

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 stationary point 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 propose a distributed algorithm, DSGTD-GD, which combines the distributed stochastic gradient tracking descent method with the greedy deployment scheme. Under a 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 a diminishing step size policy, we show that the iterates generated by DSGTD-GD converge to the performative stable solution with a rate of $\mathcal{O}(\frac{1}{k^a})$, where $a \in (\frac{1}{2}, 1]$. 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 stationary point, we propose a distributed algorithm, DSGTD-AG, which combines the distributed stochastic gradient tracking descent method with the adaptive gradient scheme. Under a constant step size, we show that the iterates generated by DSGTD-AG converge to a stationary point with a rate of $\mathcal{O}(\frac{\ln K}{K^{\frac{2}{3}}})$, where K is the number of iterates. 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 considerable attention in recent years due to their wide range of applications, including large-scale machine learning [4, 5], sensor networks [3, 9], and parameter estimation [56]. Traditional stochastic optimization problems typically rely on the assumption that data follow a static distribution. However, in many real-world applications, the data can be influenced by the decision [13, 18], which may be characterized by the following distributed stochastic optimization problem with decision-dependent

*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)

distributions (distributed SO-DD)

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

where $\ell_i : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a smooth function of the decision vector x and the random vector ξ_i , the distribution map $\mathcal{D}_i(\cdot) : \mathbb{R}^d \rightarrow \mathcal{P}(\mathbb{R}^m)$ maps the decision variable x to a probability distribution and $\mathbf{E}[\cdot]$ denotes the expectation with respect to the distribution $\mathcal{D}_i(x)$ supported on $\Xi \subseteq \mathbb{R}^m$. Different from traditional distributed stochastic optimization problems, there is an interactive mechanism between the decision vectors and data: the i -th agent draws samples from $\mathcal{D}_i(x)$ to update the decision variable, while the distribution $\mathcal{D}_i(x)$ itself depends on the decision variable.

The study of SO-DD dates back to early stochastic programming [1, 16, 26]. We refer the interested reader to [19] for a more comprehensive overview on the taxonomy and various models of decision-dependent uncertainties. The dependence of the distribution on the decision variable makes SO-DD substantially more challenging than conventional stochastic optimization and may lead to a non-convex objective. Recently, [46] propose the concept of two solutions, namely, the performative stable solution

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

and the optimal solution

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

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 seek the performative stable solution.

Since the seminal work [46], there is a growing literature [6, 14, 25, 29, 35, 38, 39, 53, 58] in studying SO-DD. Most existing works [6, 14, 29, 38, 53, 58] focus on the performative stable solution. Building on the repeated gradient descent method [46], [38] propose the greedy deployment method and the lazy deployment method. The algorithm with greedy deployment draws a sample from the current distribution as independent samples and performs an update at each iteration. Different from 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, [38] show that the algorithm with greedy deployment converges to performative stable solution at a rate of $\mathcal{O}(\frac{1}{k})$, while the one with lazy deployment achieves a rate of $\mathcal{O}(\frac{1}{k^a})$ provided that $\mathcal{O}(k^{1.1a})$ samples are drawn between the k -th and $(k+1)$ -th iteration, where $a > 0$. [14] show that typical stochastic algorithms originally tailored for static problems can be directly applied for finding the performative stable solution with little loss in efficiency. [58] 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 solution. As opposed to the setting in [38], where samples are independently drawn from the distribution induced by the current decision, [29] consider a state-dependent setting in which the samples depend on both the learner's and the agents' 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 at a rate of $\mathcal{O}(\frac{1}{k})$. [6] propose a theoretical framework in which the population-provided samples are modeled as a function of both the deployed decision and the current state (distribution) of the population. They establish the necessary and sufficient conditions for the convergence of repeated risk minimization method to the performative stable solution. [53] study decision-dependent risk minimization in a dynamic environment where data distribution is decision-dependent and evolves dynamically with time according to a geometric decay process. They 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.

Under standard assumptions, the performative stable solution and the optimal solution lie within a small neighborhood of each other [46]. In general, however, the performative stable solution may be far from the optimal solution [25]. Therefore, several recent studies propose algorithms to seek the optimal solution. [25] assume the distribution can be approximately expressed to a mixture of Gaussian distribution and introduce a performative gradient descent algorithm with a finite difference method to estimate the part of the gradient with distribution shift. [39] assume that the distribution map is a global linear structure with respect to the decision variable and 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. [35] consider SO-DD without imposing a global parametric structure on the distribution map $\mathcal{D}(x)$. They develop a coupled learning-enabled optimization method in which the distribution map is locally approximated using linear regression around the current iterate and a candidate solution is then sought by a trust region method. Without imposing structural assumptions on the distribution map, zeroth-order methods provide a feasible approach for seeking the optimal solution [20–22]. [22] consider non-convex stochastic optimization with unknown decision-dependent distributions and propose zeroth-order methods with one-point and two-point gradient estimators. [20] further develop guided zeroth-order methods that exploit partial gradient information from the known loss function to generate more effective search directions. Moreover, [21] provide a unified sample complexity analysis of zeroth-order methods with different gradient estimators and show that averaging over multiple random directions can improve the sample complexity.

More recently, several studies consider the multi-agent stochastic optimization problems with decision-dependent distributions [30, 42, 47, 57] to address the substantial computational demands induced by large-scale datasets and provide privacy guarantees. [30] consider the case where n agents collaboratively seek a common decision vector that minimizes the sum of loss functions over an undirected and connected communication network, while the agents draw samples from local distribution that depends on the decision deployed by the corresponding agent, which is precisely the distributed SO-DD (1) considered in this paper. They propose a distributed stochastic gradient descent method with greedy deployment (DSGD-GD) and show that the proposed method with a diminishing step size achieves convergence rate of $\mathcal{O}(\frac{1}{k})$ to the performative stable solution of distributed SO-DD. On the other hand, [42], [57] and [47] consider a game with decision-dependent distributions. [42] consider the distribution map of each player is a global linear structure and propose a stochastic gradient descent algorithm with

the adaptive gradient method, which alternately runs an adaptive gradient step for learning the distribution map and a stochastic gradient descent step for updating the decision variable at each iteration, to seek the Nash equilibrium of the game with decision-dependent distributions. [47] consider a slightly different setting in which samples that react to the global decisions are drawn from the same distribution. They introduce learning dynamics to study the dynamical behavior of the system, where the players use reinforcement learning algorithms. Moreover, [57] study the game with decision-dependent distributions over multiplex networks, where agents cooperate over an agent network while each local distribution can be influenced by the decisions of neighboring agents over a population network, and propose a stochastic gradient method with greedy deployment for seeking the performative stable equilibrium.

Compared with standard distributed gradient methods, the gradient tracking method introduces agent-based auxiliary variables to dynamically track the averaged stochastic gradient, thereby incorporating global gradient information into local updates, which is more robust to heterogeneous data and maintains the stability of the performance. In recent years, the gradient tracking technique has been extensively studied in distributed optimization [28, 36, 43, 45, 49, 52]. This motivates us to propose two gradient tracking-based methods to seek the performative stable solution and a stationary point of distributed SO-DD, respectively. The setting with decision-dependent distributions introduces additional technical challenges beyond those arising in standard distributed stochastic optimization. The contributions of the paper can be summarized as follows.

- We propose a distributed stochastic gradient tracking descent method with the greedy deployment scheme, termed DSGTD-GD, to seek the performative stable solution. The key difficulty lies in the endogenous feedback between the optimization dynamics and the data-generating process. Unlike standard distributed stochastic optimization, the stochastic gradient available to each agent is generated from a local decision-dependent distribution. Consequently, the algorithm does not merely suffer from stochastic noise and network disagreement; changes in the iterates simultaneously induce changes in the decision-dependent distributions, which in turn perturb the stochastic gradients and feed back into the optimization dynamics. This introduces a fundamentally different error-propagation structure in which the optimization error, consensus error, gradient tracking error are coupled with an additional error induced by distribution shift. To overcome this challenge, we bound the endogenous distribution-shift effect in terms of the evolution of the agents' decisions by leveraging the Lipschitz regularity of the decision-dependent distribution maps, allowing the resulting coupled system to be closed under appropriate network and problem conditions. Moreover, we derive a unified system of recursive inequalities that jointly tracks the four error items and characterize how distributional perturbations propagate through the gradient tracking dynamics.
- Under a constant step size, we show that the iterates generated by DSGTD-GD achieve a linear convergence rate to a neighborhood of the performative stable solution, where the neighborhood size depends on both the step size and the connectivity of the communication network. Under a diminishing step size policy, we show that the expected squared distances between the iterates generated by DSGTD-GD and the performative stable solution converge to zero at a rate of $\mathcal{O}(\frac{1}{k^a})$, where $a \in (\frac{1}{2}, 1]$. These results demonstrate

that, under both constant and diminishing step size policies, DSGTD-GD achieves convergence rates of the same order as those established for standard distributed stochastic gradient tracking methods in [49], despite the additional distribution-shift error arising from the decision-dependent distributions. 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. To the best of our knowledge, this provides the first asymptotic normality for a stochastic-approximation-based method in distributed SO-DD, thereby extending the theoretical scope from optimization guarantees to asymptotic statistical characterization.

- We propose a distributed stochastic gradient tracking descent method with the adaptive gradient, termed DSGTD-AG, to seek a stationary point of non-convex distributed SO-DD. The main difficulty is that the optimization error, the consensus error, and the gradient tracking error are coupled with the additional estimation error induced by learning the parameters of the distribution map. To address this difficulty, we incorporate the estimation error into the descent recursions to jointly control these coupled errors, thereby establishing a convergence guarantee for non-convex distributed SO-DD. Under a constant step size policy, we show that the iterates of DSGTD-AG converge to a stationary point at a rate of $\mathcal{O}(\frac{\ln K}{K^{\frac{2}{3}}})$, where K is the number of iterates. Compared with the rate of $\mathcal{O}(\frac{1}{K^{\frac{2}{3}}})$ in [27], the additional logarithmic factor arises from learning the parameters of the distribution maps. The effectiveness of DSGTD-GD and DSGTD-AG is further demonstrated numerically with synthetic and real-world datasets.

The rest of this paper is organized as follows. Section 2 introduces DSGTD-GD and DSGTD-AG for the distributed SO-DD. Section 3 studies the convergence of DSGTD-GD, where Subsection 3.1 focuses on DSGTD-GD with a constant step size and Subsection 3.2 focuses on DSGTD-GD with the diminishing step sizes. The convergence analysis of DSGTD-AG is presented in Section 4. Numerical experiments are provided in Section 5. In Section 6, we provide concluding remarks. Some auxiliary results can be found in the Appendix A.

Throughout this paper, vectors are column vectors unless otherwise specified. $\mathbb{R}^d$ denotes the d -dimensional Euclidean space endowed with norm $\|x\| = \sqrt{\langle x, x \rangle}$. The vectors $\mathbf{1}$ and $\mathbf{0}$ denote the all-ones and all-zeros vectors, respectively, whose dimensions are clear from context. $\mathbf{I} \in \mathbb{R}^{n \times n}$ denotes the identity matrix. For two matrices A and B of the same size, the inner product is denoted by $\langle A, B \rangle$. For matrices, $\|\cdot\|$ and $\|\cdot\|_2$ denote the Frobenius norm and the spectral norm, respectively. Denote $f_i(x; x') := \mathbf{E}_{\xi_i \sim \mathcal{D}_i(x')} [\ell_i(x; \xi_i)]$, $f(x; x') := \sum_{i=1}^n f_i(x; x')$, $f_i(x) := \mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} [\ell_i(x; \xi_i)]$ and $f(x) := \sum_{i=1}^n f_i(x)$. For clarity, $\nabla \ell_i(x; \xi)$, $\nabla f_i(x; x')$ and $\nabla f(x; x')$ denote the gradients with respect to x , respectively, while $\nabla_{\xi} \ell_i(x; \xi)$ denotes the gradient with respect to ξ . Moreover, define $\nabla_{x, \xi} \ell_i(x; \xi) := [\nabla \ell_i(x; \xi), \nabla_{\xi} \ell_i(x; \xi)]^\top$, and $\nabla f(x)$ and $\nabla f_i(x)$ denote the gradients with respect to x , respectively. For the given σ -algebra $\mathcal{F}$, $\mathbf{E}[\cdot | \mathcal{F}]$ is conditional expectation given $\mathcal{F}$. For a sequence of random vectors $\{\xi_k\}$ and a random vector ξ , $\xi_k \xrightarrow{d} \xi$ denotes the convergence in distribution and $\text{Cov}(\xi)$ denotes the covariance matrix of the random vector ξ . $\mathcal{N}(z, \Sigma)$ denotes the normal distribution with mean z and covariance matrix Σ . Finally, we use bold letters to represent the matrix with rows z_i , i.e. $\mathbf{z} := [z_1, \dots, z_n]^\top \in \mathbb{R}^{n \times d}$, and denote $\bar{z} := \frac{1}{n} \sum_{i=1}^n z_i$.

2 Preliminaries and algorithms

In this section, we introduce DSGTD-GD and DSGTD-AG in Subsections 2.1 and 2.2, respectively. Throughout this paper, we consider a set of agents $\mathcal{V} = \{1, \dots, n\}$ connected over a communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the set of communication links among the agents. The corresponding weight matrix $\mathbf{W} = [w_{ij}]_{n \times n}$ satisfies that $w_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $w_{ij} = 0$ otherwise for $i \neq j$. We assume the communication network and the weight matrix satisfy the following condition.

Assumption 1. [49] **[Networks and weight matrices]** The network is undirected and connected. The nonnegative matrix $\mathbf{W}$ is doubly stochastic, i.e., $\mathbf{W}\mathbf{1} = \mathbf{W}^\top \mathbf{1} = \mathbf{1}$, and the diagonal entries of $\mathbf{W}$ are positive.

Assumption 1 is a standard condition on the communication network and the matrix $\mathbf{W}$ [24, 49, 59] and implies that $\mathbf{1}\mathbf{1}^\top \mathbf{W} = \mathbf{W}\mathbf{1}\mathbf{1}^\top = \mathbf{1}\mathbf{1}^\top$, the spectral norm ρ of the matrix $\mathbf{W} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ satisfies $\rho \in (0, 1)$ [49, Lemma 1].

2.1 DSGTD-GD

DSGTD-GD is designed to seek the performative stable solution x_{ps} satisfying

$$x_{ps} := \arg \min_{x \in \mathbb{R}^d} \sum_{i=1}^n \mathbf{E}_{\xi_i \sim \mathcal{D}_i(x_{ps})} [\ell_i(x; \xi_i)].$$

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

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

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

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

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

$$x_{i,k+1} = \sum_{j=1}^n w_{ij} (x_{j,k} - \gamma_k y_{j,k}). \quad (4)$$

3: **The i -th 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 $i \in \mathcal{V}$,

$$y_{i,k+1} = \sum_{j=1}^n w_{ij} y_{j,k} + \nabla \ell_i(x_{i,k+1}; \xi_{i,k+2}) - \nabla \ell_i(x_{i,k}; \xi_{i,k+1}). \quad (5)$$

5: **end for**

At each iteration k , DSGTD-GD consists of three steps: (i) the i -th agent updates its decision 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 neighbors of the i -th agent; (ii) the i -th agent draws a new sample $\xi_{i,k+2} \sim \mathcal{D}_i(x_{i,k+1})$; (iii) each agent updates its gradient tracking variable $y_{i,k+1}$ by combining its own and its neighbors' gradient tracking variables $y_{j,k}$, $j \in \{i\} \cup \mathcal{N}_i$, with specific weights and adding a stochastic gradient-difference term $\nabla \ell_i(x_{i,k+1}; \xi_{i,k+2}) - \nabla \ell_i(x_{i,k}; \xi_{i,k+1})$, where $\nabla \ell_i(x_{i,k+1}; \xi_{i,k+2})$ is evaluated using the updated decision vector and the newly drawn sample. Since the stochastic gradient evaluated at the previous iterate is already available from the preceding iteration and is reused in the gradient tracking update, each agent draws only one fresh sample per iteration. Therefore, over K iterations, DSGTD-GD requires K fresh samples per agent and nK fresh samples across the network.

For ease of presentation, we write Algorithm 1 in the following compact form:

$$\begin{aligned}\mathbf{x}_{k+1} &:= \mathbf{W}(\mathbf{x}_k - \gamma_k \mathbf{y}_k), \\ \mathbf{y}_{k+1} &:= \mathbf{W}\mathbf{y}_k + \mathbf{L}(\mathbf{x}_{k+1}; \xi_{k+2}) - \mathbf{L}(\mathbf{x}_k; \xi_{k+1}),\end{aligned}\tag{6}$$

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

$$\mathbf{L}(\mathbf{x}_k; \xi_{k+1}) := \begin{bmatrix} \nabla \ell_1(x_{1,k}; \xi_{1,k+1})^\top \\ \vdots \\ \nabla \ell_n(x_{n,k}; \xi_{n,k+1})^\top \end{bmatrix}.$$

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

$$\bar{x}_{k+1} = \bar{x}_k - \gamma_k \bar{y}_k, \quad \bar{y}_k = \frac{1}{n} \sum_{i=1}^n \nabla \ell_i(x_{i,k}; \xi_{i,k+1}),\tag{7}$$

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 tailored for seeking a stationary point x_{sp} of problem (1) satisfying

$$\nabla f(x_{sp}) = 0,$$

as detailed in Algorithm 2.

At each iteration k , DSGTD-AG consists of four steps: (i) based on the current estimation $\mathcal{D}_{i,k}(\cdot)$, the decision variables $x_{i,k}$ and a sample generated from the real distribution map $\mathcal{D}_i(x_{i,k})$, the i -th agent forms a new estimation of the distribution map $\mathcal{D}_{i,k+1}(\cdot)$ via a generic distribution-map update operator π , which will be specified later; (ii) the i -th agent updates its vector $x_{i,k}$ using the same gradient-tracking descent step as in Algorithm 1; (iii) the i -th agent draws a new sample $\xi_{i,k+2} \sim \mathcal{D}_i(x_{i,k+1})$; (iv) the i -th agent updates its gradient tracking variable $y_{i,k+1}$ by combining its own and its neighbors' gradient tracking variables $y_{j,k}$, $j \in \{i\} \cup \mathcal{N}_i$ with

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

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

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

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

$$\mathcal{D}_{i,k+1}(\cdot) = \pi(\mathcal{D}_{i,k}(\cdot), x_{i,k}, \xi_{i,k+1}). \quad (8)$$

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

$$x_{i,k+1} = \sum_{j=1}^n w_{ij} (x_{j,k} - \gamma y_{j,k}). \quad (9)$$

4: **The i -th 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 $i \in \mathcal{V}$,

$$y_{i,k+1} = \sum_{j=1}^n w_{ij} y_{j,k} + v_{i,k+1} - v_{i,k}. \quad (10)$$

6: **end for**

specific weights and adding the adaptive gradient difference $v_{i,k+1} - v_{i,k}$, where $v_{i,k}$ denotes an adaptive gradient constructed using the current estimate of distribution map $\mathcal{D}_{i,k}(\cdot)$, whose formal definition is given in (50). Under the adaptive-gradient implementation specified below, each agent draws two fresh samples per iteration: one is used for estimating the parameter of the distribution map, and the other is $\xi_{i,k+2}$ drawn from the distribution corresponding to the updated decision. The sample available from the previous iteration is reused in the adaptive gradient and gradient tracking updates. Hence, over K iterations, DSGTD-AG requires $2K$ fresh samples per agent and $2nK$ fresh samples across the network.

Similar to (6) and (7), we write (9) and (10) in Algorithm 2 in the following compact form:

$$\begin{aligned} \mathbf{x}_{k+1} &:= \mathbf{W}(\mathbf{x}_k - \gamma \mathbf{y}_k), \\ \mathbf{y}_{k+1} &:= \mathbf{W} \mathbf{y}_k + \mathbf{v}_{k+1} - \mathbf{v}_k, \end{aligned} \quad (11)$$

and then

$$\bar{x}_{k+1} = \bar{x}_k - \gamma \bar{y}_k, \quad \bar{y}_k = \frac{1}{n} \sum_{i=1}^n v_{i,k}, \quad (12)$$

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

For both algorithms, the total number of fresh samples scales linearly with the number of agents, while the per-agent sampling requirement at each iteration is independent of the network size and sampling can be performed in parallel across agents. Therefore, over K iterates, the proposed algorithms require $\mathcal{O}(K)$ samples per agent and $\mathcal{O}(nK)$ samples across the entire network.

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 a convergence rate $\mathcal{O}(\frac{1}{k^a})$, where $a \in (\frac{1}{2}, 1]$, and establish the asymptotic normality of the deviation between the averaged iterates of DSGTD-GD and the performative stable solution.

To study the convergence of Algorithm 1, we first impose the following assumptions.

Assumption 2. [30] [**Objective function and stochastic gradient**] Let x_{ps} be the performative stable solution of problem (1).

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

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

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

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

- (c) There exists a constant $\sigma > 0$ such that, for all $i \in \mathcal{V}$ and $x \in \mathbb{R}^d$,

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} [\|\nabla \ell_i(x; \xi_i) - \nabla f_i(x; x)\|^2] \leq \sigma^2(1 + \|x - x_{ps}\|^2).$$

- (d) There exist constants $p > 2$ and $c_g > 0$ such that, for all $i \in \mathcal{V}$,

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} [\|\nabla \ell_i(x; \xi_i)\|^p] \leq c_g^{\frac{p}{2}}(1 + \|x - x_{ps}\|^2).$$

- (e) $\nabla R(x_{ps})$ is positively stable, i.e., every eigenvalue of $\nabla R(x_{ps})$ has a strictly positive real part, where $R(x) := \sum_{i=1}^n \nabla f_i(x; x)$.

Assumption 3. [30] [**Lipschitz distribution**] For any $i \in \mathcal{V}$, there exists a constant $\epsilon_i > 0$ such that

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

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

Assumption 4. [11] [**Joint smoothness**] Let x_{ps} be the performative stable solution of problem (1). The map $x \mapsto \nabla f(x; x)$ is smooth and has a Lipschitz continuous Jacobian on a neighborhood of x_{ps} .

In Assumption 2, conditions (a) and (b) are widely used for the SO-DD [14, 30, 38] and require the loss functions to be strongly convex and smooth; condition (c) bounds the variance of the stochastic gradient $\nabla \ell_i(x; \xi_i)$; conditions (d) and (e) are for establishing the asymptotic normality of the averaged iterates generated by DSGTD-GD for each agent [11, 48]. Assumption

3 implies that the amount of distribution shift resulting from the reaction of data to the decision increases linearly with the difference in decision, where ϵ_i characterizes the sensitivity of the decision-dependent distributions for the i -th agent [30, 38, 46]. Under Assumptions 2 (a), (b) and 3, a sufficiently small distribution sensitivity, e.g., $\epsilon_{\text{avg}} < \mu/L$ with $\epsilon_{\text{avg}} := \frac{1}{n} \sum_{i=1}^n \epsilon_i$, guarantees the existence and uniqueness of the performative stable solution [30]. We should note that the strong convexity in Assumption 2 is not only sufficient for finding the unique performative stable solution, but also essential for the convergence of DSGTD-GD [30, 46]. Again, Assumption 4 is for studying the asymptotic normality of DSGTD-GD [11, 48].

Next, we 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_{ps} be the performative stable solution of problem (1). Denote*

$$\begin{aligned} \epsilon_{\max} &:= \max_{i \in \mathcal{V}} \epsilon_i, & \epsilon_{\text{avg}} &:= \frac{1}{n} \sum_{i=1}^n \epsilon_i, & \bar{\mu} &:= \mu - (1 + \varrho) \epsilon_{\text{avg}} L, \\ c_1 &:= 2L^2(1 + \epsilon_{\max})^2 + \frac{2\varrho^2}{n}, & c_2 &:= \frac{L}{2n\varrho\epsilon_{\text{avg}}}(1 + \epsilon_{\max})^2, & c_3 &:= L(1 + \epsilon_{\max}), \\ m_k &:= \frac{\sigma^2}{1 - \rho} [c_3^2\gamma_k^2 + 2n\gamma_k L(1 + \epsilon_{\text{avg}}) + 2n + 2\sigma^2\gamma_k^2], \end{aligned} \quad (13)$$

where ϱ is a positive constant. Suppose that $\epsilon_{\text{avg}} < \frac{\mu}{(1+\varrho)L}$ and $\gamma_k \leq \frac{\bar{\mu}}{c_1}$, then under Assumptions 1–3,

(i):

$$\mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2 | \mathcal{F}_k] \leq (1 - \bar{\mu}\gamma_k) \|\bar{x}_k - x_{ps}\|^2 + \frac{c_2\gamma_k + c_1\gamma_k^2}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma_k^2}{n}. \quad (14)$$

(ii):

$$\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 \leq \frac{1 + \rho^2}{2} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \gamma_k^2 \frac{(1 + \rho^2)\rho^2}{1 - \rho^2} \|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2. \quad (15)$$

(iii):

$$\begin{aligned} & \mathbf{E} [\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2 | \mathcal{F}_k] \\ & \leq \frac{2}{1 - \rho} [c_3^2\gamma_k^2(\sigma^2 + c_3^2) + 2n\gamma_k c_3\sigma^2 + n\sigma^2(2 - \bar{\mu}\gamma_k)] \|\bar{x}_k - x_{ps}\|^2 + \frac{1}{1 - \rho} [(3 + \rho^2)\sigma^2 \\ & \quad + 2\gamma_k\sigma^2(2c_3 + c_2) + c_3^2(1 + \frac{1}{\beta}) \|\mathbf{W} - \mathbf{I}\|_2^2 + \frac{2\gamma_k^2(nc_1\sigma^2 + c_3^2(\sigma^2 + c_3^2))}{n}] \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 \\ & \quad + \left[\frac{\rho^2\gamma_k^2}{1 - \rho} ((1 + \beta)c_3^2 + 2\sigma^2 \frac{1 + \rho^2}{1 - \rho^2}) + \rho \right] \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 | \mathcal{F}_k] + m_k, \end{aligned} \quad (16)$$

where β is any positive constant.

Proof. For ease of analysis, we denote,

$$f_i(x; y) := \mathbf{E}_{\xi_i \sim \mathcal{D}_i(y)} [\ell_i(x; \xi_i)], \quad h(\mathbf{x}_k) := \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_{i,k}; x_{i,k}), \quad (17)$$

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

Part (i). By the updating recursion (7),

$$\mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2 | \mathcal{F}_k] = \|\bar{x}_k - x_{ps}\|^2 - 2\gamma_k \mathbf{E} [\langle \bar{x}_k - x_{ps}, \bar{y}_k \rangle | \mathcal{F}_k] + \gamma_k^2 \mathbf{E} [\|\bar{y}_k\|^2 | \mathcal{F}_k], \quad (18)$$

where $\mathcal{F}_k$ denotes the σ -algebra generated by $\{\xi_1, \dots, \xi_k\}$. According to the definition of $h(\mathbf{x}_k)$ in (17),

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

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

$$\begin{aligned} & \mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2 | \mathcal{F}_k] \\ &= \|\bar{x}_k - x_{ps}\|^2 - 2\gamma_k \langle \bar{x}_k - x_{ps}, h(\mathbf{x}_k) \rangle + \gamma_k^2 (\mathbf{E} [\|\bar{y}_k - h(\mathbf{x}_k)\|^2 | \mathcal{F}_k] + \|h(\mathbf{x}_k)\|^2). \end{aligned} \quad (19)$$

For the third term on the right-hand side of (19),

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

where the equality follows from the fact that $\bar{y}_k = \frac{1}{n} \sum_{i=1}^n \nabla \ell_i(x_{i,k}; \xi_{i,k+1})$ in (7) and the definition of $h(\mathbf{x}_k)$, the inequality is due to Assumption 2 (d). On the other hand,

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

where the equality and the last inequality follow from the fact that $\frac{1}{n} \sum_{i=1}^n \nabla f_i(x_{ps}; x_{ps}) = 0$ and [30, Lemma 2], respectively. Then

$$\begin{aligned} & \mathbf{E} [\|\bar{y}_k - h(\mathbf{x}_k)\|^2 | \mathcal{F}_k] + \|h(\mathbf{x}_k)\|^2 \\ &\leq \frac{\sigma^2}{n} + \frac{\sigma^2}{n^2} \sum_{i=1}^n \|x_{i,k} - x_{ps}\|^2 + \frac{L^2}{n} \sum_{i=1}^n (1 + \epsilon_i)^2 \|x_{i,k} - x_{ps}\|^2 \\ &\leq \frac{\sigma^2}{n} + \frac{2(\sigma^2 + c_3^2)}{n^2} [n\|\bar{x}_k - x_{ps}\|^2 + \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2], \end{aligned} \quad (20)$$

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

For the inner product term on the right-hand side of (19),

$$\begin{aligned}
& \langle \bar{x}_k - x_{ps}, h(\mathbf{x}_k) \rangle \\
&= \frac{1}{n} \sum_{i=1}^n \langle \bar{x}_k - x_{ps}, \nabla f_i(x_{i,k}; x_{i,k}) \rangle \\
&= \frac{1}{n} \sum_{i=1}^n (\langle \bar{x}_k - x_{ps}, \nabla f_i(x_{i,k}; x_{i,k}) - \nabla f_i(\bar{x}_k; x_{ps}) \rangle + \langle \bar{x}_k - x_{ps}, \nabla f_i(\bar{x}_k; x_{ps}) - \nabla f_i(x_{ps}; x_{ps}) \rangle) \\
&\geq -\frac{L\|\bar{x}_k - x_{ps}\|}{n} \sum_{i=1}^n (\|x_{i,k} - \bar{x}_k\| + \epsilon_i \|x_{i,k} - x_{ps}\|) + \mu \|\bar{x}_k - x_{ps}\|^2 \\
&\geq (\mu - L\epsilon_{avg}) \|\bar{x}_k - x_{ps}\|^2 - \frac{L}{n} (1 + \epsilon_{max}) \sum_{i=1}^n \|\bar{x}_k - x_{ps}\| \|x_{i,k} - \bar{x}_k\| \\
&\geq \left[\mu - L\epsilon_{avg} - \frac{\alpha L}{2n} (1 + \epsilon_{max}) \right] \|\bar{x}_k - x_{ps}\|^2 - \frac{L}{2n\alpha} (1 + \epsilon_{max}) \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2, \tag{21}
\end{aligned}$$

where α is any positive constant, the first inequality follows from the Cauchy-Schwarz inequality and [14, Lemma 2.1], the last inequality follows from the Young's inequality.

Substituting (20) and (21) back to the equation (19) gives us the following desired result,

$$\begin{aligned}
& \mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2 | \mathcal{F}_k] \\
&\leq \left[1 - 2\gamma_k (\mu - L\epsilon_{avg} - \frac{\alpha L}{2n} (1 + \epsilon_{max})) + c_1 \gamma_k^2 \right] \|\bar{x}_k - x_{ps}\|^2 \\
&\quad + \left[\frac{\gamma_k L}{n\alpha} (1 + \epsilon_{max}) + \frac{c_1 \gamma_k^2}{n} \right] \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2 \gamma_k^2}{n} \\
&\leq (1 - \bar{\mu} \gamma_k) \|\bar{x}_k - x_{ps}\|^2 + \frac{c_2 \gamma_k + c_1 \gamma_k^2}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2 \gamma_k^2}{n}, \tag{22}
\end{aligned}$$

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

Part (ii). By the definition of $\mathbf{x}_{k+1}$ in (6) and the update recursion (7),

$$\begin{aligned}
& \|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2 \\
&= \|\mathbf{W}\mathbf{x}_k - \gamma_k \mathbf{W}\mathbf{y}_k - \bar{\mathbf{x}}_k + \gamma_k \bar{\mathbf{y}}_k\|^2 \\
&\leq \rho^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + 2\gamma_k \rho^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \|\mathbf{W}\mathbf{y}_k - \mathbf{1}\bar{y}_k\| + \gamma_k^2 \rho^2 \|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 \\
&\leq \rho^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \gamma_k \rho^2 \left[\frac{1 - \rho^2}{2\gamma_k \rho^2} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{2\gamma_k \rho^2}{1 - \rho^2} \|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 \right] + \gamma_k^2 \rho^2 \|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 \\
&= \frac{1 + \rho^2}{2} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \gamma_k^2 \frac{(1 + \rho^2)\rho^2}{1 - \rho^2} \|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2, \tag{23}
\end{aligned}$$

where $\rho \in (0, 1)$, the first inequality and the last inequality follow from [49, Lemma 1] and the Young's inequality, respectively.

Part (iii). Denote

$$\mathbf{L}_k := \mathbf{L}(\mathbf{x}_k; \xi_{k+1}), \quad \mathbf{F}_k := [\nabla f_1(x_{1,k}; x_{1,k}), \dots, \nabla f_n(x_{n,k}; x_{n,k})]^\top, \tag{24}$$

we have by the definition of $\mathbf{y}_{k+1}$ in (6) that

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

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

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

where the first inequality is due to the Young's inequality, the second inequality is due to [49, Lemma 1], the equality follows by setting $\eta = \frac{1-\rho}{\rho}$ and the fact that $\rho \in (0, 1)$ by [49, Lemma 1].

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

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

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

$$\begin{aligned}
& \mathbf{E} [\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2 | \mathcal{F}_k] \\
& \leq \rho \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 | \mathcal{F}_k] + \frac{1}{1 - \rho} \left[2\mathbf{E} [\langle \mathbf{F}_{k+1}, -\mathbf{L}_k + \mathbf{F}_k \rangle | \mathcal{F}_k] + \mathbf{E} [\|\mathbf{F}_{k+1} - \mathbf{F}_k\|^2 | \mathcal{F}_k] \right. \\
& \quad \left. + \sigma^2 (2n + \sum_{i=1}^n \mathbf{E} [\|x_{i,k+1} - x_{ps}\|^2 | \mathcal{F}_k] + \sum_{i=1}^n \|x_{i,k} - x_{ps}\|^2) \right]. \tag{25}
\end{aligned}$$

For the second term on the right-hand side of (25),

$$\mathbf{E} [\langle \mathbf{F}_{k+1}, -\mathbf{L}_k + \mathbf{F}_k \rangle | \mathcal{F}_k]$$

$$\begin{aligned}
&= \sum_{i=1}^n \mathbf{E} [\langle \nabla f_i(x_{i,k+1}; x_{i,k+1}), -\nabla \ell_i(x_{i,k}; \xi_{i,k+1}) + \nabla f_i(x_{i,k}; x_{i,k}) \rangle | \mathcal{F}_k] \\
&= \sum_{i=1}^n \mathbf{E} \left[\left\langle \nabla f_i \left(\sum_{j=1}^n w_{ij} x_{j,k} - \gamma_k \sum_{j=1}^n w_{ij} \sum_{l=1}^n w_{jl} y_{l,k-1} - \gamma_k \sum_{j=1}^n w_{ij} \ell_j(x_{j,k}; \xi_{j,k+1}) \right. \right. \right. \\
&\quad \left. \left. + \gamma_k \sum_{j=1}^n w_{ij} \ell_j(x_{j,k-1}; \xi_{j,k}); x_{i,k+1} \right) - \nabla f_i \left(\sum_{j=1}^n w_{ij} x_{j,k} - \gamma_k \sum_{j=1}^n w_{ij} \sum_{l=1}^n w_{jl} y_{l,k-1} \right. \right. \\
&\quad \left. \left. - \gamma_k \sum_{j \neq i}^n w_{ij} \ell_j(x_{j,k}; \xi_{j,k+1}) - \gamma_k w_{ii} \nabla f_i(x_{i,k}; x_{i,k}) + \gamma_k \sum_{j=1}^n w_{ij} \ell_j(x_{j,k-1}; \xi_{j,k}); x_{i,k+1} \right) \right. \\
&\quad \left. - \nabla \ell_i(x_{i,k}; \xi_{i,k+1}) + \nabla f_i(x_{i,k}; x_{i,k}) \right\rangle | \mathcal{F}_k] \\
&\leq \sum_{i=1}^n \gamma_k L (1 + \epsilon_i) \mathbf{E} [\| \nabla \ell_i(x_{i,k}; \xi_{i,k+1}) - \nabla f_i(x_{i,k}; x_{i,k}) \|^2 | \mathcal{F}_k] \\
&\leq \gamma_k \sigma^2 L \sum_{i=1}^n (1 + \epsilon_i) (1 + \|x_{i,k} - x_{ps}\|^2) \\
&\leq 2\gamma_k \sigma^2 L (1 + \epsilon_{max}) [n \|\bar{x}_k - x_{ps}\|^2 + \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] + \gamma_k n \sigma^2 L (1 + \epsilon_{avg}), \tag{26}
\end{aligned}$$

where the second equality follows from the formulas (4) and (5) in Algorithm 1, the first inequality is due to [14, 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 (6) and c_3 in (13), for any $\beta > 0$, we have

$$\begin{aligned}
&\|\mathbf{F}_{k+1} - \mathbf{F}_k\|^2 \\
&= \sum_{i=1}^n \|\nabla f_i(x_{i,k+1}; x_{i,k+1}) - \nabla f_i(x_{i,k}; x_{i,k})\|^2 \\
&\leq c_3^2 \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\
&= c_3^2 \|(\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \bar{\mathbf{x}}_k) - \gamma_k \mathbf{W} \mathbf{y}_k\|^2 \\
&= c_3^2 [\|\mathbf{W} - \mathbf{I}\|_2^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 - 2\gamma_k \langle (\mathbf{W} - \mathbf{I})(\mathbf{x}_k - \bar{\mathbf{x}}_k), \mathbf{W} \mathbf{y}_k - \bar{\mathbf{y}}_k \rangle \\
&\quad + \gamma_k^2 (\|\mathbf{W} \mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 + n \|\bar{y}_k\|^2)] \\
&\leq c_3^2 \left[\left(1 + \frac{1}{\beta}\right) \|\mathbf{W} - \mathbf{I}\|_2^2 \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + (1 + \beta) \gamma_k^2 \rho^2 \|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2 + \gamma_k^2 n \|\bar{y}_k\|^2 \right], \tag{27}
\end{aligned}$$

where the first inequality follows from [14, Lemma 2.1] and the last inequality follows from [49, Lemma 1] and the Young's inequality.

Combining (20), (22), (23), (26), (27) with the inequality (25), we arrive at (16). The proof is complete. $\square$

3.1 DSGTD-GD with constant step size

The stochastic gradient descent (SGD) method with constant step size is widely used in practice as it may achieve an exponentially fast convergence rate [12, 54]. 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_{ps} be the performative stable solution of problem (1). Denote*

$$\begin{aligned} \mathbf{A} &:= \begin{bmatrix} 1 - \gamma\bar{\mu} & \frac{c_2\gamma + c_1\gamma^2}{n} & 0 \\ 0 & \frac{1+\rho^2}{2} & \frac{(1+\rho^2)\rho^2\gamma^2}{1-\rho^2} \\ a_{31} & a_{32} & \rho + ((1+\hat{\beta})c_3^2 + 2\sigma^2\frac{1+\rho^2}{1-\rho^2})\frac{\rho^2\gamma^2}{1-\rho} \end{bmatrix}, \\ c_W &:= 2\sigma^2\frac{1+\rho^2}{1-\rho^2} + c_3^2, \quad \hat{\beta} := \frac{\frac{(1-\rho)^2}{2} - c_W\gamma^2\rho^2}{\gamma^2c_3^2\rho^2}, \\ m_\sigma &:= \frac{\sigma^2}{1-\rho} [c_3^2\gamma^2 + 2n\gamma L(1 + \epsilon_{avg}) + 2n + 2\sigma^2\gamma^2] \\ a_{31} &:= \frac{2}{1-\rho} [\gamma^2c_3^2(\sigma^2 + c_3^2) + 2n\gamma c_3\sigma^2 + n\sigma^2(2 - \bar{\mu}\gamma)], \\ a_{32} &:= \frac{1}{1-\rho} \left[c_3^2(1 + \frac{1}{\hat{\beta}})\|\mathbf{W} - \mathbf{I}\|_2^2 + \frac{2\gamma^2(nc_1\sigma^2 + (\sigma^2 + c_3^2)c_3^2)}{n} \right. \\ &\quad \left. + 2\gamma(2c_3 + c_2)\sigma^2 + (3 + \rho^2)\sigma^2 \right], \end{aligned} \quad (28)$$

where $\bar{\mu}$, c_1 , c_2 and c_3 are defined in (13). Suppose that the constant step size

$$\begin{aligned} \gamma \leq \min & \left\{ \frac{(1-\rho)^2(1+\rho)}{4 \left[\frac{(1-\rho)^2c_3^2(\sigma^2 + c_3^2)}{2c_W} + 2n\sigma^2 + \frac{n\sigma^2(1-\rho)}{\sqrt{2c_W}}(2nc_3 - \bar{\mu}) \right]^{\frac{1}{2}}} \left[\frac{n\bar{\mu}(\kappa - 1)}{\kappa(\kappa + 1)(c_2 + \frac{c_1(1-\rho)}{\sqrt{2c_W}})} \right]^{\frac{1}{2}}, \right. \\ & \left. \frac{(1-\rho^2)(1-\rho)}{2\sqrt{2\kappa} \max \left\{ \frac{\sqrt{4\sigma^2 + c_3^2}c_3\|\mathbf{W} - \mathbf{I}\|_2}{\sigma}, (1-\rho)\sqrt{\frac{(nc_1\sigma^2 + c_3^2(\sigma^2 + c_3^2))}{nc_W}}, \sqrt{\frac{8\sigma^2(1-\rho)(2c_3 + c_2)}{\sqrt{2c_W}}}, 4\sigma \right\}}, \frac{1-\rho}{\sqrt{2c_W}} \right\}, \end{aligned} \quad (29)$$

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

(i):

$$\sup_{t \geq k} \mathbf{E} [\|\bar{x}_t - x_{ps}\|^2], \quad \sup_{t \geq k} \mathbf{E} [\|\mathbf{x}_t - \bar{\mathbf{x}}_t\|^2] \quad (30)$$

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

(ii):

$$\limsup_{k \rightarrow \infty} \mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] \leq \frac{\gamma\sigma^2(\kappa + 1)}{\kappa n \bar{\mu}} + \frac{4\gamma^2\rho^2 m_\sigma(\kappa + 1)(c_2 + \gamma c_1)(1 + \rho^2)}{n\bar{\mu}(\kappa - 1)(1 - \rho)(1 - \rho^2)^2}, \quad (31)$$

$$\limsup_{k \rightarrow \infty} \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \leq \frac{4\gamma^2\rho^2(\kappa + 1)(1 + \rho^2)(a_{31}\gamma\sigma^2 + \bar{\mu}nm_\sigma)}{n\bar{\mu}(\kappa - 1)(1 - \rho)(1 - \rho^2)^2}. \quad (32)$$

Proof. Part (i). By the definition of c_W in (28),

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

where the first inequality follows from the fact that $\gamma \leq \frac{1-\rho}{\sqrt{2}c_W}$ in (29). Taking expectation on both sides of (14)-(16) with $\beta = \hat{\beta}$ and constant step size γ , we obtain the following linear system of inequalities

$$\begin{bmatrix} \mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2] \\ \mathbf{E} [\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2] \\ \mathbf{E} [\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2] \end{bmatrix} \leq \mathbf{A} \begin{bmatrix} \mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] \\ \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \\ \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2] \end{bmatrix} + \begin{bmatrix} \frac{\gamma^2 \sigma^2}{n} \\ 0 \\ m_\sigma \end{bmatrix}, \quad (33)$$

where m_σ and the matrix $\mathbf{A} = (a_{ij})_{3 \times 3}$ are defined in (13) and (28), respectively. Then,

$$\begin{bmatrix} \mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] \\ \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \\ \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2] \end{bmatrix} \leq \mathbf{A}^k \begin{bmatrix} \mathbf{E} [\|\bar{x}_0 - x_{ps}\|^2] \\ \mathbf{E} [\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2] \\ \mathbf{E} [\|\mathbf{y}_0 - \bar{\mathbf{y}}_0\|^2] \end{bmatrix} + \sum_{l=0}^{k-1} \mathbf{A}^l \begin{bmatrix} \frac{\gamma^2 \sigma^2}{n} \\ 0 \\ m_\sigma \end{bmatrix}. \quad (34)$$

By some calculations,

$$\left\| \sum_{l=0}^{k-1} \mathbf{A}^l \mathbf{B} \right\| \leq \sum_{l=0}^{k-1} \rho(\mathbf{A})^l \|\mathbf{B}\| \leq \sum_{l=0}^{\infty} \rho(\mathbf{A})^l \|\mathbf{B}\| = \lim_{k \rightarrow \infty} \frac{(1 - \rho(\mathbf{A})^k) \|\mathbf{B}\|}{1 - \rho(\mathbf{A})},$$

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

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

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

Next, we employ Lemma 4 in Appendix to show $\rho(\mathbf{A}) < 1$, 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

$$a_{33} = \rho + \frac{\gamma^2 \rho^2}{1 - \rho} \left[(1 + \hat{\beta}) c_3^2 + 2\sigma^2 \frac{1 + \rho^2}{1 - \rho^2} \right] = \frac{1 + \rho}{2} < 1. \quad (35)$$

By (29),

$$\begin{aligned} & a_{23} a_{32} \\ &= \frac{\gamma^2 \rho^2 (1 + \rho^2)}{(1 - \rho^2)(1 - \rho)} \left[c_3^2 \left(1 + \frac{1}{\hat{\beta}}\right) \|\mathbf{W} - \mathbf{I}\|_2^2 + \frac{2\gamma^2 (c_1 n \sigma^2 + c_3^2 (\sigma^2 + c_3^2))}{n} \right. \\ & \quad \left. + 2\gamma \sigma^2 (2c_3 + c_2) + (3 + \rho^2) \sigma^2 \right] \\ &\leq \frac{\gamma^2 \rho^2 (1 + \rho^2)}{(1 - \rho^2)(1 - \rho)} \left[\frac{c_3^2 \|\mathbf{W} - \mathbf{I}\|_2^2 (4\sigma^2 + c_3^2)}{4\sigma^2} + \frac{(1 - \rho)^2 (c_1 n \sigma^2 + c_3^2 (\sigma^2 + c_3^2))}{n c_W} \right. \\ & \quad \left. + \frac{2\sigma^2 (1 - \rho) (2c_3 + c_2)}{\sqrt{2} c_W} \right] \\ &\leq \frac{(1 - a_{22})(1 - a_{33})}{\kappa}, \end{aligned} \quad (36)$$

where c_1, c_2, c_3 are defined in (13), $\kappa > 1$ is arbitrarily chosen, and the first inequality follows from the fact that $\hat{\beta} > \frac{4\sigma^2}{c_3^2}$ and $\gamma \leq \frac{1-\rho}{\sqrt{2}c_W}$. Then

$$\begin{aligned}
& a_{12}a_{23}a_{31} \\
&= \frac{2\rho^2\gamma^3(c_2 + \gamma c_1)(1 + \rho^2) [\gamma^2 c_3^2(\sigma^2 + c_3^2) + 2\gamma c_3\sigma^2 n + n\sigma^2(2 - \bar{\mu}\gamma)]}{n(1 + \rho)(1 - \rho)^2} \\
&\leq \frac{\bar{\mu}\gamma(1 - \rho)^2(1 + \rho)(\kappa - 1)}{4\kappa(\kappa + 1)} \\
&= \frac{(1 - a_{11})(1 - a_{22})(1 - a_{33})(\kappa - 1)}{\kappa(\kappa + 1)} \\
&\leq \frac{(1 - a_{11})[(1 - a_{22})(1 - a_{33}) - a_{23}a_{32}]}{\kappa + 1}, \tag{37}
\end{aligned}$$

where $\bar{\mu}, c_1, c_2, c_3$ are defined in (13), $\rho \in (0, 1)$ due to [49, Lemma 1], the first inequality and the last inequality follow from (29) and (36), respectively. Subsequently, by (35), (36) and (37), we have

$$\begin{aligned}
& \det(\mathbf{I} - \mathbf{A}) \\
&= (1 - a_{11})(1 - a_{22})(1 - a_{33}) - (1 - a_{11})a_{23}a_{32} - a_{12}a_{23}a_{31} \\
&\geq \frac{\kappa}{\kappa + 1}(1 - a_{11})[(1 - a_{22})(1 - a_{33}) - a_{23}a_{32}] \\
&\geq \frac{\kappa - 1}{\kappa + 1}(1 - a_{11})(1 - a_{22})(1 - a_{33}) > 0. \tag{38}
\end{aligned}$$

Summarizing the 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 $\mathbf{E} [\|\bar{x}_k - x_{ps}\|^2]$ and $\mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2]$. Denoting $[(\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}]_j$ as the j th element of the vector $[(\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}]$, we have

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

where $\bar{\mu}, c_1, c_2, m_\sigma$ are defined in (13), γ is defined in (29), the first inequality follows from (33) and the second inequality follows from (36) and (38). Similarly, we have

$$\begin{aligned}
& \limsup_{k \rightarrow \infty} \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \\
&\leq [(\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}]_2 \\
&= \frac{\frac{a_{31}a_{23}\gamma^2\sigma^2}{n} + a_{23}(1 - a_{11})m_\sigma}{\det(\mathbf{I} - \mathbf{A})}
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{a_{23}(\kappa+1)}{(\kappa-1)(1-a_{11})(1-a_{22})(1-a_{33})} \left[\frac{\gamma^2 \sigma^2 a_{31}}{n} + (1-a_{11})m_\sigma \right] \\
&= \frac{4\gamma\rho^2(\kappa+1)(1+\rho^2)(a_{31}\sigma^2\gamma^2 + \gamma\bar{\mu}nm_\sigma)}{n\bar{\mu}(\kappa-1)(1-\rho^2)^2(1-\rho)}.
\end{aligned}$$

The proof is complete. $\square$

Theorem 1 achieves a convergence rate similar to that established in [49], that is, the iterates generated by DSGTD-GD converge to a neighborhood of the performative stable solution at the linear rate when the step size satisfies (29). By the definitions of c_3 and $\bar{\mu}$ in (13),

$$2nc_3 - \bar{\mu} = L(2n(1 + \epsilon_{\max}) + (1 + \varrho)\epsilon_{\text{avg}}) - \mu > 0.$$

Combined with the fact that $\rho \in (0, 1)$, all the terms in (29) are positive, which means the step size γ is obtainable. On the other hand, in light of (31) and (32),

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

which means the distance $\mathbf{E} [\|\mathbf{x}_k - \mathbf{x}_{ps}\|^2]$ is proportional to the step size and the spectral norm ρ . In other words, the smaller step size and 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 [33, 41]. In what follows, we study DSGTD-GD with a diminishing step size policy.

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

$$\begin{cases} 2r^2\sigma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2} + r^2c_3^2\rho^2 < \frac{b^{2a}(1-\rho)^2}{2} \\ r > \frac{b^a - (b-1)^a}{b^a - (b-1)^a} \\ \frac{b^{2a}(1-\rho^2)\bar{\mu}}{2(1+\rho^2)\rho^2} \left(\frac{b^{2a}}{(b+1)^{2a}} - \frac{1+\rho^2}{2} \right) > \frac{K_1r^3(c_2b^a + rc_1)}{b^an(\bar{\mu}r + (b-1)^a - b^a)} + K_2r^2 \end{cases}, \quad (39)$$

with

$$\begin{aligned}
\hat{\beta}_k &:= \frac{\frac{(1-\rho)^2}{2} - 2\gamma_k^2\sigma^2\frac{(1+\rho^2)\rho^2}{1-\rho^2} - \gamma_k^2c_3^2\rho^2}{\gamma_k^2c_3^2\rho^2}, \\
K_1 &:= \frac{2}{(1-\rho)b^{2a}} [r^2c_3^2b^a(\sigma^2 + c_3^2) + 2rc_3\sigma^2n + 2nb^{2a}\sigma^2], \\
K_2 &:= \frac{1}{1-\rho} \left[c_3^2\left(1 + \frac{1}{\hat{\beta}_0}\right) \|\mathbf{W} - \mathbf{I}\|_2^2 + \frac{2r^2(c_1n\sigma^2 + c_3^2(\sigma^2 + c_3^2))}{nb^{2a}} + \frac{2r\sigma^2(2c_3 + c_2)}{b^a} \right. \\
&\quad \left. + (3 + \rho^2)\sigma^2 \right]
\end{aligned} \quad (40)$$

and $\bar{\mu}$, c_1 , c_2 , and c_3 as defined in (13). Then under the conditions of Lemma 1,

$$\mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] = \mathcal{O}(\gamma_k), \quad \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] = \mathcal{O}(\gamma_k^2). \quad (41)$$

Moreover, $\bar{x}_k$ converges to x_{ps} almost surely.

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

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

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 (39). Taking expectation on both sides of (14)-(16) with $\beta = \hat{\beta}_k$ and $\gamma = \gamma_k$, we have

$$\begin{bmatrix} \mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2] \\ \mathbf{E} [\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2] \\ \mathbf{E} [\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2] \end{bmatrix} \leq \mathbf{A}_k \begin{bmatrix} \mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] \\ \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \\ \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2] \end{bmatrix} + \begin{bmatrix} \frac{\gamma_k^2 \sigma^2}{n} \\ 0 \\ m_k \end{bmatrix}, \quad (42)$$

where

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

and $\bar{\mu}$, c_1 , c_2 , c_3 are defined in (13).

Before formally proving the results, we first denote

$$\begin{aligned} \Delta_1 &:= \frac{1-\rho^2}{(1+\rho^2)\rho^2} \left(\frac{b^{2a}}{(b+1)^{2a}} - \frac{1+\rho^2}{2} \right), & \Delta_2 &:= \frac{2}{(1-\rho)b^{2a}} [K_1 b^{2a} \|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2 + b^{2a} m_0], \\ \Delta_3 &:= \frac{2K_2 r^2}{(1-\rho)b^{2a}}, & \Delta_4 &:= \frac{2}{1-\rho} \left[\frac{K_1 r^2 \sigma^2}{b^a n (\bar{\mu} r + (b-1)^a - b^a)} + m_0 \right], \\ \Delta_5 &:= \frac{2r^2}{(1-\rho)b^{2a}} \left[\frac{K_1 r (c_2 b^a + r c_1)}{b^a n (\bar{\mu} r + (b-1)^a - b^a)} + K_2 \right], \\ N_1 &:= \max \left\{ \frac{r [(rc_2 b^a + r^2 c_1) N_2 + \sigma^2 b^{2a}]}{b^{2a} n (\bar{\mu} r + (b-1)^a - b^a)}, \frac{b^a \|\bar{x}_0 - x_{ps}\|^2}{r} \right\}, \\ N_2 &:= \max \left\{ \frac{b^{2a} \|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2}{r^2}, \frac{\mathbf{E} [\|\mathbf{y}_0 - \bar{\mathbf{y}}_0\|^2]}{\Delta_1}, \frac{\Delta_2}{\Delta_1 - \Delta_3}, \frac{\Delta_4}{\Delta_1 - \Delta_5} \right\}, \\ N_3 &:= \max \left\{ \mathbf{E} [\|\mathbf{y}_0 - \bar{\mathbf{y}}_0\|^2], \frac{1-\rho}{2} \left(\frac{r K_1 N_1}{b^a} + \frac{K_2 N_2 r^2}{b^{2a}} + m_0 \right) \right\}, \end{aligned} \quad (43)$$

where $\bar{\mu}$, c_1 , c_2 are defined in (13), K_1 , K_2 are defined in (40), the relations between r , a , b are defined in (39).

In what follows, we show that

$$\mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] \leq N_1 \gamma_k, \quad \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] \leq N_2 \gamma_k^2, \quad \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2] \leq N_3 \quad (44)$$

by induction. Obviously, when $k = 0$, (44) holds. Next, we show that (44) holds for $k + 1$ provided that (44) holds for some $k > 0$.

By (42) 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

$$\mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2] \leq (1 - \bar{\mu}\gamma_k) \mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] + \gamma_k \frac{(c_2 + c_1 \gamma_k)}{n} \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] + \frac{\gamma_k^2 \sigma^2}{n},$$

$$\mathbf{E} [\|\mathbf{x}_{k+1} - \bar{\mathbf{x}}_{k+1}\|^2] \leq \frac{1 + \rho^2}{2} \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] + \gamma_k^2 \frac{(1 + \rho^2) \rho^2}{1 - \rho^2} \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2],$$

$$\mathbf{E} [\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2] \leq a'_{31} \mathbf{E} [\|\bar{x}_k - x_{ps}\|^2] + a'_{32} \mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2] + \frac{1 + \rho}{2} \mathbf{E} [\|\mathbf{y}_k - \bar{\mathbf{y}}_k\|^2] + m_k.$$

Then

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

where r , a and b are defined in (39) and the first inequality, the second inequality and the last inequality follow from the fact that (44) holds for some $k > 0$, the definition of γ_k and the definition of N_1 in (43), respectively. Similarly, we have

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

and

$$\begin{aligned} & \mathbf{E} [\|\mathbf{y}_{k+1} - \bar{\mathbf{y}}_{k+1}\|^2] \\ & \leq \gamma_k K_1 N_1 + \gamma_k^2 K_2 N_2 + \frac{(1 + \rho) N_3}{2} + m_0 \\ & \leq \frac{r K_1 N_1}{b^a} + \frac{K_2 N_2 r^2}{b^{2a}} + \frac{(\rho - 1) N_3}{2} + m_0 + N_3 \\ & \leq N_3, \end{aligned}$$

where $m_0 = \frac{\sigma^2}{(1-\rho)b^{2a}} [r^2(2\sigma^2 + c_3^2) + 2rnL(1 + \epsilon_{avg})b^a + 2nb^{2a}]$, K_1, K_2 are defined in (40).

Summarizing the above, we can conclude that (44) holds for any $k \geq 0$. Obviously, N_1, N_2 , and N_3 are bounded constants and then $\mathbf{E} [\|\bar{x}_k - x_{ps}\|^2]$ and $\mathbf{E} [\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2]$ 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_{ps} almost surely.

Choosing $\gamma = \gamma_k$ in (14), we have

$$\mathbf{E} [\|\bar{x}_{k+1} - x_{ps}\|^2 | \mathcal{F}_k] \leq (1 - \bar{\mu}\gamma_k) \|\bar{x}_k - x_{ps}\|^2 + \frac{c_2\gamma_k + c_1\gamma_k^2}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma_k^2}{n}. \quad (45)$$

Denote

$$C_k = \|\bar{x}_k - x_{ps}\|^2, \quad D_k = 0, \quad E_k = \frac{c_2\gamma_k + c_1\gamma_k^2}{n} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2\gamma_k^2}{n}, \quad F_k = \bar{\mu}\gamma_k \|\bar{x}_k - x_{ps}\|^2,$$

(45) falls into the setting of Lemma 5.

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

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

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

where the first inequality and the second inequality follow from (44) and the fact that $\gamma_k = r(k+b)^{-a}$ with $a \in (\frac{1}{2}, 1)$, respectively. Monotone convergence theorem implies $\sum_{k=0}^{\infty} E_k < \infty$. Then the conditions of Lemma 5 hold, which implies that there is a non-negative finite random variable C_{∞} such that $C_k \rightarrow C_{\infty}$ and $\sum_k F_k < \infty$ almost surely which implies

$$\|\bar{x}_k - x_{ps}\|^2 \rightarrow C_{\infty}, \quad \sum_{k=0}^{\infty} \gamma_k \|\bar{x}_k - x_{ps}\|^2 < \infty$$

almost surely. Noting that $\sum_{k=0}^{\infty} \gamma_k$ diverges, we conclude that $\lim_{k \rightarrow \infty} \|\bar{x}_k - x_{ps}\|^2 = 0$ almost surely. The proof is complete. $\square$

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 the consensus errors of iterates converge to 0 at rate $\mathcal{O}(\gamma_k^2)$, which achieves the similar convergence rate as [49, 61]. Moreover, we have by (44) that

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

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

Next, we establish the asymptotic normality of the time-averaged iterates generated by DSGTD-GD.

Theorem 3. Let x_{ps} be the performative stable solution of problem (1). Under the conditions of Theorem 2 with $a \in (\frac{1}{2}, 1)$ and Assumption 4, the average iterates $\frac{1}{k} \sum_{t=0}^{k-1} x_{i,t}$ generated by DSGTD-GD satisfy

$$\sqrt{k} \left(\frac{1}{k} \sum_{t=0}^{k-1} x_{i,t} - x_{ps} \right) \xrightarrow{d} \mathcal{N}(0, \nabla R(x_{ps})^{-1} \Sigma \nabla R(x_{ps})^{-\top}), \quad \forall i \in \mathcal{V}, \quad (46)$$

where $R(x) = \sum_{i=1}^n \nabla f_i(x; x)$ and $\Sigma := \text{Cov}(\sum_{i=1}^n \nabla \ell_i(x_{ps}; \xi_i))$.

Proof. By (44), we have

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

By the definition of γ_k and Kronecker lemma, $\frac{\sqrt{N_2}}{\sqrt{k}} \sum_{t=0}^{k-1} \gamma_t$ converges to 0. Then Slutsky's theorem implies (46) if

$$\sqrt{k} \left(\frac{1}{k} \sum_{t=0}^{k-1} \bar{x}_t - x_{ps} \right) \xrightarrow{d} N(0, \nabla R(x_{ps})^{-1} \Sigma \nabla R(x_{ps})^{-\top}). \quad (47)$$

Next, we show (47) by Lemma 3 in Appendix.

Firstly, we rewrite the recursion $\bar{x}_k - x_{ps}$ in the form of (67) in Lemma 3. By the update recursion in (7) and the fact $\bar{y}_k = \frac{1}{n} \sum_{i=1}^n \nabla \ell_i(x_{i,k}; \xi_{i,k+1})$, we have

$$\begin{aligned} & \bar{x}_{k+1} \\ &= \bar{x}_k - \gamma_k \bar{y}_k \\ &= \bar{x}_k - \frac{\gamma_k}{n} R(\bar{x}_k) - \gamma_k \left[\frac{1}{n} \sum_{i=1}^n (\nabla \ell_i(x_{i,k}; \xi_{i,k+1}) - \nabla f_i(x_{i,k}; x_{i,k})) \right] \\ &\quad - \gamma_k \left[\frac{1}{n} \sum_{i=1}^n (\nabla f_i(x_{i,k}; x_{i,k}) - \nabla f_i(\bar{x}_k; \bar{x}_k)) \right]. \end{aligned} \quad (48)$$

Denote

$$e_k := \frac{1}{n} \sum_{i=1}^n (\nabla \ell_i(x_{i,k}; \xi_{i,k+1}) - \nabla f_i(x_{i,k}; x_{i,k})), \quad d_k := \frac{1}{n} \sum_{i=1}^n (\nabla f_i(x_{i,k}; x_{i,k}) - \nabla f_i(\bar{x}_k; \bar{x}_k)),$$

the recursion (48) can be rewritten as

$$\bar{x}_{k+1} = \bar{x}_k - \gamma_k \left(\frac{1}{n} R(\bar{x}_k) + e_k + d_k \right),$$

which falls into the setting of Lemma 3.

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

By the definition of e_k , $\{e_k\}$ is a martingale difference sequence and thus condition (a) Lemma 3 hold. Moreover, by Assumption 2 (c),

$$\begin{aligned} & \mathbf{E} [\|e_k\|^2 \mid \mathcal{F}_k] \\ &= \frac{1}{n^2} \sum_{i=1}^n \mathbf{E} [\|\nabla \ell_i(x_{i,k}; \xi_{i,k+1}) - \nabla f_i(x_{i,k}; x_{i,k})\|^2 \mid \mathcal{F}_k] \\ &\leq \frac{\sigma^2}{n} + \frac{\sigma^2}{n^2} \sum_{i=1}^n \|x_{i,k} - x_{ps}\|^2. \end{aligned}$$

Taking expectations on both sides, we obtain

$$\mathbf{E} \|e_k\|^2 \leq \frac{\sigma^2}{n} + \frac{\sigma^2}{n^2} \mathbf{E} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 + \frac{\sigma^2}{n} \mathbf{E} \|\bar{x}_k - x_{ps}\|^2,$$

which implies that $\sup_{k \geq 0} \mathbf{E} \|e_k\|^2 < \infty$ by Theorem 2. Then condition (b) of Lemma 3 holds.

Next, we verify condition (c) of Lemma 3. Define

$$V_i(x) := \text{Cov}(\nabla \ell_i(x; \xi_i)), \quad \mathbf{E} [e_k e_k^\top \mid \mathcal{F}_k] = \frac{1}{n^2} \sum_{i=1}^n V_i(x_{i,k}).$$

By the fact that $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ and (41), we have $\sum_{k=0}^{\infty} \mathbf{E} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 < \infty$. Tonelli's theorem further implies that $\sum_{k=0}^{\infty} \|\mathbf{x}_k - \bar{\mathbf{x}}_k\|^2 < \infty$ almost surely, and then we have $\|\mathbf{x}_k - \bar{\mathbf{x}}_k\| \rightarrow 0$ almost surely. Together with $\bar{x}_k \rightarrow x_{ps}$ almost surely from Theorem 2, this yields that $x_{i,k} \rightarrow x_{ps}$ almost surely, for all $i \in \mathcal{V}$. Therefore,

$$\mathbf{E} [e_k e_k^\top \mid \mathcal{F}_k] \longrightarrow \frac{1}{n^2} \Sigma, \quad a.s.$$

where $\Sigma = \text{Cov}(\sum_{i=1}^n \nabla \ell_i(x_{ps}; \xi_i))$. Consequently, by Cesàro's lemma,

$$\frac{1}{k} \sum_{t=0}^{k-1} \mathbf{E} [e_t e_t^\top \mid \mathcal{F}_t] \longrightarrow \frac{1}{n^2} \Sigma, \quad a.s.$$

Moreover, since Assumption 2 (d) holds for some $p > 2$, the conditional p -th moments of e_k are uniformly bounded along the sample paths for all sufficiently large k . Hence, for any $\varepsilon > 0$,

$$\frac{1}{k} \sum_{t=0}^{k-1} \mathbf{E} [\|e_t\|^2 \mathbf{1}_{\{\|e_t\| \geq \varepsilon \sqrt{k}\}} \mid \mathcal{F}_t] \leq \frac{1}{\varepsilon^{p-2} k^{(p-2)/2}} \frac{1}{k} \sum_{t=0}^{k-1} \mathbf{E} [\|e_t\|^p \mid \mathcal{F}_t] \longrightarrow 0.$$

Then condition (c) of Lemma 3 holds.

On the other hand, by [14, Lemma 2.1] and Theorem 2,

$$\mathbf{E}[\|d_k\|] \leq \frac{L(1 + \epsilon_{\max})}{\sqrt{n}} \mathbf{E}[\|\mathbf{x}_k - \bar{\mathbf{x}}_k\|] = \mathcal{O}(\gamma_k). \quad (49)$$

Then,

$$\mathbf{E} \left[\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \|d_t\| \right] = \mathcal{O} \left(\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \gamma_t \right) = \mathcal{O}(k^{1/2-a}) \longrightarrow 0$$

Markov's inequality further implies that $\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \|d_t\| \xrightarrow{P} 0$, thereby condition (d) of Lemma 3 holds.

According to the definition of $\nabla R(x)$ and γ_k , conditions (e) of Lemma 3 hold with $G := \nabla R(x_{ps})/n$.

Finally, we verify condition (f) of Lemma 3. By Assumption 4 and $R(x_{ps}) = 0$, there exists a constant $c > 0$ such that, in a neighborhood of x_{ps} ,

$$\left\| \frac{1}{n} R(x) - G(x - x_{ps}) \right\| \leq c \|x - x_{ps}\|^2.$$

Then,

$$\mathbf{E} \left[\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \left\| \frac{1}{n} R(\bar{x}_t) - G(\bar{x}_t - x_{ps}) \right\| \right] \leq \frac{c}{\sqrt{k}} \sum_{t=0}^{k-1} \mathbf{E} \|\bar{x}_t - x_{ps}\|^2 = \mathcal{O} \left(\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \gamma_t \right) = o(1).$$

Thus, condition (f) of Lemma 3 holds by Markov inequality.

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

$$\sqrt{k} \left(\frac{1}{k} \sum_{t=0}^{k-1} x_{i,t} - x_{ps} \right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \nabla R(x_{ps})^{-1} \Sigma \nabla R(x_{ps})^{-\top}).$$

The proof is complete. $\square$

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 $\nabla R(x_{ps})^{-1} \Sigma \nabla R(x_{ps})^{-\top}$, where

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

Comparing with the asymptotic normality results of stochastic approximation method for standard stochastic optimization problems [7, 10, 17, 48], the second term $\sum_{i=1}^n \frac{d}{dy} \nabla f_i(x_{ps}; y)|_{y=x_{ps}}$ 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 stationary point based on the technical lemma.

To study the convergence of Algorithm 2, we first impose the following assumptions.

Assumption 5. [42] **[Objective function and gradient]**

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

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

- (b) There exists a constant $\delta > 0$ such that, for all $i \in \mathcal{V}$ and $x \in \mathbb{R}^d$,

$$\mathbf{E}_{\xi_i \sim \mathcal{D}_i(x)} [\|\nabla_{\xi_i} \ell_i(x; \xi_i)\|] \leq \delta.$$

- (c) There exists a constant $\sigma > 0$ such that, for all $i \in \mathcal{V}$ and $x \in \mathbb{R}^d$,

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

Assumption 6. [42] **[Distribution map]** For any $i \in \mathcal{V}$, there exists a probability measure $\mathcal{P}_i$ such that

$$\xi_i \sim \mathcal{D}_i(x) \iff \xi_i = A_i x + \zeta_i, \quad \zeta_i \sim \mathcal{P}_i,$$

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

In Assumption 5, condition (a) requires the objective functions to be smooth, while conditions (b) and (c) bound the expected norm of the stochastic gradient and its variance, respectively. Assumption 6 imposes an additive linear structure on how the distribution changes with the decision variable, rather than restricting the random variable ξ_i to a particular distributional family. Specifically, $\zeta_i \sim \mathcal{P}_i$ represents the decision-independent random variable, while $A_i x$ characterizes the distributional shift induced by the decision variable x . For example, if $\zeta_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, then under Assumption 6, we have $\xi_i = A_i x + \zeta_i$ and consequently $\xi_i \sim \mathcal{N}(A_i x + \mu_i, \Sigma_i)$. Under this Assumption, problem (1) can be rewritten as the following standard stochastic optimization problem:

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

which may be non-convex, and then by using the chain rule, we can derive the gradient of $f_i(x)$,

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

Again, under the linear structure of the distribution map, we may estimate the distribution map $\mathcal{D}_i(\cdot)$ by estimating the parameter A_i with $A_{i,k}$ dynamically using the adaptive gradient method in [42]. Specifically, at the k -th iteration, the i -th agent updates its $A_{i,k}$ with

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

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 $u_{i,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. Furthermore, an unbiased stochastic gradient of f_i at $x_{i,k}$ can be computed as

$$\nabla \ell_i(x_{i,k}; \xi_{i,k+1}) + A_i^\top \nabla_{\xi_{i,k+1}} \ell_i(x_{i,k}; \xi_{i,k+1}),$$

where $\xi_{i,k+1} \sim \mathcal{D}_i(x_{i,k})$, and with initial $A_{i,0} \in \mathbb{R}^{m \times d}$ and the estimation $A_{i,k}$, we form its estimator as

$$v_{i,k} = \nabla \ell_i(x_{i,k}; \xi_{i,k+1}) + A_{i,k}^\top \nabla_{\xi_{i,k+1}} \ell_i(x_{i,k}; \xi_{i,k+1}), \quad (50)$$

which is used as a biased stochastic gradient by DSGTD-AG.

Next, we state a 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 iterates $\bar{x}_k$ generated by DSGTD-AG satisfy*

$$\begin{aligned} & \mathbf{E}[f(\bar{x}_{k+1}) | \mathcal{F}_k] \\ & \leq f(\bar{x}_k) - \frac{\gamma}{4n} \|\nabla f(\bar{x}_k)\|^2 + \frac{3\gamma L^2}{2} \sum_{i=1}^n \|\bar{x}_k - x_{i,k}\|^2 \\ & \quad + \left(\frac{3L\delta^2\gamma^2}{n} + \gamma\delta^2 + \frac{3L\sigma^2\gamma^2}{n} \right) \|A_k - A\|^2 + \frac{3L\sigma^2\gamma^2}{n} \|A\|^2 + \frac{3L\sigma^2\gamma^2}{2n}, \end{aligned} \quad (51)$$

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

Proof. By the updating recursion (12) and the definition of $\mathbf{v}_k$,

$$\begin{aligned} & \mathbf{E}[f(\bar{x}_{k+1}) | \mathcal{F}_k] \\ & = \mathbf{E} \left[f(\bar{x}_k - \frac{\gamma}{n} \sum_{i=1}^n v_{i,k}) | \mathcal{F}_k \right] \\ & \leq f(\bar{x}_k) - \mathbf{E} \left[\langle \nabla f(\bar{x}_k), \frac{\gamma}{n} \sum_{i=1}^n v_{i,k} \rangle | \mathcal{F}_k \right] + \frac{L\gamma^2}{2} \mathbf{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n v_{i,k} \right\|^2 | \mathcal{F}_k \right] \\ & = f(\bar{x}_k) - \frac{\gamma}{n} \|\nabla f(\bar{x}_k)\|^2 + \frac{\gamma}{n} \mathbf{E} \left[\langle \nabla f(\bar{x}_k), \nabla f(\bar{x}_k) - \sum_{i=1}^n v_{i,k} \rangle | \mathcal{F}_k \right] + \frac{L\gamma^2}{2} \mathbf{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n v_{i,k} \right\|^2 | \mathcal{F}_k \right] \\ & \leq f(\bar{x}_k) - \frac{\gamma}{2n} \|\nabla f(\bar{x}_k)\|^2 + \frac{\gamma}{2} \sum_{i=1}^n \|\nabla f_i(\bar{x}_k) - \mathbf{E}[v_{i,k} | \mathcal{F}_k]\|^2 + \frac{L\gamma^2}{2} \mathbf{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n v_{i,k} \right\|^2 | \mathcal{F}_k \right], \end{aligned} \quad (52)$$

where $L = \sum_{i=1}^n L_i$, $\mathcal{F}_k$ denotes the σ -algebra generated by $(x_{i,l}, u_{i,l})_{l=0, \dots, k-1}, i \in \mathcal{V}$, 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 (52),

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

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

where the last inequality is due to the definition of $v_{i,k}$ in (50) and Assumption 5 (c). Then

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

Again, by the definition of $v_{i,k}$ in (50),

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

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

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

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. $\square$

Different from the descent lemma [27, Lemma 27], the last term on the right-hand side of (51) 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 $C_1 = 2560$, $C_2 = 12000$, $c_l := \max\{7\|A_0 - A\|^2, 8\sum_{i=1}^n \text{tr}(\Sigma_i)\}$, $f^* := \min_{x \in \mathbb{R}^d} f(x)$ and denote

$$\begin{aligned} c_4 &:= 8n, \quad c_5 := 12nL^2\tau^2 \left[\frac{c_l C_2 (2\sigma^2 + \delta^2) + 2C_2\sigma^2\|A\|^2 + C_2\sigma^2}{(1-\rho)^2} + C_1\sigma^2 + 2C_1\sigma^2\|A\|^2 \right], \\ c_6 &:= 24c_l L(\delta^2 + \sigma^2) + 24L\sigma^2\|A\|^2 + 12L\sigma^2, \quad c_7 = \frac{512\sqrt{10}F_0}{(1-\rho)} + 768\phi_0 L, \\ \tau &:= \frac{2}{1-\rho} \log\left(\frac{50}{1-\rho} \left(1 + \log \frac{1}{1-\rho}\right)\right) + 1, \quad F_0 := f(\bar{x}_0) - f^*, \\ \phi_0 &:= 16\|\mathbf{x}_0 - \bar{\mathbf{x}}_0\|^2 + \frac{24\gamma^2}{(1-\rho)^2} \|\mathbf{y}_0 - \bar{\mathbf{y}}_0\|^2. \end{aligned} \tag{53}$$

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

$$\gamma := \min \left\{ \left(\frac{c_4 F_0}{c_5 (K+1)} \right)^{\frac{1}{3}}, \left(\frac{c_4 F_0}{c_6 (K+1)} \right)^{\frac{2}{3}}, \frac{1-\rho}{64\sqrt{10}L\tau} \right\}. \tag{54}$$

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

$$\begin{aligned} & \frac{1}{K+1} \sum_{k=0}^K \mathbf{E} \|\nabla f(\bar{x}_k)\|^2 \\ & \leq 2c_5^{\frac{1}{3}} \left(\frac{c_4 F_0}{K+1} \right)^{\frac{2}{3}} + 2 \left(\frac{c_4 c_6 F_0}{K+1} \right)^{\frac{2}{3}} + \frac{n\tau L c_7}{K+1} + \frac{8nc_l \delta^2 \ln(K+6d)}{K+1} + \frac{16(c_5 c_4^2 F_0^2)^{\frac{1}{3}} \ln(K+6d)}{75(K+1)^{\frac{2}{3}}}. \end{aligned}$$

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

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

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

$$\begin{aligned} & \frac{1}{K+1} \sum_{k=0}^K \mathbf{E} \|\nabla f(\bar{x}_k)\|^2 \\ & \leq \frac{4nF_0}{\gamma(K+1)} + 12L\gamma\sigma^2\|A\|^2 + 6L\gamma\sigma^2 \\ & \quad + \frac{4(3L\gamma\delta^2 + n\delta^2 + 3L\gamma\sigma^2)}{K+1} \sum_{k=0}^K \mathbf{E} [\|A_k - A\|^2] + \frac{6nL^2}{K+1} \sum_{k=0}^K \mathbf{E} \|\psi_k\|^2, \end{aligned} \tag{55}$$

where $F_0 := f(\bar{x}_0) - f^*$. Note that the first three terms on the right-hand side of (55) do not depend on iteration k , we may provide the bound for them by selecting the step size, that is,

$$\frac{4nF_0}{\gamma(K+1)} + 12L\gamma\sigma^2\|A\|^2 + 6L\gamma\sigma^2 \leq 6L\sigma^2 (2\|A\|^2 + 1) \left(\frac{c_4 F_0}{c_6 (K+1)} \right)^{\frac{2}{3}} + \frac{32\sqrt{10}c_4 L\tau F_0}{(1-\rho)(K+1)}, \tag{56}$$

where c_4 and c_6 are defined in (53), the inequality is due to the definition of γ in (54). Similarly, as the fourth term on the right-hand side of (55) only depends on the error of $\|A_k - A\|^2$, we may bound it with [42, Lemma 21],

$$\begin{aligned} & \frac{4(3L\gamma\delta^2 + n\delta^2 + 3L\gamma\sigma^2)}{K+1} \sum_{k=0}^K \mathbf{E} [\|A_k - A\|^2] \\ & \leq \frac{4c_l(3L\gamma\delta^2 + n\delta^2 + 3L\gamma\sigma^2)}{K+1} \sum_{k=0}^K \frac{1}{k+6d} \\ & \leq \frac{12c_l L\gamma(\delta^2 + \sigma^2) \ln(K+6d)}{K+1} + \frac{4nc_l\delta^2 \ln(K+6d)}{K+1}, \end{aligned} \quad (57)$$

where $c_l := \max\{7\|A_0 - A\|^2, 8\sum_{i=1}^n \text{tr}(\Sigma_i)\}$ and the last inequality follows from the definition of γ in (54).

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

Define

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

we may rewrite Algorithm 2 in the form as

$$\psi_{k+1} = \psi_k J + \gamma \Xi_k, \quad (58)$$

where $\psi_k = \left[\mathbf{x}_k - \bar{\mathbf{x}}_k, \gamma(\mathbf{y}_k - \bar{\mathbf{y}}_k) \right]$. By unrolling (58) 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,

$$\begin{aligned} & \mathbf{E} \|\psi_{k+t}\|^2 \\ & \leq \frac{3}{4} \mathbf{E} \|\psi_k\|^2 + 5\gamma^2 \mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m} - \mathbf{v}_{k+m-1}) \tilde{\mathbf{W}}^{t-m} \right\|^2 \\ & \quad + 5\gamma^2 \mathbf{E} \left\| \sum_{m=1}^{t-1} (\mathbf{v}_{k+m} - \mathbf{v}_{k+m-1})(t-m) \tilde{\mathbf{W}}^{t-m} \right\|^2, \end{aligned} \quad (59)$$

where the inequality follows from Young's inequality, Lemma 8 and the fact that $\|\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\|_2 \leq 1$.

In what follows, we give the upper bounds for the last two items on the right-hand side of (59).

Denote

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

For the second term on the right-hand side of (59),

$$\begin{aligned}
& \mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m} - \mathbf{v}_{k+m-1}) \tilde{\mathbf{W}}^{t-m} \right\|^2 \\
& \leq 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m} - \mathbf{F}_{k+m}) \right\|^2 + 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m-1} - \mathbf{F}_{k+m-1}) \right\|^2 + 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{F}_{k+m} - \mathbf{F}_{k+m-1}) \right\|^2 \\
& \leq 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m} - \mathbf{F}_{k+m}) \right\|^2 + 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m-1} - \mathbf{F}_{k+m-1}) \right\|^2 + 3t \sum_{m=1}^t \mathbf{E} \left\| \mathbf{F}_{k+m} - \mathbf{F}_{k+m-1} \right\|^2 \\
& \leq 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m} - \mathbf{F}_{k+m}) \right\|^2 + 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m-1} - \mathbf{F}_{k+m-1}) \right\|^2 + 9t \sum_{m=1}^t \mathbf{E} \left\| \mathbf{F}_{k+m} - \bar{\mathbf{F}}_{k+m} \right\|^2 \\
& \quad + 9t \sum_{m=1}^t \mathbf{E} \left\| \mathbf{F}_{k+m-1} - \bar{\mathbf{F}}_{k+m-1} \right\|^2 + 9t \sum_{m=1}^t \mathbf{E} \left\| \bar{\mathbf{F}}_{k+m} - \bar{\mathbf{F}}_{k+m-1} \right\|^2 \\
& \leq 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m} - \mathbf{F}_{k+m}) \right\|^2 + 3\mathbf{E} \left\| \sum_{m=1}^t (\mathbf{v}_{k+m-1} - \mathbf{F}_{k+m-1}) \right\|^2 + 9tL^2 \sum_{m=1}^t \mathbf{E} \left\| \mathbf{x}_{k+m} - \bar{\mathbf{x}}_{k+m} \right\|^2 \\
& \quad + 9tL^2 \sum_{m=1}^t \mathbf{E} \left\| \mathbf{x}_{k+m-1} - \bar{\mathbf{x}}_{k+m-1} \right\|^2 + 9tnL^2 \sum_{m=1}^t \mathbf{E} \left\| \bar{x}_{k+m} - \bar{x}_{k+m-1} \right\|^2, \tag{60}
\end{aligned}$$

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 (59),

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

$$\begin{aligned}
& + 3t \sum_{m=1}^{t-2} \|(\mathbf{v}_{k+m} - \mathbf{F}_{k+m}) \left[(t-m)\tilde{\mathbf{W}}^{t-m} - (t-m-1)\tilde{\mathbf{W}}^{t-m-1} \right]\|^2 \\
& + 3t\mathbf{E}\|(\mathbf{F}_{k+t-1} - \bar{\mathbf{F}}_{k+t-1})\tilde{\mathbf{W}}\|^2 + 3t\mathbf{E}\|(\mathbf{F}_k - \bar{\mathbf{F}}_k)(t-1)\tilde{\mathbf{W}}^{t-1}\|^2 \\
& + 3t \sum_{m=1}^{t-2} \|(\mathbf{F}_{k+m} - \bar{\mathbf{F}}_{k+m}) \left[(t-m)\tilde{\mathbf{W}}^{t-m} - (t-m-1)\tilde{\mathbf{W}}^{t-m-1} \right]\|^2 \\
& + 3\mathbf{E}\| \sum_{m=1}^{t-1} (\bar{\mathbf{F}}_{k+m} - \bar{\mathbf{F}}_{k+m-1})(t-m)\tilde{\mathbf{W}}^{t-m} \|^2 \\
& \leq \frac{12t}{(1-\rho)^2} \sum_{m=1}^{t-1} \mathbf{E}\|\mathbf{v}_{k+m} - \mathbf{F}_{k+m}\|^2 + \frac{12t}{(1-\rho)^2} \sum_{m=1}^{t-1} \mathbf{E}\|\mathbf{F}_{k+m} - \bar{\mathbf{F}}_{k+m}\|^2 \\
& \quad + \frac{3(t-1)}{(1-\rho)^2} \sum_{m=1}^{t-1} \mathbf{E}\|\bar{\mathbf{F}}_{k+m} - \bar{\mathbf{F}}_{k+m-1}\|^2 \\
& \leq \frac{12t}{(1-\rho)^2} \sum_{m=1}^{t-1} \mathbf{E}\|\mathbf{v}_{k+m} - \mathbf{F}_{k+m}\|^2 + \frac{12tL^2}{(1-\rho)^2} \sum_{m=1}^{t-1} \mathbf{E}\|\mathbf{x}_{k+m} - \bar{\mathbf{x}}_{k+m}\|^2 \\
& \quad + \frac{3n(t-1)L^2}{(1-\rho)^2} \sum_{m=1}^{t-1} \mathbf{E}\|\bar{x}_{k+m} - \bar{x}_{k+m-1}\|^2, \tag{61}
\end{aligned}$$

where the third inequality is due to Lemma 6, Lemma 7, the fact that $\|\tilde{\mathbf{W}}^j\| \leq 1$, for $\forall j \geq 1$, and the last inequality is due to Assumption 5 (a). By the updating recursion in (12),

$$\begin{aligned}
& \mathbf{E}\|\bar{x}_{k+m} - \bar{x}_{k+m-1}\|^2 \\
& = \gamma^2 \mathbf{E}\left\| \frac{1}{n} \sum_{i=1}^n v_{i,k+m-1} \right\|^2 \\
& = \gamma^2 \mathbf{E}\left\| \frac{1}{n} \sum_{i=1}^n v_{i,k+m-1} - \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_{i,k+m-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_{i,k+m-1}) \right. \\
& \quad \left. - \frac{1}{n} \sum_{i=1}^n \nabla f_i(\bar{x}_{k+m-1}) + \frac{1}{n} \sum_{i=1}^n \nabla f_i(\bar{x}_{k+m-1}) \right\|^2 \\
& \leq 3\gamma^2 \mathbf{E}\left\| \frac{1}{n} \sum_{i=1}^n v_{i,k+m-1} - \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_{i,k+m-1}) \right\|^2 \\
& \quad + \frac{3\gamma^2}{n} \sum_{i=1}^n \|\nabla f_i(x_{i,k+m-1}) - \nabla f_i(\bar{x}_{k+m-1})\|^2 + \frac{3\gamma^2}{n^2} \|\nabla f(\bar{x}_{k+m-1})\|^2 \\
& \leq 3\gamma^2 \mathbf{E}\left\| \frac{1}{n} \sum_{i=1}^n v_{i,k+m-1} - \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_{i,k+m-1}) \right\|^2 \\
& \quad + \frac{3L^2\gamma^2}{n} \sum_{i=1}^n \|x_{i,k+m-1} - \bar{x}_{i,k+m-1}\|^2 + \frac{3\gamma^2}{n^2} \|\nabla f(\bar{x}_{k+m-1})\|^2, \tag{62}
\end{aligned}$$

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

Combining (60), (61), (62) with the inequality (59), we have

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

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

$$\begin{aligned}
& \mathbf{E}\|\mathbf{v}_{k+m} - \mathbf{F}_{k+m}\|^2 \\
& = \mathbf{E}\|\mathbf{v}_{k+m} - \mathbf{E}\mathbf{v}_{k+m}\|^2 + \|\mathbf{E}\mathbf{v}_{k+m} - \mathbf{F}_{k+m}\|^2 \\
& \leq (2\sigma^2 + \delta^2)\mathbf{E}\left[\|A_{k+m} - A\|^2\right] + 2\sigma^2\|A\|^2 + \sigma^2 \\
& \leq \frac{c_l(2\sigma^2 + \delta^2)}{k+m+6d} + 2\sigma^2\|A\|^2 + \sigma^2,
\end{aligned} \tag{63}$$

where $c_l := \max\{7\|A_0 - A\|^2, 8\sum_{i=1}^n \text{tr}(\Sigma_i)\}$, the first inequality follows from Assumption 5 (b), (c), the second inequality follows from the fact $\max\{1, \|A_k\|^2\} \leq 1 + \|A_k\|^2$ and $\|A_k\|^2 \leq 2\|A\|^2 + 2\|A_k - A\|^2$, and the last inequality follows from [42, Lemma 21]. Then, by the fact that $t < 2\tau$,

$$\begin{aligned}
& \mathbf{E}\|\psi_{k+t}\|^2 \\
& \leq \frac{7}{8}\mathbf{E}\|\psi_k\|^2 + \frac{1}{128\tau} \sum_{m=0}^{t-1} \mathbf{E}\|\psi_{k+m}\|^2 + \frac{2\tau\gamma^2}{n} \sum_{m=0}^{t-1} \|\nabla f(\bar{x}_{k+m})\|^2 \\
& \quad + 128\tau^2\gamma^2 \left[\frac{c_l(2\sigma^2 + \delta^2)}{k+6d} + 2\sigma^2\|A\|^2 + \sigma^2 \right] + \frac{600\tau^2\gamma^2}{(1-\rho)^2} \left[\frac{c_l(2\sigma^2 + \delta^2)}{k+6d} + 2\sigma^2\|A\|^2 + \sigma^2 \right],
\end{aligned}$$

which falls into the similar form to the recursion in [27, Lemma 7]. Subsequently, unrolling this recursion with the technique in [27, Lemma 8], we have

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

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

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

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

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

Combining (56), (57), (64) with (55) and rearranging, we have

$$\begin{aligned} & \frac{1}{K+1} \sum_{k=0}^K \mathbf{E}\|\nabla f(\bar{x}_k)\|^2 \\ & \leq 12L\sigma^2 (2\|A\|^2 + 1) \left(\frac{c_4F_0}{c_6(K+1)} \right)^{\frac{2}{3}} + \frac{64\sqrt{10}c_4L\tau F_0}{n(1-\rho)(K+1)} + 24L(\delta^2 + \sigma^2) \left(\frac{c_4F_0}{c_6(K+1)} \right)^{\frac{2}{3}} \\ & \quad + \frac{8nc_l\delta^2 \ln(K+6d)}{K+1} + \frac{12nL^2}{K+1} \left[64\tau\phi_0 + C_1\tau^2\gamma^2 \sum_{k=0}^K \left[\frac{c_l(2\sigma^2 + \delta^2)}{k + 6d} + 2\sigma^2\|A\|^2 + \sigma^2 \right] \right. \\ & \quad \left. + \frac{C_2\tau^2\gamma^2}{(1-\rho)^2} \sum_{k=0}^K \left[\frac{c_l(2\sigma^2 + \delta^2)}{k + 6d} + 2\sigma^2\|A\|^2 + \sigma^2 \right] \right] \\ & \leq 12L\sigma^2 (2\|A\|^2 + 1) \left(\frac{c_4F_0}{c_6(K+1)} \right)^{\frac{2}{3}} + \frac{64\sqrt{10}c_4L\tau F_0}{(1-\rho)(K+1)} + 24L(\delta^2 + \sigma^2) \left(\frac{c_4F_0}{c_6(K+1)} \right)^{\frac{2}{3}} \end{aligned}$$

$$\begin{aligned}
& + \frac{8nc_l\delta^2 \ln(K+6d)}{K+1} + \frac{768n\tau\phi_0L^2}{K+1} + 12nc_lC_1L^2\tau^2\gamma^2(2\sigma^2 + \delta^2) \ln(K+6d) \\
& + \frac{12C_2nc_lL^2\tau^2\gamma^2(2\sigma^2 + \delta^2)}{(1-\rho)^2} + \frac{12C_2nL^2\tau^2\gamma^2(2\sigma^2\|A\|^2 + \sigma^2)}{(1-\rho)^2} \\
& + 12C_1nL^2\tau^2\gamma^2(2\sigma^2\|A\|^2 + \sigma^2) \\
& \leq 2c_5^{\frac{1}{3}} \left(\frac{c_4F_0}{K+1} \right)^{\frac{2}{3}} + 2 \left(\frac{c_4c_6F_0}{K+1} \right)^{\frac{2}{3}} + \frac{n\tau Lc_7}{K+1} + \frac{8nc_l\delta^2 \ln(K+6d)}{K+1} + \frac{16(c_5c_4^2F_0^2)^{\frac{1}{3}} \ln(K+6d)}{75(K+1)^{\frac{2}{3}}},
\end{aligned}$$

where c_4, c_5, c_6, c_7, F_0 are defined in (53), the first inequality and the last inequality are due to [42, Lemma 21] and the definition of γ in (54), respectively. The proof is complete. $\square$

Theorem 4 shows that the averaged iterates generated by DSGTD-AG converge to the stationary point with rate $\mathcal{O}(\frac{\ln K}{K^{\frac{2}{3}}})$, where K is the number of iterates. Notably, compared to the rate of $\mathcal{O}(\frac{1}{K^{\frac{2}{3}}})$ in [27], the rate in Theorem 4 includes an additional logarithmic factor due to learning the parameters of the distribution maps. On the other hand, all the terms in (54) are positive as $\rho \in (0, 1)$, which means the step size γ is obtainable.

5 Numerical experiments

To show the performance of the proposed algorithms, we conduct experiments on two synthetic multi-agent problems with Gaussian decision-dependent distributions and a strategic binary classification problem with a real dataset [23]. The two synthetic experiments provide controlled settings with closed-form solutions, which facilitate comparison of the convergence behaviors of the algorithms. Specifically, the first experiment considers a case where the performative stable solution coincides with the optimal solution, while the second considers a case where the two solutions differ. The strategic binary classification experiment further evaluates the algorithms on a real-world data.

In all experiments, we compare the proposed algorithms with four distributed variants based on DSGD [31, 50] and EXTRA [40, 55]. Specifically, we combine each framework with the greedy deployment scheme [38] and the adaptive gradient [42], resulting in DSGD-GD [30], DSGD-AG, EXTRA-GD and EXTRA-AG. To the best of our knowledge, theoretical convergence guarantees for DSGD-AG, EXTRA-GD, and EXTRA-AG in the distributed SO-DD have not been established in the existing literature. These algorithms represent three classes of distributed optimization methods: stochastic-gradient type methods, primal-dual type methods, and gradient-tracking type methods; see Table 1. For the greedy-deployment variants, stochastic gradients are evaluated using samples drawn from the current local decision-dependent distributions under the greedy deployment scheme. For the adaptive-gradient variants, the adaptive gradient is employed to account for the effect of the decision variable on the distribution. We consider a ring communication network with $n = 25$ agents, where the mixing weight $\mathbf{W}$ is given by $w_{ij} = \frac{1}{3}$ for all $(i, j) \in \mathcal{E}$ and $w_{ij} = 0$ otherwise. For EXTRA-GD and EXTRA-AG, the second mixing matrix is chosen as $\frac{\mathbf{I} + \mathbf{W}}{2}$. All algorithms are initialized from the same starting point. In all figures, the orange and green lines denote DSGTD-GD with constant and diminish-

Table 1: Distributed algorithms compared in the numerical experiments.

Method class	Greedy-deployment variant	Adaptive-gradient variant
Stochastic-gradient type	DSGD-GD [30]	DSGD-AG
Primal-dual type	EXTRA-GD	EXTRA-AG
Gradient-tracking type	DSGTD-GD (our algorithm)	DSGTD-AG (our algorithm)

ing step sizes, respectively, the red line denotes DSGTD-AG, the blue and magenta dashed lines denote DSGD-GD and DSGD-AG, respectively, and the brown and cyan dashed lines denote EXTRA-GD and EXTRA-AG, respectively.

5.1 Multi-agent Gaussian Mean Estimation I

Consider the following distributed stochastic optimization problem with decision-dependent distributions [30]

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

where for all $i \in [25]$, $\mathcal{D}_i(x_i) = \mathcal{N}(\bar{z} + \epsilon_i x_i, \sigma^2)$, $\bar{z} = 10$, $\sigma^2 = 10$, $\epsilon_i = \epsilon_{avg} = 0.1$. Obviously, the objective value can be computed as 125 with the performative stable solution and the optimal solution $x_{ps} = x_{op} = 100/9$.

In the experiment, we run all algorithms with $k = 10^6$ iterates, where the i -th agent draws one sample from $\mathcal{D}_i(x_{i,k})$ 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 $\gamma = 0.001$. In particular, we set the step sizes for DSGD-GD and DSGD-AG to $\gamma_k = 50/(10000 + k)$, as described in [30], the step size for EXTRA-GD to $\gamma = 0.01$ and the step size for EXTRA-AG to $\gamma = 0.001$. All the parameters of the setting are chosen the same as that in [30]. Moreover, we do 100 simulations for the test on asymptotic normality.

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. 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 iterates, which matches the conclusion of Theorem 1 that DSGTD-GD with constant step size converges to a neighborhood of the performative stable solution. 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 solution and coincides exactly with Theorem 2. EXTRA-GD, using the same constant step size as DSGTD-GD, exhibits a comparable convergence rate and also approaches a neighborhood of the performative stable solution. This stable convergence behavior under a constant step size is consistent with that of EXTRA in deterministic decentralized optimization [55]. On the other

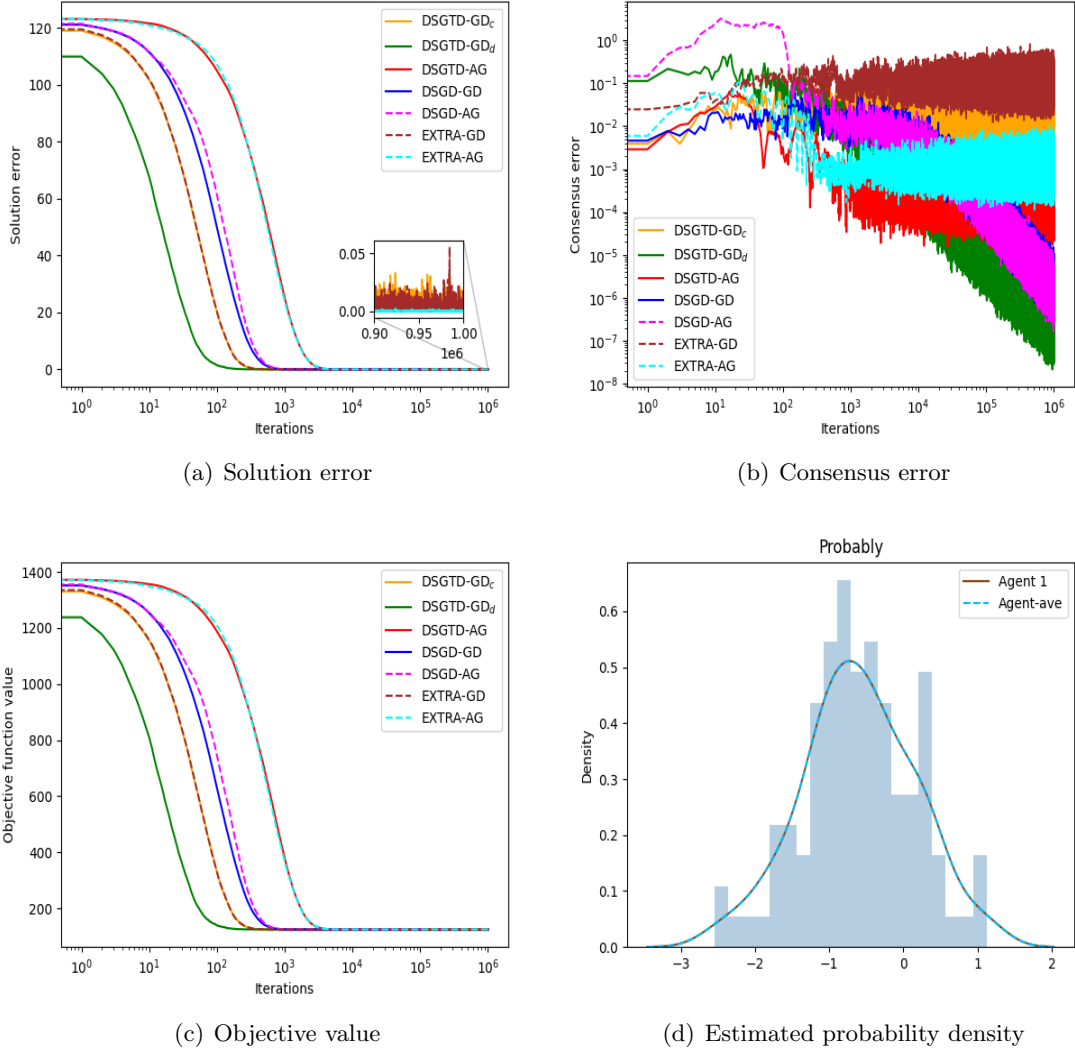


Figure 1: Multi-agent Gaussian Mean Estimation under the ring graph.

hand, the solution error of DSGTD-AG tends to 0, which implies DSGTD-AG can converge to the optimal solution as demonstrated in Theorem 4 when A_k sufficiently tend to A . DSGD-AG and EXTRA-AG exhibit the same convergence behavior, with their solution errors also tending to 0. These results indicate that DSGTD-GD and DSGTD-AG have comparable rates to those of the other algorithms. From Figure 1 (b), we may observe that the consensus errors of DSGTD-GD with constant step size, EXTRA-GD and EXTRA-AG consistently stabilize around 10^{-2} , 10^{-1} and 10^{-3} , respectively. The underlying reason may be that DSGTD-GD with constant step size, EXTRA-GD and EXTRA-AG converge to the neighborhoods of the performative stable solution or the optimal solution. 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, DSGD-AG. From Figure 1 (c), we can observe that the objective value of all algorithms reach similar values, which verifies the performative stable solution is close to the optimal solution. Overall, the comparable convergence behavior of DSGTD-GD with a constant

step size and EXTRA-GD is consistent with the close relationship between gradient-tracking and EXTRA-type methods, which can be represented as different special cases of a unified decentralized optimization framework [2, 60]. Similar convergence behavior is also observed between DSGTD-AG and EXTRA-AG. 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 $\mathcal{O}(\frac{1}{k^{2a}})$ in Theorem 2, where $a \in (\frac{1}{2}, 1]$.

5.2 Multi-agent Gaussian Mean Estimation II

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

$$\min_{x \in \mathbb{R}} \sum_{i=1}^{25} \mathbf{E}_{\omega_i \sim \mathcal{D}_i(x)} [\omega_i x], \quad (65)$$

where for all $i \in [25]$, $\mathcal{D}_i(x) = \mathcal{N}(\bar{z} + \epsilon_i x, \sigma^2)$, $\bar{z} = 10$, $\sigma^2 = 10$, $\epsilon_i = \epsilon_{avg} = 0.1$. The performative stable solution and the optimal solution can be computed as $x_{ps} = -100$ and $x_{op} = -50$, respectively, with the corresponding objective values 0 and -6250 .

In experiments, we run all algorithms with $k = 10^6$ iterates. Because of the similar setting of the experiment, we adopt the same parameters as in the first experiment.

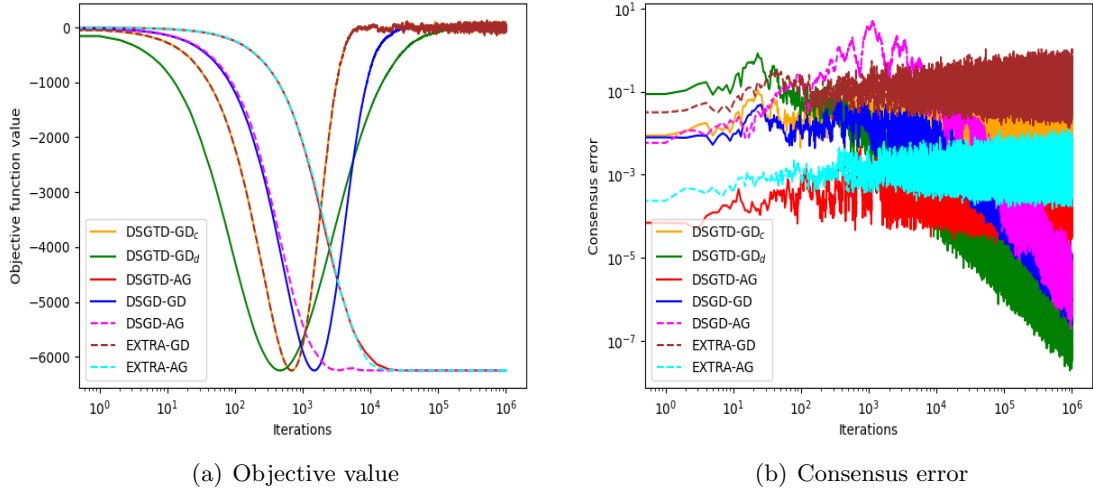


Figure 2: Multi-agent Gaussian Mean Estimation under the ring graph.

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. From Figure 2 (a), we can observe that the objective function value of DSGTD-AG, DSGD-AG and EXTRA-AG approaches optimal. In contrast, the objective function values generated by DSGTD-GD, DSGD-GD and EXTRA-GD first decrease and then increase as the iterates, eventually approach

the objective value associated with the performative stable solution. Moreover, this verifies that the performative stable solution may be far from optimal without strong convexity. From Figure 2 (b), we may observe that DSGTD-GD with diminishing step sizes, DSGTD-AG, DSGD-GD and DSGD-AG asymptotically reach consensus, while the consensus errors of DSGTD-GD with constant step size, EXTRA-GD and EXTRA-AG consistently stabilize around 10^{-2} , 10^{-1} and 10^{-3} , respectively. Similar to the first experiment, DSGTD-GD with constant step size and EXTRA-GD exhibit comparable convergence behavior, as do DSGTD-AG and EXTRA-AG, which is consistent with the structural relationship discussed above.

5.3 Logistic Regression

Consider the following logistic regression problem:

$$\min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^N \ell(x; a_{ij}, b_{ij}) + \frac{n\beta}{2} \|x\|^2, \quad (66)$$

with

$$\ell(x; a_{ij}, b_{ij}) = \log(1 + \exp(\langle a_{ij}, x \rangle)) - b_{ij} \langle a_{ij}, x \rangle,$$

where $\ell(\cdot)$ is the logistic loss function, x represents the classifier parameter, $\beta > 0$ is a regularization parameter, N is the number of training samples for each agent, a_{ij} is the feature vector of the j -th sample for the i -th agent and $b_{ij} \in \{0, 1\}$ is the corresponding label.

In what follows, we conduct the experiments on the spambase, a dataset [23] with $m = 4601$ samples and $d = 57$ features, which contains a collection of spam and valid emails. Given the base dataset $S_i^0 = \{a_{ij}^0, b_{ij}^0\}_{j=1}^N$ for the i -th agent, following the linear utility formulation in [30], we set $a_{ij} = a_{ij}^0 + \epsilon_i x_i$ and $b_{ij} = b_{ij}^0$. We note that ϵ_i corresponds to the special case with $A_i = \epsilon_i \mathbf{I}$ of the general distribution map considered in Assumption 6. 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 (66) with $\beta = 10^{-4}$. The sensitivity parameters are set as $\epsilon_i \in \{0.4\epsilon_{avg}, 0.45\epsilon_{avg}, \dots, 1.6\epsilon_{avg}\}$ with $\epsilon_{avg} = 0.1$. In experiments, we run all algorithms with $k = 6 \times 10^6$ iterates where the i -th server draws one sample 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 and EXTRA-AG is $\gamma = 0.0001$, the step sizes of DSGD-GD and DSGD-AG are $\gamma_k = 50/(100000 + k)$, the step size of EXTRA-GD is $\gamma = 0.001$.

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. From Figure 3 (a), we can observe that the train gradient norms of DSGTD-GD and DSGTD-AG tend to 0 at a comparable rate to those of the other algorithms. From Figure 3 (b), we may observe the consensus errors of DSGTD-GD with constant step size, DSGTD-AG, EXTRA-GD and EXTRA-AG consistently stabilize around 10^{-4} , 10^{-6} , 10^{-3} and 10^{-5} , 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 and DSGD-AG. Figure 3 (c), we can

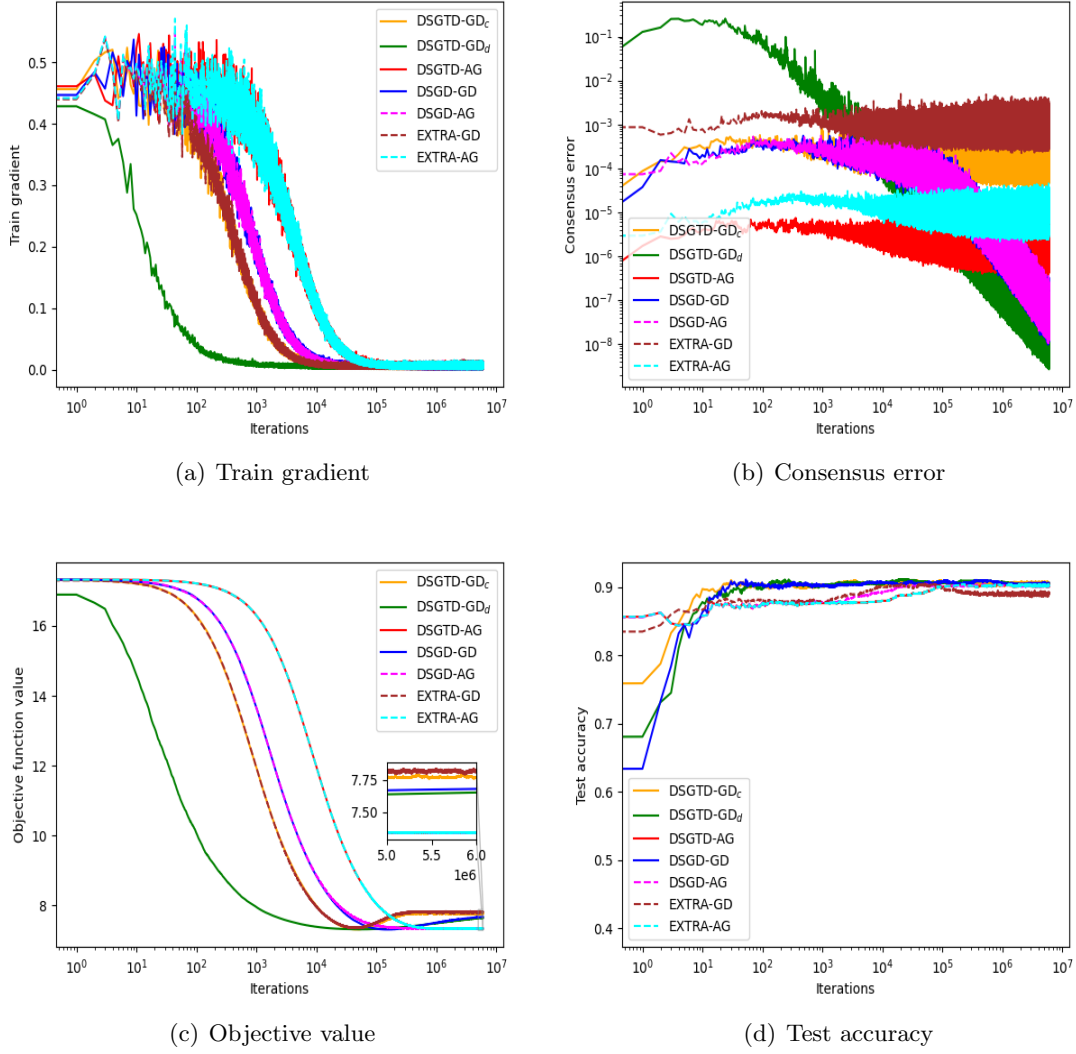


Figure 3: Spam Email Classification under the ring graph.

observe that the objective values of DSGTD-GD, DSGD-GD and EXTRA-GD increase and then decrease as the iteration, reaching slightly bigger values than those of DSGTD-AG, DSGD-AG and EXTRA-AG. This further verifies that DSGTD-AG, DSGD-AG, EXTRA-AG converges to the stationary point and DSGTD-GD, DSGD-GD, EXTRA-GD converge to the performative stable solution, where the performative stable solution may be far from the stationary 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 increases as iterates k and achieves the comparable level of accuracy with DSGD-GD, DSGD-AG and EXTRA-AG. In contrast, EXTRA-GD attains a noticeably lower test accuracy than the other methods, which is consistent with its relatively large consensus error observed in Figure 3 (b). Consistent with the observations in the synthetic experiments, DSGTD-GD with constant step size and EXTRA-GD exhibit comparable convergence behavior, as do DSGTD-AG and EXTRA-AG. Overall, these results indicate that the adaptive gradient method can be effectively

incorporated into different distributed optimization algorithms for distributed SO-DD, with the corresponding AG variants empirically converging to the optimal solution.

6 Conclusion

In this paper, we propose two stochastic gradient tracking descent methods, DSGTD-GD and DSGTD-AG, to seek the performative stable solution and the stationary point of distributed SO-DD. For DSGTD-GD, we establish its convergence to a neighborhood of the performative stable solution under a constant step size and to the performative stable solution under diminishing step sizes. We further establish the asymptotic normality of the averaged iterates, characterizing the effect of decision-dependent distributions. For DSGTD-AG, we establish its convergence to a stationary point of non-convex distributed SO-DD under a constant step size. Moreover, we compare the proposed algorithms with representative distributed methods in two synthetic-data experiments and one real-data experiment. Across the numerical experiments, the stochastic-gradient type and primal-dual type methods exhibit empirical convergence rates comparable to those of our algorithms under their corresponding greedy-deployment and adaptive-gradient variants. These results demonstrate that the greedy deployment scheme and adaptive gradient can be effectively incorporated into different distributed optimization algorithms for distributed SO-DD. Meanwhile, the three classes of methods differ in their algorithmic structures and communication requirements. Specifically, EXTRA-based methods employ two compatible symmetric mixing matrices, whereas gradient-tracking type methods introduce an auxiliary variable to track the averaged stochastic gradient and do not inherently require symmetric mixing matrices, which allows gradient-tracking methods to be extended to a broader range of network structures, including directed and time-varying networks. A systematic theoretical comparison of different distributed optimization algorithms for distributed SO-DD therefore remains an interesting direction for future research.

Beyond the development of other distributed optimization algorithms, several other interesting directions remain for future research. Building on the asymptotic normality established for DSGTD-GD, it would be valuable to develop statistical inference for distributed SO-DD. Consistent estimators of the asymptotic covariance matrix could be constructed to obtain confidence regions for the performative stable solution [8, 61]. Another promising direction is to develop zeroth-order methods for distributed SO-DD. Existing zeroth-order methods can solve centralized SO-DD without requiring an explicit structural model of the distribution map [20–22], and extending these methods to distributed settings may provide a useful approach for seeking stationary or optimal solutions under more general decision-dependent distributions. In addition, incorporating mini-batch strategies, which have been studied in distributed stochastic optimization [32, 34], into the proposed algorithms is expected to reduce the variance of the stochastic gradient and further improve the convergence performance of algorithms for distributed SO-DD. Finally, exploring distributed distributionally robust optimization with decision-dependent ambiguity sets is also a worthwhile research direction. Several studies [37, 44, 51] extend distributionally robust optimization to decision-dependent settings by allowing empirical distributions or radii of ambiguity set to depend on the decision variable and derive tractable reformulations. Extending these formulations to distributed settings could improve robustness against distribu-

tional uncertainty in distributed SO-DD. We leave these extensions for future research.

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.A. Alghunaim and K. Yuan. A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing*, 70:3264–3279, 2022.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] H.-F. Chen. *Stochastic approximation and its applications*, volume 64. Springer Science & Business Media, 2005.
- [8] X. Chen, J.D. Lee, X.-T. Tong, and Y.-C. Zhang. Statistical inference for model parameters in stochastic gradient descent. *The Annals of Statistics*, 48:251 – 273, 2020.
- [9] 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.
- [10] K.L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.
- [11] 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.
- [12] 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.

- [13] 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.
- [14] D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.
- [15] M. Duflo. *Random iterative models*, volume 34. Springer Science & Business Media, 2013.
- [16] J. Dupacová. Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*, 2006.
- [17] V. Fabian. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics*, pages 1327–1332, 1968.
- [18] 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.
- [19] L. Hellemo, P.I. Barton, and A. Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3):369–395, 2018.
- [20] Y. Hikima, H. Sawada, and A. Fujino. Guided zeroth-order methods for stochastic non-convex problems with decision-dependent distributions. In *Forty-second International Conference on Machine Learning*, 2025.
- [21] Y. Hikima and A. Takeda. Zeroth-order gradient estimators for stochastic problems with decision-dependent distributions. *arXiv preprint arXiv:2510.24929*, 2025.
- [22] Y. Hikima and A. Takeda. Zeroth-order methods for nonconvex stochastic problems with decision-dependent distributions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):17195–17203, 2025.
- [23] M. Hopkins, E. Reeber, G. Forman, and J. Suermondt. Spambase. UCI Machine Learning Repository, 1999. DOI: <https://doi.org/10.24432/C53G6X>.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] A. Koloskova, T. Lin, and S.U. Stich. An improved analysis of gradient tracking for decentralized machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 11422–11435, 2021.

- [28] 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.
- [29] 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.
- [30] 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.
- [31] X.-R. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [32] X.-R. Lian, W. Zhang, C. Zhang, and J. Liu. Asynchronous decentralized parallel stochastic gradient descent. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3043–3052. PMLR, 2018.
- [33] H.-D. Lim and D. Lee. Finite-time analysis of asynchronous q-learning under diminishing step-size from control-theoretic view. *IEEE Access*, 12:149916–149939, 2024.
- [34] Z.-W. Lin, J.-F. Xia, Q. Deng, and L. Luo. Decentralized gradient-free methods for stochastic non-smooth non-convex optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17477–17486, 2024.
- [35] 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.
- [36] 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.
- [37] F.-Q. Luo and S. Mehrotra. Distributionally robust optimization with decision dependent ambiguity sets. *Optimization Letters*, 14(8):2565–2594, 2020.
- [38] C. Mendler-Dünnér, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.
- [39] 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.
- [40] A. Mokhtari and A. Ribeiro. Dsa: Decentralized double stochastic averaging gradient algorithm. *Journal of Machine Learning Research*, 17(61):1–35, 2016.
- [41] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

- [42] 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.
- [43] 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.
- [44] N. Noyan, G. Rudolf, and M. Lejeune. Distributionally robust optimization under a decision-dependent ambiguity set with applications to machine scheduling and humanitarian logistics. *INFORMS Journal on Computing*, 34(2):729–751, 2022.
- [45] Y. Pang and G. Hu. Nash equilibrium seeking in n-coalition games via a gradient-free method. *Automatica*, 136:110013, 2022.
- [46] 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.
- [47] 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.
- [48] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.
- [49] S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.
- [50] S. Pu, A. Olshevsky, and I.C. Paschalidis. A sharp estimate on the transient time of distributed stochastic gradient descent. *IEEE Transactions on Automatic Control*, 67(11):5900–5915, 2022.
- [51] C.-R. Qu, H.-W. Jia, and P.-C. You. Decision-dependent distributionally robust optimization with application to dynamic pricing. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pages 693–698, 2025.
- [52] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.
- [53] 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.
- [54] M. Schmidt, N. Le Rouxand, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.
- [55] W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.
- [56] 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.

- [57] X.-L. Wang, C.-Y. Yau, and H.-T. Wai. Network effects in performative prediction games. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 36514–36540. PMLR, 2023.
- [58] 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.
- [59] 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.
- [60] J.-M. Xu, Y. Tian, Y. Sun, and G. Scutari. Distributed algorithms for composite optimization: Unified framework and convergence analysis. *IEEE Transactions on Signal Processing*, 69:3555–3570, 2021.
- [61] S.-C. Zhao and Y.-C. Liu. Confidence region for distributed stochastic optimization problem via stochastic gradient tracking method. *Automatica*, 159:111352, 2024.

Appendix

A Preliminary Results and Technical Lemmas

In this appendix, we provide some classic results on stochastic approximation and technical lemmas.

Lemma 3. [11, Theorem 26] *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with a non-decreasing sequence of σ -algebras $(\mathcal{F}_k)$. Let $\{x_k\}$, $\{\xi_k\}$, and $\{\zeta_k\}$ be sequences of random vectors in $\mathbb{R}^d$, and let $\{\eta_k\}$ be a deterministic sequence of positive step sizes. Consider the recursion*

$$x_{k+1} = x_k - \eta_k (R(x_k) + \xi_k + \zeta_k), \quad (67)$$

where $R: \mathbb{R}^d \rightarrow \mathbb{R}^d$ satisfies $R(x^*) = 0$. Suppose that the following conditions hold:

- (a) x_k is $\mathcal{F}_k$ -measurable, while ξ_k and ζ_k are $\mathcal{F}_{k+1}$ -measurable. Moreover, $\{\xi_k\}$ is a martingale difference sequence satisfying $\mathbf{E}[\xi_k \mid \mathcal{F}_k] = 0$.
- (b) The noise sequence $\{\xi_k\}$ is L^2 -bounded, i.e., $\sup_{k \geq 0} \mathbf{E} \|\xi_k\|^2 < \infty$.
- (c) There exists a deterministic positive semidefinite matrix Σ such that

$$\frac{1}{k} \sum_{t=0}^{k-1} \mathbf{E} [\xi_t \xi_t^\top \mid \mathcal{F}_t] \xrightarrow{p} \Sigma.$$

Moreover, for all $\varepsilon > 0$,

$$\frac{1}{k} \sum_{t=0}^{k-1} \mathbf{E} [\|\xi_t\|^2 \mathbf{1}_{\{\|\xi_t\| \geq \varepsilon \sqrt{k}\}} \mid \mathcal{F}_t] \xrightarrow{p} 0.$$

- (d) The residual sequence $\{\zeta_k\}$ satisfies $\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \|\zeta_t\| \xrightarrow{p} 0$.
- (e) There exists a positively stable matrix $G \in \mathbb{R}^{d \times d}$, and the step-size sequence satisfies

$$\eta_k = o(1), \quad \frac{\eta_k - \eta_{k+1}}{\eta_k} = o(\eta_k).$$

- (f) The sequence $\{x_t\}$ satisfies the linear approximation condition

$$\frac{1}{\sqrt{k}} \sum_{t=0}^{k-1} \|R(x_t) - G(x_t - x^*)\| \xrightarrow{p} 0.$$

Then, we have

$$\sqrt{k}(\bar{x}_k - x^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, G^{-1} \Sigma G^{-\top}),$$

where $\bar{x}_k := \frac{1}{k} \sum_{t=1}^k x_t$.

Lemma 4. [49, Lemma 5] Let $S = [s_{ij}] \in \mathbb{R}^{3 \times 3}$ be a non-negative, irreducible matrix with $s_{ii} < \lambda^*$ for some $\lambda^* > 0$ for all $i = 1, 2, 3$. Then $\rho(S) < \lambda^*$ iff $\det(\lambda^* \mathbf{I} - S) > 0$.

Lemma 5. [15, 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

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

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 $A_k \rightarrow A_\infty$ 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 $\mathbf{W} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top$, respectively. Then, for $\forall k \geq 0$, $\|k(\mathbf{W} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top)^k\|^2 \leq \frac{1}{(1-\rho)^2}$.

Proof. By the definition of ρ , we have

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

where $\rho \in (0, 1)$ is due to [49, Lemma 1]. The proof is complete. $\square$

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

Proof. Let $\tilde{\mathbf{W}} := \mathbf{W} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top$. By the triangle inequality and the submultiplicativity of the spectral norm, we have

$$\|(k+1)\tilde{\mathbf{W}}^{k+1} - k\tilde{\mathbf{W}}^k\|_2 \leq (k+1)\|\tilde{\mathbf{W}}^{k+1}\|_2 + k\|\tilde{\mathbf{W}}^k\|_2 \leq (k+1)\rho^{k+1} + k\rho^k \leq \frac{2}{1-\rho},$$

where the last inequality follows from Lemma 6. Therefore, $\|(k+1)\tilde{\mathbf{W}}^{k+1} - k\tilde{\mathbf{W}}^k\|_2^2 \leq \frac{4}{(1-\rho)^2}$. The proof is complete. $\square$

Lemma 8. For $\tau = \frac{2}{1-\rho} \log\left(\frac{50}{1-\rho} \left(1 + \log \frac{1}{1-\rho}\right)\right) + 1$, we have

$$\|J^t\|_2^2 \leq \frac{1}{2}, \quad \tau \leq t \leq 2\tau,$$

where

$$\tilde{\mathbf{W}} := \mathbf{W} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top, \quad J := \begin{bmatrix} \tilde{\mathbf{W}} & 0 \\ -\tilde{\mathbf{W}} & \tilde{\mathbf{W}} \end{bmatrix},$$

and $\rho = \|\tilde{\mathbf{W}}\|_2 \in (0, 1)$.

Proof. For any integer $t \geq 1$, direct calculation gives

$$J^t = \begin{bmatrix} \tilde{\mathbf{W}}^t & 0 \\ -t\tilde{\mathbf{W}}^t & \tilde{\mathbf{W}}^t \end{bmatrix}.$$

Since $\|\tilde{\mathbf{W}}^t\|_2 \leq \|\tilde{\mathbf{W}}\|_2^t = \rho^t$, we have $\|J^t\|_2 \leq (t+1)\rho^t$. The definition of τ guarantees that $(t+1)^2 \rho^{2t} \leq 1/2$ for $\tau \leq t \leq 2\tau$. Hence, we have $\|J^t\|_2^2 \leq \frac{1}{2}$. The proof is complete. $\square$