SCIENCE CHINA
Information Sciences

• LETTER •

August 2023, Vol. 66 189204:1–189204:2
https://doi.org/10.1007/s11432-021-3603-0

A novel policy iteration algorithm for solving the optimal consensus control problem of a discrete-time multiagent system with unknown dynamics

Wenkai XU[1,2], Li WANG[1,2*], Shiwen SUN[1,2], Chengyi XIA[1,2] & Zengqiang CHEN[3]

[1]*Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin* 300384*, China;*
[2]*The Engineering Research Center of Learning-Based Intelligent System, Ministry of Education, Tianjin* 300384*, China;*
[3]*College of Artificial Intelligence, Nankai University, Tianjin* 300350*, China*

Received 10 August 2021/Accepted 21 July 2022/Published online 22 February 2023

Citation Xu W K, Wang L, Sun S W, et al. A novel policy iteration algorithm for solving the optimal consensus control problem of a discrete-time multiagent system with unknown dynamics. Sci China Inf Sci, 2023, 66(8): 189204, https://doi.org/10.1007/s11432-021-3603-0

At present, an increasing number of researchers have noticed the importance of optimal consensus control (OCC) of multiagent systems (MASs) because of their rich practical applications in various areas [1–4]. To accomplish OCC, a common method is to solve the coupled Hamilton-Jacobi-Bellman (HJB) equation, which barely obtains the analytical solution and requires the complete mathematical model of an MAS. Because of the limitations in solving the HJB equation, we need to find more effective methods for overcoming these challenges. Fortunately, adaptive dynamic programming (ADP) is an efficacious way to address distributed control problems. Therefore, we focus on solving an OCC problem of an MAS with unknown dynamics by combining ADP approaches and the reinforcement learning (RL) method. We design a novel policy iteration-based ADP (PI-ADP) method called the β-PI algorithm, in which we can learn the distributed optimal control policies (OCPs) by relying only on the agent's and its neighbors' state information rather than an accurate mathematical model. In addition, the β-PI algorithm fully uses current iterative control policies to expedite convergence during training.

We research the OCC problem of a homogeneous DT-MAS, which includes N follower agents and one leader agent. Each follower agent's dynamics is given as

$$x_p(i+1) = Ax_p(i) + B_p u_p(i), \quad p = 1, 2, \ldots, N-1, N, \tag{1}$$

where $x_p(i) \in \mathbb{R}^n$ and $u_p(i) \in \mathbb{R}^{m_p}$ represent each follower agent's state and the control input, respectively. Both $A \in \mathbb{R}^{(n\times n)}$ and $B_p \in \mathbb{R}^{(n\times m_p)}$ are system matrices that are considered unknown.

The leader with dynamics is given as

$$x_0(i+1) = Ax_0(i), \tag{2}$$

where $x_0(i) \in \mathbb{R}^n$ is the leader's state.

Then, we write the neighbor tracking error for agent p by its state and the neighbors' state information as follows:

$$\mathrm{err}_p(i) = \sum_{q\in\Omega_p} a_{pq}(x_p(i) - x_q(i)) + b_p(x_p(i) - x_0(i)), \tag{3}$$

where Ω_p denotes the set of neighbors of agent p, $b_p \geqslant 0$ represents the relationship between a follower and the leader agent, $b_p > 0$ means that agent p can obtain the leader agent's state directly, and $b_p = 0$ means otherwise.

The global tracking error is given in Appendix A.1, and from (1), (2), and (3), we obtain the dynamics of tracking error err_p, which are described in the following equation:

$$\begin{aligned}\mathrm{err}_p(i+1) = A\mathrm{err}_p(i) &- \sum_{q\in\Omega_p} (a_{pq}B_q u_q(i)) \\ &+ (d_{pp} + b_p)B_p u_p(i).\end{aligned} \tag{4}$$

In addition, we also consider the OCC of a grouped, heterogeneous DT-MAS, in which the N follower agents are divided into two groups, $g1$ and $g2$. The follower agents have the same dynamics when they belong to the same group; otherwise, their dynamics differ. The dynamics of each follower agent can be depicted as $x_p(i+1) = A_{gs}x_p(i) + B_p u_p(i), A_{gs} \in \{A_{g1}, A_{g2}\}$, and the dynamics of the leader agent is identical to that in (2). The system matrix A differs from the matrices A_{g1} and A_{g2}. We give the local tracking error $\mathrm{err}_p(i)$ as (3), similar to the homogeneous case. However, $\mathrm{err}_p(i+1)$ cannot be written in an iterative form similar to that of (4) because it contains follower agents with different dynamics, but it can be given by function (3) when letting $T = i+1, \mathrm{err}_p(T) = \sum_{q\in\Omega_p} a_{pq}(x_p(T) - x_q(T)) + b_p(x_p(T) - x_0(T))$, where T is the time step; that is, we can obtain the dynamics of the tracking errors regardless of whether the dynamics of the agents are homogeneous or heterogeneous.

* Corresponding author (email: wltjut08@126.com)

Remark 1. We consider two groups of followers as a whole system; i.e., the communication topology of all follower agents constitutes a directed graph.

Assumption 1. The directed graph of the MAS communication network includes a spanning tree.

We can define the same performance index functions (PIFs) [4] for both homogeneous and grouped heterogeneous DT-MASs because the local tracking error definitions are identical. As a result, we rewrite the time step i as D and define the PIF for each follower agent as follows:

$$\begin{aligned} &Q_p(\mathrm{err}_p(D), u_p(D), u_{\Omega_p}(D)) \\ &\quad = \sum_{t=D}^{\infty} \theta^{t-D} g_p(\mathrm{err}_p(D), u_p(D), u_{\Omega_p}(D)), \end{aligned} \tag{5}$$

where $g_p(\mathrm{err}_p(D), u_p(D), u_{\Omega_p}(D)) = \mathrm{err}_p^{\mathrm{T}}(D) P_{pp} \mathrm{err}_p(D) + u_p^{\mathrm{T}}(D) S_{pp} u_p(D) + \sum_{q\in\Omega_p} u_q^{\mathrm{T}}(D) R_{pq} u_q(D)$ represents the evaluation function, $u_{\Omega_p}(D)$ signifies a set of control policies, $\{u_q(D)|q \in \Omega_p\}$, $\theta \in (0, 1]$ represents the discount factor of the PIF equation, and P_{pp}, S_{pp}, and R_{pq} are positive definite symmetric weighting matrices, where $P_{pp} \geqslant 0, S_{pp} \geqslant 0, R_{pq} \geqslant 0$.

Definition 1 (Admissible control [5])**.** The admissible control policies $u_p(D)$ can stabilize system (4) and guarantee that PIF (5) is finite.

Presenting the admissible control polices $u_p(D)$ for followers, Eq. (5) can be rewritten as $Q_p(\mathrm{err}_p(D)) = g_p(\mathrm{err}_p(D), u_p(D), u_{\Omega_p}(D)) + \theta Q_p(\mathrm{err}_p(D+1))$.

Then, we define $Q_p^*(\mathrm{err}_p(D)) = \min_{u_p(D)} Q_p(\mathrm{err}_p(D), u_p(D), u_{\Omega_p}(D))$ as the optimal performance index function (OPIF). From the Bellman-Optimality principle, Q_p^* satisfies the DT-HJB equation as follows: $Q_p^*(\mathrm{err}_p(D)) = \min_{u_p(D)}\{g_p(\mathrm{err}_p(D), u_p(D), u_{\Omega_p}(D)) + \theta Q_p^*(\mathrm{err}_p(D+1))\}$.

Remark 2. The final target of our research is to obtain the distributed OCP $u_p^*(i)$, which can minimize the PIF. We present a β-PI algorithm to obtain the iterative policies and iterative PIFs for solving the OCC problem of a DT-MAS with unknown dynamics. Let $Q_p^l(\mathrm{err}_p(D))$ and $u_p^l(D)$ represent the iterative PIF and iterative control policy, respectively, which are optimized with an iteration index l that increases in real time. We define β_p^l as the parameter β for the pth agent with index l; then, the algorithm is given as the Algorithm 1 in Appendix A.2.

Theorem 1 (Convergence of the β-PI algorithm)**.** For any p and l, the iterative control policy $u_p^l(D)$ and iterative PIF $Q_p^l(\mathrm{err}_p(D))$ are computed using Algorithm 1, where the initial control policies $u_p^0(D)$ are admissible. The parameter β_p^l is computed using Algorithm 1. When $l \to \infty$, $Q_p^l(\mathrm{err}_p(D))$ can converge to the OPIF $Q_p^*(\mathrm{err}_p(D))$ and $u_p^l(D)$ can converge to the OCP $u_p^*(D)$, which means that $\lim_{l\to\infty} Q_p^l(\mathrm{err}_p(D)) = Q_p^*(\mathrm{err}_p(D))$, $\lim_{l\to\infty} u_p^l(D) = u_p^*(D)$. The complete proof is in Appendix B.2.

Theorem 2 (Stability analysis)**.** Assume that Assumption 1 holds. If the OPIF $Q_p^*(\mathrm{err}_p(D))$ meets the DT-HJB equation and the OCP $u_p^*(D)$ satisfies the definition, then the tracking error system of (4) must be asymptotically stable. The complete proof is in Appendix B.3.

Then, we implement the β-PI algorithm by training the neural networks (NNs). We use two different three-layer back propagation (BP) NNs (the function is in Appendix C) as the actor NN and critic NN to compute the iterative PIF $Q_p^l(\mathrm{err}_p(D))$ and the iterative control policy $u_p^l(D)$.

The critic NN is used to compute the iterative PIF $Q_p^l(\mathrm{err}_p(D))$. It is given as $\hat{Q}_p(D) = \hat{W}_{cp}^{\mathrm{T}} \Phi_{cp}(Y_{cp}^{\mathrm{T}} z_{cp}(D))$. Then, we give the critic NN training BP error as $\sigma_{cp}(D) = g_p(D-1) + \theta \hat{W}_{cp}^{\mathrm{T}} \Phi_{cp}(Y_{cp}^{\mathrm{T}} z_{cp}(D)) - \hat{W}_{cp}^{\mathrm{T}} \Phi_{cp}(Y_{cp}^{\mathrm{T}} z_{cp}(D-1))$.

For training the critic NN, we define the loss function as $E_{cp}(D) = \frac{1}{2}\sigma_{cp}^{\mathrm{T}}(D)\sigma_{cp}(D)$, and we update the weight matrix of the NN by solving gradient descending of the loss function $E_{cp}(D)$, i.e., $\hat{W}_{cp}(l+1) = \hat{W}_{cp}(l) - lr_{cp} \frac{\partial E_{cp}(D)}{\partial \sigma_{cp}(D)} \frac{\partial \sigma_{cp}(D)}{\partial \hat{W}_{cp}(l)}$, where lr_{cp} represents the updating step size of the critic NN.

The actor NN is used to compute the iterative control policies $u_p^l(D)$, and it is described as $\hat{u}_p(D) = \hat{W}_{ap}^{\mathrm{T}} \Phi_{ap}(Y_{ap}^{\mathrm{T}} z_{ap}(D))$. We set $\sigma_{ap}(D)$ as the BP NN error of the actor NN and $\sigma_{ap}(D) = \hat{Q}_p(D) - U_p(D)$, where $U_p(D)$ is the final objective cost function and is usually set to zero; i.e., $\sigma_{ap}(D) = \hat{Q}_p(D)$.

The loss function of the training actor NN is defined as $E_{ap}(D) = \frac{1}{2}\sigma_{ap}^{\mathrm{T}}(D)\sigma_{ap}(D)$. We update $\hat{W}_{ap}(l+1)$ by $\hat{W}_{ap}(l+1) = \hat{W}_{ap}(l) - lr_{ap} \frac{\partial E_{ap}(D)}{\partial \sigma_{ap}(D)} \frac{\partial \sigma_{ap}(D)}{\partial \hat{Q}_p(D)} \frac{\partial \hat{Q}_p(D)}{\partial \hat{u}_p(D)} \frac{\partial \hat{u}_p(D)}{\partial \hat{W}_{ap}(l)}$, where lr_{ap} is the learning rate of the actor NN.

Remark 3. By approximating the value of the iterative PIF $Q_p^l(\mathrm{err}_p(D))$ by $\hat{Q}_p(D)$ and the iterative control policies $u_p^l(D)$ by $\hat{u}_p(D)$, the OCP $u_p^*(D)$ can be obtained by the β-policy iteration algorithm; that is, the OCC problem is resolved by only using the error information between itself and its neighbors' states.

Remark 4. It is important to obtain the admissible control policies in the β-PI algorithm initialization phase, i.e., to find the admissible control policies $u_p^0(D)$. A viable method for approximating the admissible policies is to repeat experiments by training the actor NN until the initial control policy is found to stabilize the $\mathrm{err}_p(i+1)$ or $\mathrm{err}_p(T)$.

Simulation. We provide two experiments to demonstrate the abovementioned theoretical analyses. The results show that the proposed algorithm can not only solve the OCC problem but also improve the speed of convergence and reduce fluctuation compared with traditional algorithms. The figures of the two experiments' results are shown in Appendix C.3.

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61403280, 61773286). Li WANG acknowledged the support from 131 Innovative Talents Program of Tianjin.

References

1 Peng Z N, Zhao Y Y, Hu J P, et al. Data-driven optimal tracking control of discrete-time multi-agent systems with two-stage policy iteration algorithm. Inf Sci, 2019, 481: 189–202
2 Zhou D H, Qin L G, He X, et al. Distributed sensor fault diagnosis for a formation system with unknown constant time delays. Sci China Inf Sci, 2018, 61: 112205
3 Yu W W, Li C J, Yu X H, et al. Economic power dispatch in smart grids: a framework for distributed optimization and consensus dynamics. Sci China Inf Sci, 2018, 61: 012204
4 Wang D, Liu D R, Li H, et al. Neural-network-based robust optimal control design for a class of uncertain nonlinear systems via adaptive dynamic programming. Inf Sci, 2014, 282: 167–179
5 Zhang H G, Jiang H, Luo Y H, et al. Data-driven optimal consensus control for discrete-time multi-agent systems with unknown dynamics using reinforcement learning method. IEEE Trans Ind Electron, 2017, 64: 4091–4100