Hindawi
Security and Communication Networks
Volume 2021, Article ID 1020466, 7 pages
https://doi.org/10.1155/2021/1020466

Research Article

Distributed Policy Evaluation with Fractional Order Dynamics in Multiagent Reinforcement Learning

Wei Dai,[1] **Wei Wang**,[2] **Zhongtian Mao**,[1] **Ruwen Jiang**,[1] **Fudong Nian**,[3] **and Teng Li**[1]

[1] *Anhui Engineering Laboratory of Human-Robot Integration System and Intelligent Equipment,*
School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China
[2] *Center for Assessment and Demonstration Research, Academy of Military Science, Beijing 100091, China*
[3] *Hefei University, Hefei 230601, China*

Correspondence should be addressed to Wei Wang; wwei009@mail.ustc.edu.cn

Received 15 July 2021; Revised 15 August 2021; Accepted 17 August 2021; Published 6 September 2021

Academic Editor: Zhenhua Tan

The main objective of multiagent reinforcement learning is to achieve a global optimal policy. It is difficult to evaluate the value function with high-dimensional state space. Therefore, we transfer the problem of multiagent reinforcement learning into a distributed optimization problem with constraint terms. In this problem, all agents share the space of states and actions, but each agent only obtains its own local reward. Then, we propose a distributed optimization with fractional order dynamics to solve this problem. Moreover, we prove the convergence of the proposed algorithm and illustrate its effectiveness with a numerical example.

1. Introduction

In recent years, reinforcement learning [1] has received much attention from the society and succeeded remarkably in many areas such as machine learning and artificial intelligence [2]. As we all know, in reinforcement learning, an agent determines the optimal strategy under the feedback of rewards via constantly interacting with the environment. The function of the policy maps possible states to possible actions. Although reinforcement learning has made great achievements in single agent, it remains challenging in the application of multiagent [3]. The goal of the multiagent system is to enable several agents with simple intelligence, but it is easy to manage and control to realize complex intelligence through mutual cooperation. While reducing the complexity of system modeling, the robustness, reliability, and flexibility of the system should be improved [4, 5].

In this paper, the objective of this paper is to investigate multiagent reinforcement learning (MARL), where each agent exchanges information with their neighbors in network systems [6]. All agents share the state space and action except local rewards. The purpose of the MARL is to determine the global optimal policy, and a feasible way is to construct a central controller, where each agent must exchange information with the central controller [7], which makes decisions for all of them. However, with the increase of state dimensions, the computation of the central controller becomes extensively heavy. The whole system would collapse if the central controller was attacked.

Then, we try to replace the centralized algorithm mentioned above with distributed control [8, 9]. Consistency protocol based on design enables all agents to achieve the same state [10–13]. In [14], Zhang et al. proposed a continuous-time distributed version of the gradient algorithm. As far as we know, most of the gradient methods use integer order iteration. In fact, fractional order has been developed for 300 years and used to solve many kinds of problems such as control applications and systems' theory [15–17]. In comparison with the traditional integer order algorithm, the fractional order algorithm has more design freedom and potential to obtain better convergence performance [18, 19].

Hereinafter, the contributions of the paper are listed:

(1) We transform the multiagent strategy evaluation problem into a distributed optimization problem with a consensus constraint

(2) We construct the fractional order dynamics and prove the convergence of the algorithm

(3) We take a numerical example to verify the superiority of the proposed fractional order algorithm

The rest organization of this paper is listed as follows. Section 2 introduces some problems of formulation on MARL and fractional order calculus. Section 3 transforms the multiagent strategy evaluation problem into the optimization problem with a consensus constraint, proposes an algorithm with fractional order dynamics, and proves that the algorithm asymptotically converge to an exact solution. Section 4 presents a simulation example, and we summarize the work in Section 5.

2. Problem Formulation

2.1. Notations. Let $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$ represent the real number set, n-dimensional real column vector set, and $n\times m$ real matrix set, respectively. AT represents the transpose of A. $\|A\| = (\sum_{i=1}^n \sum_{i=1}^n a_{ij})^{1/2}$, $\|X\|_G = \sqrt{X^T GX}$, and $\langle A, B\rangle = A^T B$. $(\mathcal{S}, \{\mathcal{A}\}_{i=1}^n, \mathrm{P}, \{R_i\}_{i=1}^n, \gamma)$ represents a multiagent Markov decision process (MDP), where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the joint action space. $\mathbf{P}^{\mathrm{a}}$ is the probability of transition from s_t to s_{t+1} when the agent takes the joint action a and $[\mathrm{P}^\pi]_{s,s'} = \mathbb{E}_{a\sim\pi(\cdot|\mathrm{s})}[\mathrm{P}^a]_{s,s'}$, $\mathrm{R}_i(\mathbf{s},\mathbf{a})$ is the local reward when agent i takes joint action a at state s and $\gamma\in(0,1)$ is a discount parameter. $\pi(\mathbf{a}|\mathbf{s})$ represents the condition of probability when the agent takes joint action a at state s. The reward function of agent i is defined when follows a joint policy π at state s as follows:

$$R_i^\pi(\mathbf{s}) = \mathbb{E}_{a\sim\pi(\cdot|\mathrm{s})}[\mathrm{R}_i(\mathbf{s},\mathbf{a})], \tag{1}$$

where the right-hand side of the equation means that there is a probability for all possible choices of action a, and we calculate the expected value for all rewards of agent i:

$$R_c^\pi(s) = \frac{1}{n}\sum_{i=1}^n R_i^\pi(s), \tag{2}$$

where $R_c^\pi(s)$ represents the average of the local rewards.

2.2. Graph theory. The graph is expressed as $\mathcal{G}(\mathcal{V},\mathcal{E})$, where $\mathcal{G}$ represents a graph, $\mathcal{V}$ is the set of vertices, and $\mathcal{E}$ is the set of edges in $\mathcal{G}$. If any edge in the graph is undirected, the graph is named as undirected graph [20]. In graph, $\mathbf{A} = [a_{\mathrm{ij}}]\in\mathbb{R}^{n\times n}$ is the adjacency matrix with $a_{\mathrm{ij}}\neq 0$ if $\{i,j\}\in\mathcal{E}$, $a_{\mathrm{ij}} = 0$ otherwise. $\mathbf{D} = \mathrm{diag}[\mathrm{d}_1, d_2, \ldots, d_3]$ is the degree matrix with $d_i = \sum_{j=1}^n a_{\mathrm{ij}}$ and Laplacian matrix is $\mathbf{L} = \mathbf{D} - \mathbf{A}$. Moreover, if the graph is connected, $\mathbf{L}$ has the following two properties:

(1) Laplacian matrix is a semipositive definite matrix

(2) The minimum eigenvalue is 0 because the sum of every row of the Laplace matrix is 0

The minimum nonzero eigenvalue is defined as the algebraic connectivity of the graph.

Assumption 1. The undirected graph mentioned in the following text is connected.

Lemma 1 (see [21]). *The frequency distributed model is defined for a fractional order system $\mathcal{D}^\alpha x(t) = u(t)$, where $\alpha\in(0,1)$ as follows:*

$$\begin{cases} \dfrac{\partial \mathrm{z}(\omega,t)}{\partial z} = -\omega \mathrm{z}(\omega,t) + u(t), \\ \\ y(t) = \displaystyle\int_0^\infty \mu_\alpha(\omega)\mathrm{z}(\omega,t)\mathrm{d}\omega, \end{cases} \tag{3}$$

where $\mu_\alpha = \sin(\alpha\pi)/\omega^\alpha\pi$.

Definition 1 (see [22]). The αth order Caputo derivative is

$$\mathcal{D}^\alpha f(t) = \frac{1}{\Gamma(n-\alpha)}\int_0^t (t-\tau)^{n-1-\alpha} f^{(n)}(\tau)\mathrm{d}\tau, \tag{4}$$

where $\alpha\in(n-1,n)$, $n\in\mathbb{N}$, $\Gamma(t) = \int_0^\infty \tau^{t-1}e^{-\tau}\mathrm{d}\tau$ is Gamma function, and $f^n(t)$ is the nth order derivative of $f(t)$.

2.3. Policy Evaluation. To measure the benefits of agents in its current state, we establish the following value function, which represents the value of the cumulative return obtained by agents starting from the state s_t, adopting a certain strategy π:

$$V^\pi(\mathbf{s}) = \mathbb{E}_\pi\left[\sum_{m=1}^\infty \gamma^m R_c^\pi(s_{t+m+1})|s_t = \mathbf{s}\right]. \tag{5}$$

We construct Bellman equation based on $\mathbf{V}^\pi\in\mathbb{R}^{|S|}$ and $\mathbf{R}_c^\pi\in\mathbb{R}^{|S|}$:

$$\mathbf{V}^\pi = \mathbf{R}_c^\pi + \gamma\mathbf{P}^\pi\mathbf{V}^\pi. \tag{6}$$

It is difficult to evaluate $\mathbf{V}^\pi$ directly if the dimension of the state space is very large. Therefore, we use $V_\theta(\mathrm{s}) = \phi^T(\mathrm{s})\theta$ to approximate $\mathbf{V}^\pi$, where $\theta\in\mathbb{R}^d$ is the vector and $\phi(\mathrm{s})$: $\mathcal{S}\longrightarrow\mathbb{R}^d$, which is a particular function for state s. Indeed, solving equation (6) is equivalent to obtain the vector θ via $V_\theta\approx V^\pi$. In other words, it means to minimize the mean square error about $1/2\|\mathbf{V}_\theta - \mathbf{V}^\pi\|_{\mathrm{D}}^2$, where $\mathbf{D} = \mathrm{diag}\{\mu^\pi(\mathrm{s}), \mathrm{s}\in S\}$, $\in\in\mathbb{R}^{\|S\|\times\|S\|}$ is a diagonal matrix determined by the stationary distribution. We construct the equation as follows:

$$f(\theta) = \frac{1}{2}\left\|\mathbf{\Pi}_{\mathbf{\Phi}}(\mathbf{V}_\theta - \gamma\mathbf{P}^\pi\mathbf{V}_\theta - \mathbf{R}_{\mathrm{c}}^\pi)\right\|_D^2 + \frac{\rho}{2}\|\theta\|^2, \tag{7}$$

where ρ is a regularization parameter and Π_Φ is a projection operator in the column subspace of Φ. It is not difficult to rewrite Π_Φ as $\Pi_\Phi = \Phi(\Phi^T\mathbf{D\Phi})^{-1}\Phi^T\mathbf{D}$ substituting Π_Φ into (7):

$$\begin{aligned} f(\boldsymbol{\theta}) &= \frac{\rho}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{2}\left\|\boldsymbol{\Pi}_{\boldsymbol{\Phi}}\left(\mathbf{V}_{\boldsymbol{\theta}} - \gamma\mathbf{P}^{\pi}\mathbf{V}_{\boldsymbol{\theta}} - \mathbf{R}_c^{\pi}\right)\right\|_D^2 \\ &= \frac{\rho}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{2}\left(\mathbf{V}_{\boldsymbol{\theta}} - \gamma\mathbf{P}^{\pi}\mathbf{V}_{\boldsymbol{\theta}} - \mathbf{R}_c^{\pi}\right)^T\boldsymbol{\Pi}_{\boldsymbol{\Phi}}^{\mathrm{T}}\mathbf{D}\boldsymbol{\Pi}_{\boldsymbol{\Phi}} \times \left(\mathbf{V}_{\boldsymbol{\theta}} - \gamma\mathbf{P}^{\pi}\mathbf{V}_{\boldsymbol{\theta}} - \mathbf{R}_c^{\pi}\right) \\ &= \frac{\rho}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{2}\left(\mathbf{V}_{\theta} - \gamma\mathbf{P}^{\pi}\mathbf{V}_{\boldsymbol{\theta}} - \mathbf{R}_c^{\pi}\right)^T\mathbf{D}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^T\mathbf{D}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{D} \times \left(\mathbf{V}_{\boldsymbol{\theta}} - \gamma\mathbf{P}^{\pi}\mathbf{V}_{\boldsymbol{\theta}} - \mathbf{R}_c^{\pi}\right) \\ &= \frac{\rho}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{2}\left\|\boldsymbol{\Phi}^T\mathbf{D}\left(\mathbf{V}_{\boldsymbol{\theta}} - \gamma\mathbf{P}^{\pi}\mathbf{V}_{\boldsymbol{\theta}} - \mathbf{R}_c^{\pi}\right)\right\|_{(\boldsymbol{\Phi}^T\mathbf{D}\boldsymbol{\Phi})^{-1}}^2 \\ &= \frac{\rho}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{2}\left\|\boldsymbol{\Phi}^T\mathbf{D}\left(\boldsymbol{\Phi} - \gamma\mathbf{P}^{\pi}\boldsymbol{\Phi}\right)\boldsymbol{\theta} - \boldsymbol{\Phi}^T\mathbf{D}R_c^{\pi}\right\|_{(\boldsymbol{\Phi}^T\mathbf{D}\boldsymbol{\Phi})^{-1}}^2 \\ &= \frac{\rho}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{2}\|\mathbf{A}\boldsymbol{\theta} - \mathbf{b}\|_{\mathbf{C}^{-1}}^2, \end{aligned} \tag{8}$$

where $\mathbf{A}= \Phi^T\mathbf{D}(\Phi - \gamma\mathbf{P}^{\pi}\Phi) = \mathbb{E}_{s\sim\mu^{\pi}}[\phi(\mathbf{s})(\phi(\mathbf{s}) - \gamma\phi(\mathbf{s}'))^T]$, $\mathbf{C} = \Phi^T\mathbf{D}\boldsymbol{\Phi} = \mathbb{E}_{s\sim\mu^{\pi}}[\phi(\mathbf{s})\phi^T(\mathbf{s})]$, and $\mathbf{b} = \Phi^T\mathbf{D}\mathbf{R}_c^{\pi} = \mathbb{E}_{s\sim\mu^{\pi}}[R_c^{\pi}(\mathbf{s})\phi(\mathbf{s})]$.

The minimum value of θ in equation (8) is unique if A is a full rank matrix and C is a positive definite matrix. In practice, it is difficult to get the expectations in the compact form when the distribution is unknown. We replace expectation with the average as follows:

$$\begin{aligned} \widehat{\mathbf{A}} &= \frac{1}{p}\sum_{t=1}^{P} A_t, \\ \widehat{\mathbf{b}} &= \frac{1}{p}\sum_{t=1}^{P} b_t, \\ \widehat{\mathbf{C}} &= \frac{1}{p}\sum_{t=1}^{P} C_t, \end{aligned} \tag{9}$$

where $\mathbf{A}_t = \phi(s_t)\varphi^T(s_t), \varphi(s_t)(\phi(s_t)\phi(s_{\mathrm{t+1}}))^T$, $\mathbf{C}_t = \phi(s_t)\phi^T(s_t)$, and $\mathbf{b}_t = \mathbf{R}_c^{\pi}(s_t)\phi(s_t)$.

We assume that the sample size p approaches infinity to make sure its confidence level. In these sequences, each state is attached at least once. Then, we reconstruct equation (8) as follows:

$$f(\boldsymbol{\theta})= \frac{1}{2}\|\widehat{\mathbf{A}}\boldsymbol{\theta} - \widehat{\mathbf{b}}\|_{\widehat{\mathbf{C}}^{-1}}^2 + \frac{\rho}{2}\|\boldsymbol{\theta}\|^2. \tag{10}$$

Noteworthy, in a shared space, the agent observes the states and actions of the neighbors, but only observes the local rewards of its own. In other words, we get $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{C}}$ except $\widehat{\mathbf{b}}$. So, we define $\widehat{\mathbf{b}}_i = (1/p)\sum_{t=1}^{p} b_{\mathrm{t,i}}$ with $\mathbf{b}_{t,i} = R_i^{\pi}(\mathbf{s}_{\mathbf{t}}, \mathbf{a}_{\mathbf{t}})\phi(\mathbf{s}_t)$. Then, we rewrite equation (10) as follows:

$$\min_{\theta\in R^d} \frac{1}{n}\sum_{t=1}^{n}\frac{1}{2}\left\|\widehat{\mathbf{A}}\boldsymbol{\theta} - \widehat{\mathbf{b}}_i\right\|_{C^{-1}}^2 + \frac{\rho}{2}\|\boldsymbol{\theta}\|^2. \tag{11}$$

3. Fractional Order Dynamics for Policy Evaluation

Hereinbefore, the aim of policy evaluation becomes to minimize the object function. Now, we rewrite (11) as follows:

$$\begin{cases} \min_{\theta_i} & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{1}{2}\left\|\widehat{\mathbf{A}}\theta_i - \widehat{\mathbf{b}}_i\right\|_{\widehat{\mathrm{C}}^{-1}}^2 + \frac{\rho}{2}\left\|\boldsymbol{\theta}_i\right\|^2, \\ \text{s.t.} & \boldsymbol{\theta}_i= \boldsymbol{\theta}_j. \end{cases} \tag{12}$$

We define $\overline{\theta} \in \mathbb{R}^{n\,d}$ as a factor concatenating all θ_i: $\overline{\theta} = [\theta_1^T, \theta_2^T, \ldots \theta_{\mathrm{n}T}]^T \in \mathbb{R}^{n\,d}$ and the aggregative function f as $f(\overline{\theta}) = \sum_{i=1}^{n} f(\theta_i)$. As we all know, the consensus constraint (12) is expressed as

$$\begin{cases} \min_{\overline{\theta}} & \dfrac{1}{2}\left\|\overline{\widehat{\mathbf{A}}}\overline{\boldsymbol{\theta}} - \overline{\widehat{\mathbf{b}}}_i\right\|_{\overline{\widehat{C}}^{-1}}^2 + \dfrac{\rho}{2}\|\overline{\boldsymbol{\theta}}\|^2, \\ \text{s.t.} & \overline{\mathbf{L}}\overline{\boldsymbol{\theta}} = \mathbf{0}, \end{cases} \tag{13}$$

where $\overline{\widehat{\mathbf{b}}} = [\widehat{\mathbf{b}}_1^T, \widehat{\mathbf{b}}_2^T, \ldots \widehat{\mathbf{b}}_n^T]^T \in \mathbb{R}^{\mathrm{nd}}$, $L \in \mathbb{R}^{n\times n}$, $\overline{\mathbf{L}} = \mathbf{L}\otimes\mathbf{I}_d \in \mathbb{R}^{\mathrm{nd}\times\mathrm{nd}}$, $\overline{\widehat{\mathbf{A}}}= \widehat{\mathbf{A}}\otimes\mathbf{I}_n \in \mathbb{R}^{\mathrm{nd}\times\mathrm{nd}}$, and $\overline{\widehat{\mathbf{C}}}= \widehat{\mathbf{C}}\otimes\mathbf{I}_n \in\in\mathbb{R}^{\mathrm{nd}\times\mathrm{nd}}$. Based on (13), we formulate the following the augmented Lagrangian:

$$\mathscr{L}(\overline{\boldsymbol{\theta}}, \boldsymbol{\lambda}) = f(\overline{\boldsymbol{\theta}}) + \langle\boldsymbol{\lambda}, \overline{\mathbf{L}}\overline{\boldsymbol{\theta}}\rangle + \frac{1}{2}\overline{\boldsymbol{\theta}}^T\overline{\mathbf{L}}\overline{\boldsymbol{\theta}}, \tag{14}$$

where $\lambda \in \mathbb{R}^{n\,d}$ is the Lagrange multiplier.

It is feasible to design a fractional order continuous-time optimization algorithm from primal-dual viewpoint, gradient descend for primal variable $\overline{\theta}$, and gradient ascent for dual variable λ via (14). Both of them are updated according to the fractional order law:

$$\begin{cases} \mathscr{D}^{\alpha_1}\overline{\boldsymbol{\theta}}(\mathrm{t}) = -\nabla_{\overline{\theta}(\mathrm{t})}\mathscr{L}(\overline{\boldsymbol{\theta}}(\mathrm{t}), \boldsymbol{\lambda}(\mathrm{t})), \\ \mathscr{D}^{\alpha_2}\boldsymbol{\lambda}(\mathrm{t}) = \nabla_{\lambda(\mathrm{t})}\mathscr{L}(\overline{\boldsymbol{\theta}}(\mathrm{t}), \boldsymbol{\lambda}(\mathrm{t})), \end{cases} \tag{15}$$

where $0 < \alpha_1 < 2$, $0 < \alpha_2 < 1$, $\nabla_{\overline{\theta}(t)}\mathscr{L}(\overline{\theta}(t), \lambda(t))$, and $\nabla_{\lambda(t)}\mathscr{L}(\overline{\theta}(t), \lambda(t))$ are gradient of $(\overline{\theta}(t), \lambda(t))$ on variables $\overline{\theta}(t)$ and $\lambda(t)$, respectively. We express the detail of (15) in Algorithm 1.

The aim of the distributed algorithm is to obtain the solution of the value function. The proposed algorithm has more potential to get better convergence performance and design freedom than the conventional integer order. Hereinafter, we provide the following convergence conclusion.

Theorem 1. *Under Assumption 1, let $\overline{\theta}(t)$ and $\lambda(t)$ be generated according to Algorithm 1. If $0 < \alpha_1, \alpha_2 < 1$, then $\overline{\theta}(t)$ asymptotically converges to the optimal solution.*

Proof. We obtain the detailed dynamics of $\overline{\theta}(\mathrm{t})$ and $\lambda(t)$:

$$\begin{cases} \mathscr{D}^{\alpha_1}\overline{\boldsymbol{\theta}}(\mathrm{t}) = -\left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\overline{\boldsymbol{\theta}}(\mathrm{t}) + \overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\mathbf{b}} - \overline{\mathbf{L}}\boldsymbol{\lambda}(\mathrm{t}), \\ \mathscr{D}^{\alpha_2}\boldsymbol{\lambda}(\mathrm{t}) = \overline{\mathbf{L}}\overline{\boldsymbol{\theta}}((\mathrm{t}), \end{cases} \tag{16}$$

where $\mathbf{I}$ is an identity matrix. We consider the equilibrium of (16):

$$\begin{cases} \mathbf{0} = -\left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\overline{\theta}^{*} + \overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\mathbf{b}} - \overline{\mathbf{L}}\boldsymbol{\lambda}^{*}, \\ \mathbf{0} = \overline{\mathbf{L}}\overline{\boldsymbol{\theta}}^{*}. \end{cases} \tag{17}$$

Then, we combine (16) and (17), and according to the facts $\mathscr{D}^{\alpha_1}\overline{\theta}^{*} = 0$, $\mathscr{D}^{\alpha_2}\lambda^{*} = 0$,

$$\begin{cases} \mathscr{D}^{\alpha_1}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) = -\left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) - \overline{\mathbf{L}}\tilde{\boldsymbol{\lambda}}(\mathrm{t}), \\ \mathscr{D}^{\alpha_2}\tilde{\boldsymbol{\lambda}}(\mathrm{t}) = \overline{\mathbf{L}}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}). \end{cases} \tag{18}$$

Through Lemma 1, we reconstruct (18) as follows:

$$\begin{cases} \dfrac{\partial \mathrm{z}_1(\omega, \mathrm{t})}{\partial t} = -\omega\mathbf{z}_1(\omega, \mathrm{t}) - \left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) - \overline{\mathbf{L}}\tilde{\boldsymbol{\lambda}}(\mathrm{t}), \\ \tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) = \displaystyle\int_0^\infty \mu_{\alpha_1}(\omega)\mathrm{z}_1(\omega, \mathrm{t})\mathrm{d}\omega \end{cases} \tag{19}$$

and

$$\pi\begin{cases} \dfrac{\partial \mathrm{z}_2(\omega, \mathrm{t})}{\partial t} = -\omega\mathrm{z}_2(\omega, \mathrm{t}) + \overline{\mathbf{L}}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}), \\ \tilde{\lambda}(\mathrm{t}) = \displaystyle\int_0^\infty \mu_{\alpha_2}\mathrm{z}_2(\omega, \mathrm{t})\mathrm{d}\omega. \end{cases} \tag{20}$$

We construct the Lyapunov function as follows:

$$V_1 = \frac{1}{2}\int_0^\infty \sum_{i=1}^{2}\mu_{\alpha_i}(\omega)\left\|\mathrm{z}_\mathrm{i}(\omega, \mathrm{t})\right\|^2\mathrm{d}\omega. \tag{21}$$

Then,

$$\begin{aligned} \dot{V}_1 &= \int_0^\infty \sum_{i=1}^{2}\mu_{\alpha_i}(\omega)\langle \mathrm{z}_\mathrm{i}(\omega, t), \frac{\partial \mathrm{z}_\mathrm{i}(\omega, t)}{\partial t}\rangle\mathrm{d}\omega \\ &= \int_0^\infty \mu_{\alpha_1}(\omega)\langle \mathrm{z}_1(\omega, t), -\omega\mathrm{z}_1(\omega, t) - \left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) - \overline{\mathbf{L}}\tilde{\boldsymbol{\lambda}}(\mathrm{t})\rangle 0\omega + \int_0^\infty \mu_{\alpha_2}(\omega)\langle \mathrm{z}_2(\omega, t) - \omega\mathrm{z}_2(\omega, t) + \overline{\mathbf{L}}\tilde{\overline{\theta}}(\mathrm{t})\rangle 0\omega \\ &= -\int_0^\infty \sum_{i=1}^{2}\mu_{\alpha_i}(\omega)\left\|\mathrm{z}_\mathrm{i}(\omega, t)\right\|^2\mathrm{d}\omega + \langle\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}), -\left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\overline{\theta}}(\mathrm{t}) - \overline{\mathbf{L}}\tilde{\boldsymbol{\lambda}}(\mathrm{t})\rangle + \langle\tilde{\lambda}(\mathrm{t}), \overline{\mathbf{L}}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t})\rangle \\ &= -\int_0^\infty \sum_{i=1}^{2}\mu_{\alpha_i}(\omega)\left\|\mathrm{z}_\mathrm{i}(\omega, t)\right\|^2\mathrm{d}\omega - \tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t})^{T}\left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) \le \mathbf{0}. \end{aligned} \tag{22}$$

We obtain the result according to the Lasalle invariance principle.

Hereinafter, we improve the convergence conclusion of Theorem 1 by extending α_1 from (0,1) to (1,2).

Theorem 2. *Under Assumption 1, let $\overline{\theta}(t)$ and $\lambda(t)$ be generated according to Algorithm 1. If $1 < \alpha_1 < 2$, $\alpha_1 + \alpha_2 = 2$, then $\overline{\theta}(t)$ asymptotically converges to the optimal solution.*

Proof. Under the condition $\alpha_1 = 1 + \overline{\alpha}_1$, we rewrite the dynamics with the condition of Theorem 1 as follows:

$$\begin{cases} \dot{\tilde{\overline{\theta}}}_a(\mathrm{t}) = -\left(\overline{\hat{\mathbf{A}}}^{T}\overline{\hat{\mathbf{C}}}^{-1}\overline{\hat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) - \overline{L}\tilde{\lambda}(\mathrm{t}), \\ \mathscr{D}^{\overline{\alpha}_1}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) = \tilde{\overline{\boldsymbol{\theta}}}_a(\mathrm{t}), \\ \mathscr{D}^{\alpha_2}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}) = \overline{\mathbf{L}}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}). \end{cases} \tag{23}$$

Due to $\alpha_1 = 1 + \overline{\alpha}_1$ and $\alpha_1 + \alpha_2 = 2$,

$$\dot{\tilde{\lambda}}(\mathrm{t}) = \mathscr{D}^{\overline{\alpha_1}}\mathscr{D}^{\alpha_2}\tilde{\boldsymbol{\lambda}}(\mathrm{t}) = \mathscr{D}^{\overline{\alpha_1}}[\overline{\mathbf{L}}\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t})] = \overline{\mathbf{L}}_a\tilde{\overline{\boldsymbol{\theta}}}(\mathrm{t}). \tag{24}$$

Under the condition of (23) and (24), we obtain the frequency distributed model by Lemma 1 as follows:

```
Initialization: $\theta_i = \mathbf{0} \in \mathbb{R}^d$, $\lambda_i = \mathbf{0} \in \mathbb{R}^d$.
Update
  For $t \le 50$
    $\mathscr{D}^{\alpha_1}\theta_i(\mathrm{t}) = -((\widehat{\mathbf{A}}^T\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{A}} + \rho\mathbf{I})\theta_i(\mathrm{t}) - \widehat{\mathbf{A}}^T\widehat{\mathbf{C}}^{-1}\widehat{\mathbf{b}}_i + \overline{\mathbf{L}}_i\lambda(\mathrm{t}) + \overline{\mathbf{L}}_i\overline{\theta}(\mathrm{t}))$, $\mathscr{D}^{\alpha_2}\lambda_i(\mathrm{t}) = \overline{\mathbf{L}}_i\overline{\theta}(\mathrm{t})$
  End
Return $\theta$
```

Algorithm 1

$$\begin{cases} \dot{\tilde{\bar{\theta}}}_a(\mathrm{t}) = -\left(\overline{\widehat{\mathbf{A}}}^T\overline{\widehat{\mathbf{C}}}^{-1}\overline{\widehat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\bar{\boldsymbol{\theta}}}(\mathrm{t}) - \overline{\mathbf{L}}\tilde{\boldsymbol{\lambda}}(\mathrm{t}), \\ \dfrac{\partial \mathrm{z}_1(\omega,\mathrm{t})}{\partial t} = -\omega\mathbf{z}_1(\omega,\mathrm{t}) + \tilde{\bar{\theta}}_a(\mathrm{t}), \\ \tilde{\bar{\boldsymbol{\theta}}}(\mathrm{t}) = \displaystyle\int_0^\infty \mu_{\overline{\alpha}_1}(\omega)\mathrm{z}_1(\omega,t)\mathrm{d}\omega, \\ \dot{\tilde{\lambda}}(\mathrm{t}) = \overline{\mathbf{L}}\tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t}). \end{cases} \tag{25}$$

We construct the Lyapunov function:

$$V_2 = \frac{1}{2}\left\|\tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t})\right\|^2 + \frac{1}{2}\|\tilde{\boldsymbol{\lambda}}(\mathrm{t})\|^2. \tag{26}$$

Then,

$$\begin{aligned} \dot{V}_2 &= \langle\tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t}), \tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t})\rangle + \langle\tilde{\boldsymbol{\lambda}}(\mathrm{t}), \dot{\tilde{\lambda}}(\mathrm{t})\rangle \\ &= \langle\tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t}), -\left(\overline{\widehat{\mathbf{A}}}^T\overline{\widehat{\mathbf{C}}}^{-1}\overline{\widehat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t}) - \overline{\mathbf{L}}\tilde{\boldsymbol{\lambda}}(\mathrm{t})\rangle \\ &\quad + \langle\tilde{\boldsymbol{\lambda}}(\mathrm{t}), \overline{\mathbf{L}}\tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t})\rangle \\ &= -\tilde{\bar{\boldsymbol{\theta}}}_a(\mathrm{t})^T\left(\overline{\widehat{\mathbf{A}}}^T\overline{\widehat{\mathbf{C}}}^{-1}\overline{\widehat{\mathbf{b}}} + \rho\mathbf{I} + \overline{\mathbf{L}}\right)\tilde{\bar{\boldsymbol{\theta}}}(\mathrm{t}) \le \mathbf{0}. \end{aligned} \tag{27}$$

Through the LaSalle invariance principle, we obtain the result.

4. Experimental Simulation

In this section, we provide an example to illustrate the effectiveness of the proposed algorithm. There are 20 states in the multiagent reinforcement learning. We set $d = 5$, regularization parameter $\rho = 0.1$, and discount parameter $\gamma = 0.5$. There are 4 agents in the connected network in Figure 1. State s is a randomly generated 5-dimensional column vector, the dimension of $\phi(s)$ is a cosine function, and P is a randomly generated 5-dimensional matrix.

Then, we randomly generate the matrices $\widehat{\mathbf{A}}, \widehat{\mathbf{C}}, \widehat{\mathbf{b}}_i$ as follows:

$$\begin{aligned} \widehat{\mathbf{A}} &= \begin{bmatrix} 0.9797 & 0.5949 & 0.1174 & 0.0855 & 0.7303 \\ 0.4389 & 0.2622 & 0.2967 & 0.2625 & 0.4886 \\ 0.1111 & 0.6028 & 0.3188 & 0.8010 & 0.5785 \\ 0.2581 & 0.7112 & 0.4242 & 0.0292 & 0.2373 \\ 0.4087 & 0.2217 & 0.5079 & 0.9289 & 0.4588 \end{bmatrix}, \\ \widehat{\mathbf{C}} &= \begin{bmatrix} 0.2500 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.2500 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.2500 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.2500 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.2500 \end{bmatrix}, \\ \widehat{\mathbf{b}}_1 &= [0.9631, 0.5468, 0.5211, 0.2316, 0.4889]^T, \\ \widehat{\mathbf{b}}_2 &= [0.6241, 0.6791, 0.3955, 0.3674, 0.9880]^T, \\ \widehat{\mathbf{b}}_3 &= [0.0377, 0.8852, 0.9133, 0.7962, 0.0987]^T, \\ \widehat{\mathbf{b}}_4 &= [0.2619, 0.3354, 0.6797, 0.1366, 0.7212]^T. \end{aligned} \tag{28}$$

Before the simulation, it is necessary to obtain the solution of the multiagent reinforcement learning:

$$\boldsymbol{\theta}^* = [-0.0756, 0.0211, 0.5362, 0.0508, 0.6956]^T. \tag{29}$$

We show the comparison about the fractional order algorithm with the conventional integer order one. In Figures 2 and 3, the curve illustrates almost the same convergence performance as the conventional integer order when α is 0.995. In Figures 4 and 5, the fractional order algorithm achieves a faster convergent rate than that of the integer order algorithm. Simulation results illustrate the convergence about the integer order and the fractional order. Furthermore, the proposed distributed

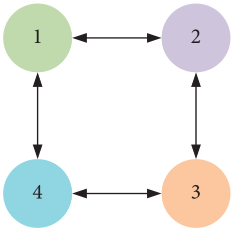

Figure 1: Undirected communication network graph.

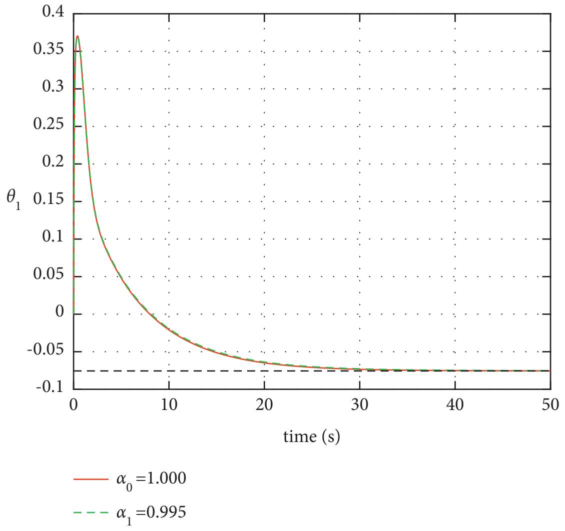

Figure 2: The trajectory of estimation of θ_1.

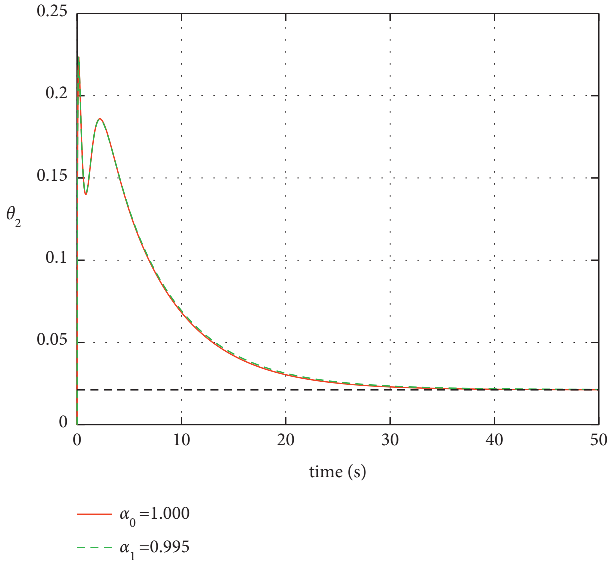

Figure 3: The trajectory of estimation of θ_2.

algorithm with fractional order dynamics has more design freedom to achieve a better performance than that of the conventional first-order algorithm.

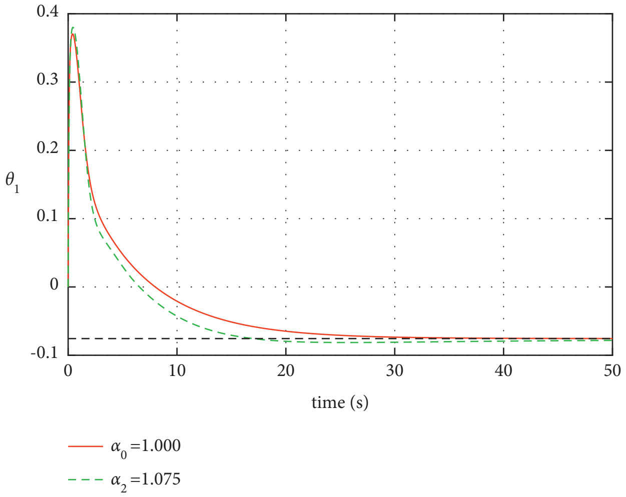

Figure 4: The trajectory of estimation of θ_1.

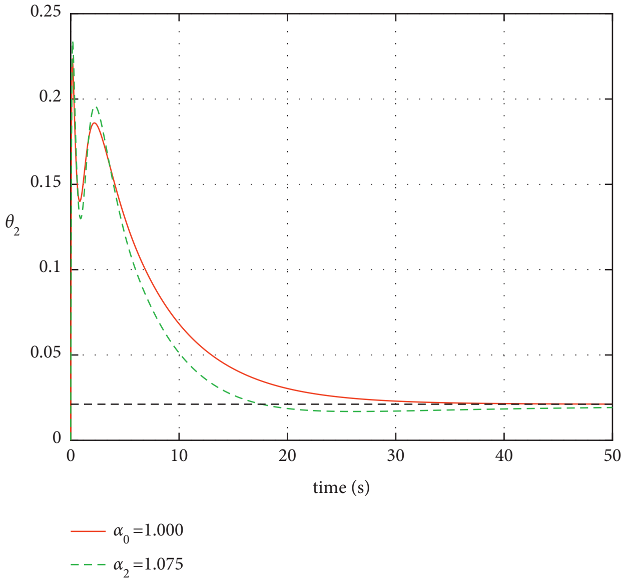

Figure 5: The trajectory of estimation of θ_2.

5. Conclusion

In this paper, the value function problem of the multiagent reinforcement learning was transformed as a distributed optimization problem with a consensus constraint. Then, we proposed a distributed algorithm with fractional order dynamics to solve this problem. Besides, we proved the asymptotic convergence of the algorithm by Lyapunov functions and illustrated the effectiveness of the proposed algorithm with an example. In the future, we will consider applying reinforcement learning to the recommendation system, so as to get better results [23].

Data Availability

The .m and .slx data used to support the findings of this study have been deposited in the Github repository (97weiD/data_DPEFOD).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61973002 and 61902104), and the Anhui Provincial Natural Science Foundation (2008085J32 and 2008085QF295).

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press Cambridge, Cambridge, MA, USA, 1998.

[2] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: a survey,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.

[3] L. Busoniu, R. De Schutter, and B. D. Schutter, “A comprehensive survey of m reinforcement learning,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.

[4] X. Wang, G. Wang, and S. Li, “Distributed finite-time optimization for integrator chain m systems with disturbances,” *IEEE Transactions on Automatic Control*, vol. 65, no. 12, pp. 5296–5311, 2020.

[5] X. Wang, S. Li, X. Yu, and J. Yang, “Distributed active anti-disturbance consensus for leader-follower higher-order multi-agent systems with mismatched disturbances,” *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5795–5801, 2017.

[6] K. Zhang, Z. Yang, and T. Basar, “Networked multi-agent reinforcement learning in continuous spaces,” in *Proceedings of the 57th IEEE Conference on Decision and Control*, pp. 2771–2776, Fontainebleau in Miami Beach, FL, USA, December 2018.

[7] J. K. Gupta, M. Egorov, and M. Kochenderfer, “Cooperative multi-agent control using deep reinforcement learning,” in *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 66–83, São Paulo, Brazil, May 2017.

[8] X. Zhao, P. Yi, and L. Li, “Distributed policy evaluation via inexact ADMM in multi-agent reinforcement learning,” *Control Theory and Technology*, vol. 18, no. 4, pp. 362–378, 2020.

[9] A. Nedic and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[10] X. Wang, S. Li, and G. Wang, “Distributed optimization for disturbed second-order m systems based on active a control,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 2104–2117, 2020.

[11] X. Wang, G. Wang, and S. Li, “A distributed fixed-time optimization algorithm for multi-agent systems,” *Automatica*, vol. 122, Article ID 109289, 2020.

[12] X. Wang, S. Li, and J. Lam, “Distributed active anti-disturbance output consensus algorithms for higher-order multi-agent systems with mismatched disturbances,” *Automatica*, vol. 74, pp. 30–37, 2016.

[13] T. Yang, X. Yi, J. Wu et al., “A survey of distributed optimization,” *Annual Reviews in Control*, vol. 47, pp. 278–305, 2019.

[14] M. Zhang, X. Liu, and J. Liu, “Convergence analysis of a continuous-time distributed gradient descent algorithm,” *IEEE Control Systems Letters*, vol. 5, no. 4, pp. 1339–1344, 2021.

[15] J.-G. Luo and Y.-Q. Chen, “Robust stability and stabilization of fractional-order interval systems with the fractional order α: the $0 \ll \alpha \ll 1$ case,” *IEEE Transactions on Automatic Control*, vol. 55, no. 1, pp. 152–158, 2010.

[16] Y.-Q. Wei, D.-Y. Liu, and D. Boutat, “Innovative fractional derivative estimation of the pseudo-state for a class of fractional order linear systems,” *Automatica*, vol. 99, pp. 157–166, 2019.

[17] Y. Wei, Y. Chen, J. Wang, and Y. Wang, “Analysis and description of the infinite-dimensional nature for nabla discrete fractional order systems,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 72, pp. 472–492, 2019.

[18] S. Cheng, Y. Wei, Y. Chen, Y. Li, and Y. Wang, “An innovative fractional order LMS based on variable initial value and gradient order,” *Signal Processing*, vol. 133, pp. 260–269, 2017.

[19] S. Cheng, S. Liang, and Y. Fan, “Distributed solving sylvester equations with fractional order dynamics,” *Control Theory and Technology*, vol. 19, no. 1, pp. 249–259, 2021.

[20] A. Torres and G. Anders, “Spectral graph theory and network dependability,” in *Proceedings of the 2009 4th International Conference on Dependability of Computer Systems*, Brunow, Poland, July 2009.

[21] J. C. Trigeassou, N. Maamri, J. Sabatier, and A. Oustaloup, “Transients of fractional-order integrator and derivatives,” *Signal, Image and Video Processing*, vol. 6, no. 3, pp. 359–372, 2012.

[22] C. Monje, Y. Chen, B. Vinagre, D. Xue, and V. Feliu-Batlle, *Fractionalorder Systems and Controls: Fundamentals and Applications*, Springer, New York, NY, USA, 2010.

[23] F. Xue, X. He, X. Wang, J. Xu, K. Lia, and R. Hong, “Deep item-based collaborative filtering for top-n recommendation,” *ACM Transactions on Information Systems*, vol. 37, no. 3, pp. 33–25, 2019.