Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2023, 12(4), 373-381
Published Online November 2023 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2023.124040

Hans汉斯

基于奇异摄动强化学习的时变系统线性二次零和博弈研究

刘明相

广东工业大学自动化学院，广东 广州

收稿日期：2023年9月18日；录用日期：2023年11月21日；发布日期：2023年11月29日

摘　要

本研究探讨了时变系统中的线性二次零和博弈问题，与以往依赖系统模型的方法有所不同。本文提出了一种无模型的强化学习算法，用于寻找纳什均衡解。首先，通过奇异摄动理论，将时变动态博弈问题转化为两个定常系统的博弈问题。接着，利用无模型的强化学习算法，确定这两个定常系统的纳什均衡，进而近似求解了时变系统的纳什均衡解。本文提出的算法框架将为处理基于强化学习的时变系统鲁棒控制问题或信息物理系统的弹性控制问题提供新的研究思路。

关键词

强化学习，时变系统，博弈论，线性二次优化

Singular Perturbation-Based Reinforcement Learning for Time-Varying Linear Quadratic Zero-Sum Games

Mingxiang Liu

School of Automation, Guangdong University of Technology, Guangzhou Guangdong

Received: Sep. 18th, 2023; accepted: Nov. 21st, 2023; published: Nov. 29th, 2023

Abstract

This paper tackles the challenge of linear quadratic zero-sum games within dynamic systems that evolve over time. In contrast to previous methods that heavily rely on system models, this paper introduces a novel model-free reinforcement learning algorithm to determine Nash equilibrium

文章引用: 刘明相. 基于奇异摄动强化学习的时变系统线性二次零和博弈研究[J]. 人工智能与机器人研究, 2023, 12(4): 373-381. DOI: 10.12677/airr.2023.124040

solutions. To begin, the paper employs the singular perturbation theory to transform the time-varying dynamic game problem into two separate time-invariant dynamic game problems. Then, by leveraging a model-free reinforcement learning algorithm, it identifies Nash equilibria for these two time-invariant systems, effectively approximating the Nash equilibrium solution for the original time-varying system. The algorithm framework proposed in this paper introduces a fresh perspective for addressing robust control problems in dynamic systems with time variations. Additionally, it opens up new possibilities for robust control problems in time-varying systems or achieving resilient control in cyber-physical systems by harnessing the power of reinforcement learning.

Keywords

Reinforcement Learning, Time-Varying Systems, Game Theory, Linear Quadratic Optimization

Copyright © 2023 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

二人零和微分博弈研究涉及微分方程驱动的系统中的冲突问题[1]。这种类型的博弈经常涉及到追捕－逃避博弈，其中一方(通常是追捕者)试图在尽可能短的时间内将另一方(逃避者)引导到一个特定的目标位置，这类问题在航空航天等领域中经常出现[2]。另一个例子是带有干扰的最优控制问题，其中第二个参与者被视为干扰源，控制器必须努力优化系统的性能，同时考虑到干扰的存在。这种情况下，通常需要研究最坏情况设计，以确保系统在最不利的情况下仍然能够正常运行[3]，在经济学文献中，这种情况有时也被称为奈特氏不确定性[4]。

对于连续时间线性系统，解决线性二次零和微分博弈问题通常需要求解广义博弈代数里卡提方程[1]。通过将克莱曼算法扩展到连续时间线性双人零和博弈中[5]，可以近似地离线求解博弈问题。在过去的研究中，一些学者采用了不同的方法来解决这个问题。Lewis 和 Feng 等人在处理第一个控制器(即第一个玩家的动作)时使用了带有迭代的内部循环[6] [7]。而 Van der Schaft 以及 Abu-Khalaf 等人则设计了一种在处理第二个控制器(即第二个玩家的动作)时进行迭代的内部循环[8] [9]。尽管这些迭代算法可以近似求解博弈代数里卡提方程，但通常情况下，每个迭代步骤的代价函数仍然难以精确求解。以上研究均考虑定常系统，在过去的十年里，时变(动态)系统的控制和分析得到了广泛的研究。火箭着陆[10]，电子电路中的能量节约[11]等应用可归类为有限时域的时变系统。求解这类问题需要求解线性时变系统的时变里卡提方程或非线性系统的时变哈密顿－雅可比方程。与定常方程相比，求解时变方程则更加复杂和困难。此外，以上的算法都需要了解系统的全部或部分动力学知识。

强化学习最近在几个突出的决策问题上取得了令人瞩目的进展[12]，例如下围棋[13] [14]和玩实时策略游戏[15]。有趣的是，所有这些问题都可以表述为涉及两个对手或团队的零和马尔可夫博弈。强化学习被广泛应用于解决最优控制问题，如[16] [17] [18] [19]，它已被证明是处理具有未知动力学的线性或非线性系统的零和博弈问题的有效方法。文献[20]针对离散系统提出了对应的强化学习算法。文献[21]则针对连续系统提出了一种不使用任何系统动力学先验知识的线性二次零和博弈的强化学习算法。在[20]和[21]中提出的解决方案适用于在假设状态的全部知识可供反馈的情况下。文献[22] [23]提出基于输出反馈的强化学习算法用于处理线性二次零和博弈。值得注意的是，以上提到的强化学习算法都针对时不变系统，

在处理时变系统是则不适用。在文献[24]中，作者提出了一种针对离散时变系统的策略迭代法来寻找控制器。而文献[25]则关注了连续时间周期系统，文献[26]则是专注于无限时域情况下设计基于值迭代的学习控制器。此外，文献[27]针对时变的连续系统提出了强化学习方法。然而，对于有限时域内时变系统的博弈问题，目前尚未进行充分的相关研究。

为了解决上述问题，本文借鉴了开创性工作[28]中解决具有给定边界条件的有限视界时变问题的思想。其核心理念是：当成本函数需要在较长时间段内进行优化时，时变系统呈现出双时间尺度的特性。文献[28]的研究表明，这种时变系统可以被简化成两个定常系统。最终，原始系统的结果可以通过叠加解决初始边界问题和终端边界问题的结果来逼近。进一步地，我们应用了离线学习的思想[29]来估计这两个边界问题的纳什均衡。通过将原系统的复杂性简化成两个相对简单的问题，最终成功地无模型地学习了这两个独立的纳什均衡。该方法使得处理有限视界时变系统的博弈问题变得更加可行和高效。

综上所述，我们提出的模型的主要贡献如下

针对时变系统的线性二次动态博弈问题，基于奇异摄动的强化学习方法近似地学习到了纳什均衡的次优解。

为后续研究时变系统的 H_∞ 控制问题，时变的信息物理系统的弹性控制问题提供了强化学习求解的新思路。

本文的剩余部分结构安排如下：第二节描述时变系统线性二次动态博弈的问题。在这一节中，将详细讨论时变系统的性质以及线性二次动态博弈的背景和关键问题。第三节利用奇异摄动方法将原问题描述为两个双时间尺度的子问题。本节将介绍奇异摄动方法的应用，将复杂的问题分解为两个不同时间尺度下的子问题，以便更好地理解和解决。第四节概述了估计系统纳什均衡的强化学习算法。在这一节中，将介绍用于估计系统纳什均衡的强化学习算法的关键原理和方法。第五节给出本文的结论和未来进一步研究方向。

这一结构安排清晰地指导了读者在文章中的导向，使他们能够逐步理解问题、方法和结论，并为未来的研究提供了一个有益的参考点。

2. 问题描述

考虑由线性动力系统表示的时变系统如下

$$\frac{\mathrm{d}x}{\mathrm{d}t} = A(t)x(t) + B(t)u(t) + D(t)w(t) \tag{1}$$

其中，$x(t) \in R^n$ 为系统状态，$u(t) \in R^m$ 和 $w(t) \in R^d$ 分别表示玩家一和玩家二的控制输入，矩阵 $A(t) \in \mathbb{R}^{n\times n}$，$B(t) \in \mathbb{R}^{n\times m}$ 和 $D(t) \in \mathbb{R}^{n\times d}$ 分别表示关于时间 $t \in [0,T]$ 的光滑的函数且矩阵信息未知。玩家一(玩家二)的目标为最小化(最大化)如下所示有限时域的问题

$$J = \int_0^T x^T(t)Q(t)x(t) + u^T(t)R^u(t)u(t) - w^T(t)R^w(t)w(t)dt \tag{2}$$

其中，权值矩阵 $Q(t)$ 半正定，$R^u(t)$ 和 $R^w(t)$ 正定，且均为关于时间 $t \in [0,T]$ 的光滑的函数。$x(0) = x_0$ 和 $x(T) = x_T$ 分别表示初始状态和终端状态。动态零和博弈的问题为寻找鞍点 (u^*, w^*) 满足如下所示的纳什均衡不等式

$$J(x(0) \quad u^* \quad w) \le J(x(0) \quad u^* \quad w^*) \le J(x(0) \quad u \quad w^*) \tag{3}$$

假设 $(A(t);B(t))$ 可控，$(A(t);\sqrt{Q(t)})$ 可观[30]，T 相对较大，即相对于控制目标，系统变化较慢。

为求解该问题，定义如下的哈密尔顿函数

$$H = x^{\mathrm{T}}(t)Q(t)x(t) + u^{\mathrm{T}}(t)R^{u}(t)u(t) - w^{\mathrm{T}}(t)R^{w}(t)w(t) + \lambda^{\mathrm{T}}(t)\left(A(t)x(t) + B(t)u(t) + D(t)w(t)\right) \tag{4}$$

其中，$\lambda(t) \in \mathbb{R}^{n}$ 为协态变量，且满足如下方程

$$\dot{\lambda}(t) = -\nabla_{x}H = -Q(t)x(t) - A^{\mathrm{T}}(t)\lambda(t) \tag{5}$$

最小化目标函数(2)的动态系统(1)的纳什均衡解 $\left(u(t), w(t)\right)$ 由 $\nabla_{u}H = 0$ 和 $\nabla_{w}H = 0$ 决定

$$\begin{cases} u(t) = -\left(R^{u}(t)\right)^{-1}B(t)\lambda(t) \\ w(t) = \left(R^{w}(t)\right)^{-1}D(t)\lambda(t) \end{cases} \tag{6}$$

本文的目标为不需要知道系统矩阵 $A(t)$ 和 $B(t)$ 学习由(6)给出的时变系统(1)的纳什均衡解 $\left(u(t), w(t)\right)$。

3. 奇异摄动的设计

本节将根据文献[27] [28]的结果，利用奇异摄动的方法将时变系统的求解问题转化为初始边界和终端边界的博弈问题，当 T 足够长且系统矩阵 $A(t)$ 和 $B(t)$ 已知时，控制问题可实现次优解。这为下一节推导无模型的强化学习算法做铺垫。

通过引入缩放变量 τ，将时间段 0 到 T 归一化为区间 $[0,1]$

$$\tau = \frac{t}{T} \tag{7}$$

定义

$$\varepsilon = \frac{1}{T} \tag{8}$$

结合公式(1) (5)可得如下系统

$$\varepsilon\begin{bmatrix} \frac{dx}{d\tau} \\ \frac{d\lambda}{d\tau} \end{bmatrix} = \begin{bmatrix} A(\tau) & 0 \\ -Q(\tau) & -A^{T}(\tau) \end{bmatrix}\begin{bmatrix} x \\ \lambda \end{bmatrix} + \begin{bmatrix} B(\tau) \\ 0 \end{bmatrix}u + \begin{bmatrix} D(\tau) \\ 0 \end{bmatrix}w \tag{9}$$

对应的代价函数为

$$J = T * \int_{0}^{1} x^{\mathrm{T}}(\tau)Q(\tau)x(\tau) + u^{\mathrm{T}}(\tau)R^{u}(\tau)u(\tau) - w^{\mathrm{T}}(\tau)R^{w}(\tau)w(\tau)\mathrm{d}t \tag{10}$$

对应的鞍点表达式为

$$\begin{cases} u(\tau) = -\left(R^{u}(\tau)\right)^{-1}B(\tau)\lambda(\tau) \\ w(\tau) = \left(R^{w}(\tau)\right)^{-1}D(\tau)\lambda(\tau) \end{cases} \tag{11}$$

定义哈密尔顿矩阵

$$H_{M} = \begin{bmatrix} A(\tau) & -B\left(R^{u}(\tau)\right)^{-1}B^{\mathrm{T}}(\tau) - D\left(R^{w}(\tau)\right)^{-1}D^{\mathrm{T}}(\tau) \\ -Q(\tau) & -A^{\mathrm{T}}(\tau) \end{bmatrix} \tag{12}$$

假设 H_{M} 对任意时间 $\tau \in [0,1]$ 均偏离虚轴。

参考文献[31]，对系统(9)进行解耦，结果如下所示

$$\begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} I & I \\ P_a(\tau,\varepsilon) & P_b(\tau,\varepsilon) \end{bmatrix} \begin{bmatrix} x_a \\ x_b \end{bmatrix} \tag{13}$$

[引理 2.3，[31]]中表明，在本文的假设下，对于足够小的 ε，矩阵(12)是非奇异的。因此，结合(12)，系统(9)可以转换为奇异摄动系统如下

$$\varepsilon \frac{\mathrm{d}x_a}{\mathrm{d}\tau} = A(\tau)x_a + B(\tau)u_a + D(\tau)w_a \tag{14}$$

$$\varepsilon \frac{\mathrm{d}x_b}{\mathrm{d}\tau} = A(\tau)x_b + B(\tau)u_b + D(\tau)w_b \tag{15}$$

其中，

$$u_a = -\left(R^u(\tau)\right)^{-1} B(\tau)P_a(\tau,\varepsilon)x_a \tag{16}$$

$$u_b = -\left(R^u(\tau)\right)^{-1} B(\tau)P_b(\tau,\varepsilon)x_b \tag{17}$$

$$w_a = \left(R^w(\tau)\right)^{-1} D(\tau)P_a(\tau,\varepsilon)x_a \tag{18}$$

$$w_b = \left(R^w(\tau)\right)^{-1} D(\tau)P_b(\tau,\varepsilon)x_b \tag{19}$$

此处的 $P_a(\tau,\varepsilon) \geq 0, P_b(\tau,\varepsilon) \leq 0$ 为以下微分里卡提方程的两个根

$$\varepsilon\dot{P} = -A^{\mathrm{T}}(\tau)P - PA(\tau) + Q(\tau) + P\left(B(\tau)\left(R^u(\tau)\right)^{-1}B^{\mathrm{T}}(\tau) - D(\tau)\left(R^w(\tau)\right)^{-1}D^{\mathrm{T}}(\tau)\right)P \tag{20}$$

接下来，考虑时间尺度的变化，将奇异摄动系统(14)~(15)转化为边界系统，定义变量如下

$$\gamma = \frac{\tau}{\varepsilon}, \beta = \frac{1-\tau}{\varepsilon} \tag{21}$$

结合(21)，令奇异摄动系统中的(14)~(20)中参数 ε 趋于零，可得初始边界系统如下

$$\frac{\mathrm{d}x_a}{\mathrm{d}\gamma} = A(0)x_a + B(0)u_a + D(0)w_a \tag{22}$$

对应的反馈形式解为

$$u_a(\gamma) = K_a x_a = -\left(R^u(0)\right)^{-1} B^{\mathrm{T}}(0)P_a(0)x_a(\gamma) \tag{23}$$

$$w_a(\gamma) = L_a x_a = \left(R^w(0)\right)^{-1} D^{\mathrm{T}}(0)P_a(0)x_a(\gamma) \tag{24}$$

对应的代价函数为

$$J(x_a,u_a,w_a) = \int_0^\infty x_a^{\mathrm{T}}Q(0)x_a + u_a^{\mathrm{T}}R^u(0)u_a - w_a^{\mathrm{T}}R^w(0)w_a\mathrm{d}\gamma \tag{25}$$

同理可得终端边界系统如下

$$\frac{\mathrm{d}x_b}{\mathrm{d}\beta} = A(1)x_b + B(1)u_b + D(1)w_b \tag{26}$$

对应的反馈形式解为

$$u_b(\beta) = K_b x_b = -\left(R^u(1)\right)^{-1} B^{\mathrm{T}}(1)P_b(1)x_b(\beta) \tag{27}$$

$$w_b(\beta) = L_b x_b = \left(R^w(1)\right)^{-1} D^{\mathrm{T}}(1)P_b(1)x_b(\beta) \tag{28}$$

对应的代价函数为

$$J\left(x_b,u_b,w_b\right)=\int_0^\infty x_b^{\mathrm{T}}Q(1)x_b+u_b^{\mathrm{T}}R^u(1)u_b-w_b^{\mathrm{T}}R^w(1)w_b\mathrm{d}\beta \tag{29}$$

文献[31]中的定理 2.1 表明，如果解决了初始和终端这两个线性定常系统的博弈问题，当 ε 足够小时，两个边值问题解将近似于原始博弈问题(1)~(5)的解如下

$$x(\tau)=x_a(\gamma)+x_b(\beta)+\mathrm{O}(\varepsilon) \tag{30}$$

$$\lambda(\tau)=P_a(0)x_a(\gamma)+P_b(1)x_b(\beta)+\mathrm{O}(\varepsilon) \tag{31}$$

$$u(\tau)=K_ax_a(\gamma)+K_bx_b(\beta)+\mathrm{O}(\varepsilon) \tag{32}$$

$$w(\tau)=L_ax_a(\gamma)+L_bx_b(\beta)+\mathrm{O}(\varepsilon) \tag{33}$$

其中 $\mathrm{O}(\varepsilon)$ 为高阶项，在下一节中，我们将依赖于以上的结果，开发强化学习算法在无需系统矩阵 $A(t)$ 和 $B(t)$ 信息的情况下来求解初始和终端的博弈问题(22)~(29)，进而逼近时变系统的纳什均衡解。

4. 强化学习算法

在本节中，我们将提出强化学习算法以无模型的方式分别求解两个初始边界系统和终端边界系统对应的纳什策略，并结合文献[31]中的定理 2.1，逼近时变系统的纳什均衡解。

4.1. 初始边界的博弈问题

本节目标是在不了解系统动力学的情况下，学习(22)~(25)中所述系统的纳什均衡，从而纳什均衡优化了(25)中描述的成本函数。注意到 $P_a(0)$ 是代数 *Riccati* 方程的解：

$$0=-A^{\mathrm{T}}(0)P_a(0)-P_a(0)A(0)+Q(0)+P_a(0)\left(B(0)\left(R^u(0)\right)^{-1}B^{\mathrm{T}}(0)-D(0)\left(R^w(0)\right)^{-1}D^{\mathrm{T}}(0)\right)P_a(0) \tag{34}$$

首先回顾状态反馈积分强化学习方程如下[21] [29]

$$\begin{aligned}x_a^{\mathrm{T}}(t)P_a(0)x_a(t)=&\int_t^{t+\delta t}x_a^{\mathrm{T}}Q(0)x_a+\left(u_a+e_1\right)^{\mathrm{T}}R^u(0)\left(u_a+e_1\right)-\left(w_a+e_1\right)^{\mathrm{T}}R^w(0)\left(w_a+e_1\right)\mathrm{d}\upsilon\\&+x_a^{\mathrm{T}}(t+\delta t)P_a(0)x_a(t+\delta t)\end{aligned} \tag{35}$$

其中 e_1 和 e_2 为具有边界的探测噪声。

将上式表示为克罗内克积形式($\otimes$)如下

$$\psi^{\mathrm{T}}\begin{bmatrix}\left[vec\left(P_a^k\right)\right]\\\left[vec\left(K_a^{k+1}\right)\right]\\\left[vec\left(L_a^{k+1}\right)\right]\end{bmatrix}=\theta \tag{36}$$

其中，$\theta=\int_t^{t+\delta t}x_a^{\mathrm{T}}Q(0)x_a+u_a^{\mathrm{T}}R^u(0)u_a-w_a^{\mathrm{T}}R^w(0)w_a\mathrm{d}\upsilon$

$$\psi=\left[x_a^{\mathrm{T}}\otimes x_a^{\mathrm{T}}\quad 2\int_t^{t+\delta t}\left(x_a\otimes e_1\right)^{\mathrm{T}}\mathrm{d}\upsilon\left(I_n\otimes R^u(0)\right)\quad 2\int_t^{t+\delta t}\left(x_a\otimes e_2\right)^{\mathrm{T}}\mathrm{d}\upsilon\left(I_n\otimes R^w(0)\right)\right]^{\mathrm{T}}$$

$$\left[x_a^{\mathrm{T}}\otimes x_a^{\mathrm{T}}\Big|_t^{t+\delta t}\right]\left[vec\left(P_a^k\right)\right]=\left[x_a^{\mathrm{T}}\otimes x_a^{\mathrm{T}}\Big|_t^{t+\delta t}\right]\left[vec\left(K_a^k\right)\right]$$

由于方程(36)为一维方程，所以无法保证解的唯一性。本文将使用最小二乘法来解决系统参数未知的问题，对任意正整数 N，$\Phi=\left[\psi_1,\cdots,\psi_N\right]$，$\Theta=\left[\theta_1,\ \ldots,\ \theta_N\right]^T$，可得如下 N 维方程

$$\Phi^{\mathrm{T}}\begin{bmatrix}\left[vec\left(P_a^k\right)\right]\\ \left[vec\left(K_a^{k+1}\right)\right]\\ \left[vec\left(L_a^{k+1}\right)\right]\end{bmatrix}=\Theta \tag{37}$$

当样本数量 $N\geq\frac{n(n+1)}{2}+nm_1+nm_2$ 时， Φ^{T} 列满秩，则可以得到参数如下所示

$$\begin{bmatrix}\left[vec\left(P_a^k\right)\right]\\ \left[vec\left(K_a^{k+1}\right)\right]\\ \left[vec\left(L_a^{k+1}\right)\right]\end{bmatrix}=\left(\Phi\Phi^{\mathrm{T}}\right)^{-1}\Phi\Theta \tag{38}$$

在学习过程结束时(即收敛结束时)，反馈增益为 $K_a=K_a^{k+1}$ 和 $L_a=L_a^{k+1}$，则初始边界系统(22)的纳什均衡为 $\left(K_a x_a, L_a x_a\right)$。

4.2. 终端边界的博弈问题

遵循与初始边界的博弈问题相同的步骤，反馈增益矩阵的初始化值应该满足 $K_b>0, L_b<0$。则终端边界系统(26)~(29)对应的反馈纳什均衡为 $\left(K_b x_b, L_b x_b\right)$。

综合以上阐述，提出强化学习算法如下所示

针对时变 LQ 零和博弈的强化学习策略迭代纳什均衡搜索

选择容许的策略 $K_a<0, L_a>0$。

初始化：玩家一的策略 $u_a^0=K_a^0 x_a+e_1$，玩家二的策略 $w_a^0=L_a^0 x_a+e_2$

收集数据：作用于终端边界系统，在 $\gamma\in\left[\gamma_0,\gamma_N\right]$，其中 N 为在采样周期 $\gamma_{j+1}-\gamma_j=\delta$ 下的采样数据量，其中 $j=1,2,\cdots,N$，构造数据矩阵 Φ,Θ。

循环

通过求解问题(37)得到 P_a^k， K_a^{k+1}， L_a^{k+1}

如果 $\left\|P_a^k-P_a^{k-1}\right\|<\mu$，其中 μ 为判定阈值，则

P_a^k 作为最优的解 P_a^*

否则

$k\leftarrow k+1$

结束

结束循环

$u_a=K_a^* x_a$， $w_a=L_a^* x_a$

选择容许的策略 $K_b>0, L_b<0$。

初始化：玩家一的策略 $u_b^0=K_b^0 x_b+e_1$，玩家二的策略 $w_b^0=L_b^0 x_b+e_2$

收集数据：作用于初始边界系统，在 $\gamma\in\left[\gamma_0,\gamma_N\right]$，其中 N 为在采样周期 $\gamma_{j+1}-\gamma_j=\delta$ 下的采样数据量，其中 $j=1,2,\cdots,N$，构造数据矩阵 Φ,Θ。

Continued

循环

通过求解问题(37)得到 P_b^k， K_b^{k+1}， L_b^{k+1}

如果 $\left\|P_b^k - P_b^{k-1}\right\| < \mu$，其中 μ 为判定阈值，则

P_b^k 作为最优的解 P_b^*

否则

$k \leftarrow k+1$

结束

结束循环

$u_b = K_b^* x_b$， $w_b = L_b^* x_b$

本节中描述的学习过程，结合文献[31]中的定理 2.1 中的结果，只要 ε 足够小或 T 足够大，强化学习学习得到的最优策略 $u_{learned} = u_a + u_b$ 和 $w_{learned} = w_a + w_b$，该结果可近似原始时变系统的纳什均衡。

5. 结论与展望

本研究针对有限时间内的时变线性二次动态博弈问题，提出了一种无模型的强化学习算法。首先，利用奇异摄动理论，将有限时间内的时变系统转化为两个时间尺度的无限时间内的定常系统。随后，我们引入策略迭代的强化学习算法，分别求解这两个动态系统的对应纳什均衡解。最终，通过奇异摄动理论的应用，逼近原始系统的纳什均衡解。未来的研究方向包括将本文所提出的算法框架用于解决时变系统的 H_∞ 控制问题或信息物理系统的弹性控制问题。

参考文献

[1] Başar, T. and Olsder, G.J. (1998) Dynamic Noncooperative Game Theory. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611971132

[2] Ho, Y., Bryson, A. and Baron, S. (1965) Differential Games and Optimal Pursuit-Evasion Strategies. *IEEE Transactions on Automatic Control*, **10**, 385-389. https://doi.org/10.1109/TAC.1965.1098197

[3] Başar, T. and Bernhard, P. (2008) *H*∞-Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach. Birkhäuser, Boston. https://doi.org/10.1007/978-0-8176-4757-5

[4] Dow, J. and Werlang, S.R.D.C. (1994) Nash Equilibrium under Knightian Uncertainty: Breaking down Backward Induction. *Journal of Economic Theory*, **64**, 305-324. https://doi.org/10.1006/jeth.1994.1071

[5] Kleinman, D. (1968) On an Iterative Technique for Riccati Equation Computations. *IEEE Transactions on Automatic Control*, **13**, 114-115. https://doi.org/10.1109/TAC.1968.1098829

[6] Feng, Y., Anderson, B.D. and Rotkowitz, M. (2009) A Game Theoretic Algorithm to Compute Local Stabilizing Solutions to HJBI Equations in Nonlinear H_∞ Control. *Automatica*, **45**, 881-888. https://doi.org/10.1016/j.automatica.2008.11.006

[7] Vamvoudakis, K.G. and Lewis, F.L. (2012) Online Solution of Nonlinear Two-Player Zero-Sum Games Using Synchronous Policy Iteration. *International Journal of Robust and Nonlinear Control*, **22**, 1460-1483. https://doi.org/10.1002/rnc.1760

[8] Van Der Schaft, A.J. (1992) L/Sub 2/-Gain Analysis of Nonlinear Systems and Nonlinear State-Feedback H/Sub Infinity/Control. *IEEE Transactions on Automatic Control*, **37**, 770-784. https://doi.org/10.1109/9.256331

[9] Abu-Khalaf, M., Lewis, F.L. and Huang, J. (2006) Policy Iterations on the Hamilton-Jacobi-Isaacs Equation for H_∞ State Feedback Control with Input Saturation. *IEEE Transactions on Automatic Control*, **51**, 1989-1995. https://doi.org/10.1109/TAC.2006.884959

[10] Szmuk, M. and Acikmese, B. (2018) Successive Convexification for 6-DoF Mars Rocket Powered Landing with Free-Final-Time. 2018 *AIAA Guidance, Navigation, and Control Conference*, Kissimmee, 8-12 January 2018, 617-630. https://doi.org/10.2514/6.2018-0617

[11] Mahdavi, J., Emaadi, A., Bellar, M.D. and Ehsani, M. (1997) Analysis of Power Electronic Converters Using the Generalized State-Space Averaging Approach. *IEEE Transactions on Circuits and Systems I*: *Fundamental Theory and Applications*, **44**, 767-770. https://doi.org/10.1109/81.611275

[12] Sutton, R.S. and Barto, A.G. (2018) Reinforcement Learning: An Introduction. MIT Press, Cambridge.

[13] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Hassabis, D., *et al*. (2016) Mastering the Game of Go with Deep Neural Networks and tree Search. *Nature*, **529**, 484-489. https://doi.org/10.1038/nature16961

[14] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hassabis, D., *et al*. (2017) Mastering the Game of Go without Human Knowledge. *Nature*, **550**, 354-359. https://doi.org/10.1038/nature24270

[15] Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W.M. and Ewalds, T. (2019) AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. DeepMind Blog. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

[16] Vrabie, D., Pastravanu, O., Abu-Khalaf, M. and Lewis, F.L. (2009) Adaptive Optimal Control for Continuous-Time Linear Systems Based on Policy Iteration. *Automatica*, **45**, 477-484. https://doi.org/10.1016/j.automatica.2008.08.017

[17] Zhang, H., Luo, Y. and Liu, D. (2009) Neural-Network-Based Near-Optimal Control for a Class of Discrete-Time Affine Nonlinear Systems with Control Constraints. *IEEE Transactions on Neural Networks*, **20**, 1490-1503. https://doi.org/10.1109/TNN.2009.2027233

[18] Jiang, Y. and Jiang, Z.P. (2012) Computational Adaptive Optimal Control for Continuous-Time Linear Systems with Completely Unknown Dynamics. *Automatica*, **48**, 2699-2704. https://doi.org/10.1016/j.automatica.2012.06.096

[19] Jiang, Y., Shi, D., Fan, J., Chai, T. and Chen, T. (2022) Event-Triggered Model Reference Adaptive Control for Linear Partially Time-Variant Continuous-Time Systems with Nonlinear Parametric Uncertainty. *IEEE Transactions on Automatic Control*, **68**, 1878-1885. https://doi.org/10.1109/TAC.2022.3169847

[20] Al-Tamimi, A., Lewis, F.L. and Abu-Khalaf, M. (2007) Model-Free Q-Learning Designs for Linear Discrete-Time Zero-Sum Games with Application to H-Infinity Control. *Automatica*, **43**, 473-481. https://doi.org/10.1016/j.automatica.2006.09.019

[21] Li, H., Liu, D. and Wang, D. (2014) Integral Reinforcement Learning for Linear Continuous-Time Zero-Sum Games with Completely Unknown Dynamics. *IEEE Transactions on Automation Science and Engineering*, **11**, 706-714. https://doi.org/10.1109/TASE.2014.2300532

[22] Rizvi, S.A.A. and Lin, Z. (2018) Output Feedback Q-Learning for Discrete-Time Linear Zero-Sum Games with Application to the H-Infinity Control. *Automatica*, **95**, 213-221. https://doi.org/10.1016/j.automatica.2018.05.027

[23] Rizvi, S.A.A. and Lin, Z. (2020) Output Feedback Adaptive Dynamic Programming for Linear Differential Zero-Sum Games. *Automatica*, **122**, Article ID: 109272. https://doi.org/10.1016/j.automatica.2020.109272

[24] Pang, B., Bian, T. and Jiang, Z.P. (2019) Adaptive Dynamic Programming for Finite-Horizon Optimal Control of Linear Time-Varying Discrete-Time Systems. *Control Theory and Technology*, **17**, 73-84. https://doi.org/10.1007/s11768-019-8168-8

[25] Pang, B., Jiang, Z.P. and Mareels, I. (2020) Reinforcement Learning for Adaptive Optimal Control of Continuous-Time Linear Periodic Systems. *Automatica*, **118**, Article ID: 109035. https://doi.org/10.1016/j.automatica.2020.109035

[26] Pang, B. and Jiang, Z.P. (2020) Adaptive Optimal Control of Linear Periodic Systems: An Off-Policy Value Iteration Approach. *IEEE Transactions on Automatic Control*, **66**, 888-894. https://doi.org/10.1109/TAC.2020.2987313

[27] Reddy, V., Eldardiry, H. and Boker, A. (2022) Singular Perturbation-Based Reinforcement Learning of Two-Point Boundary Optimal Control Systems. 2022 *American Control Conference* (*ACC*), Atlanta, 8-10 June 2022, 3323-3328. https://doi.org/10.23919/ACC53348.2022.9867376

[28] Wilde, R. and Kokotovic, P. (1972) A Dichotomy in Linear Control Theory. *IEEE Transactions on Automatic control*, **17**, 382-383. https://doi.org/10.1109/TAC.1972.1099976

[29] Jiang, Y. and Jiang, Z.P. (2012) Robust Adaptive Dynamic Programming. In: Lewis, F.L. and Liu, D., Eds., *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, Wiley-IEEE Press, New York, 281-302. https://doi.org/10.1002/9781118453988.ch13

[30] Lewis, F.L., Vrabie, D. and Syrmos, V.L. (2012) Optimal Control. John Wiley & Sons, New York. https://doi.org/10.1002/9781118122631

[31] Kokotović, P., Khalil, H.K. and O'reilly, J. (1999) Singular Perturbation Methods in Control: Analysis and Design. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611971118