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

Licheng Deng, Yan Gao, Yongchao Liu

***Abstract*—This paper proposes a distributed asynchronous adaptive gradient tracking method, DASYAGT, to solve the distributed stochastic optimization problems with decision-dependent distributions over directed graphs. DASYAGT employs the local adaptive gradient to estimate the gradient of the objective function and introduces the auxiliary running-sum variable to handle asynchrony. We show that the iterates generated by DASYAGT converge, in expectation, to a stationary solution with a rate of $\mathcal{O}\left(\frac{\ln K}{\sqrt{K}}\right)$. The effectiveness of DASYAGT is further demonstrated numerically with synthetic and real-world data.**

***Index Terms*—Stochastic optimization with decision-dependent distributions, adaptive gradient tracking method, asynchrony, directed graphs.**

I. Introduction

DISTRIBUTED stochastic optimization problems have been widely studied in recent years due to its applications in large-scale machine learning [1, 2], sensor networks [3, 4] and parameter estimation [5, 6]. In many real-world applications, it may be the case that the distributions of stochastic elements depend on or shift in reaction to decision variables. For example, demand depends on price [7, 8], traffic predictions from navigation systems for route planning influence traffic patterns [9, 10], and predictions of credit default risk influence interest rate assignments and hence default rates [11, 12]. The corresponding distributed stochastic optimization problems with decision-dependent distributions (distributed SO-DD) [13] can be formulated as follows:

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

where n is the number of agents, $\mathbf{E}_{\xi_i\sim\mathcal{D}_i(x)}\left[\ell_i(x;\xi_i)\right]$ is the local objective function of the i-th agent and $\mathcal{D}(\cdot):\mathbb{R}^d\to\mathcal{P}(\mathbb{R}^m)$ is a distribution map.

SO-DD can be traced to early works [14–16]. More recently, Perdomo et al. [10] introduce the notion of performative stable point, which has motivated the research on SO-DD such as models [17–19] and algorithms [20–22]. We refer the interested readers to the surveys [23, 24] for more developments on SO-DD. To address substantial computational demands arising from large datasets and provide privacy guarantees, people consider the multi-agent stochastic optimization problems with decision-dependent distributions [13, 25, 26]. Li et al. [13] propose a distributed stochastic gradient descent algorithm and show that the proposed algorithm achieves a convergence rate of $\mathcal{O}(\frac{1}{K})$ to the performative stable solution of distributed SO-DD. Narang et al. [25] study a decision-dependent game, where the distribution map of each player is a global linear structure with respect to the decision variable. They propose an adaptive gradient algorithm that adaptively estimates the parametric description of the distribution map and uses the current estimate of the parameters to compute an approximate stochastic gradient. They show that the proposed algorithm achieves a convergence rate of $\mathcal{O}(\frac{1}{K})$ to the Nash equilibrium of the decision-dependent game. Motivated by [13, 25], Deng and Liu [26] propose the distributed stochastic gradient tracking algorithm and the distributed adaptive gradient tracking algorithm to seek the performative stable solution and the optimal solution of distributed SO-DD, respectively. For the performative stable solution, they show that the proposed algorithm achieves the convergence rate of $\mathcal{O}(\frac{1}{K})$. For the optimal solution, they show that the proposed algorithm achieves the convergence rate of $\mathcal{O}(\frac{\ln K}{\sqrt{K}})$.

In the multi-agent optimization problems, the distributed synchronous algorithms may require waiting for the slowest agent to complete its task before all agents can proceed to the next one. Therefore, the asynchronous algorithms [27–31] are extensively studied to address such case. Recently, Tian et al. [32] propose the asynchronous SONATA, which combines the gradient tracking mechanism with the push-sum strategy in the asynchronous setting. They show that, over strongly connected directed graphs, the proposed algorithm achieves a linear convergence rate and a convergence rate of $\mathcal{O}(\frac{1}{K})$ for the strongly convex and the non-convex objective function, respectively. Subsequently, Kungurtsev et al. [33] propose a stochastic version of asynchronous SONATA and show that, over strongly connected directed graphs, the proposed algorithm achieves a convergence rate of $o(\frac{1}{K^a})$, where $a\in(0,\frac{1}{2})$, for non-convex objective function. Moreover, Zhu et al. [34] study the convergence for the stochastic version of asynchronous SONATA on the more general directed graphs and show that it achieves linear convergence rate for strongly convex objective function and converges to a stationary point with a rate of $\mathcal{O}(\frac{1}{\sqrt{K}})$ for non-convex objective function. We refer the interested readers to the surveys [35, 36] on more distributed asynchronous algorithms.

Indeed, as stated in [25, Remark 12], the agents for solving the multi-agent stochastic optimization problems with

*This work was supported by National Key R&D Program of China No. 2023YFA1009200, the NSFC #12471283 and Fundamental Research Funds for the Central Universities DUT24LK001. (Corresponding author: Yongchao Liu)

The authors are with the School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China (e-mail: denglicheng@mail.dlut.edu.cn; gydllg123@mail.dlut.edu.cn; lyc@dlut.edu.cn).

decision-dependent distributions may have to learn the decisions of other agents in practice, resulting in that the agents may observe local data or the decisions of others asynchronously. This motivates us to propose a distributed asynchronous gradient tracking-based algorithm to seek the optimal solution of non-convex distributed SO-DD over the directed graph. As far as we are concerned, the contribution of the paper can be summarized as follows.

- We provide a distributed asynchronous adaptive gradient tracking method (DASYAGT), which generalizes the distributed gradient tracking-based algorithms in [26, 34]. Compared to the stochastic version of asynchronous SONATA in [34], DASYAGT employs auxiliary tracking variables to track the average adaptive gradient. Compared to the distributed adaptive gradient tracking algorithm in [26], DASYAGT introduces the auxiliary running-sum variables to handle asynchrony.
- We show that the iterates generated by DASYAGT converge to a stationary point at a rate of $\mathcal{O}\left(\frac{\ln K}{\sqrt{K}}\right)$, which differs from the rate of $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ in [34] by a logarithmic factor. The underlying reason is that DASYAGT needs to learn the parameter of the distribution map. To the best of our knowledge, the convergence of DASYAGT seems to be the first rigorous result on the convergence of the distributed asynchronous algorithm for solving the distributed SO-DD. The effectiveness of DASYAGT is further demonstrated numerically with synthetic and real-world data.

The rest of this paper is organized as follows. Section II introduces DASYAGT and presents some standard assumptions. Section III studies the convergence of DASYAGT. Numerical experiments are provided in Section IV. In Section V, we provide concluding remarks. Moreover, the proof can be found in the Appendix.

Throughout this paper, vectors default to columns if not otherwise specified. $\mathbb{R}^d$ denotes the d-dimension Euclidean space endowed with norm $\|x\| = \sqrt{\langle x, x\rangle}$. Denote $\mathbf{1} := (1\,1\ldots1)^\intercal \in \mathbb{R}^d$ and $\mathbf{0} := (0\,0\ldots0)^\intercal \in \mathbb{R}^d$. $\mathbf{I} \in \mathbb{R}^{n\times n}$ stands for the identity matrix. The inner product of two matrices A, B is denoted by $\langle A, B\rangle$. Denote $F(x) := \sum_{i=1}^n f_i(x)$, where $f_i(x) := \mathbf{E}_{\xi_i\sim\mathcal{D}_i(x)}\left[\ell_i(x;\xi_i)\right]$. To clarify the expression, we denote $\nabla\ell_i(x;\xi)$ as the gradient taken with respect to the first argument x, $\nabla_\xi\ell_i(x;\xi)$ as the gradient taken with respect to the second argument ξ. We declare that given the sequence $\{G_t\}_{t=s}^k$ with $k \geq s$, $G_{k:s} := G_kG_{k-1}\cdots G_{s+1}G_s$ if $k \geq s$ and $G_{k:s} := G_s$ otherwise. In addition, we consider a set of agents $\mathcal{V} = \{1, \cdots, n\}$ connected on a communication network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$ represents links or edges among the agents. If $(i, j) \in \mathcal{E}$ it means that i-th agent can send information to j-th agent. We use $\mathcal{N}_i^{\text{in}} \triangleq \{j \in \mathcal{V} \mid (j, i) \in \mathcal{E}\}$ and $\mathcal{N}_i^{\text{out}} \triangleq \{j \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}$ to denote the sets of in-neighbors and out-neighbors that i-th agent can communicate with.

II. DASYAGT

In this section, we first present DASYAGT for seeking the optimal solution of the non-convex distributed SO-DD (1), which reads as follows.

Algorithm 1 Distributed ASYnchronous Adaptive Gradient Tracking (DASYAGT):

Require: 1. For any $i \in \mathcal{V}$, initial values $x_{i,0} \in \mathbb{R}^d$, distribution $\hat{\mathcal{D}}_{i,0}(\cdot)$, $\xi_{i,0} \sim \mathcal{D}_i(x_{i,0})$, $v_{i,0} = \tilde{\rho}_{ij,0} = 0$; $\tau_{i_kj,-1} = -D$, $\rho_{ij,l} = v_{i,l} = 0$, $\forall l \in \{-D, -D+1, \cdots, 0\}$; $z_{i,0} = g_{i,0} = G(x_{i,0}, \xi_{i,0}, \hat{\mathcal{D}}_{i,0}(\cdot))$; step sizes $\gamma_k, \nu_k > 0$; weight matrices $\mathbf{W} := [w_{ij}]_{n\times n}$ and $\mathbf{M} := [m_{ij}]_{n\times n}$.
2. Preprocessing the distribution map for any $i \in \mathcal{V}$:

$$\hat{\mathcal{D}}_{i,1}(\cdot) = \pi\left(\hat{\mathcal{D}}_{i,0}(\cdot), x_{i,0}, \xi_{i,0}\right).$$

1: **for** $k = 0, 1, 2, \cdots$ **do**
2: Agent i_k wakes up: pick the delay $d_{j,k}$ and set $\tau_{i_kj,k} = \max\{\tau_{i_kj,k-1}, k - d_{j,k}\}$, $\forall j \in \mathcal{N}_{i_k}^{in}$.
3: Perform local descent: $v_{i_k,k+1} = x_{i_k,k} - \gamma_k z_{i_k,k}$.
4: Update decision:

$$x_{i_k,k+1} = w_{i_ki_k}v_{i_k,k+1} + \sum_{j\in\mathcal{N}_{i_k}^{in}} w_{i_kj}v_{j,\tau_{i_kj,k}}.$$

5: Draw samples: $\xi_{i_k,k+1} \sim \mathcal{D}_i(x_{i_k,k+1})$.
6: Compute local adaptive gradient:

$$g_{i_k,k+1} = G(x_{i_k,k+1}, \xi_{i_k,k+1}, \hat{\mathcal{D}}_{i_k,k+1}(\cdot)).$$

7: Update gradient tracking:

$$z_{i_k,k+\frac{1}{2}} = z_{i_k,k} + \sum_{j\in\mathcal{N}_{i_k}^{in}}\left(\rho_{i_kj,\tau_{i_kj,k}} - \tilde{\rho}_{i_kj,k}\right) + g_{i_k,k+1} - g_{i_k,k}.$$

8: Process messages: $z_{i_k,k+1} = m_{i_ki_k}z_{i_k,k+\frac{1}{2}}$, $\rho_{ji_k,k+1} = \rho_{ji_k,k} + m_{ji_k}z_{i_k,k+\frac{1}{2}}$, $\forall j \in \mathcal{N}_{i_k}^{out}$ (Send $\rho_{ji_k,k+1}$ to every $j \in \mathcal{N}_{i_k}^{out}$).
9: Update buffer: $\tilde{\rho}_{i_kj,k+1} = \rho_{i_kj,\tau_{i_kj,k}}$, $\forall j \in \mathcal{N}_{i_k}^{in}$.
10: Update the distribution map:

$$\hat{\mathcal{D}}_{i_k,k+2}(\cdot) = \pi\left(\hat{\mathcal{D}}_{i_k,k+1}(\cdot), x_{i_k,k+1}, \xi_{i_k,k+1}\right).$$

11: Untouched state variables shift to state $k+1$ while keeping the same value.
12: **end for**

In Algorithm 1, at each iteration k, only the i_k-th agent wakes up and performs: (i) Local computations. With $z_{i_k,k}$ being a proxy to the global adaptive gradient, the i_k-th agent first performs an approximate stochastic gradient descent on $x_{i_k,k}$ and generates the intermediate result $v_{i_k,k+1}$ at Step 3. (ii) Local communication for consensus and calculation of local adaptive gradient. The i_k-th agent performs a consensus step on the x-variables with possibly outdated information $v_{j,\tau_{i_kj,k}}$ from their in-neighbors at Step 4, and then calculates the local adaptive gradient $G(x_{i_k,k+1}, \xi_{i_k,k+1}, \hat{\mathcal{D}}_{i_k,k+1}(\cdot))$ at Step 6, where $\hat{\mathcal{D}}_{i_k,k}(\cdot)$ represents an estimate of $\mathcal{D}_i(\cdot)$ and $\xi_{i_k,k+1} \sim \mathcal{D}_i(x_{i_k,k+1})$. (iii) Local communication for gradient-tracking. The i_k-th agent forms the local estimate $z_{i_k,k+\frac{1}{2}}$ based on the current cumulative mass variables $\rho_{i_kj,\tau_{i_kj,k}}$ and buffer variables $\tilde{\rho}_{i_kj,k}$ from its in-neighbors, and local adaptive gradient-difference term $g_{i_k,k+1} - g_{i_k,k}$ at Step 7, where the local buffer $\tilde{\rho}_{i_kj,k}$ stores the value of ρ_{i_kj} that the i_k-th agent

used in its last update. (iv) Update the buffer variables. After transmitting information to its out-neighbors at Step 8, the i_k-th agent updates the buffer variable $\tilde{\rho}_{i_kj}$ to match the most recently consumed ρ_{i_kj} variable at Step 9. (v) Update the distribution map. Based on $x_{i_k,k+1}$ and $\hat{\mathcal{D}}_{i_k,k+1}(\cdot)$, the i_k-th agent updates the distribution map $\hat{\mathcal{D}}_{i_k,k+2}(\cdot)$ at Step 10.

In what follows, we make following conditions on the communication network, the weight matrix, asynchronous model, the objection function and the distribution map to guarantee the convergence of Algorithm 1.

Assumption 1 (Weight matrices and network). Let $\mathcal{G}(W)$ and $\mathcal{G}(M^\intercal)$ be graphs induced by matrixs $\mathbf{W}$ and $\mathbf{M}^\intercal$, respectively. Suppose that

(i) $\mathbf{W}$ is row-stochastic and $\mathbf{A}$ is column-stochastic, i.e., $\mathbf{W1} = \mathbf{1}$ and $\mathbf{1}^\intercal\mathbf{M} = \mathbf{1}^\intercal$. In addition, the weight matrices satisfy: there exists $g > 0$ such that $w_{ii} \geq g$ and $m_{ii} \geq g$, for $\forall i \in \mathcal{V}$; and $w_{ij} \geq g$ and $m_{ij} \geq g$, for $\forall (j,i) \in \mathcal{E}$; $w_{ij} = 0$ and $m_{ij} = 0$, otherwise;

(ii) the graphs $\mathcal{G}(W)$ and $\mathcal{G}(M^\intercal)$ each contain at least one spanning tree. Moreover, there exists at least one node that is a root of spanning trees for both $\mathcal{G}(W)$ and $\mathcal{G}(M^\intercal)$, i.e., $\mathcal{R} := \mathbf{R}_W \cup \mathbf{R}_{M^\intercal} \neq \emptyset$, where $\mathbf{R}_W$ (resp. $\mathbf{R}_{M^\intercal}$) denotes the set of roots of all possible spanning trees in the graph $\mathcal{G}(W)$ (resp. $\mathcal{G}(M^\intercal)$). Define $r := |\mathcal{R}|$.

Assumption 1 (i) is a standard condition on the directed network and the matrix [32–34, 37]. As commented in [37], Assumption 1 (ii) is weaker than requiring strong connectivity for both $\mathcal{G}(W)$ and $\mathcal{G}(M^\intercal)$. Assumption 1 (ii) allows flexible design of the underlying network topology, including popular structures such as Ring and Gossip structures. Building on consensus with non-doubly stochastic matrices, the AB/push–pull method has been applied on reinforcement learning [38] and economic dispatch problems [39].

Assumption 2 (Asynchronous model). Suppose that

(i) there exists $T \geq n$ such that $\cup_{t=k}^{k+T-1} i_t = \mathcal{V}$, for $\forall k \in \mathbb{N}$;

(ii) there exists $D \in \mathbb{N}$ such that $0 \leq d_{j,k} \leq D$, for $\forall j \in \mathcal{N}_{i_k}^{in}$ and $\forall k \in \mathbb{N}$.

Assumption 2 has been well used in the distributed asynchronous optimization [32–34] and excludes scenarios where some agents remain inactive indefinitely and some communication links fail for infinitely long time.

Assumption 3 (Objective function and gradient). For any $i \in \mathcal{V}$,

(i) there exists a positive constant L such that

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

(ii) there exists $\delta > 0$ such that

$$\mathbf{E}_{\xi\sim\mathcal{D}_i(x)}\left[\|\nabla \ell_i(x;\xi)\|\right] \leq \delta,\ \forall x \in \mathbb{R}^d;$$

(iii) there exists $\sigma > 0$ such that

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

In Assumption 3, condition (i) requires the objective functions to be smooth, while conditions (ii) and (iii) bound the norm of the stochastic gradient and its variance, respectively.

Assumption 4 (Distribution map). There exists a probability measure $\mathcal{P}_i$ and $A_i \in \mathbb{R}^{p\times d}$ such that

$$\xi_i \sim \mathcal{D}_i(x) \Longleftrightarrow \xi_i = A_i x + \zeta_i,\ \zeta_i \sim \mathcal{P}_i,\ \forall i \in \mathcal{V}.$$

Under Assumption 4, the gradient of $f_i(\cdot)$ can be derived as

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

and the corresponding stochastic gradient of $f_i(\cdot)$ at $x_{i,k}$ can be derived as

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

where $\xi_{i,k} \sim \mathcal{D}_i(x_{i,k})$. In Algorithm 1, the learning step 10 provides an estimate of A_i, and then we have an adaptive gradient of $f_i(\cdot)$ at $x_{i,k}$,

$$G(x_{i,k},\xi_{i,k},\hat{\mathcal{D}}_{i,k}(\cdot)) = \nabla \ell_i(x_{i,k};\xi_{i,k}) + A_{i,k}^\intercal \nabla_\xi \ell_i(x_{i,k};\xi_{i,k}),$$

where $\xi_{i,k} \sim \mathcal{D}_i(x_{i,k})$ and $A_{i,k}$ may be updated dynamically with

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

where $\nu_k = \frac{2}{k+6d}$, $q_{i,k} \sim \mathcal{D}_i(x_{i,k}+u_{i,k})$ and $u_{i,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

III. Convergence of DASYAGT

In this section, we study the convergence of DASYAGT.

For ease of analysis, we first reduce the asynchronous agent system to a synchronous augmented one with no delays by adding virtual agents, which virtually store the value of delayed variables during transmission of information between adjacent agents, to the graph $\mathcal{G}(W)$ and $\mathcal{G}(M^\intercal)$. For the augmented graph of $\mathcal{G}(W)$, we add $D+1$ virtual agents for each agent i, denoted by $i[0]$, $i[1]$,$\cdots$,$i[D]$, to store delayed information $v_{i,k}$, $v_{i,k-1}$, $\cdots$,$v_{i,k-D}$. For the augmented graph of $\mathcal{G}(M^\intercal)$, we add $D+1$ virtual agents for each edge $(j,i) \in \mathcal{E}(M)$, denoted by $(j,i)^0$, $(j,i)^1$,$\cdots$,$(j,i)^D$, to store the delayed information $z_{(j,i)^0,k}$, $z_{(j,i)^1,k}$,$\cdots$, $z_{(j,i)^D,k}$. Then, we define the set of real and virtual agents as $\hat{\mathcal{V}} := \mathcal{V} \cup \{(j,i)^d | (j,i) \in \mathcal{E}(M), d = 0,1,\cdots,D\}$ and its cardinality as $S := |\hat{\mathcal{V}}| = n + (D+1)|\mathcal{E}(M)|$. Following from the above augmented system, we define the augmented matrix $\hat{\mathbf{W}}_k \in \mathbb{R}^{n(D+2)\times n(D+2)}$ and $\hat{\mathbf{M}}_k \in \mathbb{R}^{S\times S}$ as

$$\hat{W}_{rl,k} := \begin{cases} w_{i_ki_k}, & \text{if } r = l = i_k; \\ w_{i_kj}, & \text{if } r = i_k,\ l = j + (d_{j,k}+1)n; \\ 1, & \text{if } r = l \in \{1,2,\ldots,2n\} \setminus \{i_k, i_k+n\}; \\ 1, & \text{if } r \in \{2n+1, 2n+2, \ldots, (D+2)n\} \\ & \cup\{i_k+n\} \text{ and } l = r-n; \\ 0, & \text{otherwise,} \end{cases}$$

and $\hat{\mathbf{M}}_k := \mathbf{P}_k\mathbf{S}_k$, where

$$S_{hl,k} := \begin{cases} 1, & \text{if } l \in \{(j,i_k)^d \mid d_{j,k} \leq d \leq D\} \\ & \text{and } h = i_k; \\ 1, & \text{if } l \in \hat{\mathcal{V}} \setminus \{(j,i_k)^d \mid d_{j,k} \leq d \leq D\} \\ & \text{and } h = l; \\ 0, & \text{otherwise,} \end{cases}$$

and

$$P_{hl,k} := \begin{cases} m_{ji_k}, & \text{if } l = i_k \text{ and } h = (j,i_k)^0, j \in N_{i_k}^{\text{out}}(M); \\ m_{i_ki_k}, & \text{if } l = h = i_k; \\ 1, & \text{if } l = h \in \mathcal{V} \setminus i_k; \\ 1, & \text{if } l = (i,j)^d, h = (i,j)^{d+1}, (i,j) \in \mathcal{E}(M), \\ & 0 \leq d \leq D-1; \\ 1, & \text{if } l = h = (i,j)^D, (i,j) \in \mathcal{E}(M); \\ 0, & \text{otherwise.} \end{cases}$$

With the augmented matrix $\hat{\mathbf{W}}_k$ and $\hat{\mathbf{M}}_k$, we write Algorithm 1 of the augmented system in a compact form as

$$\mathbf{h}_{k+1} = \hat{\mathbf{W}}_k \left(\mathbf{h}_k - \gamma_k \mathbf{e}_{i_k} z_{i_k,k}^{\intercal}\right), \tag{2}$$

$$\hat{\mathbf{z}}_{k+1} = \hat{\mathbf{M}}_k \hat{\mathbf{z}}_k + \mathbf{P}_k \mathbf{e}_{i_k} (g_{i_k,k+1} - g_{i_k,k})^{\intercal}, \tag{3}$$

where

$$\begin{aligned} &g_{i,k} = \nabla \ell_i(x_{i,k}, \xi_{i,k}) + (A_{i,k})^{\intercal} \nabla_{\xi} \ell_i(x_{i,k}, \xi_{i,k}), \\ &\mathbf{h}_k := \begin{bmatrix} \mathbf{x}_k \\ \mathbf{v}_k \\ \vdots \\ \mathbf{v}_{k-D} \end{bmatrix}, \quad \hat{\mathbf{z}}_k := \begin{bmatrix} \mathbf{z}_k \\ \mathbf{z}_{\mathcal{E}(M)^0,k} \\ \vdots \\ \mathbf{z}_{\mathcal{E}(M)^D,k} \end{bmatrix}, \\ &\mathbf{z}_{\mathcal{E}(\mathbf{A})^d,k} := \begin{bmatrix} z_{1^d,k} \\ z_{2^d,k} \\ \vdots \\ z_{|\mathcal{E}(M)|^d,k} \end{bmatrix}, \ d = 0, \cdots, D, \end{aligned} \tag{4}$$

and $z_{s^d,k}$ denotes $z_{(j,i)^d,k}$ if (j,i) is the s-th edge of $\mathcal{E}(M)$.

In what follows, we present the asymptotic behaviors of $\hat{\mathbf{W}}_k$ and $\hat{\mathbf{M}}_k$ over spanning-tree graphs.

Lemma 1. *[34, Lemma 1] Suppose Assumptions 1 and 2 hold. We have for any $k \geq t \geq 0$, $\hat{\mathbf{W}}_k$ is row stochastic and there exists a sequence of stochastic vectors $\{\psi_k\}_{k\geq 0}$ such that*

$$\|\hat{\mathbf{W}}_{k:t} - \mathbf{1}\psi_t^{\intercal}\| \leq c_1 \kappa^{k-t},$$

where $\psi_{i,k} \geq \eta := g^{(2n-1)T+nD}$, for $\forall i \in \mathcal{R}_W$, $c_1 := \frac{2\sqrt{n(D+2)(1+\eta)}}{1-\eta}$, $\kappa := (1-\eta)^{\frac{1}{(2n-1)T+nD}} \in (0,1)$.

Lemma 2. *[34, Lemma 2] Suppose Assumptions 1 and 2 hold. We have for any $k \geq t \geq 0$, $\hat{\mathbf{M}}_k$ is column stochastic and there exists a sequence of stochastic vectors $\{\phi_k\}_{k\geq 0}$ such that for $\forall i, j \in \{1, \cdots, S\}$,*

$$|\left(\hat{\mathbf{M}}_{k:t}\right)_{i,j} - \phi_{i,k}| \leq c_2 \kappa^{k-t}, \quad \forall i, j \in \{1, \cdots, S\}$$

where $\phi_{i,k} \geq \eta$, for $\forall i \in \mathcal{R}_{M^{\intercal}}$, $c_2 := \frac{2(1+\eta^{-1})}{1-\eta}$.

Moreover, by Lemma 1, we have that

$$\mathbf{1}\psi_t^{\intercal} = \lim_{k\to\infty} \hat{\mathbf{W}}_{k:t} = \left(\lim_{k\to\infty} \hat{\mathbf{W}}_{k:t+1}\right) \hat{\mathbf{W}}_t = \mathbf{1}\psi_{t+1}^{\intercal} \hat{\mathbf{W}}_t,$$

and thus $\psi_t^{\intercal} = \psi_{t+1}^{\intercal} \hat{\mathbf{W}}_t$, for $\forall t \geq 0$. Then, by (2), we have

$$x_{k+1}^{\psi} = x_k^{\psi} - \gamma_k \psi_k^{\intercal} \mathbf{e}_{i_k} z_{i_k,k}^{\intercal}, \tag{5}$$

where $x_k^{\psi} := \psi_k^{\intercal} \mathbf{h}_k$.

Obviously, by Lemma 2 and the definition of $\mathbf{P}_k$, we have

$$\mathbf{1}^{\intercal} \hat{\mathbf{z}}_k = \left(\sum_{i=1}^{n} g_{i,k}\right)^{\intercal}. \tag{6}$$

In addition, we introduce an auxiliary sequence $\{z'_{i,k}\}_{i\in\mathcal{V}}$, initialized as $z'_{i,0} = \mathbf{E}[g_{i,0}]$ for $i \in \mathcal{V}$, which resembles the recursion of tracking variable $\{z_{i,k}\}_{i\in\mathcal{V}}$. We denote $\mathbf{z}'_k$ as the augmented auxiliary variable corresponding to the tracking variables $\hat{\mathbf{z}}_k$. Then, similar to (3), the recursion of the auxiliary variable $\mathbf{z}'_k$ can be rewritten in a compact form as follows:

$$\mathbf{z}'_{k+1} = \hat{\mathbf{M}}_k \mathbf{z}'_k + \mathbf{P}_k \mathbf{e}_{i_k} (\hat{g}_{i_k,k+1} - \hat{g}_{i_k,k})^{\intercal}, \tag{7}$$

where $\hat{g}_{i,k} = \mathbf{E}[g_{i,k}]$ and $g_{i,k}$ is defined in (4). Again, by the column stochasticity of $\hat{\mathbf{M}}_k$ and $\mathbf{P}_k$, we have

$$\mathbf{1}^{\intercal} \mathbf{z}'_k = \left(\sum_{i=1}^{n} \hat{g}_{i,k}\right)^{\intercal}. \tag{8}$$

We first provide a technical result that plays a key role in analyzing the convergence of DASYAGT.

Proposition 1. *Suppose that Assumptions 1 - 4 hold. Then (i):*

$$\begin{aligned} &\mathbf{E}\left[\|\mathbf{h}_{k+1} - \mathbf{1}x_{k+1}^{\psi}\|^2\right] \\ \leq& 2c_1^2 \kappa^{2k} \|\mathbf{h}_0 - \mathbf{1}x_0^{\psi}\|^2 + \frac{4c_1^2}{1-\kappa} \sum_{l=0}^{k} \kappa^{k-l} \gamma_l^2 \mathbf{E}\left[\|z'_{i_l,l}\|^2\right] \\ &+ \frac{8nc_1^2\sigma^2}{1-\kappa} \sum_{l=0}^{k} \kappa^{k-l} \gamma_l^2 \mathbf{E}\left[\|A_{l-\tau_l} - A\|^2\right] \\ &+ \frac{4nc_1^2\sigma^2(1+2\|A\|^2)}{1-\kappa} \sum_{l=0}^{k} \kappa^{k-l} \gamma_l^2. \end{aligned} \tag{9}$$

(ii):

$$\begin{aligned} &\mathbf{E}\left[\|z'_{i_{k+1},k+1} - \phi_{i_{k+1},k} \mathbf{1}^{\intercal} \mathbf{z}'_{k+1}\|^2\right] \\ \leq& 4Sc_2^2 \kappa^{2k} \|\mathbf{z}'_0\|^2 + \frac{72Sc_2^2L^2}{(1-\kappa)\kappa^2} \sum_{l=0}^{k} \kappa^{k-l} \mathbf{E}\left[\|\mathbf{h}_l - \mathbf{1}x_l^{\psi}\|^2\right] \\ &+ \frac{72Sc_2^2L^2}{(1-\kappa)\kappa^2} \sum_{l=0}^{k} \kappa^{k-l} \gamma_l^2 \mathbf{E}\left[\|z'_{i_l,l}\|^2\right] \\ &+ \frac{144nSc_2^2\sigma^2L^2}{(1-\kappa)\kappa^2} \sum_{l=0}^{k} \kappa^{k-l} \gamma_l^2 \mathbf{E}\left[\|A_{l-\tau_l} - A\|^2\right] \\ &+ \frac{32Sc_2^2\delta^2}{(1-\kappa)\kappa^2} \sum_{l=0}^{k+1} \kappa^{k+1-l} \mathbf{E}\left[\|A_{l-\tau_l} - A\|^2\right] \\ &+ \frac{72nSc_2^2L^2\sigma^2(1+2\|A\|^2)}{(1-\kappa)\kappa^2} \sum_{l=0}^{k} \kappa^{k-l} \gamma_l^2, \end{aligned} \tag{10}$$

where $c_1 := \frac{2\sqrt{n(D+2)}(1+\eta)}{1-\eta}$, $c_2 := \frac{2(1+\eta^{-1})}{1-\eta}$, $\kappa := (1-\eta)^{\frac{1}{(2n-1)T+nD}}$, $\eta := g^{(2n-1)T+nD}$ and τ_k is the number of iterations agent i_k has skipped since it's last update.

Proof. Part (i):

Applying (2) recursively, we have

$$\mathbf{h}_{k+1} = \hat{\mathbf{W}}_{k:0} \mathbf{h}_0 - \sum_{l=0}^{k} \gamma_l \hat{\mathbf{W}}_{k:l} \mathbf{e}_{i_l} z_{i_l,l}^{\intercal}. \tag{11}$$

Then, by left multiplying $\psi_{k+1}^{\intercal}$ on both side of (11) and the fact that $\psi_t^{\intercal} = \psi_{t+1}^{\intercal} \hat{\mathbf{W}}_t$, we have

$$x_{k+1}^{\psi} = x_0^{\psi} - \sum_{l=0}^{k} \gamma_l \psi_l^{\intercal} \mathbf{e}_{i_l} z_{i_l,l}^{\intercal}, \tag{12}$$

where $x_k^{\psi} = \psi_k^{\intercal} \mathbf{h}_k$.

Using (11) and (12), we have

$$\begin{aligned} &\|\mathbf{h}_{k+1} - \mathbf{1}x_{k+1}^{\psi}\|^2 \\ \leq& 2c_1^2 \kappa^{2k} \|\mathbf{h}_0 - \mathbf{1}x_0^{\psi}\|^2 + \frac{2c_1^2}{1-\kappa} \sum_{l=0}^{k} \kappa^{k-l} \gamma_l^2 \|z_{i_l,l}\|^2, \end{aligned}$$

where the inequality follows from Lemma 1 and [34, Lemma 8]. Taking expectation on both sides of the above inequality, we have

$$\begin{aligned}&\mathbf{E}\left[\|\mathbf{h}_{k+1}-\mathbf{1}x_{k+1}^{\psi}\|^2\right]\\ \le&2c_1^2\kappa^{2k}\|\mathbf{h}_0-\mathbf{1}x_0^{\psi}\|^2+\frac{4c_1^2}{1-\kappa}\sum_{l=0}^{k}\kappa^{k-l}\gamma_l^2\mathbf{E}\left[\|z'_{i_l,l}\|^2\right]\\ &+\frac{4c_1^2}{1-\kappa}\sum_{l=0}^{k}\kappa^{k-l}\gamma_l^2\mathbf{E}\left[\|z_{i_l,l}-z'_{i_l,l}\|^2\right]. \end{aligned}\tag{13}$$

For the last term on the right-hand side of (13),

$$\begin{aligned}&\mathbf{E}\left[\|z_{i_k,k}-z'_{i_k,k}\|^2\right]\\ \le&\mathbf{E}\left[\|\mathbf{1}^{\intercal}\hat{\mathbf{z}}_k-\mathbf{1}^{\intercal}\mathbf{z}'_k\|^2\right]\\ =&\mathbf{E}\left[\|\sum_{i=1}^{n}g_{i,k}-\sum_{i=1}^{n}\hat{g}_{i,k}\|^2\right]\\ \le&n\sigma^2(1+2\|A\|^2)+2n\sigma^2\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right], \end{aligned}\tag{14}$$

where τ_k is the number of iterations agent i_k has skipped since it's last update, the equality follows from (6) and (8), the last inequality follows from Assumption 3 (ii) and (iii). Substituting (14) into (13), we arrive at (9).

Part (ii):

Applying (7) recursively, we have

$$\begin{aligned}\mathbf{z}'_{k+1}=&\hat{\mathbf{M}}_{k:0}\mathbf{z}'_0+\sum_{l=1}^{k}\hat{\mathbf{M}}_{k:l}\mathbf{P}_{l-1}\mathbf{e}_{i_{l-1}}(\hat{g}_{i_{l-1},l}-\hat{g}_{i_{l-1},l-1})^{\intercal}\\ &+\mathbf{P}_k\mathbf{e}_{i_k}(\hat{g}_{i_k,k+1}-\hat{g}_{i_k,k})^{\intercal}.\end{aligned}$$

Then

$$\begin{aligned}&\|z'_{i_{k+1},k+1}-\phi_{i_{k+1},k}\mathbf{1}^{\intercal}\mathbf{z}'_{k+1}\|\\ =&\|\left(\hat{\mathbf{M}}_{k:0}\right)_{i_k,:}\mathbf{z}'_0+P_{i_k,k}(\hat{g}_{i_k,k+1}-\hat{g}_{i_k,k})^{\intercal}-\phi_{i_{k+1},k}\mathbf{1}^{\intercal}\mathbf{z}'_0\\ &+\sum_{l=1}^{k}\left(\hat{\mathbf{M}}_{k:l}\right)_{i_k,:}\mathbf{P}_{l-1}\mathbf{e}_{i_{l-1}}(\hat{g}_{i_{l-1},l}-\hat{g}_{i_{l-1},l-1})^{\intercal}\\ &-\sum_{l=0}^{k}\phi_{i_{k+1},k}\mathbf{1}^{\intercal}\mathbf{P}_l\mathbf{e}_{i_l}(\hat{g}_{i_l,l+1}-\hat{g}_{i_l,l})^{\intercal}\|\\ \le&c_2\sqrt{S}\kappa^k\|\mathbf{z}'_0\|+\frac{c_2\sqrt{2S}}{\kappa}\sum_{l=0}^{k}\kappa^{k-l}\|\hat{g}_{i_l,l+1}-\hat{g}_{i_l,l}\|,\end{aligned}\tag{15}$$

where the inequality follows from Lemma 2.

For the last term on the right-hand side of (15),

$$\begin{aligned}&\|\hat{g}_{i_l,l+1}-\hat{g}_{i_l,l}\|\\ \le&\|\hat{g}_{i_l,l+1}-\nabla f_{i_l}(x_{i_l,l+1})\|+\|\hat{g}_{i_l,l}-\nabla f_{i_l}(x_{i_l,l})\|\\ &+\|\nabla f_{i_l}(x_{i_l,l+1})-\nabla f_{i_l}(x_{i_l,l})\|\\ \le&\delta\|A_{l+1-\tau_{l+1}}-A\|+\delta\|A_{l-\tau_l}-A\|+L\|x_{i_l,l+1}-x_{i_l,l}\|,\end{aligned}$$

where the last inequality is due to Assumption 3 (i) and (ii). Then

$$\begin{aligned}&\mathbf{E}\left[\|z'_{i_{k+1},k+1}-\phi_{i_{k+1},k}\mathbf{1}^{\intercal}\mathbf{z}'_{k+1}\|^2\right]\\ \le&4Sc_2^2\kappa^{2k}\|\mathbf{z}'_0\|^2+\frac{72Sc_2^2L^2}{(1-\kappa)\kappa^2}\sum_{l=0}^{k}\kappa^{k-l}\mathbf{E}\left[\|\mathbf{h}_l-\mathbf{1}x_l^{\psi}\|^2\right]\\ &+\frac{72Sc_2^2L^2}{(1-\kappa)\kappa^2}\sum_{l=0}^{k}\kappa^{k-l}\gamma_l^2\mathbf{E}\left[\|z_{i_l,l}\|^2\right]\\ &+\frac{32Sc_2^2\delta^2}{(1-\kappa)\kappa^2}\sum_{l=0}^{k+1}\kappa^{k+1-l}\mathbf{E}\left[\|A_{l-\tau_l}-A\|^2\right]\\ \le&4Sc_2^2\kappa^{2k}\|\mathbf{z}'_0\|^2+\frac{72Sc_2^2L^2}{(1-\kappa)\kappa^2}\sum_{l=0}^{k}\kappa^{k-l}\mathbf{E}\left[\|\mathbf{h}_l-\mathbf{1}x_l^{\psi}\|^2\right]\\ &+\frac{72Sc_2^2L^2}{(1-\kappa)\kappa^2}\sum_{l=0}^{k}\kappa^{k-l}\gamma_l^2\mathbf{E}\left[\|z'_{i_l,l}\|^2\right]\\ &+\frac{72Sc_2^2L^2}{(1-\kappa)\kappa^2}\sum_{l=0}^{k}\kappa^{k-l}\gamma_l^2\mathbf{E}\left[\|z_{i_l,l}-z'_{i_l,l}\|^2\right]\\ &+\frac{32Sc_2^2\delta^2}{(1-\kappa)\kappa^2}\sum_{l=0}^{k+1}\kappa^{k+1-l}\mathbf{E}\left[\|A_{l-\tau_l}-A\|^2\right],\end{aligned}\tag{16}$$

where the first inequality follows form [34, Lemma 8]. Substituting (14) into (16), we arrive at (10). The proof is complete. □

With Proposition 1, we provide the upper bounds for the accumulative consensus error $\sum_{l=0}^{k}\mathbf{E}\left[\|\mathbf{h}_l-\mathbf{1}x_l^{\psi}\|^2\right]$ and tracking error $\sum_{l=0}^{k}\mathbf{E}\left[\|z'_{i_l,l}-\phi_{i_l,l-1}\mathbf{1}^{\intercal}\mathbf{z}'_l\|^2\right]$ with $\sum_{l=0}^{k}\mathbf{E}\left[\|\nabla F(x_l^{\psi})\|^2\right]$.

Lemma 3. *Suppose that Assumptions 1 - 4 hold and the step size satisfies* $\gamma_k=\gamma<\frac{1}{\sqrt{4nc_3L^2+27c_4}}$, *where* $c_3:=\frac{4c_1^2}{(1-\kappa)^2}$ *and* $c_4:=\frac{4Sc_2^2L^2\left[4c_1^2+(1-\kappa)^2\right]}{\kappa^2(1-\kappa)^4}$. *Then*

(i):

$$\begin{aligned}&\sum_{l=0}^{k}\mathbf{E}\left[\|\mathbf{h}_l-\mathbf{1}x_l^{\psi}\|^2\right]\\ \le&\frac{4c_3\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\sum_{l=0}^{k}\mathbf{E}\left[\|\nabla F(x_l^{\psi})\|^2\right]\\ &+\frac{c_h\left(1-72c_4\gamma^2\right)}{1-(4nc_3L^2+27c_4)\gamma^2}+\frac{4c_zc_3\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\\ &+\left[\frac{128Sc_3c_2^2\delta^2\gamma^2}{\kappa^2\left(1-(4nc_3L^2+27c_4)\gamma^2\right)(1-\kappa)^2}\right.\\ &\left.+\frac{2nc_3\gamma^2\left[(\sigma^2+2\delta^2)-36c_4\sigma^2\gamma^2\right]}{1-(4nc_3L^2+27c_4)\gamma^2}\right]\sum_{l=0}^{k}\mathbf{E}\left[\|A_{l-\tau_l}-A\|^2\right]\\ &+\frac{nc_3\sigma^2\gamma^2(k+1)(1+2\|A\|^2)}{1-(4nc_3L^2+27c_4)\gamma^2}.\end{aligned}\tag{17}$$

(ii):

$$\begin{aligned}&\sum_{l=0}^{k}\mathbf{E}\left[\|z'_{i_l,l}-\phi_{i_l,l-1}\mathbf{1}^{\intercal}\mathbf{z}'_l\|^2\right]\\ \le&\frac{72c_4\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\sum_{l=0}^{k}\mathbf{E}\left[\|\nabla F(x_l^{\psi})\|^2\right]\\ &+\frac{c_z\left[1-4nc_3L^2\gamma^2\right]}{1-(4nc_3L^2+27c_4)\gamma^2}+\frac{72nc_hc_4L^2\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\\ &+\left[\frac{32Sc_2^2\delta^2\left[1-4nc_3L^2\gamma^2\right]}{\kappa^2\left(1-(4nc_3L^2+27c_4)\gamma^2\right)(1-\kappa)^2}\right.\\ &+\frac{18nc_4\gamma^2(\sigma^2+4\delta^2)}{1-(4nc_3L^2+27c_4)\gamma^2}\\ &\left.+\frac{72nc_3c_4\sigma^2L^2\gamma^4}{1-(4nc_3L^2+27c_4)\gamma^2}\right]\sum_{l=0}^{k}\mathbf{E}\left[\|A_{l-\tau_l}-A\|^2\right]\\ &+\frac{18nc_4\sigma^2\gamma^2(k+1)(1+2\|A\|^2)}{1-(4nc_3L^2+27c_4)\gamma^2},\end{aligned}\tag{18}$$

where $c_h := \left(1+\frac{2c_1^2}{1-\kappa^2}\right)\|\mathbf{h}_0 - \mathbf{1}x_0^\psi\|^2$ *and* $c_z := \|z'_{i_0,0} - \phi_{i_0,-1}\mathbf{1}^\intercal\mathbf{z}'_0\|^2 + \frac{4Sc_2^2\|\mathbf{z}'_0\|^2}{1-\kappa^2} + \frac{72Sc_2^2L^2}{(1-\kappa)^2\kappa^2}\left(1+\frac{2c_1^2}{1-\kappa^2}\right)\|\mathbf{h}_0 - \mathbf{1}x_0^\psi\|^2$.

Proof. See Appendix A for the detailed proof. □

Due to Assumption 1 (ii), the activation of non-root node cannot guarantee a sufficient descent towards stationary point, leading to that the vanilla descent lemma in [32, 33] no longer can be established at every global iteration. Therefore, we move to establish the following lemma based on two-time-scale techniques which provides an upper bound of $\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|\nabla F(x_k^\psi)\|^2\right]$.

Lemma 4. *Suppose that Assumptions 1 - 4 hold and the step size satisfy* $\gamma_k = \gamma \le \min\{\frac{2}{nL(2T^2+rT\eta^2)}, \frac{1}{8nL}\}$. *Then, for* $\forall k' \ge 1$, *we have*

$$\begin{aligned}
&\left(\frac{r\eta^2}{8T}-\gamma\right)\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|\nabla F(x_k^\psi)\|^2\right]\\
\le&\frac{F(x_0^\psi)-F^\star}{\gamma}+nL^2\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|\mathbf{h}_k-\mathbf{1}x_k^\psi\|^2\right]\\
&+\frac{1+\gamma}{2\gamma}\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|z'_{i_k,k}-\phi_{i_k,k-1}\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]+\left(rT\eta^2\sigma^2L^2n^3\gamma^2\right.\\
&\left.+n\delta^2+2L\gamma n^2\sigma^2+2\sigma^2T^2L^2n^3\gamma^2\right)\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right]\\
&+\frac{rk'\eta^2\sigma^2L^2T^2n^3\gamma^2(1+2\|A\|^2)}{2}+Lk'T\gamma n^2\sigma^2(1+2\|A\|^2)\\
&+k'\sigma^2L^2\gamma^2T^3n^3(1+2\|A\|^2),
\end{aligned}\tag{19}$$

where $F^\star := \min_{x\in\mathbb{R}^d}F(x)$.

Proof. See Appendix B for the detailed proof. □

With the above supporting lemmas, we are ready to study the convergence rate of DASYAGT.

Theorem 1. *Suppose that Assumptions 1 - 4 hold. For* $\forall K > 0$ *being a multiple of* T *defined in Assumption 2 (i), the step size* $\gamma_k = \gamma = \frac{1}{\sigma\sqrt{rK}+\bar\gamma^{-1}}$ *and* $\nu_k = \frac{2}{k+6dT}$, *we have*

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}\left[\|\nabla F(\bar x_k)\|^2\right]\\
\le&\frac{64T\sigma(F(x_0^\psi)-F^\star+c_z)+C_3\sigma\eta^2}{\eta^2\sqrt{rK}}\\
&+\frac{2048c_a\sigma Sc_2^2\delta^2}{\eta^2\kappa^2(c_q-1)(1-\kappa)^2\sqrt{rK}}+\frac{2048c_aSc_2^2\delta^2}{r\eta^2\bar\gamma\kappa^2(c_q-1)(1-\kappa)^2K}\\
&+\frac{64T\bar\gamma^{-1}(F(x_0^\psi)-F^\star+c_z)+C_1rT\eta^2+C_4\eta^2}{r\eta^2K}\\
&+\frac{2048STc_2^2\delta^2\bar\gamma^{-1}+C_2rT\eta^2\kappa^2(1-\kappa)^2}{rK\eta^2\kappa^2(1-\kappa)^2}\ln\left(K+(c_q-1)T\right)\\
&+\frac{2048STc_2^2\delta^2\sigma}{\eta^2\kappa^2(1-\kappa)^2\sqrt{rK}}\ln\left(K+(c_q-1)T\right),
\end{aligned}\tag{20}$$

where $\bar x_k := \frac{1}{n}\mathbf{1}^\intercal\mathbf{x}_k$, $F^\star := \min_{x\in\mathbb{R}^d}F(x)$,

$$\begin{aligned}
\bar\gamma :=& \min\left\{\frac{1+18c_4}{2nc_3L^2+18c_4}, \frac{r\eta^2}{32T(1+18c_4)}, \frac{2}{nL(2T^2+rT\eta^2)},\right.\\
&\left.\frac{1}{8nL}, \frac{1}{\sqrt{2(4nc_3L^2+27c_4)}}\right\},\\
c_z :=& \|z'_{i_0,0}-\phi_{i_0,-1}\mathbf{1}^\intercal\mathbf{z}'_0\|^2+\frac{4Sc_2^2\|\mathbf{z}'_0\|^2}{1-\kappa^2}\\
&+\frac{72Sc_2^2L^2}{(1-\kappa)^2\kappa^2}\left(1+\frac{2c_1^2}{1-\kappa^2}\right)\|\mathbf{h}_0-\mathbf{1}x_0^\psi\|^2,\\
c_3 :=& \frac{4c_1^2}{(1-\kappa)^2}, \qquad c_4 := \frac{4Sc_2^2L^2\left[4c_1^2+(1-\kappa)^2\right]}{\kappa^2(1-\kappa)^4},\\
c_a :=& \max\{(1+6d)\|A_0-A\|^2, 8d\sum_{i=1}^n tr(\Sigma_i)\}, \quad c_q := 6d,\\
C_1 :=& \frac{4nc_hL^2}{T}+\frac{c_zc_3}{4nT}+\frac{64(2c_hnL^2+9c_4c_hL+c_z)}{r\eta^2}\\
&+\frac{c_3c_a(\sigma^2+2\delta^2)}{8(c_q-1)}+\frac{nc_a\sigma^2}{c_q-1}+\frac{8(c_zc_3+9c_4c_h)}{nr\eta^2}\\
&+\frac{8c_ac_3Sc_2^2\delta^2}{n\kappa^2(c_q-1)(1-\kappa)^2}+\frac{c_a\sigma^2(4c_3+16n+2nT^2)}{Tr\eta^2(c_q-1)}\\
&+\frac{c_4c_a\sigma^2(144Ln^3+9nc_3L+18n^2+2c_3)}{Tr\eta^2L^2n^3(c_q-1)}\\
&+\left[\frac{576c_4}{L}+\frac{72c_4}{nL^2}+\frac{256c_3Sc_2^2}{n\kappa^2(1-\kappa)^2}+\frac{2048Sc_2^2}{\kappa^2(1-\kappa)^2}\right.\\
&\left.+16c_3+64n\right]\frac{c_a\delta^2}{Tr\eta^2(c_q-1)},\\
C_2 :=& \frac{8Sc_3c_2^2\delta^2}{nT\kappa^2(1-\kappa)^2}+\frac{c_3(\sigma^2+2\delta^2)}{8T}+nT\sigma^2\\
&+\left[\frac{144c_4}{L}+\frac{9c_3c_4}{Ln^2}+\frac{2c_3c_4}{L^2n^3}+\frac{18c_4}{nL^2}+2nT^2\right.\\
&\left.+4c_3+16n\right]\frac{\sigma^2}{r\eta^2}+\left[\frac{576c_4}{L}+\frac{72c_4}{nL^2}+\frac{256c_3Sc_2^2}{n\kappa^2(1-\kappa)^2}\right.\\
&\left.+\frac{2048Sc_2^2}{\kappa^2(1-\kappa)^2}+64n+8c_3\right]\frac{\delta^2}{r\eta^2},\\
C_3 :=& \frac{64nT(1+2\|A\|^2)(18c_4+Ln)}{r\eta^2},\\
C_4 :=& \frac{64nT(1+2\|A\|^2)(18c_4+2c_3nL^2+n^2T^2L^2)}{r\eta^2}\\
&+4n^2L^2(c_3+8nT^2)(1+2\|A\|^2),
\end{aligned}\tag{21}$$

and Σ_i *is the covariance matrix of* ζ_i *defined in Assumption 4,* c_1, c_2, κ, η *are defined in Proposition 1.*

Proof. For $\forall k' \ge 1$, we have

$$\begin{aligned}
&\frac{1}{k'T}\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|\nabla F(\bar x_k)\|^2\right]\\
\le&\frac{2}{k'T}\sum_{k=0}^{k'T-1}\left[\mathbf{E}\left[\|\nabla F(x_k^\psi)\|^2\right]+\mathbf{E}\left[\|\nabla F(\bar x_k)-\nabla F(x_k^\psi)\|^2\right]\right]\\
\le&\frac{2}{k'T}\sum_{k=0}^{k'T-1}\left[\mathbf{E}\left[\|\nabla F(x_k^\psi)\|^2\right]+n^2L^2\mathbf{E}\left[\|\bar x_k-x_k^\psi\|^2\right]\right]\\
\le&\frac{2}{k'T}\sum_{k=0}^{k'T-1}\left[\mathbf{E}\left[\|\nabla F(x_k^\psi)\|^2\right]+nL^2\mathbf{E}\left[\|\mathbf{h}_k-\mathbf{1}x_k^\psi\|^2\right]\right],
\end{aligned}\tag{22}$$

where the second inequality follows from Assumption 3 (i) and the last inequality follows from the fact that $\|\frac{\mathbf{1}\mathbf{1}^\intercal}{n}\| = 1$.

By substituting (17), (18) into (19) and rearranging the terms yields, we have

$$\left(\frac{r\eta^2}{8}-T\gamma-\frac{4nc_3TL^2\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\right.$$

$$
\begin{aligned}
&-\frac{36c_4T\gamma(1+\gamma)}{1-(4nc_3L^2+27c_4)\gamma^2}\Big)\frac{1}{k'T}\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]\\
\leq&\frac{1}{k'}\left[\frac{F(x_0^{\psi})-F^{\star}}{\gamma}+\frac{4nc_3c_zL^2\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\right.\\
&+\frac{c_hnL^2\left(1-72c_4\gamma^2\right)}{1-(4nc_3L^2+27c_4)\gamma^2}+\frac{36nc_hc_4\gamma L^2(1+\gamma)}{1-(4nc_3L^2+27c_4)\gamma^2}\\
&\left.+\frac{c_z(1+\gamma)\left(1-4nc_3L^2\gamma^2\right)}{2\gamma\left(1-(4nc_3L^2+27c_4)\gamma^2\right)}\right]\\
&+\left[\frac{9nc_4\gamma(1+\gamma)\left[(\sigma^2+4\delta^2)+4c_3\sigma^2L^2\gamma^2\right]}{1-(4nc_3L^2+27c_4)\gamma^2}\right.\\
&+\frac{128nSc_3c_2^2\delta^2L^2\gamma^2}{\kappa^2\left(1-(4nc_3L^2+27c_4)\gamma^2\right)(1-\kappa)^2}\\
&+\frac{2c_3n^2L^2\gamma^2\left[(\sigma^2+2\delta^2)-36c_4\sigma^2\gamma^2\right]}{1-(4nc_3L^2+27c_4)\gamma^2}\\
&+\frac{16Sc_2^2\delta^2(1+\gamma)\left[1-4nc_3L^2\gamma^2\right]}{\gamma\kappa^2\left(1-(4nc_3L^2+27c_4)\gamma^2\right)(1-\kappa)^2}\\
&+2L\gamma n^2\sigma^2+n\delta^2+2\sigma^2T^2L^2\gamma^2n^3\\
&\left.+rT\eta^2\sigma^2L^2\gamma^2n^3\right]\frac{1}{k'}\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right]\\
&+LT\gamma n^2\sigma^2(1+2\|A\|^2)+\sigma^2L^2\gamma^2T^3n^3(1+2\|A\|^2)\\
&+\frac{r\eta^2\sigma^2L^2T^2n^3\gamma^2(1+2\|A\|^2)}{2}+\frac{c_3Tn^2\sigma^2L^2\gamma^2(1+2\|A\|^2)}{1-(4nc_3L^2+27c_4)\gamma^2}\\
&+\frac{9nTc_4\gamma\sigma^2(1+\gamma)(1+2\|A\|^2)}{(1-(4nc_3L^2+27c_4)\gamma^2)}. \qquad (23)
\end{aligned}
$$

On the other hand, by the fact that $\gamma\leq\bar{\gamma}$, we have

$$
\begin{aligned}
&\frac{r\eta^2}{8}-T\gamma-\frac{4nc_3TL^2\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\\
&-\frac{36c_4T\gamma(1+\gamma)}{1-(4nc_3L^2+27c_4)\gamma^2}\geq\frac{r\eta^2}{16}. \qquad (24)
\end{aligned}
$$

Then, by (23) and (24), we can bound the first term on the right-hand side of (22), that is,

$$
\begin{aligned}
&\frac{1}{k'T}\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]\\
\leq&\frac{16}{k'r\eta^2}\left[\frac{F(x_0^{\psi})-F^{\star}}{\gamma}+\frac{4nc_zc_3L^2\gamma^2}{1-(4nc_3L^2+27c_4)\gamma^2}\right.\\
&+\frac{c_hnL^2\left(1-72c_4\gamma^2\right)}{1-(4nc_3L^2+27c_4)\gamma^2}+\frac{36nc_4c_h\gamma L^2(1+\gamma)}{1-(4nc_3L^2+27c_4)\gamma^2}\\
&\left.+\frac{c_z(1+\gamma)\left(1-4nc_3L^2\gamma^2\right)}{2\gamma\left(1-(4nc_3L^2+27c_4)\gamma^2\right)}\right]\\
&+\left[\frac{9nc_4\gamma(1+\gamma)\left[(\sigma^2+4\delta^2)+4c_3\sigma^2L^2\gamma^2\right]}{1-(4nc_3L^2+27c_4)\gamma^2}\right.\\
&+\frac{128nc_3Sc_2^2\delta^2L^2\gamma^2}{\kappa^2\left(1-(4nc_3L^2+27c_4)\gamma^2\right)(1-\kappa)^2}\\
&+\frac{2c_3n^2L^2\gamma^2\left[(\sigma^2+2\delta^2)-36c_4\sigma^2\gamma^2\right]}{1-(4nc_3L^2+27c_4)\gamma^2}\\
&+\frac{16Sc_2^2\delta^2(1+\gamma)\left[1-4nc_3L^2\gamma^2\right]}{\gamma\kappa^2\left(1-(4nc_3L^2+27c_4)\gamma^2\right)(1-\kappa)^2}\\
&+2L\gamma n^2\sigma^2+n\delta^2+2\sigma^2T^2L^2\gamma^2n^3\\
&\left.+rT\eta^2\sigma^2L^2\gamma^2n^3\right]\frac{16}{k'r\eta^2}\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right]\\
&+\frac{16}{r\eta^2}\left[\frac{9c_4nT\gamma\sigma^2(1+\gamma)(1+2\|A\|^2)}{1-(4nc_3L^2+27c_4)\gamma^2}\right.\\
&+\frac{r\eta^2\sigma^2L^2T^2n^3\gamma^2(1+2\|A\|^2)}{2}\\
&+\frac{c_3Tn^2\sigma^2L^2\gamma^2(1+2\|A\|^2)}{1-(4nc_3L^2+27c_4)\gamma^2}+\sigma^2L^2\gamma^2T^3n^3(1+2\|A\|^2)\\
&\left.+LT\gamma n^2\sigma^2(1+2\|A\|^2)\right]. \qquad (25)
\end{aligned}
$$

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

$$
\begin{aligned}
&\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right]\\
\leq&\sum_{k=0}^{k'T-1}\frac{c_a}{k+(c_q-1)T}\\
\leq&\frac{c_a}{T(c_q-1)}+\ln\left(k'T+(c_q-1)T\right). \qquad (26)
\end{aligned}
$$

where $c_a=\max\{(1+6d)\|A_0-A\|^2,8\sum_{i=1}^n d\Sigma_i\}$, $c_q=6d$ and the first inequality follows from [25, Lemma 21].

Combining (17), (25) and (26) with (22), we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}\left[\|\nabla F(\bar{x}_k)\|^2\right]\\
\leq&\frac{64T(F(x_0^{\psi})-F^{\star}+c_z)}{r\eta^2K\gamma}+\frac{2048c_aSc_2^2\delta^2}{rK\gamma\eta^2\kappa^2(c_q-1)(1-\kappa)^2}\\
&+\frac{2048STc_2^2\delta^2}{\gamma Kr\eta^2\kappa^2(1-\kappa)^2}\ln\left(K+(c_q-1)T\right)\\
&+\frac{C_1T}{K}+\frac{C_2T\ln\left(K+(c_q-1)T\right)}{K}+C_3\sigma^2\gamma+C_4\sigma^2\gamma^2,
\end{aligned}
$$

where $K=k'T$ and C_1, C_2, C_3, C_4 are defined in (21). Then, choosing $\gamma=\frac{1}{\sigma\sqrt{rK}+\bar{\gamma}^{-1}}$, we arrive at (20). The proof is complete. □

Theorem 1 shows that the averaged iterates generated by DASYAGT converge to the stationary point at a rate of $\mathcal{O}\left(\frac{\ln K}{\sqrt{K}}\right)$. Notably, compared to the rate of $\mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$ in [34], the rate in Theorem 1 includes an additional logarithmic factor due to learning the parameter of the distribution map. To the best of our knowledge, Theorem 1 seems to be the first rigorous result on the convergence of the distributed asynchronous algorithm for the stationary solution of the distributed SO-DD.

IV. Experimental Results

To show the practical performance of DASYAGT, we conduct experiments on the multi-agent Gaussian mean estimation problem with synthetic data and the logistic regression problem with a Kaggle credit scoring dataset [40]. We consider three network topologies satisfying Assumption 1, i.e., binary tree, line, directed ring (c.f., Figure 3 in [34]).

In all experiments, we consider the following asynchronous model: (i) Activation lists are generated by concatenating random rounds. Within a round, we have each agent appearing exactly once, that is, the length of a round is $T=n$; (ii) Each transmitted message has a traveling time, which is sampled uniformly from the interval $[0,D]$. Moreover, in all figures, the orange line, the blue line and the red line denote binary tree, line, directed ring, respectively, and one period represents an activation period.

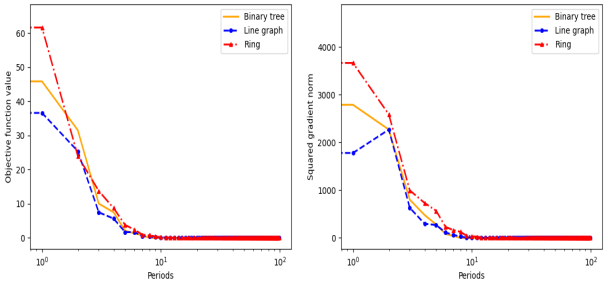
A. Multi-agent Gaussian Mean Estimation

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

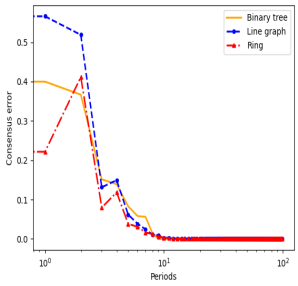
$$\min_{x\in\mathbb{R}} \sum_{i=1}^{n} \mathbf{E}_{\omega_i\sim\mathcal{D}_i(x)} \left[\omega_i x\right], \tag{27}$$

where for $i = 1, \cdots, n$, $\omega_i = 20x + \xi_i$, $\xi_i \sim \mathcal{N}(0, 1)$. Obviously, the objective value in (27) can be computed as 0 with the optimal solution $x^\star = 0$.

In the experiment, we run DASYAGT over three networks with 100 periods. The parameter of the delay model is set as $D = 10$ and the step size of DASYAGT is set as $\gamma = 0.0001$.



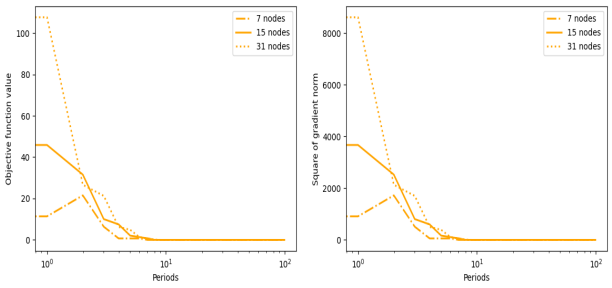
(a) Objective value (b) Squared gradient norm



(c) Consensus error

Fig. 1: Multi-agent Gaussian mean estimation over three different topologies with 15 nodes.

In Figure 1, we report the convergence of DASYAGT, where Figure 1 (a), (b) and (c) record the performance on the objective value, the squared gradient norm and consensus error over binary tree graph, line graph and directed ring graph with $n = 15$ nodes. From Figure 1 (a) and (b), we can observe that the objective value of DASYAGT reaches the optimal value and the squared gradient norm tends to 0, which matches the conclusion of Theorem 1. From Figure 1 (c), we can observe that the consensus error of DASYAGT tends to 0. In Figure 2, we record the performance of DASYAGT on the objective value and the squared gradient norm over binary tree topology with $n = 7, 15, 31$ nodes. From Figure 2 (a) and (b), we can observe that the objective value and the squared gradient norm of DASYAGT reach 0 indicating that DASYAGT can be applied to the communication networks with different size.



(a) Objective value (b) Squared gradient norm

Fig. 2: Multi-agent Gaussian mean estimation over binary tree topology with different number of nodes.

B. Logistic Regression

Consider the following logistic regression problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N} \ell(x; a_{ij}, b_{ij}) + \frac{\beta}{2}\|x\|^2, \tag{28}$$

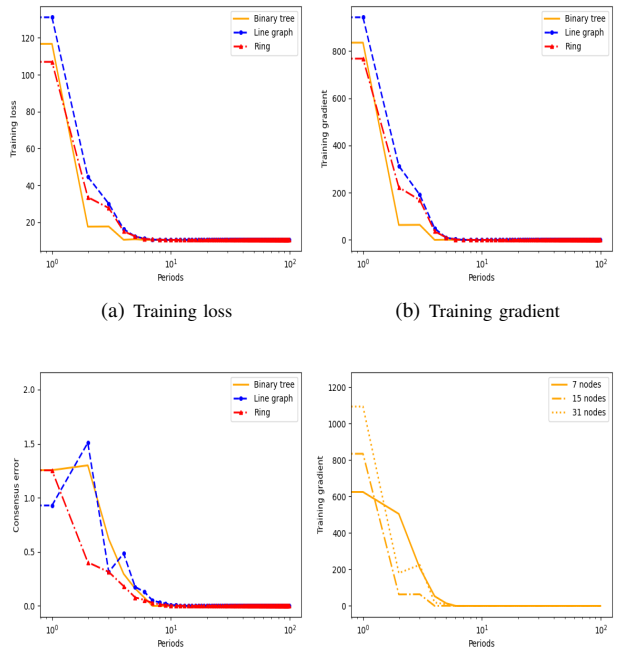
with

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

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

In what follows, we adopt the Kaggle credit scoring dataset [40] with $d = 11$ features for loan approval in bank as the base dataset, where we treat the utilization of credit lines, number of open credit lines, and number of real estate loans as strategic features, to generate the data of i-th agent $S_i = \{a_{ij}, b_{ij}\}_{j=1}^N$ that depends on the i-th decision x_i. Given the base dataset $S_i^0 = \{a_{ij}^0, b_{ij}^0\}_{j=1}^N$ for i-th agent, $a_{ij} = a_{ij}^0 + A_i x_i$ and $b_{ij} = b_{ij}^0$, where $A_i \in \mathbb{R}^{11\times 11}$ is a matrix with all entries equal to 10 except the rows corresponding to the non-strategic features. In the experiment, we set $N = 500$, $D = 5$, $\beta = 0.001$ and run DASYAGT with 100 periods, where the step size of DASYAGT is set as $\gamma = 0.0001$.

In Figure 3, we report the convergence of DASYAGT, where Figure 3 (a), (b) and (c) record the performance on the train loss, the train gradient and the consensus error over binary tree graph, line graph, directed ring graph with $n = 15$ nodes and Figure 3 (d) records the performance on the train gradient over binary tree graph with $n = 7, 15, 31$ nodes. From Figure 3 (a), we can observe that the training loss of DASYAGT reaches around 4.852. From Figure 3 (b), we may observe that the training gradient of DASYAGT tends to 0, which again matches the conclusion of Theorem 1. From Figure 3 (c), we may observe that the consensus error of DASYAGT tends to 0. From Figure 3 (d), we can observe that the training gradient of DASYAGT tends to 0, which is similar to the result of synthetic data and indicates that DASYAGT may be applied to the communication networks with different size in practical issues.



(a) Training loss (b) Training gradient

(c) Consensus error (d) Training loss

Fig. 3: Strategic classification.

V. Conclusion

We propose a distributed asynchronous algorithm, DASYAGT, to seek the optimal solution of the distributed stochastic optimization problems with decision-dependent distributions and show that DASYAGT achieves a convergence rate of $\mathcal{O}\left(\frac{\ln K}{\sqrt{K}}\right)$ in expectation. One promising research direction could be considering the performance of the asynchronous algorithms on the decision-dependent game [25].

Appendix

A. Proof of Lemma 3

Proof. According to (9), we have that for $k \geq 0$,

$$\begin{aligned}
&\sum_{l=0}^{k}\mathbf{E}\left[\|\mathbf{h}_l-\mathbf{1}x_l^{\psi}\|^2\right]\\
\leq&\left(1+\frac{2c_1^2}{1-\kappa^2}\right)\|\mathbf{h}_0-\mathbf{1}x_0^{\psi}\|^2+nc_3\sigma^2\gamma^2(1+2\|A\|^2)(k+1)\\
&+c_3\gamma^2\sum_{t=0}^{k}\mathbf{E}\left[\|z'_{i_t,t}\|^2\right]+2nc_3\sigma^2\gamma^2\sum_{t=0}^{k}\mathbf{E}\left[\|A_{t-\tau_t}-A\|^2\right]\\
\leq&\left(1+\frac{2c_1^2}{1-\kappa^2}\right)\|\mathbf{h}_0-\mathbf{1}x_0^{\psi}\|^2+nc_3\gamma^2(1+2\|A\|^2)\sum_{l=0}^{k}\sigma^2\\
&+4c_3\gamma^2\sum_{l=0}^{k}\mathbf{E}\left[\|z'_{i_l,l}-\phi_{i_l,l-1}\mathbf{1}^{\intercal}\mathbf{z}'_l\|^2\right]\\
&+4c_3\gamma^2\sum_{l=0}^{k}\phi_{i_l,l-1}^2\mathbf{E}\left[\|\sum_{i=1}^{n}\hat{g}_{i,l}-\nabla F(x_l^{\psi})\|^2\right]\\
&+4c_3\gamma^2\sum_{l=0}^{k}\phi_{i_l,l-1}^2\mathbf{E}\left[\|\nabla F(x_l^{\psi})-\sum_{i=1}^{n}\nabla f_i(x_{i,l})\|^2\right]\\
&+4c_3\gamma^2\sum_{l=0}^{k}\phi_{i_l,l-1}^2\mathbf{E}\left[\|\nabla F(x_l^{\psi})\|^2\right]\\
&+2nc_3\sigma^2\gamma^2\sum_{l=0}^{k}\mathbf{E}\left[\|A_{l-\tau_l}-A\|^2\right]\\
\leq&c_h+4c_3\gamma^2\sum_{l=0}^{k}\mathbf{E}\left[\|z'_{i_l,l}-\phi_{i_l,l-1}\mathbf{1}^{\intercal}\mathbf{z}'_l\|^2\right]\\
&+4nc_3L^2\gamma^2\sum_{l=0}^{k}\mathbf{E}\left[\|\mathbf{h}_l-\mathbf{1}x_l^{\psi}\|^2\right]\\
&+c_3\gamma^2\sum_{l=0}^{k}\Big(4\mathbf{E}\left[\|\nabla F(x_l^{\psi})\|^2\right]+n(1+2\|A\|^2)\sigma^2\\
&+2n(\sigma^2+2\delta^2)\mathbf{E}\left[\|A_{l-\tau_l}-A\|^2\right]\Big),
\end{aligned}\tag{29}$$

where $c_3=\frac{4c_1^2}{(1-\kappa)^2}$, $c_h=\left(1+\frac{2c_1^2}{1-\kappa^2}\right)\|\mathbf{h}_0-\mathbf{1}x_0^{\psi}\|^2$, the second inequality follows (8) and the last inequality follows from Assumption 3 (i), (ii).

Similarly, by (10), we have that for $k\geq 0$,

$$\begin{aligned}
&\sum_{l=0}^{k}\mathbf{E}\left[\|z'_{i_l,l}-\phi_{i_l,l-1}\mathbf{1}^{\intercal}\mathbf{z}'_l\|^2\right]\\
\leq&c_z+72c_4\gamma^2\sum_{l=0}^{k}\mathbf{E}\left[\|z'_{i_l,l}-\phi_{i_l,l-1}\mathbf{1}^{\intercal}\mathbf{z}'_l\|^2\right]\\
&+72nc_4L^2\gamma^2\sum_{l=0}^{k}\mathbf{E}\left[\|\mathbf{h}_l-\mathbf{1}x_l^{\psi}\|^2\right]\\
&+18c_4\gamma^2\sum_{l=0}^{k}\Big(n(\sigma^2+4\delta^2)\mathbf{E}\left[\|A_{l-\tau_l}-A\|^2\right]\\
&+4\mathbf{E}\left[\|\nabla F(x_l^{\psi})\|^2\right]+n(1+2\|A\|^2)\sigma^2\Big)\\
&+\frac{32Sc_2^2\delta^2}{(1-\kappa)^2\kappa^2}\sum_{t=0}^{k}\mathbf{E}\left[\|A_{t-\tau_t}-A\|^2\right],
\end{aligned}\tag{30}$$

where $c_4=\frac{4Sc_2^2L^2\left[4c_1^2+(1-\kappa)^2\right]}{\kappa^2(1-\kappa)^4}$ and $c_z=\|z'_{i_0,0}-\phi_{i_0,-1}\mathbf{1}^{\intercal}\mathbf{z}'_0\|^2+\frac{4Sc_2^2\|\mathbf{z}'_0\|^2}{1-\kappa^2}+\frac{72Sc_2^2L^2}{(1-\kappa)^2\kappa^2}\left(1+\frac{2c_1^2}{1-\kappa^2}\right)\|\mathbf{h}_0-\mathbf{1}x_0^{\psi}\|^2$.

By some calculation on (29), (30) and the fact that $\gamma^2<\frac{1}{4nc_3L^2+27c_4}$, (17) and (18) hold. The proof is complete. □

B. Proof of Lemma 4

To prove Lemma 4, we first bound the lower and upper bound of $\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]$ within a length of activation period T.

Lemma 5. *Suppose that Assumptions 1 - 4 hold. Then, for $\forall k\geq 0$ and $t\in[1,T-1]$, $\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right]$ can be lower bounded by*

$$\begin{aligned}
&\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right]\\
\geq&\frac{1}{2}\mathbf{E}\left[\|\nabla F(x_{kT}^{\psi})\|^2\right]-2\sigma^2L^2\gamma^2T^2n^3(1+2\|A\|^2)\\
&-4Tn^2L^2\gamma^2\sum_{l=0}^{T-1}\mathbf{E}\left[\|z'_{i_{kT+l},kT+l}-\phi_{i_{kT+l},kT+l-1}\mathbf{1}^{\intercal}\mathbf{z}'_{kT+l}\|^2\right]\\
&-4Tn^2L^2\gamma^2\sum_{l=0}^{T-1}\psi_{i_{kT+l},kT+l}^2\phi_{i_{kT+l},kT+l-1}^2\mathbf{E}\left[\|\mathbf{1}^{\intercal}\mathbf{z}'_{kT+l}\|^2\right]
\end{aligned}$$

$$- 4T\sigma^2L^2\gamma^2n^3\sum_{l=0}^{T-1}\psi_{i_{kT+l},kT+l}^2\mathbf{E}\left[\|A_{kT+l-T}-A\|^2\right], \quad (31)$$

and upper bounded by

$$\begin{aligned}
&\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right]\\
\leq&2\mathbf{E}\left[\|\nabla F(x_{kT}^{\psi})\|^2\right]+4\sigma^2L^2\gamma^2T^2n^3(1+2\|A\|^2)\\
&+8Tn^2L^2\gamma^2\sum_{l=0}^{T-1}\mathbf{E}\left[\|z'_{i_{kT+l},kT+l}-\phi_{i_{kT+l},kT+l-1}\mathbf{1}^\intercal\mathbf{z}'_{kT+l}\|^2\right]\\
&+8Tn^2L^2\gamma^2\sum_{l=0}^{T-1}\psi_{i_{kT+l},kT+l}^2\phi_{i_{kT+l},kT+l-1}^2\mathbf{E}\left[\|\mathbf{1}^\intercal\mathbf{z}'_{kT+l}\|^2\right]\\
&+8T\sigma^2L^2\gamma^2n^3\sum_{l=0}^{T-1}\mathbf{E}\left[\|A_{kT+l-\tau_{kT+l}}-A\|^2\right]. \quad (32)
\end{aligned}$$

Proof. By (12), we have

$$x_{kT+t}^{\psi}=x_{kT}^{\psi}-\gamma\sum_{l=0}^{t-1}\psi_{i_{kT+l},kT+l}z_{i_{kT+l},kT+l}^\intercal. \quad (33)$$

Next, we give the lower bound and the upper bound of $\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right]$, respectively.

By AM-GM inequality, we have

$$\begin{aligned}
&\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right]\geq\frac{1}{2}\mathbf{E}\left[\|\nabla F(x_{kT}^{\psi})\|^2\right]\\
&-\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})-\nabla F(x_{kT}^{\psi})\|^2\right]. \quad (34)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right]\leq 2\mathbf{E}\left[\|\nabla F(x_{kT}^{\psi})\|^2\right]\\
&+2\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})-\nabla F(x_{kT}^{\psi})\|^2\right]. \quad (35)
\end{aligned}$$

For the second term on the right-hand side of (34) and (35),

$$\begin{aligned}
&\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})-\nabla F(x_{kT}^{\psi})\|^2\right]\\
\leq&Tn^2L^2\gamma^2\sum_{l=0}^{T-1}\psi_{i_{kT+l},kT+l}^2\mathbf{E}\left[\|z_{i_{kT+l},kT+l}\|^2\right]\\
\leq&4Tn^2L^2\gamma^2\sum_{l=0}^{T-1}\mathbf{E}\left[\|z'_{i_{kT+l},kT+l}-\phi_{i_{kT+l},kT+l-1}\mathbf{1}^\intercal\mathbf{z}'_{kT+l}\|^2\right]\\
&+4Tn^2L^2\gamma^2\sum_{l=0}^{T-1}\psi_{i_{kT+l},kT+l}^2\phi_{i_{kT+l},kT+l-1}^2\mathbf{E}\left[\|\mathbf{1}^\intercal\mathbf{z}'_{kT+l}\|^2\right]\\
&+4T\sigma^2L^2\gamma^2n^3\sum_{l=0}^{T-1}\psi_{i_{kT+l},kT+l}^2\mathbf{E}\left[\|A_{kT+l-\tau_{kT+l}}-A\|^2\right]\\
&+2\sigma^2L^2\gamma^2T^2n^3(1+2\|A\|^2), \quad (36)
\end{aligned}$$

where the first inequality follows from Assumption 3 (i), (33) and the last inequality follows from (14). Substituting (36) into (34) and (35), we arrive at (31) and (32), respectively. The proof is complete. □

With the above supporting lemma, we are ready to prove Lemma 4.

Proof. By the update recursion (12) and the fact that $\psi_t^\intercal=\psi_{t+1}^\intercal\hat{\mathbf{W}}_t$, we have

$$\begin{aligned}
&\mathbf{E}\left[F(x_{k+1}^{\psi})\right]\\
\leq&\mathbf{E}\left[F(x_k^{\psi})\right]+\gamma\psi_{i_k,k}\mathbf{E}\left[\langle\nabla F(x_k^{\psi}),-(z_{i_k,k})^\intercal\rangle\right]\\
&+\frac{nL\gamma^2\psi_{i_k,k}^2}{2}\mathbf{E}\left[\|z_{i_k,k}\|^2\right]\\
\leq&\mathbf{E}\left[F(x_k^{\psi})\right]-\gamma\psi_{i_k,k}\mathbf{E}\left[\langle\nabla F(x_k^{\psi}),(z'_{i_k,k})^\intercal-\phi_{i_k,k-1}\mathbf{1}^\intercal\mathbf{z}'_k\rangle\right]\\
&-\gamma\psi_{i_k,k}\phi_{i_k,k-1}\mathbf{E}\left[\langle\nabla F(x_k^{\psi}),\mathbf{1}^\intercal\mathbf{z}'_k\rangle\right]\\
&+\gamma\psi_{i_k,k}\mathbf{E}\left[\langle\nabla F(x_k^{\psi}),(z'_{i_k,k})^\intercal-(z_{i_k,k})^\intercal\rangle\right]\\
&+2nL\gamma^2\psi_{i_k,k}^2\mathbf{E}\left[\|z'_{i_k,k}-\phi_{i_k,k-1}\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&+2nL\gamma^2\phi_{i_k,k-1}^2\psi_{i_k,k}^2\mathbf{E}\left[\|\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&+2Ln^2\sigma^2\gamma^2\psi_{i_k,k}^2\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right]\\
&+Ln^2\sigma^2\gamma^2\psi_{i_k,k}^2(1+2\|A\|^2)\\
\leq&\mathbf{E}\left[F(x_k^{\psi})\right]-\gamma\psi_{i_k,k}\phi_{i_k,k-1}\mathbf{E}\left[\langle\nabla F(x_k^{\psi}),\mathbf{1}^\intercal\mathbf{z}'_k\rangle\right]\\
&+\gamma\psi_{i_k,k}\mathbf{E}\left[\langle\nabla F(x_k^{\psi}),(z'_{i_k,k})^\intercal-(z_{i_k,k})^\intercal\rangle\right]\\
&+\frac{\gamma^2\psi_{i_k,k}}{2}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]+Ln^2\sigma^2\gamma^2\psi_{i_k,k}^2(1+2\|A\|^2)\\
&+\frac{\psi_{i_k,k}}{2}\mathbf{E}\left[\|z'_{i_k,k}-\phi_{i_k,k-1}\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&+2nL\gamma^2\psi_{i_k,k}^2\mathbf{E}\left[\|z'_{i_k,k}-\phi_{i_k,k-1}\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&+2nL\gamma^2\phi_{i_k,k-1}^2\psi_{i_k,k}^2\mathbf{E}\left[\|\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&+2Ln^2\sigma^2\gamma^2\psi_{i_k,k}^2\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right], \quad (37)
\end{aligned}$$

where the first inequality follows from Assumption 3 (i), the last inequality follows from (14) and the Cauchy-Schwarz inequality.

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

$$\begin{aligned}
&\mathbf{E}\left[\langle\nabla F(x_k^{\psi}),\mathbf{1}^\intercal\mathbf{z}'_k\rangle\right]\\
\geq&\frac{1}{2}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]-\mathbf{E}\left[\|\nabla F(x_k^{\psi})-\sum_{i=1}^{n}\nabla f_i(x_{i,k})\|^2\right]\\
&+\frac{1}{2}\mathbf{E}\left[\|\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]-\mathbf{E}\left[\|\sum_{i=1}^{n}\nabla f_i(x_{i,k})-\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
\geq&\frac{1}{2}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]+\frac{1}{2}\mathbf{E}\left[\|\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&-nL^2\mathbf{E}\left[\|\mathbf{h}_k-\mathbf{1}x_k^{\psi}\|^2\right]-n\delta^2\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right], \quad (38)
\end{aligned}$$

where the last inequality follows from (8) and Assumption 3 (i), (ii).

Substituting (38) into (37),

$$\begin{aligned}
&\mathbf{E}\left[F(x_{k+1}^{\psi})\right]\\
\leq&\mathbf{E}\left[F(x_k^{\psi})\right]-\frac{\gamma\psi_{i_k,k}\phi_{i_k,k-1}}{2}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]\\
&+\frac{\gamma^2}{2}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]+n\gamma L^2\mathbf{E}\left[\|\mathbf{h}_k-\mathbf{1}x_k^{\psi}\|^2\right]\\
&+\left(\frac{1}{2}+2nL\gamma^2\right)\mathbf{E}\left[\|z'_{i_k,k}-\phi_{i_k,k-1}\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&+\left(2nL\gamma^2\phi_{i_k,k-1}^2\psi_{i_k,k}^2-\frac{\gamma\psi_{i_k,k}\phi_{i_k,k-1}}{2}\right)\mathbf{E}\left[\|\mathbf{1}^\intercal\mathbf{z}'_k\|^2\right]\\
&+\left(2Ln^2\sigma^2\gamma^2+n\gamma\delta^2\right)\mathbf{E}\left[\|A_{k-\tau_k}-A\|^2\right]\\
&+Ln^2\sigma^2\gamma^2(1+2\|A\|^2). \quad (39)
\end{aligned}$$

Summing (39) over k from 0 to $k'T-1$, we have

$$\begin{aligned}
&\frac{\gamma}{2}\sum_{k=0}^{k'-1}\sum_{t=0}^{T-1}\psi_{i_{kT+t},kT+t}\phi_{i_{kT+t},kT+t-1}\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right]\\
&-\gamma^2\sum_{k=0}^{k'T-1}\mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
\leq & F(x_0^{\psi}) - F^{\star} + n\gamma L^2 \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|\mathbf{h}_k - \mathbf{1}x_k^{\psi}\|^2\right] \\
& - \sum_{k=0}^{k'T-1} \left(\frac{\gamma \psi_{i_k,k}\phi_{i_k,k-1}}{2} - 2nL\gamma^2 \phi_{i_k,k-1}^2 \psi_{i_k,k}^2 \right) \mathbf{E}\left[\|\mathbf{1}^{\intercal}\mathbf{z}_k'\|^2\right] \\
& + \left(\frac{1}{2} + 2nL\gamma^2\right) \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|z_{i_k,k}' - \phi_{i_k,k-1}\mathbf{1}^{\intercal}\mathbf{z}_k'\|^2\right] \\
& + \left(2Ln^2\sigma^2\gamma^2 + n\gamma\delta^2\right) \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|A_{k-\tau_k} - A\|^2\right] \\
& + Lk'Tn^2\sigma^2\gamma^2(1 + 2\|A\|^2), \qquad (40)
\end{aligned}
$$

where $F^{\star} = \min_{x\in\mathbb{R}^d} F(x)$.

For the first term on the left-hand side of (40), we have

$$
\begin{aligned}
& \sum_{k=0}^{k'-1}\sum_{t=0}^{T-1} \psi_{i_{kT+t},kT+t}\phi_{i_{kT+t},kT+t-1}\mathbf{E}\left[\|\nabla F(x_{kT+t}^{\psi})\|^2\right] \\
\geq & \frac{r\eta^2}{2} \sum_{k=0}^{k'-1} \mathbf{E}\left[\|\nabla F(x_{kT}^{\psi})\|^2\right] - 2k'\sigma^2L^2\gamma^2T^3n^3(1+2\|A\|^2) \\
& - 4T^2n^2L^2\gamma^2 \sum_{k=0}^{k'T-1} \psi_{i_k,k}^2\phi_{i_k,k-1}^2\mathbf{E}\left[\|\mathbf{1}^{\intercal}\mathbf{z}_k'\|^2\right] \\
& - 4\sigma^2T^2L^2\gamma^2n^3 \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|A_{k-\tau_k} - A\|^2\right] \\
& - 4T^2n^2L^2\gamma^2 \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|z_{i_k,k}' - \phi_{i_k,k-1}\mathbf{1}^{\intercal}\mathbf{z}_k'\|^2\right] \\
\geq & \frac{r\eta^2}{4T} \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|\nabla F(x_k^{\psi})\|^2\right] \\
& - k'\sigma^2L^2\gamma^2n^3T^2\left(r\eta^2 + 2T\right)(1+2\|A\|^2) \\
& - 2n^2L^2\gamma^2\left(rT\eta^2 + 2T^2\right) \sum_{k=0}^{k'T-1} \psi_{i_k,k}^2\phi_{i_k,k-1}^2\mathbf{E}\left[\|\mathbf{1}^{\intercal}\mathbf{z}_k'\|^2\right] \\
& - 2n^2L^2\gamma^2\left(2T^2 + rT\eta^2\right) \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|z_{i_k,k}' - \phi_{i_k,k-1}\mathbf{1}^{\intercal}\mathbf{z}_k'\|^2\right] \\
& - 2\sigma^2L^2\gamma^2n^3\left(rT\eta^2 + 2T^2\right) \sum_{k=0}^{k'T-1} \mathbf{E}\left[\|A_{k-\tau_k} - A\|^2\right], \qquad (41)
\end{aligned}
$$

where the first inequality follows from (31), Assumption 2 and the last inequality follows from (32). Then, by substituting (41) into (40) and the fact that $\gamma \leq \min\{\frac{2}{Ln(2T^2+rT\eta^2)}, \frac{1}{8nL}\}$, (19) holds. The proof is complete. □

REFERENCES

[1] 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*, vol. 3, no. 1, pp. 1–122, 2011.

[2] 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*, vol. 112, pp. 59–67, 2018.

[3] S. Chouvardas, K. Slavakis, and S. Theodoridis, "Adaptive robust distributed learning in diffusion sensor networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4692–4707, 2011.

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

[5] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 674–690, 2011.

[6] Z. J. Towfic and A. H. Sayed, "Stability and performance limits of adaptive primal-dual networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 11, pp. 2888–2903, 2015.

[7] W. C. Cheung, D. Simchi-Levi, and H. Wang, "Dynamic pricing and demand learning with limited price experimentation," *Operations Research*, vol. 65, no. 6, pp. 1722–1731, 2017.

[8] W. L. Cooper, T. Homem-de Mello, and A. J. Kleywegt, "Models of the spiral-down effect in revenue management," *Operations Research*, vol. 54, no. 5, pp. 968–987, 2006.

[9] T. Liebig, N. Piatkowski, C. Bockermann, and K. Morik, "Dynamic route planning with real-time traffic predictions," *Information Systems*, vol. 64, pp. 258–265, 2017.

[10] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt, "Performative prediction," in *International Conference on Machine Learning*, vol. 119. PMLR, 2020, pp. 7599–7609.

[11] J. Inga and E. Sacoto-Cabrera, "Credit default risk analysis using machine learning algorithms with hyperparameter optimization," in *International Conference on Science, Technology and Innovation for Society*. Springer, 2022, pp. 81–95.

[12] N. Robinson and N. Sindhwani, "Loan default prediction using machine learning," in *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2024, pp. 1–5.

[13] 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*, vol. 35, pp. 38 449–38 460, 2022.

[14] T. W. Jonsbråten, R. J. Wets, and D. L. Woodruff, "A class of stochastic programs with decision dependent random elements," *Annals of Operations Research*, vol. 82, no. 0, pp. 83–106, 1998.

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

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

[17] K. Wood, G. Bianchin, and E. Dall'Anese, "Online projected gradient descent for stochastic optimization with

decision-dependent distributions," *IEEE Control Systems Letters*, vol. 6, pp. 1646–1651, 2022.

[18] K. Wood and E. Dall'Anese, "Stochastic saddle point problems with decision-dependent distributions," *SIAM Journal on Optimization*, vol. 33, no. 3, pp. 1943–1967, 2023.

[19] W.-J. YAN and X.-Y. Cao, "Zero-regret performative prediction under inequality constraints," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 1298–1308.

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

[21] 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*, vol. 139. PMLR, 2021, pp. 4641–4650.

[22] D. Drusvyatskiy and L. Xiao, "Stochastic optimization with decision-dependent distributions," *Mathematics of Operations Research*, vol. 48, no. 2, pp. 954–998, 2023.

[23] L. Hellemo, P. I. Barton, and A. Tomasgard, "Decision-dependent probabilities in stochastic programs with recourse," *Computational Management Science*, vol. 15, no. 3, pp. 369–395, 2018.

[24] M. Hardt and C. Mendler-Dünner, "Performative prediction: Past and future," *arXiv preprint arXiv:2310.16608*, 2023.

[25] 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*, vol. 24, no. 202, pp. 1–56, 2023.

[26] L.-C. Deng and Y.-C. Liu, "Gradient tracking methods for distributed stochastic optimization problems with decision-dependent distributions," *Optimization Online preprint https://optimization-online.org/?p=31572*, 2025.

[27] A. Nedić and A. Ozdaglar, "Convergence rate for consensus with delays," *Journal of Global Optimization*, vol. 47, no. 3, pp. 437–456, 2010.

[28] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions," *Journal of Machine Learning Research*, vol. 21, no. 58, pp. 1–47, 2020.

[29] J.-Q. Zhang and K.-Y. You, "Asyspa: An exact asynchronous algorithm for convex optimization over digraphs," *IEEE Transactions on Automatic Control*, vol. 65, no. 6, pp. 2494–2509, 2020.

[30] M. S. Assran and M. G. Rabbat, "Asynchronous gradient push," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 168–183, 2021.

[31] N. Bastianello, R. Carli, L. Schenato, and M. Todescato, "Asynchronous distributed optimization over lossy networks via relaxed admm: Stability and linear convergence," *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2620–2635, 2021.

[32] Y. Tian, Y. Sun, and G. Scutari, "Achieving linear convergence in distributed asynchronous multiagent optimization," *IEEE Transactions on Automatic Control*, vol. 65, no. 12, pp. 5264–5279, 2020.

[33] V. Kungurtsev, M. Morafah, T. Javidi, and G. Scutari, "Decentralized asynchronous nonconvex stochastic optimization on directed graphs," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 4, pp. 1796–1804, 2023.

[34] Z.-H. Zhu, Y. Tian, Y. Huang, J.-M. Xu, and S.-B. He, "R-fast: Robust fully-asynchronous stochastic gradient tracking over general topology," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 10, pp. 665–678, 2024.

[35] T. Ben-Nun and T. Hoefler, "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–43, 2019.

[36] M. Assran, A. Aytekin, H. R. Feyzmahdavian, M. Johansson, and M. G. Rabbat, "Advances in asynchronous parallel and distributed optimization," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 2013–2031, 2020.

[37] S. Pu, W. Shi, J.-M. Xu, and A. Nedić, "Push–pull gradient methods for distributed optimization in networks," *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 1–16, 2021.

[38] J.-N. Ren, J. Haupt, and Z.-H. Guo, "Communication-efficient hierarchical distributed optimization for multi-agent policy evaluation," *Journal of Computational Science*, vol. 49, p. 101280, 2021.

[39] W.-W. Wu, S. Liu, and S.-Y. Zhu, "Distributed dual gradient tracking for economic dispatch in power systems with noisy information," *Electric Power Systems Research*, vol. 211, p. 108298, 2022.

[40] Kaggle, "Give me some credit," https://www.kaggle.com/c/GiveMeSomeCredit/data, 2012.