_k)(t-1)\tilde{\mathbf{W}}^{t-1}\|^2\\
&+3t\sum_{m=1}^{t-2}\|(\mathbf{F}_{k+m}-\bar{\mathbf{F}}_{k+m})\left[(t-m)\tilde{\mathbf{W}}^{t-m}-(t-m-1)\tilde{\mathbf{W}}^{t-m-1}\right]\|^2\\
&+3\mathbf{E}\|\sum_{m=1}^{t-1}(\bar{\mathbf{F}}_{k+m}-\bar{\mathbf{F}}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m}\|^2\\
&\le\frac{12t}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\mathbf{v}_{k+m}-\mathbf{F}_{k+m}\|^2+\frac{12t}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\mathbf{F}_{k+m}-\bar{\mathbf{F}}_{k+m}\|^2\\
&+\frac{3(t-1)}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\bar{\mathbf{F}}_{k+m}-\bar{\mathbf{F}}_{k+m-1}\|^2\\
&\le\frac{12t}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\mathbf{v}_{k+m}-\mathbf{F}_{k+m}\|^2+\frac{12tL^2}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\mathbf{x}_{k+m}-\bar{\mathbf{x}}_{k+m}\|^2\\
&+\frac{3n(t-1)L^2}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\bar{x}_{k+m}-\bar{x}_{k+m-1}\|^2,
\end{aligned}
\tag{84}
$$

where the third inequality is due to Lemma 6, Lemma 7, the fact that $\|\tilde{\mathbf{W}}^j\|\le 1$, for $\forall j\ge 1$, and the last inequality is due to Assumption 5 (a). By the updating recursion in (16),

$$
\begin{aligned}
&\mathbf{E}\|\bar{x}_{k+m}-\bar{x}_{k+m-1}\|^2\\
&=\gamma^2\mathbf{E}\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k+m-1}\|^2\\
&=\gamma^2\mathbf{E}\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k+m-1}-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k+m-1})+\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k+m-1})\\
&-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}_{k+m-1})+\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\bar{x}_{k+m-1})\|^2\\
&\le 3\gamma^2\mathbf{E}\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k+m-1}-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k+m-1})\|^2\\
&+\frac{3\gamma^2}{n}\sum_{i=1}^{n}\|\nabla f_i(x_{i,k+m-1})-\nabla f_i(\bar{x}_{k+m-1})\|^2+\frac{3\gamma^2}{n^2}\|\nabla f(\bar{x}_{k+m-1})\|^2\\
&\le 3\gamma^2\mathbf{E}\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k+m-1}-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k+m-1})\|^2
\end{aligned}
$$

$$+\frac{3L^2\gamma^2}{n}\sum_{i=1}^{n}\|x_{i,k+m-1}-\bar{x}_{i,k+m-1}\|^2+\frac{3\gamma^2}{n^2}\|\nabla f(\bar{x}_{k+m-1})\|^2, \tag{85}$$

where the last inequality is due to Assumption 5 (a).

Combining (83), (84), (85) with the inequality (82), we have

$$\begin{aligned}
\mathbf{E}\|\psi_{k+t}\|^2 \le \frac{3}{4}\mathbf{E}\|\psi_k\|^2 + 5\gamma^2\Bigg[&3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m}-\mathbf{F}_{k+m})\|^2 + 3\mathbf{E}\|\sum_{m=1}^{t}(\mathbf{v}_{k+m-1}-\mathbf{F}_{k+m-1})\|^2\\
&+9tL^2\sum_{m=1}^{t}\mathbf{E}\|\mathbf{x}_{k+m}-\bar{\mathbf{x}}_{k+m}\|^2 + 9tL^2\sum_{m=1}^{t}\mathbf{E}\|\mathbf{x}_{k+m-1}-\bar{\mathbf{x}}_{k+m-1}\|^2\\
&+27tnL^2\gamma^2\sum_{m=1}^{t}\mathbf{E}\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k+m-1}-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k+m-1})\|^2\\
&+27t\gamma^2L^4\sum_{m=1}^{t}\|\mathbf{x}_{k+m-1}-\bar{\mathbf{x}}_{k+m-1}\|^2 + \frac{27tL^2\gamma^2}{n}\sum_{m=1}^{t}\|\nabla f(\bar{x}_{k+m-1})\|^2\\
&+\frac{12t}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\mathbf{v}_{k+m}-\mathbf{F}_{k+m}\|^2 + \frac{12tL^2}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\mathbf{x}_{k+m}-\bar{\mathbf{x}}_{k+m}\|^2\\
&+\frac{9n(t-1)L^2\gamma^2}{(1-\rho)^2}\sum_{m=1}^{t-1}\mathbf{E}\|\frac{1}{n}\sum_{i=1}^{n}v_{i,k+m-1}-\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_{i,k+m-1})\|^2\\
&+\frac{9(t-1)\gamma^2L^4}{(1-\rho)^2}\sum_{m=1}^{t-1}\|\mathbf{x}_{k+m-1}-\bar{\mathbf{x}}_{k+m-1}\|^2 + \frac{9(t-1)L^2\gamma^2}{n(1-\rho)^2}\sum_{m=1}^{t-1}\|\nabla f(\bar{x}_{k+m-1})\|^2\Bigg]\\
\le \frac{7}{8}\mathbf{E}\|\psi_k\|^2 &+ \frac{1}{128\tau}\sum_{m=0}^{t-1}\mathbf{E}\|\psi_{k+m}\|^2 + \frac{2\tau\gamma^2}{n}\sum_{m=0}^{t-1}\|\nabla f(\bar{x}_{k+m})\|^2\\
&+ 64\tau\gamma^2\sum_{m=0}^{t-1}\mathbf{E}\|\mathbf{v}_{k+m}-\mathbf{F}_{k+m}\|^2 + \frac{300\tau\gamma^2}{(1-\rho)^2}\sum_{m=1}^{t}\mathbf{E}\|\mathbf{v}_{k+m}-\mathbf{F}_{k+m}\|^2,
\end{aligned}$$

where the last inequality follows from the fact $\gamma \le \frac{1-\rho}{64\sqrt{10}L\tau}$ and $t < 2\tau$. By the definition of $\mathbf{v}_{k+m}$ and $\mathbf{F}_{k+m}$,

$$\begin{aligned}
\mathbf{E}\|\mathbf{v}_{k+m}-\mathbf{F}_{k+m}\|^2 &= \mathbf{E}\|\mathbf{v}_{k+m}-\mathbf{E}\mathbf{v}_{k+m}\|^2 + \|\mathbf{E}\mathbf{v}_{k+m}-\mathbf{F}_{k+m}\|^2\\
&\le (2\sigma^2+\delta^2)\mathbf{E}\left[\|A_{k+m}-A\|^2\right] + 2\sigma^2\|A\|^2+\sigma^2\\
&\le \frac{c_l(2\sigma^2+\delta^2)}{k+m+6d} + 2\sigma^2\|A\|^2+\sigma^2,
\end{aligned} \tag{86}$$

where $c_l := \max\{7\|A_1-A\|^2, 8\sum_{i=1}^{n} tr(\Sigma_i)\}$, the first inequality follows from Assumption 5 (b), (c), the second inequality follows from the fact $\max\{1,\|A_k\|^2\} \le 1+\|A_k\|^2$ and $\|A_k\|^2 \le 2\|A\|^2 + 2\|A_k-A\|^2$, and the last inequality follows from [32, Lemma 21]. Then, by the fact that $t < 2\tau$,

$$\begin{aligned}
\mathbf{E}\|\psi_{k+t}\|^2 \le &\frac{7}{8}\mathbf{E}\|\psi_k\|^2 + \frac{1}{128\tau}\sum_{m=0}^{t-1}\mathbf{E}\|\psi_{k+m}\|^2 + \frac{2\tau\gamma^2}{n}\sum_{m=0}^{t-1}\|\nabla f(\bar{x}_{k+m})\|^2\\
&+ 128\tau^2\gamma^2\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6d} + 2\sigma^2\|A\|^2+\sigma^2\right]
\end{aligned}$$

$$+\frac{600\tau^2\gamma^2}{(1-\rho)^2}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6d}+2\sigma^2\|A\|^2+\sigma^2\right],$$

which falls into the similar form to the recursion in [22, Lemma 7]. Subsequently, unrolling this recursion with the technique in [22, Lemma 8], we have

$$\mathbf{E}\|\psi_k\|^2 \le \left(1-\frac{1}{64\tau}\right)^k\phi_0+\frac{44\tau\gamma^2}{n}\sum_{m=0}^{k-1}\left(1-\frac{1}{64\tau}\right)^{k-m}\|\nabla f(\bar{x}_m)\|^2+C_1\tau^2\gamma^2\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6d}\right.$$
$$\left.+2\sigma^2\|A\|^2+\sigma^2\right]+\frac{C_2\tau^2\gamma^2}{(1-\rho)^2}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6d}+2\sigma^2\|A\|^2+\sigma^2\right], \tag{87}$$

where $C_1=2560$, $C_2=12000$ and $\phi_0=16\|\mathbf{x}_0-\bar{\mathbf{x}}_0\|^2+\frac{24\gamma^2}{(1-\rho)^2}\|\mathbf{y}_0-\bar{\mathbf{y}}_0\|^2$. Then

$$\begin{aligned}\sum_{k=0}^{K}\mathbf{E}\|\psi_k\|^2 &\le \sum_{k=0}^{K}\left(1-\frac{1}{64\tau}\right)^k\phi_0+\frac{44\tau\gamma^2}{n}\sum_{k=0}^{K}\sum_{m=0}^{k-1}\left(1-\frac{1}{64\tau}\right)^{k-m}\|\nabla f(\bar{x}_m)\|^2\\
&+C_1\tau^2\gamma^2\sum_{k=0}^{K}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6}+2\sigma^2\|A\|^2+\sigma^2\right]\\
&+\frac{C_2\tau^2\gamma^2}{(1-\rho)^2}\sum_{k=0}^{K}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+7}+2\sigma^2\|A\|^2+\sigma^2\right]\\
&\le 64\tau\phi_0+\frac{1}{12nL^2}\sum_{k=0}^{K}\|\nabla f(\bar{x}_k)\|^2+C_1\tau^2\gamma^2\sum_{k=0}^{K}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6}+2\sigma^2\|A\|^2+\sigma^2\right]\\
&+\frac{C_2\tau^2\gamma^2}{(1-\rho)^2}\sum_{k=0}^{K}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+7}+2\sigma^2\|A\|^2+\sigma^2\right],\end{aligned} \tag{88}$$

where the last inequality follows from the fact $\gamma\le\frac{1-\rho}{64\sqrt{10}L\tau}$, $\sum_{k=0}^{K}\left(1-\frac{1}{64\tau}\right)^k\le 64\tau$ and

$$\sum_{k=0}^{K}\sum_{m=0}^{k-1}\left(1-\frac{1}{64\tau}\right)^{k-m}\|\nabla f(\bar{x}_m)\|^2\le 64\tau\sum_{k=0}^{K}\|\nabla f(\bar{x}_k)\|^2.$$

Combining (79), (80), (88) with (78) and rearranging, we have

$$\begin{aligned}\frac{1}{K+1}\sum_{k=0}^{K}\mathbf{E}\|\nabla f(\bar{x}_k)\|^2 &\le 12L\sigma^2\left(2\|A\|^2+1\right)\left(\frac{c_4F_0}{c_6(K+1)}\right)^{\frac{1}{2}}+\frac{64\sqrt{10}c_4L\tau F_0}{n(1-\rho)(K+1)}\\
&+24L(\delta^2+\sigma^2)\left(\frac{c_4F_0}{c_6(K+1)}\right)^{\frac{1}{2}}+\frac{8nc_l\delta^2ln(K+6d)}{K+1}+\frac{12nL^2}{K+1}\left[64\tau\phi_0\right.\\
&+C_1\tau^2\gamma^2\sum_{k=0}^{K}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6d}+2\sigma^2\|A\|^2+\sigma^2\right]\\
&\left.+\frac{C_2\tau^2\gamma^2}{(1-\rho)^2}\sum_{k=0}^{K}\left[\frac{c_l(2\sigma^2+\delta^2)}{k+6d}+2\sigma^2\|A\|^2+\sigma^2\right]\right]\\
&\le 12L\sigma^2\left(2\|A\|^2+1\right)\left(\frac{c_4F_0}{c_6(K+1)}\right)^{\frac{1}{2}}+\frac{64\sqrt{10}c_4L\tau F_0}{(1-\rho)(K+1)}\end{aligned}$$

$$
\begin{aligned}
&+ 24L(\delta^2+\sigma^2)\left(\frac{c_4F_0}{c_6(K+1)}\right)^{\frac{1}{2}} + \frac{8nc_l\delta^2 ln(K+6d)}{K+1} + \frac{768n\tau\phi_0L^2}{K+1} \\
&+ 12nc_lC_1L^2\tau^2\gamma^2(2\sigma^2+\delta^2) + 12C_1nL^2\tau^2\gamma^2\left(2\sigma^2\|A\|^2+\sigma^2\right) \\
&+ \frac{12C_2nc_lL^2\tau^2\gamma^2(2\sigma^2+\delta^2)}{(1-\rho)^2} + \frac{12C_2nL^2\tau^2\gamma^2\left(2\sigma^2\|A\|^2+\sigma^2\right)}{(1-\rho)^2} \\
&\leq 2c_5^{\frac{1}{3}}\left(\frac{c_4F_0}{K+1}\right)^{\frac{2}{3}} + 2\left(\frac{c_4c_6F_0}{K+1}\right)^{\frac{1}{2}} + \frac{64\sqrt{10}c_4L\tau F_0}{(1-\rho)(K+1)} \\
&+ \frac{768n\tau\phi_0L^2}{K+1} + \frac{8nc_l\delta^2 ln(K+6d)}{K+1},
\end{aligned}
$$

where c_4, c_5, c_6, F_0 are defined in (74), the first inequality and the last inequality are due to [32, Lemma 21] and the definition of γ in (75), respectively. The proof is complete. □

Theorem 4 shows that the averaged iterations generated by DSGTD-AG converge to the stationary point with rate $\mathcal{O}(\frac{\ln K}{\sqrt{K}})$, where K is the number of iterations. On the other hand, all the terms in (75) are positive as $\rho \in (0,1)$, which means the step size γ is obtainable.

5 Numerical experiments

To show the performance of Algorithms, we conduct experiments on two multi-agent Gaussian mean estimation problems with synthetic data and the strategic binary classification problem with a real dataset [25]. In all experiments, we compare Algorithms with DSGD-GD [25]. We consider a ring graph with $n = 25$ agents, the mixing matrix weights as $w_{ij} = \frac{1}{3}$ for all $(i,j) \in \mathcal{E}$ and $w_{ij} = 0$ if $(i,j) \notin \mathcal{E}$. Moreover, in all figures, the green line and the orange line denote DSGTD-GD with constant step size and diminishing step size, the red line denotes DSGTD-AG and the blue line denotes DSGD-GD with diminishing step size.

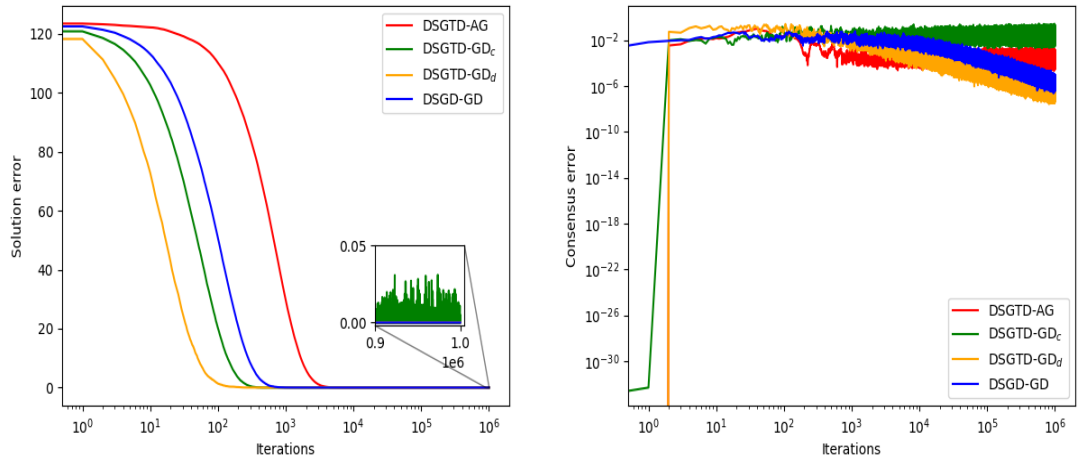
5.1 Multi-agent Gaussian Mean Estimation I

Consider the following distributed stochastic optimization problem with decision-dependent distributions [25]

$$
\min_x \sum_{i=1}^{25} \mathbf{E}_{\xi_i\sim\mathcal{D}_i(x)}\left[\frac{(x-\xi_i)^2}{2}\right],
$$

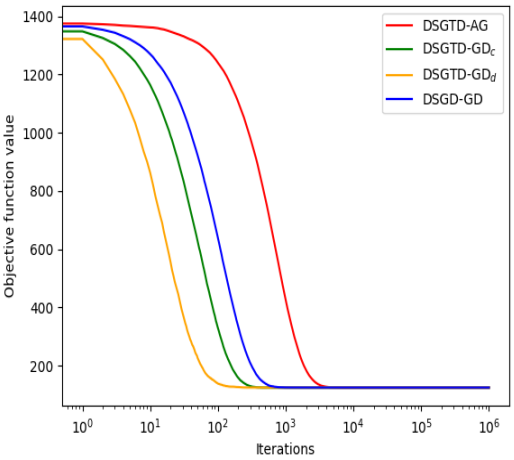
where for $i = 1, ..., 25$, $\mathcal{D}_i(x_i) = \mathcal{N}(\bar{z}+\epsilon_i x_i, \sigma^2)$, $\bar{z} = 10$, $\sigma^2 = 10$, $\epsilon_i = \epsilon_{avg} = 0.1$. When $\mu = 1$ and $L = 1$, the performative stable solution and the optimal solution can be computed in closed form as $x^\star = x^{\star\star} = \frac{\bar{z}}{(1-\epsilon_{avg})}$.

In Figure 1, we report the convergence of DSGTD-GD and DSGTD-AG, where Figure 1 (a), (b), (c), (d) record the performance on solution error, consensus error, objective value and the asymptotic normality of iterates, respectively. In the experiment, we run DSGTD-GD, DSGTD-AG and DSGD-GD with $k = 10^6$ iterations, where the ith agent draws one sample from $\mathcal{D}_i(x_{i,k})$

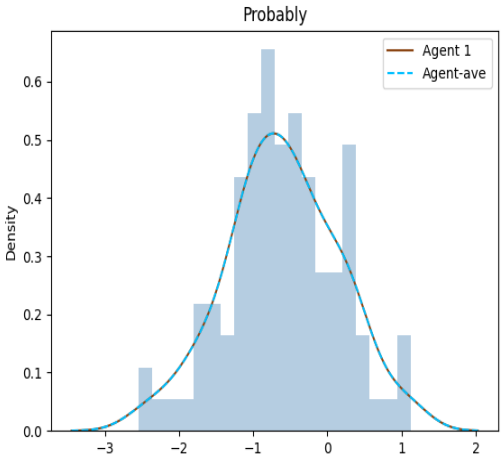


(a) Solution error

(b) Consensus error



(c) Objective value



(d) Estimated probability density

Figure 1: Multi-agent Gaussian Mean Estimation.

to calculate the gradient. We set the step sizes of DSGTD-GD as $\gamma_k = 1/(k+100)^{\frac{3}{4}}$, $\gamma = 0.01$ and the step size of DSGTD-AG as $\gamma = 0.001$, which satisfy constraints in the Theorem 1, 2 and 4 with the setting. In particular, we set the step sizes for DSGD-GD to $\gamma_k = 50/(10000+k)$, as described in [25]. All the parameters of the setting are chosen the same as that in [25]. Moreover, we do 100 simulations for the test on asymptotic normality.

From Figure 1 (a), we can observe that the performative stable error of DSGTD-GD with constant step size tends to stabilize around 0.01 after thousands of iterations, which matches the conclusion of Theorem 1 that DSGTD-GD with constant step size converges to a neighborhood of the performative stable point. At the same time, the performative stable error of DSGTD-GD with diminishing step size tends to 0, which implies under diminishing step size, DSGTD-GD converges to the performative stable point and coincides exactly with Theorem 2. On the other

hand, the solution error of DSGTD-AG tends to 0, which implies DSGTD-GD can converge to the optimal solution as demonstrated in Theorem 4 when A_k sufficiently tend to A. Moreover, DSGTD-GD and DSGTD-AG have comparable rates with DSGD-GD with diminishing step size. From Figure 1 (b), we may observe that the consensus error of DSGTD-GD with constant step size consistently stabilizes around 10^{-2}. The underlying reason may be that DSGTD-GD with constant step size converges to the neighborhoods of the performative stable solution in Theorem 1. On the other hand, DSGTD-GD with diminishing step size and DSGTD-AG reach agreement. Similarly, DSGTD-GD with diminishing step size has comparable rate with DSGD-GD. From Figure 1 (c), we can observe that the objective value of DSGTD-GD, DSGTD-AG and DSGD-GD reach similar values, which verifies the performative stable point is close to the optimal point. On the other hand, DSGTD-GD has comparable rate with DSGD-GD to reach the optimal objective value. Figure 1 (d) seems to confirm Theorem 3 since we can see that the estimated density is close to the density of a normal distribution and is also confirmed by a Kolmogorov-Smirnov test. Moreover, we can observe that the red curve closely overlaps with the blue dashed curve, which supports that DSGTD-GD with diminishing step size achieves consensus faster than $\mathcal{O}(\frac{1}{k})$ in Theorem 2.

5.2 Multi-agent Gaussian Mean Estimation II

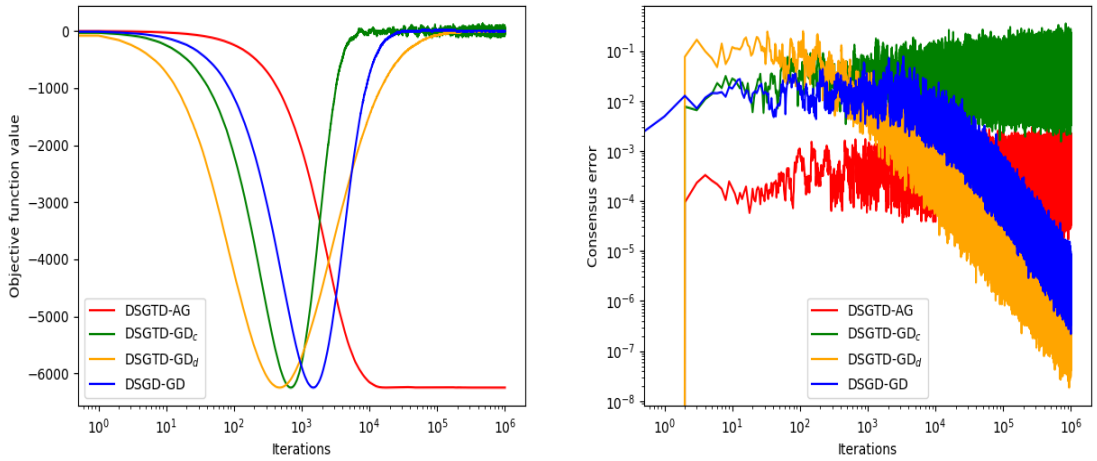
We consider the distributed stochastic optimization problem with decision-dependent distributions as follows,

$$\min_{x} \sum_{i=1}^{25} \mathbf{E}_{\omega_i \sim \mathcal{D}_i(x)} \left[\omega_i x\right], \tag{89}$$

where for $i = 1, ..., 25$, $\mathcal{D}_i(x) = \mathcal{N}(\bar{z} + \epsilon_i x, \sigma^2)$, $\bar{z} = 10$, $\sigma^2 = 10$, $\epsilon_i = \epsilon_{avg} = 0.1$. Specially, the optimal solution can be computed in closed form as $x^{\star\star} = \frac{-\bar{z}}{2\epsilon_{avg}}$. On the other hand, the performative stable solution can be computed in closed form as $x^{\star} = \frac{-\bar{z}}{\epsilon_{avg}}$, if it exists.

In Figure 2, we report the performance of DSGTD-GD and DSGTD-AG, where Figures 2 (a) and (b) record the performance on objective function value and consensus error. In experiments, we run DSGTD-GD, DSGTD-AG and DSGD-GD with $k = 10^6$ iterations. Because of the similar setting of the experiment, we adopt the same parameters as in the first experiment.

From Figure 2 (a), we can observe that the objective function value of DSGTD-AG reaches optimal, which matches the conclusion of Theorem 4 that the iterates generated by DSGTD-AG converge to the optimal point. On the other hand, the objective function values of DSGTD-GD first decrease and then increase as the iterations, which matches the conclusion of Theorem 1 and 2 that DSGTD-GD converges to the performative stable point. Moreover, this verifies that the performative stable point may be far from optimal without strong convexity. From Figure 2 (b), we may observe that DSGTD-GD with diminishing step size, DSGTD-AG and DSGD-GD reach agreement, while the consensus error of DSGTD-GD with constant step size consistently stabilizes around 10^{-2}, which is similar to the results of the first experiment.



(a) Objective value

(b) Consensus error

Figure 2: Multi-agent Gaussian Mean Estimation.

5.3 Email Spam Classification

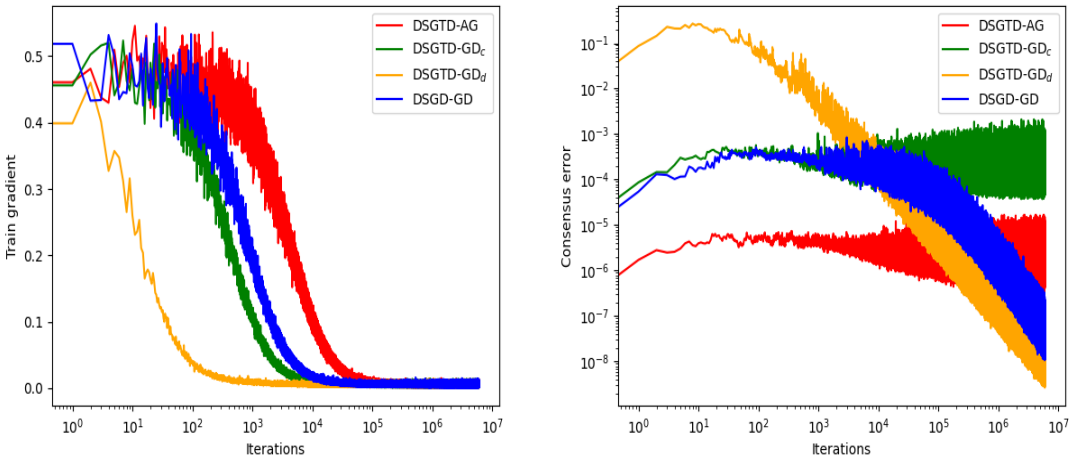
Consider the strategic binary classification problem [25], where the loss function is taken as the logistic regression function

$$l_i(x; Z_i) = log(1 + \exp(\langle Y_i^1, x\rangle)) - Y_i^2\langle Y_i^1, x\rangle + \frac{\beta}{2}\|x\|^2, \tag{90}$$

where $\beta > 0$ is a regularization parameter and data tuple $Z_i = (Y_i^1, Y_i^2) \sim \mathcal{D}_i(x_i)$ depends on the ith decision x_i. The users of the ith population provide samples that are modified via a linear utility such that $Y_i^1 = Y^1 + \epsilon_i x_i$ and $Y_i^2 = Y^2$ with $(Y^1, Y^2) \sim \mathcal{D}_i^0$, where $\mathcal{D}_i^0$ is a base data distribution of the ith population.

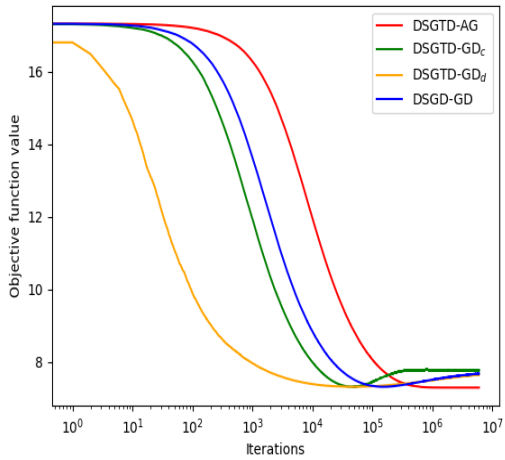
In what follows, we conduct the experiments on the spambase, a dataset [18] with $m = 4601$ samples and $d = 57$ features, which contains a collection of spam and valid emails. Each server has access to training data from $m_i = 138$ samples from spambase modeling the different sets of users; the remaining samples are taken as testing data. The servers aim to seek a common spam filter classifier via (90) with $\beta = 10^{-4}$. The sensitivity parameters are set as $\epsilon_i \in \{0.4\epsilon_{avg}, 0.45\epsilon_{avg}, ..., 1.6\epsilon_{avg}\}$ with $\epsilon_{avg} = 0.1$. In Figure 3, we report the convergence of DSGTD-GD and DSGTD-AG, where Figure 3 (a), (b), (c) and (d) record the performance on the train gradient $\|\nabla f(\bar{x}_k; \bar{x}_k)\|^2$, the consensus error $\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2$, the objective value on the training dataset, the accuracy on the testing dataset, respectively. In experiments, we run DSGTD-GD, DSGTD-AG and DSGD-GD with $k = 6\times10^6$ iterations where the ith server draws 32 samples from the corresponding dataset at each iteration, the step sizes of DSGTD-GD are $\gamma_k = 1/(k+100)^{\frac{3}{4}}$, $\gamma = 0.001$, the step size of DSGTD-AG is $\gamma = 0.0001$, and the step sizes of DSGD-GD are $\gamma_k = 50/(100000 + k)$.

From Figure 3 (a), we can observe that the train gradients of DSGTD-GD and DSGTD-AG tend to 0 at a comparable rate with DSGD-GD. From Figure 3 (b), we may observe the

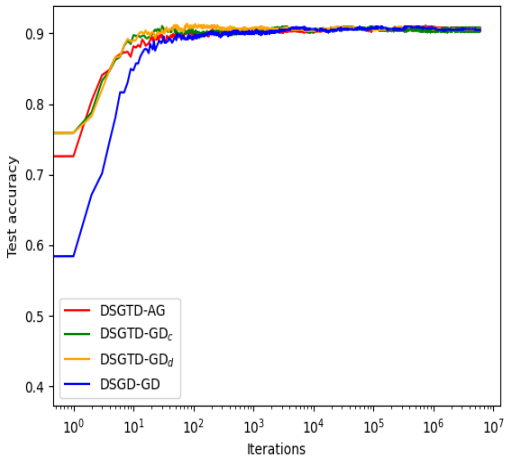


(a) Train gradient

(b) Consensus error



(c) Objective value



(d) Test accuracy

Figure 3: Spam Email Classification.

consensus error of DSGTD-GD with constant step size and DSGTD-AG consistently stabilize around 10^{-4} and 10^{-6}, which are similar to the results of synthetic data. Moreover, DSGTD-GD with diminishing step size reaches consensus at a comparable rate with DSGD-GD. Figure 3 (c), we can observe that the objective values of DSGTD-GD and DSGD-GD increase and then decrease as the iteration, reaching slightly bigger values than one of DSGTD-AG. This further verifies that DSGTD-AG converges to the optimal point and DSGTD-GD, DSGD-GD converge to the performative stable point, where the performative stable point may be far from the optimal point. In Figure 3 (d), we evaluate the performance on the trained classifier $x_{i,k}$ on the testing dataset with shifted distribution due to $x_{i,k}$. As observed, the test accuracy of DSGTD-GD and DSGTD-AG overall increases as iterates k and achieves the comparable level of accuracy with DSGD-GD.

Acknowledgment The research is supported by National Key R&D Program of China No. 2023YFA1009200, the NSFC #12471283 and Fundamental Research Funds for the Central Universities DUT24LK001.

References

[1] S. Ahmed. *Strategic planning under uncertainty: Stochastic integer programming approaches*. University of Illinois at Urbana-Champaign, 2000.

[2] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo. Distributed detection and estimation in wireless sensor networks. In *Academic Press Library in Signal Processing*, volume 2, pages 329–408. Elsevier, 2014.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[4] T.S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I.C. Paschalidis, and W. Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 112:59–67, 2018.

[5] G. Brown, S. Hod, and I. Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pages 6045–6061. PMLR, 2022.

[6] H.-F. Chen. *Stochastic approximation and its applications*, volume 64. Springer Science & Business Media, 2005.

[7] S. Chouvardas, K. Slavakis, and S. Theodoridis. Adaptive robust distributed learning in diffusion sensor networks. *IEEE Transactions on Signal Processing*, 59(10):4692–4707, 2011.

[8] K.L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

[9] J. Cutler, M. D´iaz, and D. Drusvyatskiy. Stochastic approximation with decision-dependent distributions: Asymptotic normality and optimality. *Journal of Machine Learning Research*, 25(90):1–49, 2024.

[10] A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. *The Annals of Statistics*, 48(3):1348–1382, 2020.

[11] J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z.S. Wu. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 55–70, 2018.

[12] D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.

[13] J.C. Duchi and F. Ruan. Asymptotic optimality in stochastic optimization. *The Annals of Statistics*, 49(1):21 – 48, 2021.

[14] M. Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.

[15] J. Dupacová. Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*, 2006.

[16] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.

[17] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[18] M. Hopkins, E. Reeber, G. Forman, and J. Suermondt. Spambase. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C53G6X.

[19] K. Huang, X. Li, A. Milzarek, S. Pu, and J. Qiu. Distributed random reshuffling over networks. *IEEE Transactions on Signal Processing*, 71:1143–1158, 2023.

[20] Z. Izzo, L.-X. Ying, and J. Zou. How to learn when data reacts to your model: Performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 4641–4650. PMLR, 2021.

[21] T.W. Jonsbråten, R.J. Wets, and D.L. Woodruff. A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82(0):83–106, 1998.

[22] A. Koloskova, T. Lin, and S.U. Stich. An improved analysis of gradient tracking for decentralized machine learning. volume 34, pages 11422–11435, 2021.

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

[24] Q. Li and H.-T. Wai. State dependent performative prediction with stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3164–3186. PMLR, 2022.

[25] Q. Li, C.-Y. Yau, and H.-T. Wai. Multi-agent performative prediction with greedy deployment and consensus seeking agents. *Advances in Neural Information Processing Systems*, 35:38449–38460, 2022.

[26] H.-D. Lim and D. Lee. Finite-time analysis of asynchronous q-learning under diminishing step-size from control-theoretic view. *arXiv preprint arXiv:2207.12217*, 2022.

[27] J.-Y. Liu, G.Y. Li, and S. Sen. Coupled learning enabled stochastic programming with endogenous uncertainty. *Mathematics of Operations Research*, 47(2):1681–1705, 2022.

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

[29] C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

[30] J.P. Miller, J.C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 7710–7720. PMLR, 2021.

[31] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

[32] A. Narang, E. Faulkner, D. Drusvyatskiy, M. Fazel, and L.J. Ratliff. Multiplayer performative prediction: Learning in decision-dependent games. *Journal of Machine Learning Research*, 24(202):1–56, 2023.

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

[34] Y. Pang and G. Hu. Nash equilibrium seeking in n-coalition games via a gradient-free method. *Automatica*, 136:110013, 2022.

[35] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning*, volume 119, pages 7599–7609. PMLR, 2020.

[36] G. Piliouras and F.-Y. Yu. Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1047–1074, 2023.

[37] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

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

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

[40] M. Ray, L.J. Ratliff, D. Drusvyatskiy, and M. Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8081–8088, 2022.

[41] M. Schmidt, N. Le Rouxand, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

[42] Z.J. Towfic and A.H. Sayed. Stability and performance limits of adaptive primal-dual networks. *IEEE Transactions on Signal Processing*, 63(11):2888–2903, 2015.

[43] K.R. Wood and E. Dall'Anese. Online saddle point tracking with decision-dependent data. In *Learning for Dynamics and Control Conference*, pages 1416–1428. PMLR, 2023.

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

Appendix

We recall some classic results on stochastic approximation and distributed optimization problems.

Lemma 3. [13, Proposition 2] *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(\mathcal{F}_k)$ be a non-decreasing sequence of σ-algebra. Let Θ_k, μ_k and η_k be sequences of random vectors in $\mathbb{R}^d$, α_k be the non-negative scalars. If recursion*

$$\Theta_{k+1} = (\mathbf{I} - \alpha_k \mathbf{G})\Theta_k + \alpha_k(\mu_k + \eta_k) \tag{91}$$

satisfies following conditions:

(a) $\alpha_k = \mathcal{O}(\frac{1}{k^a})$, $a \in (\frac{1}{2}, 1)$.

(b) $\mathbf{G}$ *is a positive definite matrix.*

(c) μ_k *has the decomposable structure* $\mu_k = \mu_k^{(1)} + \mu_k^{(2)}$, *where* $\mu_k^{(1)}$ *and* $\mu_k^{(2)}$ *are both martingale sequences adapted to* $\mathcal{F}_{k+1}$. *In addition,*

$$\frac{1}{\sqrt{k}} \sum_{t=1}^{k} \mu_t^{(1)} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma) \tag{92}$$

where $\mathbf{0} = (0, 0, ...0)^\intercal \in \mathbb{R}^d$, *and there exists a constant* c *such that*

$$\mathbf{E}[\|\mu_k^{(1)}\|^2 | \mathcal{F}_k] \leq c \qquad \mathbf{E}[\|\mu_k^{(2)}\|^2 | \mathcal{F}_k] \leq c\|\Theta_k\|^2$$

(d) η_k *is adapted to* $\mathcal{F}_{k+1}$ *and* $\frac{1}{\sqrt{k}} \sum_{t=1}^{k} \|\eta_k\| \to 0$ *almost surely.*

(e) Θ_k *is adapted to* $\mathcal{F}_k$, $\Theta_k \to 0$ *and* $\frac{1}{\sqrt{k}} \sum_{t=1}^{k} \|\Theta_k\|^2 \to 0$ *almost surely.*

Then

$$\sqrt{k} \sum_{t=1}^{k} \Theta_t \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{G}^{-1} \Sigma \mathbf{G}^{-1}).$$

Lemma 4. [38, Lemma 5] *Let $S = [s_{ij}] \in \mathbb{R}^{3\times 3}$ be a non-negative, irreducible matrix with $s_{ii} < \lambda^\star$ for some $\lambda^\star > 0$ for all $i = 1, 2, 3$. Then $\rho(S) < \lambda^\star$ iff $\det(\lambda^\star \mathbf{I} - S) > 0$.*

Lemma 5. [14, Theorem 1.3.12] *Let $(A_k), (B_k), (C_k), (D_k)$ be sequences of non-negative finite random variables on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$ adapted to the filtration $\mathbb{F} = (\mathcal{F}_k)$ and satisfying*

$$\mathbf{E}\left[A_{k+1} | \mathcal{F}_k\right] \leq (1 + B_k)A_k + C_k - D_k$$

for $\forall k$. Then on the event $\{\sum_k B_k < \infty, \sum_k C_k < \infty\}$, there is a non-negative finite random variable A_∞ such that $A_k \to A_\infty$ and $\sum_k D_k < \infty$ almost surely.

Lemma 6. *Let $\mathbf{W}$ and ρ be the communication weight matrix and the spectral norm of $\mathbf{W} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$, respectively. Then, for $\forall k \geq 0$, $\|k(\mathbf{W} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^k\|^2 \leq \frac{1}{(1-\rho)^2}$.*

Proof. By the definition of ρ, we have

$$\|k(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^k\|^2 \leq (k\rho^k)^2 \leq (\sum_{m=0}^{k-1}\rho^m)^2 \leq \frac{1}{(1-\rho)^2},$$

where $\rho \in (0,1)$ is due to [38, Lemma 1]. □

Lemma 7. *Under the conditions of Lemma 6, for* $\forall k \geq 0$, $\|(k+1)(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^{k+1} - k(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^k\|^2 \leq \frac{4}{(1-\rho)^2}$.

Proof. By the definition of matrix norm,

$$\|(k+1)(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^{k+1} - k(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^k\|^2 = \max_{\lambda\in\Lambda}((k+1)\lambda^{k+1}-k\lambda^k)^2,$$

where Λ denotes the set of the eigenvalues of $\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$.

Next, we derive the bound with two cases.

If the maximum is attained for a positive $\lambda > 0$, we conclude

$$((k+1)\lambda^{k+1}-k\lambda^k)^2 = (\lambda^{k+1}-k\lambda^k(1-\lambda))^2 \leq 2\lambda^{2(k+1)} + 2k^2(1-\lambda)^2\lambda^{2k} \leq 2\lambda^{2(k+1)} + 2 \leq 4,$$

where the second inequality and the last inequality follow from the fact $k\lambda^k \leq \frac{1}{1-\lambda}$ and $\lambda \leq 1$, respectively. On the other hand, if the maximum is attained for a negative $\lambda < 0$,

$$((k+1)\lambda^{k+1}-k\lambda^k)^2 \leq 2((k+1)\lambda^{k+1})^2 + 2(k\lambda^k)^2 \leq 2((k+1)\rho^{k+1})^2 + 2(k\rho^k)^2 \leq \frac{4}{(1-\rho)^2},$$

where the last inequality follows from Lemma 6.

Summing above, for $\forall k \geq 0$, we have

$$\|(k+1)(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^{k+1} - k(\mathbf{W}-\frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)^k\|^2 \leq \frac{4}{(1-\rho)^2}.$$

□