Nash Equilibrium Solution based on Safety-guarding Reinforcement Learning in Nonzero-sum Game

非零和博弈下安全保障强化学习的纳什均衡

Junkai Tan, Shuangsi Xue, Hui Cao and Huan Li

Electrical Engineering, Xi'an Jiaotong University

2023 年 7 月 9 日

1. Preliminaries and Problem Formulation
2. Safety-Guarding Controller Design
3. Online Approximation Based on RL
4. Simulation Results

1. Preliminaries and Problem Formulation
2. Safety-Guarding Controller Design
3. Online Approximation Based on RL
4. Simulation Results

Control System and Performance Index：控制系统、性能指标

Firstly, the object of study and some basic definitions are introduced: a nonlinear affine system, and the performance J:

Nonlinear Affine System（非线性仿射系统）

$$\dot{x} = f(x) + \sum_{i=1}^{N} g_i(x) u_i \tag{1}$$

Performance Index（性能指标）

$$J_i\left(x_0, u(\cdot)\right) \triangleq \int_t^{\infty} r_i(x(t), u) dt = \int_0^{\infty} (Q_i(x) + \sum_{j=1}^{N} u_j^T R_{ij} u_j) \mathrm{d}t \tag{2}$$

Problem： Finding Nonzero-sum Game Optimal Controller

The following optimization problem is established and theoretically analyzed and solved

Multi-Player nonzero-sum games：（多智能体非零和博弈）

1. Value Function（价值函数）：

$$V_i^*\left(x_0\right) \triangleq \min_{u(\cdot)\in\mathcal{S}(x_0)} J_i\left(x_0, u(\cdot)\right), \quad x_0 \in \mathbb{R}^n \tag{3}$$

2. Hamiltanian（汉密尔顿算子）：

$$H\left(x, u, V_i^{\prime\mathrm{T}}(x)\right) \triangleq r_i(x(t), u_1, ..., u_N) + (\Delta V_i^{*\ T}(f(x) + \textstyle\sum_{j=1}^{N} g_j(x) u_i) \tag{4}$$

3. Pontryagin's theory extremum condition is the optimal controller

$$u_i^\star(x) \triangleq \arg\min_{u\in\mathbb{R}^m} H\left(x, u_i, V_i^{\prime\mathrm{T}}(x)\right) = -\frac{1}{2} R_{ii}^{-1} g_i^T (\Delta V_i)^T \tag{5}$$

1. Preliminaries and Problem Formulation
2. Safety-Guarding Controller Design
3. Online Approximation Based on RL
4. Simulation Results

Safety-Guaranteed Controller: 障碍函数安全控制器

In real-world safety-critical systems, controllers that lack security are generally difficult to apply, so this research will theoretically analyze and design a safety-guaranteed controller

1. Definition: Safety State Region c （安全状态域）

Define c be the safety state region and $h(x)$ be the boundary function[1], that is, when $x \in c$ (security domain), $h(x) \geq 0$; when $x \to \partial c$ (security boundary), $h(x) \to 0$

$$
\begin{aligned}
c &= \{x \in \mathbb{R}^n \mid h(x) \geq 0\} \\
\partial c &= \{x \in \mathbb{R}^n \mid h(x) = 0\} \\
\mathrm{Int}(c) &= \{x \in \mathbb{R}^n \mid h(x) > 0\}
\end{aligned}
$$

[1] Keng Peng Tee, Shuzhi Sam Ge, and Eng Hock Tay. "Barrier Lyapunov functions for the control of output-constrained nonlinear systems". In: *Automatica* 45.4 (2009), pp. 918–927.

Safety-Guaranteed Controller: 障碍函数安全控制器

2. Definition: Barrier Function（障碍函数）

Using the function $h(x)$ associated with the safety state region c given above, the barrier function of the following specific form is designed：

$$b(x) = [\frac{1}{h(x)} - \frac{1}{h(0)}]^2$$

The barrier function has 3 properties as follows：

- For $x(t) \in Int(c)$, $|b(x)| < \infty$
- $\lim_{x \to \partial c} b(x) = \infty$
- $b(0) = 0$

Safety-Guaranteed Controller: 障碍函数安全控制器

3. Innovation: Safety-Guaranteed Controller

Improved according to the results of the article[2] to get a strictly secure controller：

$$u_b(x) = -\alpha_i g_i(x)^T \Gamma\left(\nabla b(x)^T\right) \tag{6}$$

where α_i is the gain， Γ is the mapping that prevents the gradient of the barrier function $b(x)$ from tending to infinity

[2] Max H Cohen and Calin Belta. "Safe exploration in model-based reinforcement learning using control barrier functions". In: *Automatica* 147 (2023), p. 110684.

Safety-Guaranteed Controller: 障碍函数安全控制器

For multi-player non-zero-sum game systems, the Nash equilibrium controller and the safety-guaranteed controller are summed to obtain the integrated controller

4. Integrated Design: Regular Controller+Safety-Guaranteed Controller（纳什均衡控制器 + 严格保障安全控制器）

$$u_i^* = \underset{u_i}{argmin}\, V_i = -\frac{1}{2} R_{ii}^{-1} g_i^T (\Delta V_i)^T \tag{7}$$

$$u_{b,i} = u_i(x,t) + u_b(x) \tag{8}$$

1. Preliminaries and Problem Formulation
2. Safety-Guarding Controller Design
3. Online Approximation Based on RL
4. Simulation Results

Online ADP：在线迭代的自适应动态规划方法

Approximation based on single-layer neural networks

According to Weierstrass theorem, a single-layer neural network (NN) is designed to approximate the value function and controller.

- Ideal value function：
  $V_i(x) = W_i^{\star\mathrm{T}} \phi_i(x) + \epsilon_i(x)$
- Ideal controller：
  $u_i^\star = -\frac{1}{2} R_{ii}^{-1} g_i(x)^\mathrm{T} \left(\phi_i'^{\mathrm{T}}(x) W_i^\star + \epsilon_i'^{\mathrm{T}}(x) \right)$

Optimization objective: value function

Value function $V_i(x_0)$ is the extremum of performance $J_i(x_0, u(\cdot))$:

$$V_i\left(x_0\right) \triangleq \min_{u(\cdot)\in\mathcal{S}(x_0)} J_i\left(x_0, u(\cdot)\right) = \min_{u(\cdot)\in\mathcal{S}(x_0)} \int_0^\infty (Q_i(x)+\sum_{j=1}^{N} u_j^T R_{ij} u_j)\mathrm{d}t$$

Online ADP：在线迭代的自适应动态规划方法

In order to realize the online update iteration of neural network parameters, the Hamiltonian error value is set here as the base element of the update target

Hamiltonian Error（汉密尔顿误差）

$$\delta_i = \Omega_i^{\mathrm{T}}\sigma_i + x^{\mathrm{T}}Q_i x + \sum_{j=1}^{N}\frac{1}{4}\omega_j^{\mathrm{T}}\sigma_j' G_{ij}\sigma_j'^{\mathrm{T}}\omega_j + \nabla\epsilon_i^T\Omega_i \tag{9}$$

Iteration object: normalized least squares Hamiltonian error

$$E_i = \frac{1}{2}\left[\frac{\sigma_i^2}{\left(1+\sigma_i^T\sigma_i\right)^2} + \sum_{k=1}^{M}\frac{\left(\sigma_i^k\right)^2}{\left(1+\left(\sigma_i^k\right)^T\sigma_i^k\right)^2}\right] \tag{10}$$

Online ADP：在线迭代的自适应动态规划方法

According to the paper[3] proposed Concurrent Learning, the network parameters are updated both using current data and historical data, and the update law is shown below：

the update law

$$\begin{aligned}\dot{\hat{\omega}}_i &= -\beta_i \frac{\partial E_i}{\partial \omega_i} \\ &= -\beta_i \frac{\sigma_i e_i}{\left(1+\sigma_i^T \sigma_i\right)^2} - \beta_i \sum_{k=1}^{M} \frac{\sigma_i^k e_i^k}{\left(1+(\sigma_i^k)^T \sigma_i^k\right)^2}\end{aligned} \tag{11}$$

[3] Girish Chowdhary and Eric Johnson. "Concurrent learning for convergence in adaptive control without persistency of excitation". In: *49th IEEE Conference on Decision and Control (CDC)*. IEEE. 2010, pp. 3674–3679.

Online ADP：在线迭代的自适应动态规划方法

Theorem1: Asymptotic stability

Network weights are asymptotically stable as following conditions are met:

$$\begin{cases} \bar{g}_i\bar{\phi}_j < 0 \\ \rho < 0 \\ \beta_i\left(\frac{p+1}{2} - 2\lambda_{\min}\left(\Gamma_k\right)\right) < 0 \end{cases} \tag{12}$$

where $\rho = \sum_{i=1}^{N}\left[\beta_i\frac{p+1}{2}\bar{\varepsilon}_i^2 - \left(\bar{\omega}_i\bar{\phi}_i + \bar{\epsilon}_i\right)\sum_{j=1}^{N}\left(\frac{1}{2}G_j\bar{\phi}_i\|\hat{\omega}_j\| - g_i\bar{\epsilon}_i\right)\right]$

Proof: Set the Lyapunov function

$$V_L = \sum_{i=1}^{N}\left(V_i + V_{\omega,i}\right) \tag{13}$$

其中 $V_{\omega,i} = \frac{1}{2}\tilde{\omega}_i^{\mathrm{T}}\tilde{\omega}_i$

Online ADP：在线迭代的自适应动态规划方法

According to the given assumptions, the following inequality holds:

$$\dot{V}_i \leq -r_i - \left(\overline{\omega}_i \overline{\phi}_i + \overline{\epsilon}_i\right) \sum_{j=1}^{N} \left(\frac{1}{2} G_j \overline{\phi}_i \|\hat{\omega}_j\| - g_i \overline{\epsilon}_i \right) \tag{14}$$

$$\dot{V}_{\omega,i} \leq \beta_i \left[\frac{p+1}{2} - 2\lambda_{\min}\left(\Gamma_k\right) \right] \|\tilde{\omega}_i\|^2 + \beta_i \frac{p+1}{2} \overline{\varepsilon}^2_{\text{hmax},i} \tag{15}$$

$$\begin{aligned} \dot{V} \leq & -\sum_{i=1}^{N} r_i + \rho \\ & + \sum_{i=1}^{N} \left[\overline{g}_i \overline{\phi}_i + \beta_i \left(\frac{p+1}{2} - 2\lambda_{\min}(\Gamma_k) \right) \right] \|\tilde{\omega}_i\|^2 \end{aligned} \tag{16}$$

So $\dot{V}_L \leq 0$ holds, i.e., the asymptotic stability of the weights is proved

1. Preliminaries and Problem Formulation

2. Safety-Guarding Controller Design

3. Online Approximation Based on RL

4. **Simulation Results**

Simulation Results

nonlinear affine control system for 2 players and its parameters (2 个智能体的非线性仿射控制系统)

$$\dot{x} = f(x) + g_1(x)u_1 + g_2(x)u_2 \tag{17}$$

where

$$f = \begin{bmatrix} x_2 - 2x_1 \\ -\frac{1}{2}x_1 - x_2 + \frac{1}{4}x_2\left(\cos\left(2x_1\right) + 2\right)^2 \\ +\frac{1}{4}x_2\left(\sin\left(4x_1^2\right) + 2\right)^2 \end{bmatrix}$$

$$g_1 = \begin{bmatrix} 0 \\ \cos\left(2x_1\right) + 2 \end{bmatrix} \quad g_2 = \begin{bmatrix} 0 \\ \sin\left(4x_1^2\right) + 2 \end{bmatrix}$$

Simulation Results

Selection of the parameters

- $Q_1 = 2Q_2 = 2I_2,\ R_{11} = R_{12} = 2R_{21} = 2R_{22} = 2$
- Unsafe boundary function $h(x) = px_2^2 - x_1 + 1$ ($p = -1$: convex boundary, $p = 1$: non-convex boundary)
- Initial state $x_0 = [-4, 2.2]$

Simulation

The tests are performed for the convex boundary and the non-convex boundary, respectively; the red trajectory is generated by the original controller and the yellow trajectory is generated by the safe-guaranteed controller

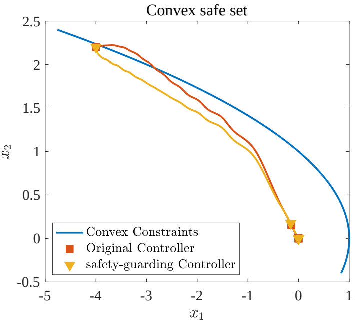

图 1: 凸性边界下的状态轨迹

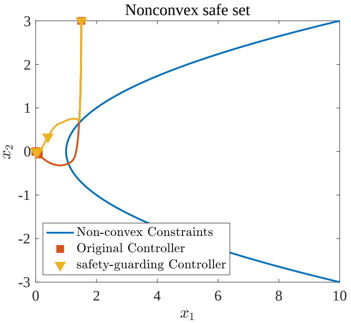

图 2: 非凸性边界下的状态轨迹

Simulation Resultss

The following figure shows the variation of network parameters for the two intelligences used for value function approximation, respectively, and the final convergence illustrates the convergence of the single-layer neural network obtained by iteration

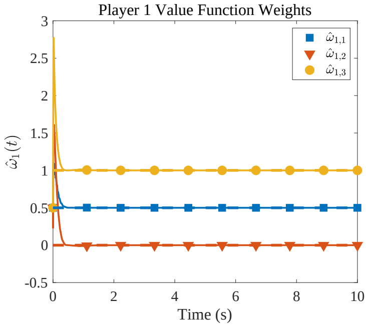

图 3: 智能体 1 的网络权重变化

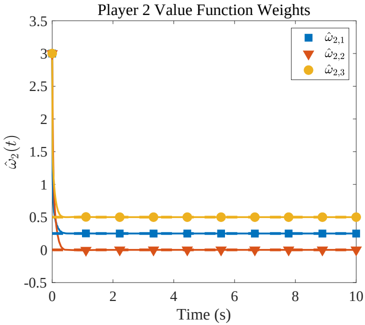

图 4: 智能体 2 的网络权重变化

参考文献 I

[1] Keng Peng Tee, Shuzhi Sam Ge, and Eng Hock Tay. “Barrier Lyapunov functions for the control of output-constrained nonlinear systems”. In: *Automatica* 45.4 (2009), pp. 918–927.

[2] Max H Cohen and Calin Belta. “Safe exploration in model-based reinforcement learning using control barrier functions”. In: *Automatica* 147 (2023), p. 110684.

[3] Girish Chowdhary and Eric Johnson. “Concurrent learning for convergence in adaptive control without persistency of excitation”. In: *49th IEEE Conference on Decision and Control (CDC)*. IEEE. 2010, pp. 3674–3679.

Thanks!

Thanks for your listening

感谢各位老师的聆听，请批评指正。