Hongyang LI, Qinglai WEI, 2022. Optimal synchronization control for multi-agent systems with input saturation: a nonzero-sum game. *Frontiers of Information Technology & Electronic Engineering,* 23(7):1010-1019.
https://doi.org/10.1631/FITEE.2200010

Optimal synchronization control for multi-agent systems with input saturation: a nonzero-sum game

Key words: Optimal synchronization control; Multi-agent systems; Nonzero-sum game; Adaptive dynamic programming; Input saturation; Off-policy reinforcement learning; Policy iteration

Corresponding author: Qinglai WEI
E-mail: qinglai.wei@ia.ac.cn
ORCID: https://orcid.org/0000-0001-7002-9800

Research problem

- Optimal synchronization control for multi-agent systems with input saturation
- Control systems: $\dot{\boldsymbol{x}}_i = \boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{B}\boldsymbol{u}_i$

$$\boldsymbol{u}_i \in \mathcal{U}_i \subset \mathbb{R}^m \qquad \dot{\boldsymbol{x}}_0 = \boldsymbol{A}\boldsymbol{x}_0$$

$$\mathcal{U}_i = \left\{ \boldsymbol{u}_i \mid \boldsymbol{u}_i \in \mathbb{R}^m, \left\| \boldsymbol{u}_i \right\|_\infty \le \lambda_i \right\}$$

Leader

- State synchronization condition: $\delta_i \to 0 \quad \delta_i = \sum_{j \in N_i} e_{ij} \left(\boldsymbol{x}_i - \boldsymbol{x}_j \right) + g_i \left(\boldsymbol{x}_i - \boldsymbol{x}_0 \right)$

Research background

- The existing synchronization condition methods for multi-agent systems based on ADP (Vamvoudakis et al., 2012; Wei et al., 2015; Qin et al., 2019)
- Contribution: The control constraint and coupled terms in the performance index functions are considered which broaden the application scope of the presented method

Multi-agent nonzero-sum game

- Performance index function

$$J_i\left(\boldsymbol{\delta}_i(0),\boldsymbol{u}_i,\boldsymbol{u}_{-i}\right)=\int_0^{\infty}\left(\boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{Q}_{ii}\boldsymbol{\delta}_i+R_i\left(\boldsymbol{u}_i\right)+\sum_{j\in N_i}R_i\left(\boldsymbol{u}_j\right)\right)\mathrm{d}t \tag{1}$$

- Nash equilibrium condition

$$J_i\left(\boldsymbol{u}_i^*,\boldsymbol{u}_{-i}^*\right)\le J_i\left(\boldsymbol{u}_i,\boldsymbol{u}_{-i}^*\right),\ i=1,2,\ldots,N \tag{2}$$

- Iterative value function

$$V_i\left(\boldsymbol{\delta}_i(t)\right)=\int_t^{\infty}\left(\boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{Q}_{ii}\boldsymbol{\delta}_i+R_i\left(\boldsymbol{u}_i\right)+\sum_{j\in N_i}R_i\left(\boldsymbol{u}_j\right)\right)\mathrm{d}\tau \tag{3}$$

V_i^* satisfies the HJB equation

Optimal control

$$\boldsymbol{u}_i^*=-\lambda_i\Psi\left(\frac{1}{2\lambda_i}\left(d_i+g_i\right)\boldsymbol{R}_{ii}^{-1}\boldsymbol{B}^{\mathrm{T}}\frac{\partial V_i^*}{\partial\boldsymbol{\delta}_i}\right) \tag{4}$$

$$\left(\frac{\partial V_i^*}{\partial\boldsymbol{\delta}_i}\right)^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{\delta}_i-\lambda_i\left(d_i+g_i\right)\boldsymbol{B}\Psi\left(\Delta_i^*\right)+\sum_{j\in N_i}\lambda_j e_{ij}\boldsymbol{B}\Psi\left(\Delta_i^*\right)\right)+\boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{Q}_{ii}\boldsymbol{\delta}_i$$
$$+R_i\left(-\lambda_i\Psi\left(\Delta_i^*\right)\right)+\sum_{j\in N_i}R_i\left(-\lambda_j\Psi\left(\Delta_j^*\right)\right)=0 \tag{5}$$

Theorem 1 Assume that the optimal control law is $\boldsymbol{u}_i^*$, and that V_i is the positive definite smooth solution to the HJB equation. Then, the tracking error is asymptotically stable. $\boldsymbol{u}_i^*, i=1,2,\ldots,N$, constitute the Nash equilibrium, and the solution V_i to the HJB equation is the optimal value of the game, i.e., $J_i^*\left(\boldsymbol{\delta}_i(0),\boldsymbol{u}_i^*,\boldsymbol{u}_{-i}^*\right)=V_i\left(\boldsymbol{\delta}_i(0)\right)$.

Policy iteration method

- Solve V_i^k

$$\left(\frac{\partial V_i^k}{\partial \boldsymbol{\delta}_i}\right)^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{\delta}_i+\left(d_i+g_i\right)\boldsymbol{B}\boldsymbol{u}_i^k-\sum_{j\in N_i}e_{ij}\boldsymbol{B}\boldsymbol{u}_j^k\right)$$
$$+\boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{Q}_{ii}\boldsymbol{\delta}_i+R_i\left(\boldsymbol{u}_i^k\right)+\sum_{j\in N_i}R_i\left(\boldsymbol{u}_j^k\right)=0 \tag{6}$$

- Update the iterative control laws

$$\boldsymbol{u}_i^{k+1}=-\lambda_i\Psi\left(\frac{1}{2\lambda_i}\left(d_i+g_i\right)\boldsymbol{R}_{ii}^{-1}\boldsymbol{B}^{\mathrm{T}}\frac{\partial V_i^k}{\partial \boldsymbol{\delta}_i}\right) \tag{7}$$

Theorem 2 Assume that agent *i* and its neighbors update their control policies according to Eqs. (6) and (7). Then, policies $\boldsymbol{u}_i^k, \boldsymbol{u}_{-i}^k$ converge to the Nash equilibrium, and V_i^k converges to V_i^*.

❑ Model-free off-policy reinforcement learning method

- Rewrite the tracking error dynamics

$$\dot{\boldsymbol{\delta}}_i = \boldsymbol{A}\boldsymbol{\delta}_i + (d_i + g_i)\boldsymbol{B}\left(\boldsymbol{u}_i - \boldsymbol{u}_i^k\right) + (d_i + g_i)\boldsymbol{B}\boldsymbol{u}_i^k - \sum_{j\in N_i} e_{ij}\boldsymbol{B}\boldsymbol{u}_j^k - \sum_{j\in N_i} e_{ij}\boldsymbol{B}\left(\boldsymbol{u}_j - \boldsymbol{u}_j^k\right) \tag{8}$$

$$\dot{V}_i^k = \left(\frac{\partial V_i^k}{\partial \boldsymbol{\delta}_i}\right)^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{\delta}_i - \sum_{j\in N_i} e_{ij}\boldsymbol{B}\boldsymbol{u}_j^k + (d_i + g_i)\boldsymbol{B}\boldsymbol{u}_i^k\right) + \left(\frac{\partial V_i^k}{\partial \boldsymbol{\delta}_i}\right)^{\mathrm{T}}(d_i + g_i)\boldsymbol{B}\left(\boldsymbol{u}_i - \boldsymbol{u}_i^k\right) - \left(\frac{\partial V_i^k}{\partial \boldsymbol{\delta}_i}\right)^{\mathrm{T}}\sum_{j\in N_i} e_{ij}\boldsymbol{B}\left(\boldsymbol{u}_j - \boldsymbol{u}_j^k\right) \tag{9}$$

$$\begin{aligned} &V_i^k\left(\boldsymbol{\delta}_i(t')\right) - V_i^k\left(\boldsymbol{\delta}_i(t)\right) \\ &= -\int_t^{t'} \boldsymbol{\delta}_i^{\mathrm{T}}\boldsymbol{Q}_{ii}\boldsymbol{\delta}_i \mathrm{d}\tau - \int_t^{t'} R_i\left(\boldsymbol{u}_i^k\right)\mathrm{d}\tau - \int_t^{t'}\sum_{j\in N_i} R_i\left(\boldsymbol{u}_j^k\right)\mathrm{d}\tau \\ &+ \int_t^{t'} 2\lambda_i\left(\Delta_i^{k+1}\right)^{\mathrm{T}}\boldsymbol{R}_{ii}\left(\boldsymbol{u}_i - \boldsymbol{u}_i^k\right)\mathrm{d}\tau - \int_t^{t'}\sum_{j\in N_i} e_{ij}\frac{2\lambda_i}{d_i + g_i}\left(\Delta_i^{k+1}\right)^{\mathrm{T}}\boldsymbol{R}_{ii}\left(\boldsymbol{u}_j - \boldsymbol{u}_j^k\right)\mathrm{d}\tau \end{aligned} \tag{10}$$

- The critic and actor NNs can be introduced

$$\hat{V}_i^k = \left(\phi_i\left(\boldsymbol{\delta}_i\right)\right)^{\mathrm{T}} \boldsymbol{W}_{vi}^k \tag{11}$$

$$\hat{\Delta}_{il_1}^k = \left(\varphi_{uil_1}\left(\boldsymbol{\delta}_i\right)\right)^{\mathrm{T}} \boldsymbol{W}_{uil_1}^k \tag{12}$$

- It can be derived that

$$\sigma_i^k = \left(\phi_i\left(\boldsymbol{\delta}_i\right) - \phi_i\left(\boldsymbol{\delta}_i'\right)\right)^{\mathrm{T}} \boldsymbol{W}_{vi}^k - \int_t^{t'} \boldsymbol{\delta}_i^{\mathrm{T}} \boldsymbol{Q}_{ii} \boldsymbol{\delta}_i \mathrm{d}\tau - \int_t^{t'} R_i\left(\hat{\boldsymbol{u}}_i^k\right)\mathrm{d}\tau - \int_t^{t'} \sum_{j\in N_i} R_i\left(\hat{\boldsymbol{u}}_j^k\right)\mathrm{d}\tau$$

$$+2\lambda_i \sum_{l_1=1}^{m}\sum_{l_2=1}^{m} r_{ii,l_1l_2} \int_t^{t'} \left(u_{il_1} + \lambda_i \Psi\left(\hat{\Delta}_{il_1}^k\right)\right)\left(\varphi_{uil_2}\left(\boldsymbol{\delta}_i\right)\right)^{\mathrm{T}} \boldsymbol{W}_{uil_2}^{k+1}\mathrm{d}\tau \tag{13}$$

$$-\frac{2\lambda_i}{d_i + g_i} \sum_{l_1=1}^{m}\sum_{l_2=1}^{m} r_{ii,l_1l_2} \int_t^{t'} \sum_{j\in N_i} e_{ij}\left(u_{jl_1} + \lambda_j \Psi\left(\hat{\Delta}_{jl_1}^k\right)\right)\left(\varphi_{uil_2}\left(\boldsymbol{\delta}_i\right)\right)^{\mathrm{T}} \boldsymbol{W}_{uil_2}^{k+1}\mathrm{d}\tau$$

Simplified form

$$\sigma_i^k = \boldsymbol{\rho}_{i,[t,t']}^k \boldsymbol{W}_i^{k+1} - \pi_{i,[t,t']}^k \tag{14}$$

$$\boldsymbol{W}_i^{k+1} = \left(\left(\boldsymbol{\Gamma}_i^k\right)^{\mathrm{T}} \boldsymbol{\Gamma}_i^k\right)^{-1} \left(\boldsymbol{\Gamma}_i^k\right)^{\mathrm{T}} \boldsymbol{\Pi}_i^k \tag{15}$$

Algorithm 1 Model-free off-policy reinforcement learning

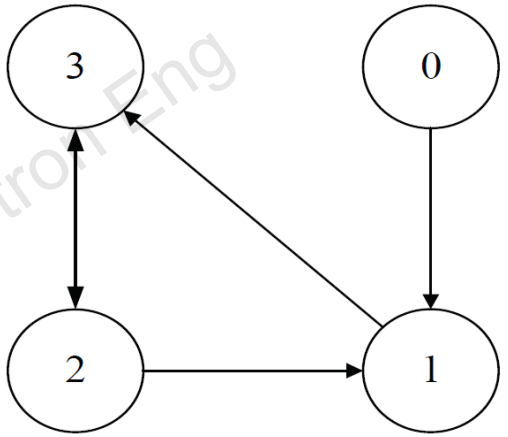
1. Choose the initial stabilizing network weights $\boldsymbol{W}_{uil_1}^0, i=1,2,\ldots,N, l_1=1,2,\ldots,m$
2. Employ the control law $\boldsymbol{u}_i, i=1,2,\ldots,N$, to the system on the time interval $[t_0,t_\upsilon]$. Collect the system data $\{\boldsymbol{\delta}_i,\boldsymbol{u}_i\}, i=1,2,\ldots,N$. Let $k=0$
3. Compute $\boldsymbol{\rho}_{i,[t_0,t_1]}^k,\ldots,\boldsymbol{\rho}_{i,[t_{\upsilon-1},t_\upsilon]}^k$ and $\pi_{i,[t_0,t_1]}^k,\ldots,\pi_{i,[t_{\upsilon-1},t_\upsilon]}^k$, and compute $\boldsymbol{W}_i^{k+1}$
4. Let iteration index $k=k+1$; repeat step 3 until $\left\|\boldsymbol{W}_i^{k+1}-\boldsymbol{W}_i^k\right\|\le\varepsilon$
5. Return $\boldsymbol{W}_i^k, i=1,2,\ldots,N$

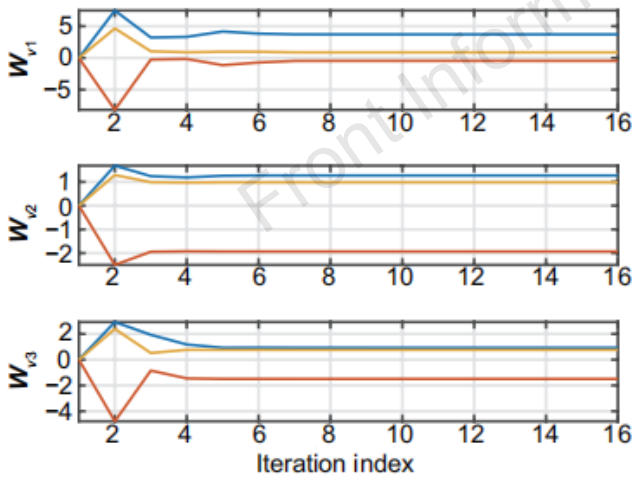
Lemma 1 If the iterative value functions and iterative control laws are designed as Eqs. (11) and (12), where $\boldsymbol{W}_{vi}^k$ and $\boldsymbol{W}_{uil_1}^k, l_1=1,2,\ldots,m$ are updated as Algorithm 1, then $\lim_{k\to\infty}\hat{\boldsymbol{u}}_i^k=\boldsymbol{u}_i^*, i=1,2,\ldots,N$.

Numerical analysis

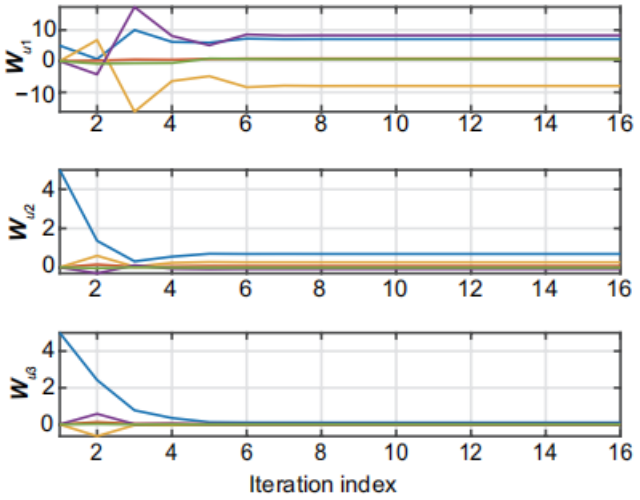
- System matrices $\boldsymbol{A}=\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}, \boldsymbol{B}=\begin{bmatrix}1\\ 1\end{bmatrix}$

- $\boldsymbol{L}=\begin{bmatrix}1 & -1 & 0\\ 0 & 1 & -1\\ -1 & -1 & 2\end{bmatrix}, \boldsymbol{G}=\begin{bmatrix}1 & & \\ & 0 & \\ & & 0\end{bmatrix}$





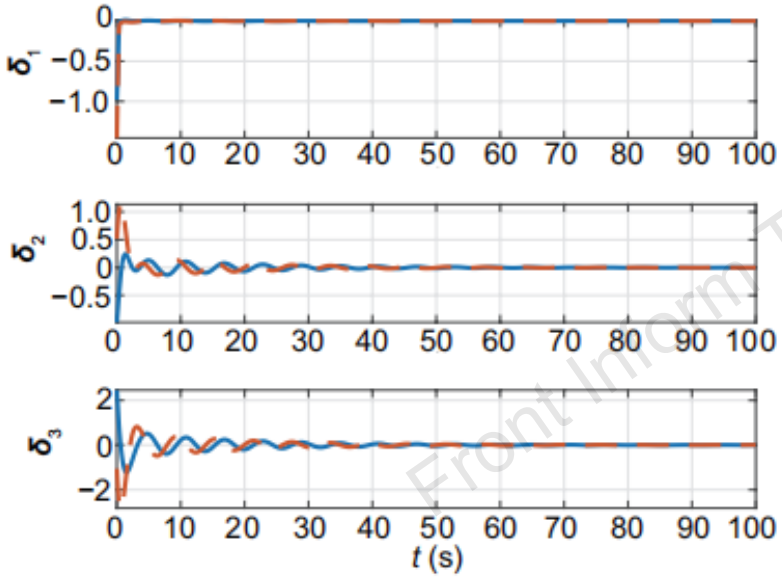
Critic NNs



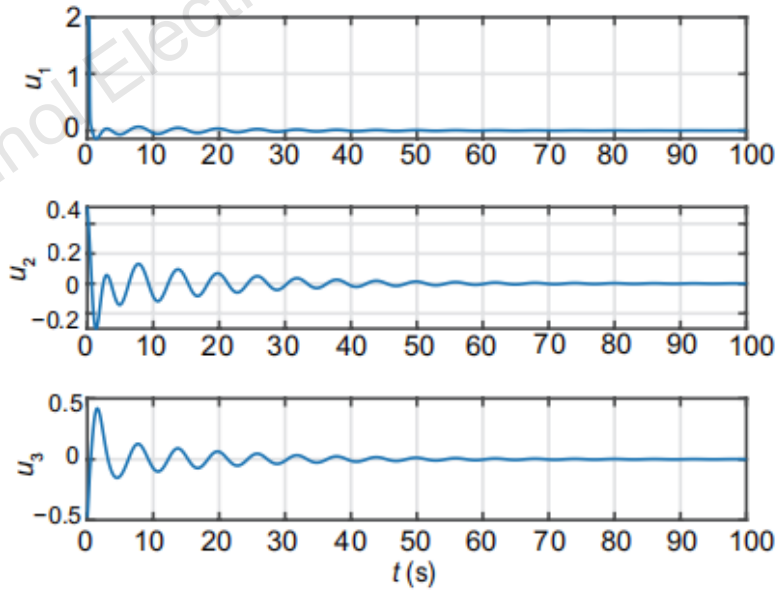
Actor NNs

Numerical analysis (Cont'd)

- $\lambda_1 = 2, \lambda_2 = 1.5, \lambda_3 = 3$



Synchronization errors



Control laws

Conclusions

The nonzero-sum game problem of multi-agent systems with input saturation has been studied based on the model-free off-policy reinforcement learning method. It is shown that the presented off-policy reinforcement learning algorithm can make the iterative control laws converge to the Nash equilibrium without the information of system models. The simulation results showed the good performance of the presented method.

Qinglai WEI received his BS degree in automation, MS degree in control theory and control engineering, and PhD degree in control theory and control engineering, from Northeastern University, Shenyang, China, in 2002, 2005, and 2008, respectively. He is an Associate Editor-in-Chief of *Neurocomputing*. He has been a vice president of the IEEE Computational Intelligence Society Beijing Chapter since 2021. His research interests include adaptive dynamic programming, neural-network-based control, optimal control, nonlinear systems, and their industrial applications.