文章编号:1671-4598(2025)12-0096-09 DOI:10.16526/j.cnki.11-4762/tp.2025.12.013 中图分类号:TP391 文献标识码:A

基于逆强化学习的奇异摄动系统最优控制算法研究

沈敏胤[1,2]，刘 飞[1,2]
(1. 江南大学 轻工过程先进控制教育部重点实验室，江苏 无锡 214122；
2. 江南大学 自动化研究所，江苏 无锡 214122)

摘要：针对具有双时间尺度特性的奇异摄动系统最优控制，给出一种基于全阶模型直接求解的逆强化学习算法，对比传统的将原始奇异摄动系统经时间尺度分离为快慢两个时间尺度的复合控制方法，降低了问题求解的复杂度；首先设计了一种基于模型的策略迭代逆强化学习算法，利用系统动力学和最优控制策略增益来重构未知成本函数；在此基础上，采用无模型 off-policy 逆强化学习算法，仅依赖于系统显示的最优行为数据，无需系统动力学模型和最优控制策略增益的先验知识，即可准确重构成本函数，使系统能够跟踪学习最优行为，同时在存在探测噪声的情况下也能实现无偏估计，仿真算例实验验证了方法的有效性。

关键词：奇异摄动系统；逆强化学习；最优控制；off-policy；数据驱动控制

Research on Optimal Control Algorithm of SPSs Based on Inverse Reinforcement Learning

SHEN Minyin[1,2], LIU Fei[1,2]
(1. Key Laboratory for Advanced Process Control of Light Industry of the Ministry of Education, Jiangnan University, Wuxi 2l4122, China;
2. Institute of Automation, Jiangnan University, Wuxi 214122, China)

Abstract: For the optimal control in singular perturbed systems (SPSs) with double time scales, a new inverse reinforcement learning algorithm based on a full-order model is proposed. the traditional composite control method is that the original SPS is divided into fast time-scale systems and slow time-scale systems. A comparison of the traditional method with the learning algorithm is made, which shows that the complexity of solving the problem is reduced. Firstly, a model-based strategy iterative inverse reinforcement learning algorithm is designed, and using the system dynamics and optimal control strategy gains to reconstruct an unknown cost function. On this basis, a model-free off-policy inverse reinforcement learning algorithm is presented, which only relies on the optimal behavior data displayed by the system and can accurately reconstruct a cost function without the prior knowledge of the system dynamics model and the gain of the optimal control strategy, Finally, the system can track optimal behavior while realizing an unbiased estimation with detection noise. The effectiveness of the method is verified by a simulation example.

Keywords: SPS; inverse reinforcement learning; optimal control; off-policy; data-driven control

0 引言

在生产制造和其他行业中，如电力系统[1]、化学过程[2]和工业生产[3]等领域中，常常会遇到一类具有复杂动态特性的系统。这些系统往往由于质量，惯量，电导，电容等小参数的存在而展现出显著的多时间尺度特征[4]。为了准确地描述这类系统动态行为，具有多时间尺度性质的系统通常被建模为奇异摄动系统（SPSs, singular perturbed systems)。SPSs 由于其涉及快速和慢速时间尺度的耦合，常在计算设计时面临数值病态问题

收稿日期:2024-11-15； **修回日期**:2024-12-24。
作者简介:沈敏胤(2000-),男,硕士。
通讯作者:刘 飞(1965-),男,博士,教授。
引用格式:沈敏胤,刘 飞. 基于逆强化学习的奇异摄动系统最优控制算法研究[J]. 计算机测量与控制,2025,33(12):96-104.

和高维难题[5]。

为有效应对这些难题，过去的几十年间，研究者们不断探索并提出了众多针对 SPSs 的最优控制方法。文献［6］通过对代数 Riccati 方程进行精确分解，将原问题转化为慢速和快速时间尺度的两个降阶方程，并且采用牛顿法分别对这两个降阶方程进行求解，从而降低了计算复杂度，文献［7］将次优控制器的设计问题转化为求解一组线性矩阵不等式的问题，使得控制器的设计过程更加简洁。针对次优输出跟踪控制问题，文献［8］通过构造矩阵不等式条件，保证了跟踪系统的渐近稳定性，并使二次型性能指标最小化。这些研究均需要使用系统详细的动力学模型，但在实际应用中，系统的精确模型往往存在不确定性，甚至难以获取。

近年来，基于强化学习（RL，reinforcement learning）的数据驱动最优控制方法得到了广泛的研究。强化学习的核心在于通过不断的试错过程，使智能体通过与环境的交互来获取奖励信号，从而逐步调整自己的行动策略，能够在充满不确定性的环境中探索并寻求最优策略[9]。这一方法已经成功地应用于多个领域，如连续时间线性系统的最优控制问题[10]、具有部分动力学未知的非线性连续系统的最优控制问题[11]、路径规划[12]等方面。在 SPSs 方面，强化学习同样展现出了巨大的潜力。为了解决两个时间尺度工业过程的运行优化控制问题，文献［3］将该双时间尺度过程建模为 SPSs，并按时间尺度不同分为两部分，分别求解分解后的两个降阶代数 Riccati 方程。文献［13］将其推广的多时间尺度大工业过程中。文献［14］考虑了过程中扰动作用，并通过扩展状态变量的方式，将扰动作为额外的状态变量纳入考量，并利用学习方法来适应这些扰动。然而，这类方法未能消除用于激励系统的探测噪声影响，且需要频繁交互。为了消除探测噪声影响，文献［15］提出了 off-policy 强化学习。off-policy 强化学习在执行一个策略时，会根据另一个策略的数据来更新价值函数，从而避免了加在当前策略探索行为上的探测噪声对价值函数更新的干扰。文献［16］使用离线 off-policy 积分强化学习与奇异摄动方法相结合，对连续双时间尺度工业过程进行跟踪控制。针对离散时间系统，文献［17］通过奇异摄动与异轨强化学习结合的方法对其进行 H_∞ 跟踪控制。虽然这些基于尺度分离方法设计的复合控制器在一定程度上降低了计算复杂度，对于原系统而言仍然是一种次优控制策略[5]，且复合控制器设计过程比较复杂[18]。为了降低设计复杂度和提高精度，文献［19］提出了一种求解全阶 SPSs 基本代数 Riccati 方程的递归算法解决线性 SPSs 的最优控制问题，但该方法依赖于系统模型。在强化学习框架下，文献［18］则通过利用成本函数参数矩阵的结构特性，将原始的代数 Riccati 方程转化为广义形式，并提出了一种基于 Kleinman 算法的模型基算法。在迭代过程中，通过构造适定的线性代数方程来求解最优控制器。归纳以上文献，强化学习控制方法均高度依赖于一个事先明确设定的性能指标（也称为代价函数），限制了强化学习方法在实际应用中的灵活性和普适性。

鉴于实际工作中人为设定的性能指标常常受限于主观偏见和认知局限，常常难以满足实际需求。为了克服这一难题，研究者们提出了一种称为逆强化学习（IRL，inverse reinforcement learning）的方法[20]。这种方法能够通过分析一组专家演示的最优行为数据，来逆向推导出未知的奖励函数，从而实现对专家最优行为的学习模仿。然而，尽管 IRL 在理论上具有巨大的潜力，但在实际应用中，它却面临着学习过程不稳定性的挑战。借鉴基于模型的逆最优控制思想[21-22]，文献［23］将逆最优控制问题作为 IRL 的子问题来解决，保证了 IRL 的稳定性，进而获得研究进展[24-29]，并且近年来开始使用了 off policy IRL[25,27,29]。然而，值得注意的是，当前 IRL 的研究聚焦于单一时间尺度的系统，对于广泛存在于现实中的 SPSs 尚缺乏有效解决方案。SPSs 快慢状态分量的常常不可观测，而 IRL 依赖于观测最优行为数据。即便尝试用可测变量组替代快慢状态分量，也需深入依赖系统模型知识，至少需掌握快时间尺度子系统的模型。这意味着在无模型场景下，传统的基于时间尺度分离的方法设计复合控制器并不适用于 IRL。因此，如何在动力学模型未知的场景下有效地利用 IRL 解决 SPSs 的最优控制问题，成为了当前研究的一大难点和热点。

本文旨在未知代价函数的条件下，运用 IRL 方法研究线性 SPSs 的最优控制新算法：首先受到文献［10-11］的启发，本文与传统时间尺度分离[3,13-14,16-17]不同，采用全阶模型，直接针对具有双时间尺度特性的 SPSs 进行求解。既简化设计过程，避免了时间尺度分解，又降低了计算复杂度。其次，本文的线性全阶 SPSs 的 IRL 算法，利，直接从系统最优行为出发，通过逆向推理重构出未知的成本函数，避免了人为设计成本函数的繁琐和不精确。由于成本函数是学习得到的，因此具有更高的灵活性和适应性。最后，本文给出了基于数据驱动的 off-policy IRL 算法，完全依赖于系统的最优行为数据，无需依赖额外的系统动力学和最优控制增益信息，适应信息有限或模型不完全可知的复杂环境。

本文其余章节安排如下：在第 2 节，阐述线性 SPSs 及与之相关的最优控制的 IRL 问题。第 3 节给出基于模型的线性全阶 SPSs 的 IRL 算法。该算法在未知的代价函数下，通过利用系统内在的动力学特性和最优

控制增益，来重构未知的代价函数，从而跟踪学习系统最优行为。在此基础上，设计一种数据驱动的 IRL 算法，该算法能够仅依赖系统的最优行为数据，而无需系统动力学和控制增益的先验知识。第四部分展示了所提出的两种算法的仿真结果。最后部分对全文进行了总结。

1 符号说明

在本篇文章中，$\|\cdot\|$ 表示向量的欧几里德范数。矩阵 $\boldsymbol{I}_{ni}$ 代表 $ni\times ni$ 的单位矩阵，即矩阵 $\boldsymbol{I}_{n1}$ 表示 $n1\times n1$ 的单位矩阵，矩阵 $\boldsymbol{I}_{n2}$ 表示 $n2\times n2$ 的单位矩阵。对于矩阵 $\boldsymbol{X}$，$\boldsymbol{Y}\in\boldsymbol{R}^{n\times n}$，矩阵 $\boldsymbol{X}>0$（$\geqslant 0$）表示矩阵 $\boldsymbol{X}$ 是（半）正定矩阵。矩阵 $\boldsymbol{X}>\boldsymbol{Y}$（$\geqslant\boldsymbol{Y}$）表示矩阵 $\boldsymbol{X}-\boldsymbol{Y}>0\boldsymbol{X}>\boldsymbol{Y}$。rank（）表示矩阵的秩。λ（）表示矩阵的特征值。克罗内克积形式变化为 $\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{X}=(\boldsymbol{X}^T\otimes\boldsymbol{X}^T)\boldsymbol{Y}$，$\otimes$ 表示克罗内克积形式算子。对于矩阵 $\boldsymbol{X}\in\boldsymbol{R}^{n\times m}$，vec（$\boldsymbol{X}$）$=[x_{11}$，…，$y_{n1}$，$y_{12}$，…，$y_{n2}$，…，$y_{mm}]$。对于方阵 $\boldsymbol{Y}\in\boldsymbol{R}^{n\times n}$，vem（$\boldsymbol{Y}$）$=[y_{11}$，…，$y_{1n}$，$y_{21}$，…，$y_{2n}$，…，$y_{(n-1)n}$，$y_{mm}]$。

2 问题表述

考虑如下一个具有最优演示行为的连续线性 SPSs：

$$\begin{bmatrix}\dot{\bar{x}}_1(t)\\ \varepsilon\dot{\bar{x}}_2(t)\end{bmatrix}=\begin{bmatrix}\boldsymbol{A}_{11}&\boldsymbol{A}_{12}\\ \boldsymbol{A}_{21}&\boldsymbol{A}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_1(t)\\ \varepsilon\boldsymbol{x}_2(t)\end{bmatrix}+\begin{bmatrix}\boldsymbol{B}_1\\ \boldsymbol{B}_2\end{bmatrix}u^*(t)\quad(1)$$

其中：$0<\varepsilon\leqslant 1$ 是奇异摄动参数，$\boldsymbol{x}^T(t)=[\boldsymbol{x}_1(t)\quad\boldsymbol{x}_2(t)]^T\in\boldsymbol{R}^n$ 是状态变量，其中 $\boldsymbol{x}_1(t)\in\boldsymbol{R}^{n1}$ 与 $\boldsymbol{x}_2(t)\in\boldsymbol{R}^{n2}$ 分别为慢状态变量与快状态变量，$\boldsymbol{u}^*(t)\in R^m$ 为控制输入。$\boldsymbol{A}_{ij}$ 和 $\boldsymbol{B}_i$（i，$j=1$，2）是具有适当维数的矩阵。$[\boldsymbol{x}(t),\boldsymbol{u}^*(t)]$ 是该连续线性 SPSs 的最优行为轨迹。令：

$$\boldsymbol{A}_\varepsilon=\begin{bmatrix}I&\\ &\varepsilon^{-1}I\end{bmatrix}$$

$$\boldsymbol{A}=\begin{bmatrix}I&\\ &\varepsilon^{-1}I\end{bmatrix}\begin{bmatrix}\boldsymbol{A}_{11}&\boldsymbol{A}_{12}\\ \boldsymbol{A}_{21}&\boldsymbol{A}_{22}\end{bmatrix}=\begin{bmatrix}\boldsymbol{A}_{11}&\boldsymbol{A}_{12}\\ \varepsilon^{-1}\boldsymbol{A}_{21}&\varepsilon^{-1}\boldsymbol{A}_{22}\end{bmatrix}\quad(2)$$

$$\boldsymbol{B}_\varepsilon=\begin{bmatrix}I&\\ &\varepsilon^{-1}I\end{bmatrix}\boldsymbol{B}=\begin{bmatrix}I&\\ &\varepsilon^{-1}I\end{bmatrix}\begin{bmatrix}\boldsymbol{B}_1\\ \boldsymbol{B}_2\end{bmatrix}=\begin{bmatrix}\boldsymbol{B}_1\\ \varepsilon^{-1}\boldsymbol{B}_2\end{bmatrix}\quad(3)$$

假设 1：系数矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是未知的，奇异摄动参数 ε 是已知的。（$\boldsymbol{A}$，$\boldsymbol{B}$）是可控的。

注 1：奇异摄动参数是作为一个衡量系统快慢差异程度的关键时间尺度常数，其值取决于快系统中的最慢特征值与慢系统中的最快特征值之比[16]。在实际应用场景中，当缺乏现成的系统模型时，这些特征值往往是未知的。然而，可以通过一系列的实验手段大致估算出系统的奇异摄动参数[16]。因此，在众多实际系统中，奇异摄动参数往往被视为一个相对确定的值[14,30]。即便在某些情况下，某一特定的奇异摄动参数存在不确定性，也可以将其视为一个常数进行处理，并将不确定性部分合并到矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 中，这样做并不会丧失问题的一般性。因此，在 SPSs 时，假设这一奇异摄动参数是已知的是合理的[18]。矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 未知是为了之后数据驱动的 off policy IRL 算法的设计。

在假设 1 下，式（1）中的最优控制 $\boldsymbol{u}^*(t)$ 能使系统的二次型代价函数最小化：

$$J[\boldsymbol{x}(t)]=\int_t^\infty(\boldsymbol{x}^T\boldsymbol{Q}^*\boldsymbol{x}+\boldsymbol{u}^{*T}\boldsymbol{R}^*\boldsymbol{u}^*)\mathrm{d}\tau=\boldsymbol{x}^T(t)\boldsymbol{P}_\varepsilon^*\boldsymbol{x}(t)\quad(4)$$

其中：$\boldsymbol{Q}^*=\boldsymbol{Q}^{*T}\in\boldsymbol{R}^{n\times n}$ 为状态处罚矩阵，$\boldsymbol{R}^*\in\boldsymbol{R}^{m\times m}>0$ 为控制加权矩阵，$\boldsymbol{P}_\varepsilon^*=\begin{bmatrix}\boldsymbol{P}_{11}^*&\varepsilon\boldsymbol{P}_{21}^{*T}\\ \varepsilon\boldsymbol{P}_{21}^*&\varepsilon\boldsymbol{P}_{22}^*\end{bmatrix}$ 为对称矩阵，其中 $\boldsymbol{P}_{11}^*\in\boldsymbol{R}^{n1\times n1}$，$\boldsymbol{P}_{21}^*\in\boldsymbol{R}^{n2\times n1}$，$\boldsymbol{P}_{22}^*\in\boldsymbol{R}^{n2\times n2}$。（$\boldsymbol{Q}^*$，$\boldsymbol{A}_\varepsilon$）是可观测的，（$\boldsymbol{A}_\varepsilon$，$\boldsymbol{B}_\varepsilon$）是可控的。

根据最优控制理论，能使代价函数（4）最小化的最优控制器为：

$$\boldsymbol{u}^*(t)=-\boldsymbol{K}_\varepsilon^*\boldsymbol{x}(t)=-(\boldsymbol{R}^*)^{-1}\boldsymbol{B}_\varepsilon^T\boldsymbol{P}_\varepsilon^*\boldsymbol{x}(t)\quad(5)$$

其中：$\boldsymbol{P}_\varepsilon^*$ 是下列连续代数 Riccati 方程的唯一稳定解：

$$\boldsymbol{A}_\varepsilon^T\boldsymbol{P}_\varepsilon^*+\boldsymbol{P}_\varepsilon^*\boldsymbol{A}_\varepsilon-\boldsymbol{P}_\varepsilon^*\boldsymbol{B}_\varepsilon\boldsymbol{R}^{*-1}\boldsymbol{B}_\varepsilon^T\boldsymbol{P}_\varepsilon^*+\boldsymbol{Q}^*=0\quad(6)$$

假设 2：公式（4）中的权重（$\boldsymbol{Q}^*$，$\boldsymbol{R}^*$）都是未知的满秩矩阵，但最优行为［$\boldsymbol{x}$（t），$\boldsymbol{u}^*$（t）］是可以观察到并且是已知。

定义 1：给定任意 $\boldsymbol{R}=\boldsymbol{R}^T\in\boldsymbol{R}^{m\times m}>0$，如果存在权重 $\boldsymbol{Q}=\boldsymbol{Q}^T\in\boldsymbol{R}^{n\times n}\geqslant 0$ 使得（$\boldsymbol{Q}$，$\boldsymbol{R}$）在连续代数 Riccati 方程（6）中生成唯一的稳定的 $\boldsymbol{P}_\varepsilon=\boldsymbol{P}_\varepsilon^T\in\boldsymbol{R}^{n\times n}\geqslant 0$，使其在（5）相同的最优控制增益 $\boldsymbol{K}_\varepsilon^*$。将这个权重 $\boldsymbol{Q}$ 称为 $\boldsymbol{Q}^*$ 的等价权重，相对应地此时 $\boldsymbol{P}_\varepsilon$ 等价于 $\boldsymbol{P}_\varepsilon^*$。于是为系统重构未知成本函数的任务就转变为（4）中的权重 $\boldsymbol{Q}^*$ 寻找一个等价权重 $\boldsymbol{Q}$。

定义 2：只要存在一个小的正常数 σ，使得 $\|\boldsymbol{X}-\boldsymbol{X}\|<\sigma$，则向量或矩阵 $\boldsymbol{X}$ 是对它的期望值 $\boldsymbol{X}^*$ 的一致近似解。

问题 1：考虑假设 1 与假设 2。选择任意权重 $\boldsymbol{R}=\boldsymbol{R}^T\in\boldsymbol{R}^{m\times m}>0$，利用观测到的行为数据据［$\boldsymbol{x}$（$t$），$\boldsymbol{u}^*$（$t$）］，为系统（1）找到一个定义 1 中的等价权重，使其能产生与（5）中的最优控制策略增益，从而使系统跟踪学习最优行为［$\boldsymbol{x}$（t），$\boldsymbol{u}^*$（t）］。

3 SPSs 最优控制问题的 IRL 算法

本部分首先提出基于模型的方法来求解问题 1，在此基础上，进一步设计了一种数据驱动的 off policy 逆强化学习算法。该算法仅利用观测到的行为数据［$\boldsymbol{x}$（t），$\boldsymbol{u}^*$（t）］，而无需依赖系统模型中的矩阵 $\boldsymbol{A}$ 与矩阵 $\boldsymbol{B}$ 以及最优控制增益等额外信息。通过这样的设计，本

文算法不仅更加灵活和实用，同时也降低了对系统内部信息的依赖，为处理复杂、不确定环境下的学习问题提供了新的解决方案。

3.1 基于模型 IRL 迭代的奇异摄动系统最优控制方法

本节给出基于模型 IRL 迭代的奇异摄动系统最优控制算法，该算法利用系统动力学 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 和最优控制增益 $\boldsymbol{K}_\varepsilon^*$，以找到与 $\boldsymbol{Q}^*$ 等效的权重 $\boldsymbol{Q}$，从而获得与最优控制增益 $\boldsymbol{K}_\varepsilon^*$ 相同的控制增益。该算法包含 3 个步骤：策略评估、使用逆最优控制进行权重改进和最优控制策略改进。

鉴于奇异摄动参数 ε 非常接近于零，直接求解可能会引发病态数值问题。传统的复合控制方法是将系统（1）按照不同的时间尺度进行分解，然后分别为两个降阶子系统的设计相应的控制算法。当 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 完全未知时，由于存在慢快耦合，所以设计复合算法将会变得更加复杂。因此，这里使用全阶模型来解决问题 1，从而避免了复合控制方法的复杂性。为了避免奇异摄动参数 ε 的影响，通过以下的 Kleinman 算法来求解 $\boldsymbol{P}_\varepsilon$。

引理 1（Kleinman 算法）[31]：设 $\boldsymbol{K}_0$ 为任意稳定的增益矩阵，$\boldsymbol{P}_\varepsilon^i$ 为如下李雅普诺夫方程的解：

$$\boldsymbol{A}_{\varepsilon i}^T\boldsymbol{P}_\varepsilon^i + P_\varepsilon^i A_{\varepsilon i} + K_i^T R^* \boldsymbol{K}_i + Q^* = 0 \tag{7}$$

其中：$A_{\varepsilon i} = A_\varepsilon - B_\varepsilon K_i$：

$$\boldsymbol{K}_i = (\boldsymbol{R}^*)^{-1}\boldsymbol{B}_\varepsilon^T\boldsymbol{P}_\varepsilon^{i-1} \tag{8}$$

则 $\lim\limits_{i\to\infty}\boldsymbol{P}_\varepsilon^i = \boldsymbol{P}_\varepsilon^*$，并且 $\boldsymbol{A}_{\varepsilon i}, i = 1\cdots$ 是 Hurwitz。

为了使算法以离线模型方式运行，由于 $\boldsymbol{P}_\varepsilon^* = \begin{bmatrix} P_{11} & \varepsilon P_{21}^T \\ \varepsilon P_{21} & \varepsilon P_{22} \end{bmatrix} = \begin{bmatrix} I_{n1} & \\ & \varepsilon I_{n2} \end{bmatrix} P^*$，因此将代数 Riccati 方程（6）写成：

$$\boldsymbol{A}^T\boldsymbol{P}^* + \boldsymbol{P}^*\boldsymbol{A} - \boldsymbol{P}^*\boldsymbol{B}\boldsymbol{R}^{*-1}\boldsymbol{B}^T\boldsymbol{P}^* + \boldsymbol{Q}^* = 0 \tag{9}$$

式中，$\boldsymbol{P}^* = \begin{bmatrix} \boldsymbol{P}_{11}^* & \varepsilon\boldsymbol{P}_{21}^{*T} \\ \boldsymbol{P}_{21}^* & \boldsymbol{P}_{22}^* \end{bmatrix}$。

则最优反馈增益可写为：

$$\boldsymbol{K}^* = (\boldsymbol{R}^*)^{-1}\boldsymbol{B}^T\boldsymbol{P}^* \tag{10}$$

根据以上的 Kleinman 算法所提出基于模型 IRL 迭代的奇异摄动系统最优控制算法如算法 1 所示。

算法 1：

第一步：初始化。假设 $i = 0$，其中 i 为迭代步长。给定初始的状态处罚矩阵 $\boldsymbol{Q}^0 \in \boldsymbol{R}^{n\times n} \geqslant 0$，对于任意 $\boldsymbol{R} = \boldsymbol{R}^T \in R^{m\times m} > 0$，给定初始增益为，调节参数为。

第二步：策略评估。通过求解（9）得 $\boldsymbol{P}^i$：

$$\boldsymbol{A}_i^T\boldsymbol{P}^i + \boldsymbol{P}^i\boldsymbol{A}_i + \boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i + \boldsymbol{Q}^i + \boldsymbol{\alpha}(\boldsymbol{K}_i - \boldsymbol{K}_\varepsilon^*)^T\boldsymbol{R}(\boldsymbol{K}_i - \boldsymbol{K}_\varepsilon^*) = 0 \tag{11}$$

第三步：使用逆最优控制改进的等价权重。通过（10）更新 $\boldsymbol{Q}^{i+1}$：

$$\boldsymbol{Q}^{i+1} = -\boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i - \boldsymbol{A}_i^T\boldsymbol{P}^i - \boldsymbol{P}^i\boldsymbol{A}_i \tag{12}$$

第四步：策略改进。根据式（11）更新控制增益 $\boldsymbol{K}_{i+1}$：

$$\boldsymbol{K}_{i+1} = (\boldsymbol{R})^{-1}\boldsymbol{B}^T\boldsymbol{P}^i \tag{13}$$

第五步：令 $i = i + 1$，转到第二步，直到对任意小的正常数 σ_1，$\|\boldsymbol{K}_i - \boldsymbol{K}_\varepsilon^*\| < \sigma_1$ 时停止。

参考文献［18］的基于全阶模型的 SPSs 强化学习最优控制算法的收敛性和稳定性证明，下面的定理 1 显示了算法 1 的收敛性和稳定性，证明之前，首先先引入一个有助于定理 1 证明的引理。

引理 2[32]：令矩阵 $\boldsymbol{C}$ 的特征值，矩阵 $\boldsymbol{D}$ 的特征值分别为矩阵 $\boldsymbol{C}\in\boldsymbol{R}^{n\times n}$ 和矩阵 $\boldsymbol{D}\in\boldsymbol{R}^{m\times m}$ 的特征值。则 $\boldsymbol{C}\otimes\boldsymbol{I}_m + \boldsymbol{D}\otimes\boldsymbol{I}_n$ 的特征值可表示为，其中 $k=1, 2, \cdots, n$，$j=1, 2, \cdots, m$。

定理 1：算法 1 将在有限次数的迭代后终止，可以得到 $\boldsymbol{Q}^*$ 近似解，即 $\boldsymbol{Q}^*$ 等价的权重 $\boldsymbol{Q}$。此外，算法 1 稳定且收敛，即 $\lim\limits_{i\to\infty}\|\boldsymbol{K}_\varepsilon^{i-1} - \boldsymbol{K}_\varepsilon^i\| = 0$，且 $\boldsymbol{A}_{\varepsilon i}, i = 1\cdots$ 是 Hurwitz。

证明：首先，将 Lyapunov 方程（7）构造为如下克罗内克积的形式：

$$\boldsymbol{\varphi}_\varepsilon \mathrm{vec}(P_\varepsilon^*) = -\mathrm{vec}(\boldsymbol{K}_i^T\boldsymbol{R}^*\boldsymbol{K}_i + \boldsymbol{Q}^*) \tag{14}$$

其中：$\boldsymbol{\varphi}_\varepsilon = (\boldsymbol{A}_{\varepsilon i}^T\otimes\boldsymbol{I}_n + \boldsymbol{I}_n\otimes\boldsymbol{A}_{\varepsilon i}^T)$ 由引理 2 可知，当 $A_{\varepsilon i}$ 为 Hurwitz 时，矩阵 $\boldsymbol{\varphi}_\varepsilon$ 是非奇异的。令 $\boldsymbol{\pi} = \begin{bmatrix} I_{n1} & \\ & \varepsilon I_{n2} \end{bmatrix}$，由此式（14）可以改写为：

$$\boldsymbol{\varphi}_\varepsilon \pi \mathrm{vec}(\boldsymbol{P}^i) = -\mathrm{vec}(\boldsymbol{K}_i^T\boldsymbol{R}^*\boldsymbol{K}_i + \boldsymbol{Q}^*) \tag{15}$$

根据式（11）可以写出如下克罗内克积的形式：

$$\boldsymbol{\varphi}\mathrm{vec}(\boldsymbol{P}^i) = -\mathrm{vec}(\boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i + \boldsymbol{Q}^i + \Delta_i) \tag{16}$$

式中，$\boldsymbol{\varphi} = (\boldsymbol{A}_i^T\otimes\boldsymbol{I}_n + \boldsymbol{I}_n\otimes\boldsymbol{A}_i^T)$，$\boldsymbol{\varphi} = \boldsymbol{\varphi}_\varepsilon\pi$，$\Delta_i = \alpha(\boldsymbol{K}_i - \boldsymbol{K}_\varepsilon^*)^T\boldsymbol{R}(\boldsymbol{K}_i - \boldsymbol{K}_\varepsilon^*)$。由于 π 是可逆的，由此 $(\boldsymbol{A}_i^T\otimes\boldsymbol{I}_n + \boldsymbol{I}_n\otimes\boldsymbol{A}_i^T)$ 也是可逆的。将式（11）代入式（12）中得：

$$\boldsymbol{Q}^{i+1} = \boldsymbol{Q}^i + \alpha(\boldsymbol{K}_i - \boldsymbol{K}_\varepsilon^*)^T\boldsymbol{R}(\boldsymbol{K}_i - \boldsymbol{K}_\varepsilon^*) \tag{17}$$

文献［25］证明了至少存在一个等价权重 $\boldsymbol{Q}$，使得通过选择 $a\in(0,1)$，$\boldsymbol{Q}^i$ 可以增加到等价权重 $\boldsymbol{Q}$ 的邻域，即通过改变 $a\in(0,1)$ 来调整 $\boldsymbol{Q}^i$ 的增量，使得增量可以任意小，使得 $\boldsymbol{Q}^i$ 成为等价权重 $\boldsymbol{Q}$ 的某个小的正常数 σ_1 的近邻。即存在一个小的阈值 σ_1，使得 $\|\boldsymbol{Q} - \boldsymbol{Q}^i\| \leqslant \sigma_1$ 成立。公式（12）能在有限迭代次数后终止，并且迭代次数 $i \leqslant ceil\left\{\dfrac{\|\boldsymbol{Q}^*\|}{\boldsymbol{\alpha}\boldsymbol{\sigma}^2\boldsymbol{\lambda}_{\min}(\boldsymbol{R})}\right\}$[25]。此时，该等价权重 $\boldsymbol{Q}$ 满足 $\boldsymbol{A}_\varepsilon^T\boldsymbol{P}^i + \boldsymbol{P}^i\boldsymbol{A}_\varepsilon - \boldsymbol{P}^i\boldsymbol{B}_\varepsilon\boldsymbol{K}_\varepsilon^* + \boldsymbol{Q} = 0$，但不唯一（这将在后面的定理 2 中证明）。由于状态惩罚权矩阵在 Riccati 方程中定义了唯一的控制增益[33]，由此当 $\boldsymbol{Q}^i$ 有限迭代次数后增加到等价权重 $\boldsymbol{Q}$ 时，$\boldsymbol{K}_i$ 近似于 $\boldsymbol{K}_\varepsilon^*$。因为 $\boldsymbol{K}_i$ 近似于 $\boldsymbol{K}_\varepsilon^*$，根据（15）与（16）及经有限迭代次数后增加到等价权重 $\boldsymbol{Q}$ 的 $\boldsymbol{Q}^i$ 可得 $\boldsymbol{P}_\varepsilon^i = \pi\boldsymbol{P}^i$。此时策略改进更新后的控制增益为：

$$\boldsymbol{K}_{i+1} = (\boldsymbol{R})^{-1}\boldsymbol{B}_\varepsilon^T\pi\boldsymbol{P}^{i-1} = (\boldsymbol{R})^{-1}\boldsymbol{B}^T\boldsymbol{P}^i \tag{18}$$

这意味着（18）在数学上等于式（8）。根据引理1，定理1的结果成立。由此 $\lim_{i\to\infty}\|\boldsymbol{P}_\varepsilon^{i-1}-\boldsymbol{P}_\varepsilon^{i}\|=0$，$\boldsymbol{A}_{\varepsilon i}$，$i=1\cdots$ 是 Hurwitz。根据公式（13）或公式（18）可知，当系统动力学已知的情况下，$\lim_{i\to\infty}\|\boldsymbol{K}_\varepsilon^{i-1}-\boldsymbol{K}_\varepsilon^{i}\|=0$。由此，算法1是稳定且收敛的。证明完毕。

算法1使用的逆最优控制改进方法存在着非唯一的解[23,25]，下一个定理将通过实际目标解与算法1所得到的收敛解之间的关系给出算法非唯一解的存在条件。

定理2（等价权重的非唯一性）：如果存在矩阵 $\boldsymbol{Q}^d\in\mathbb{R}^{n\times n}$，$\boldsymbol{P}^d\in\mathbb{R}^{n\times n}$ 满足：

$$\boldsymbol{A}_\varepsilon^T\boldsymbol{P}^d+\boldsymbol{P}^d\boldsymbol{A}_\varepsilon-\boldsymbol{P}^d\boldsymbol{B}_\varepsilon\boldsymbol{K}_\varepsilon^*+\boldsymbol{Q}^d=0 \tag{19}$$

$$\boldsymbol{R}^d\boldsymbol{K}_\varepsilon^*=\boldsymbol{B}_\varepsilon^T\boldsymbol{P}^d \tag{20}$$

算法1获得的每个收敛解是所有符合式（19）（20）的子集。

证明：参考[24]中定理4的离散系统收敛解之间的关系非唯一性证明，假设算法1收敛到收敛值 $\boldsymbol{Q}^\infty=\boldsymbol{Q}^*+\boldsymbol{Q}^d$，$\boldsymbol{P}^\infty=\boldsymbol{P}_\varepsilon^*+\boldsymbol{P}^d$，此时 $\lim_{i\to\infty}\alpha\Delta_i=0$。

充分性证明：将收敛值代入式（6）中：

$$\boldsymbol{A}_\varepsilon^T\boldsymbol{P}^\infty+\boldsymbol{P}^\infty\boldsymbol{A}_\varepsilon-\boldsymbol{P}^\infty\boldsymbol{B}_\varepsilon\boldsymbol{K}_\varepsilon^*+\boldsymbol{Q}^\infty-\boldsymbol{A}_\varepsilon^T\boldsymbol{P}^d-\boldsymbol{P}^d\boldsymbol{A}_\varepsilon+\boldsymbol{P}^d\boldsymbol{B}_\varepsilon\boldsymbol{K}_\varepsilon^*-\boldsymbol{Q}^d=0 \tag{21}$$

根据式（19）可得：

$$\boldsymbol{A}_\varepsilon^T\boldsymbol{P}^\infty+\boldsymbol{P}^\infty\boldsymbol{A}_\varepsilon-\boldsymbol{P}^\infty\boldsymbol{B}_\varepsilon\boldsymbol{K}_\varepsilon^*+\boldsymbol{Q}^\infty=0 \tag{22}$$

在式（20）两边加上 $\boldsymbol{B}_\varepsilon^T\boldsymbol{P}_\varepsilon^*$ 得：

$$\boldsymbol{B}_\varepsilon^T\boldsymbol{P}^d+\boldsymbol{B}_\varepsilon^T\boldsymbol{P}_\varepsilon^*=\boldsymbol{B}_\varepsilon^T\boldsymbol{P}^\infty=\boldsymbol{R}^d\boldsymbol{K}_\varepsilon^*+\boldsymbol{R}^*(\boldsymbol{R}^*)^{-1}\boldsymbol{B}_\varepsilon^T\boldsymbol{P}_\varepsilon^*=(\boldsymbol{R}^d+\boldsymbol{R}^*)\boldsymbol{K}_\varepsilon^* \tag{23}$$

令 $(\boldsymbol{R}^d+\boldsymbol{R}^*)=\boldsymbol{R}^\infty$ 得：

$$\boldsymbol{K}_\varepsilon^*=(\boldsymbol{R}^\infty)^{-1}\boldsymbol{B}_\varepsilon^T\boldsymbol{P}^\infty \tag{24}$$

通过式（24）可得式（13），将式（13）代入式（22）得到式（12）。充分性已证。

必要性证明：根据算法1中式（12）（13）、收敛到 $\boldsymbol{Q}^\infty$，$\boldsymbol{P}^\infty$，以及式（5）、（6）。令 $\boldsymbol{P}_\varepsilon^*=\boldsymbol{P}^\infty-\boldsymbol{P}^d$ 得：

$$\boldsymbol{R}^\infty\boldsymbol{K}_\varepsilon^*=\boldsymbol{B}_\varepsilon^T\boldsymbol{P}^\infty \tag{25}$$

$$\boldsymbol{R}^*\boldsymbol{K}_\varepsilon^*=\boldsymbol{B}_\varepsilon^T\boldsymbol{P}_\varepsilon^*=\boldsymbol{B}_\varepsilon^T(\boldsymbol{P}^\infty-\boldsymbol{P}^d) \tag{26}$$

把式（25）代入式（26），得到式（23），也可以推出式（20）。再从公式（12）中减去公式（6）得到：

$$\boldsymbol{A}_\varepsilon^T\boldsymbol{P}^\infty+\boldsymbol{P}^\infty\boldsymbol{A}_\varepsilon-\boldsymbol{P}^\infty\boldsymbol{B}_\varepsilon\boldsymbol{K}_\varepsilon^*+\boldsymbol{Q}^\infty-\boldsymbol{A}_\varepsilon^T\boldsymbol{P}_\varepsilon^*-\boldsymbol{P}_\varepsilon^*\boldsymbol{A}_\varepsilon+\boldsymbol{P}_\varepsilon^*\boldsymbol{B}_\varepsilon\boldsymbol{K}_\varepsilon^*-\boldsymbol{Q}^*=0 \tag{27}$$

将公式（27）合并同类项后得到式（19），必要性已证。证明完毕。

3.2 无模型 off policy IRL 的奇异摄动系统最优控

算法1仍然需要依赖于系统动力学模型与最优控制增益信息，本节将参考基于积分强化学习[15,18,34]，提出基于数据驱动的无模型 off policy IRL 算法，算法，该算法仅利用观测到的最优行为数据 $[\boldsymbol{x}(t),\boldsymbol{u}^*(t)]$ 来解决问题1，而无需额外的系统动力学和最优控制增益信息。此外，该算法可以消除激励探测噪声的影响，保证策略评估结果的无偏性。

使用 off-policy 技术将系统动力学方程（1）改写为以下形式：

$$\dot{\boldsymbol{x}}(t)=\boldsymbol{A}_\varepsilon\boldsymbol{x}(t)+\boldsymbol{B}_\varepsilon(\boldsymbol{u}^*(t)-\boldsymbol{u}^i(t))+\boldsymbol{B}_\varepsilon\boldsymbol{u}^i(t) \tag{28}$$

其中：$\boldsymbol{u}^i=-\boldsymbol{K}_{i+1}\boldsymbol{x}$ 为第 i 步更新的目标策略。$\boldsymbol{u}^*$ 是实际应用于系统，与目标策略 $\boldsymbol{u}^i$ 不同的行为策略。

参考文献［29］，通过式（28）得：

$$\dot{\boldsymbol{x}}^T\boldsymbol{P}_\varepsilon^i\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{P}_\varepsilon^i\dot{\boldsymbol{x}}=\boldsymbol{x}^T(\boldsymbol{A}_{\varepsilon i}^T\boldsymbol{P}_\varepsilon^i+\boldsymbol{P}_\varepsilon^i\boldsymbol{A}_{\varepsilon i})\boldsymbol{x}+2(\boldsymbol{u}^*-\boldsymbol{u}^i)\boldsymbol{B}^T\boldsymbol{P}_\varepsilon^i\boldsymbol{x} \tag{29}$$

通过公式（11）可得：

$$\boldsymbol{A}_{\varepsilon i}^T\boldsymbol{P}_\varepsilon^i+\boldsymbol{P}_\varepsilon^i\boldsymbol{A}_{\varepsilon i}=-\boldsymbol{Q}^i-\boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i-\alpha(\boldsymbol{K}_i-\boldsymbol{K}_\varepsilon^*)^T\boldsymbol{R}(\boldsymbol{K}_i-\boldsymbol{K}_\varepsilon^*) \tag{30}$$

将式（32）代入式（31）得：

$$\dot{\boldsymbol{x}}^T\boldsymbol{P}_\varepsilon^i\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{P}_\varepsilon^i\dot{\boldsymbol{x}}-2(\boldsymbol{u}^*-\boldsymbol{u}^i)\boldsymbol{B}^T\boldsymbol{P}_\varepsilon^i\boldsymbol{x}=\boldsymbol{x}^T(-\boldsymbol{Q}^i-\boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i)\boldsymbol{x}-\boldsymbol{\alpha}(\boldsymbol{u}^*-\boldsymbol{u}^i)^T\boldsymbol{R}(\boldsymbol{u}^*-\boldsymbol{u}^i) \tag{31}$$

对式（31）进行积分，积分范围从 t 到 $t+T$，可得下式：

$$\boldsymbol{x}^T(t)\boldsymbol{P}_\varepsilon^i\boldsymbol{x}(t)-\boldsymbol{x}^T(t+T)\boldsymbol{P}_\varepsilon^i x(t+T)+2\int_t^{t+T}(\boldsymbol{u}^*-\boldsymbol{u}^i)\boldsymbol{R}\boldsymbol{K}_{i+1}\boldsymbol{x}=\int_t^{t+T}\boldsymbol{x}^T(\boldsymbol{Q}^i+\boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i)\boldsymbol{x}\mathrm{d}\tau+\alpha\int_{t+T}^{t+T}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)^T\boldsymbol{R}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)\mathrm{d}\tau \tag{32}$$

通过计算系统状态的测量值与输入之间的差值，并对其进行积分，可以求出最优控制增益 $\boldsymbol{K}_{i+1}$。

在 SPSs 中，慢状态变量 $\boldsymbol{x}_1$ 的变化速率远低于快状态变量 $\boldsymbol{x}_2$，两者的速率之比约为 ε。为准确捕捉快状态变量 $\boldsymbol{x}_2$ 的动态，需要采用极小采样间隔，但这导致慢状态变量 $\boldsymbol{x}_1$ 在采样间隔内的变化量极小[18]。在这种情况下，矩阵 $\boldsymbol{P}_\varepsilon^i$ 和 $\boldsymbol{K}_{i+1}$ 对计算与测量的误差高度敏感，容易导致收敛范围缩小、精度下降，甚至产生数值病态问题[35]。为此，参考文献［18］的处理过程，将公式（29）中的 $x^TP_\varepsilon^ix$ 转化为：

$$\boldsymbol{x}^T(t)\boldsymbol{P}_\varepsilon^i\boldsymbol{x}(t)=\boldsymbol{x}^T(t)\otimes\boldsymbol{x}^T(t)\boldsymbol{\pi}_\varepsilon\mathrm{vecm}(\boldsymbol{P}_\varepsilon^i) \tag{33}$$

其中：$\boldsymbol{\pi}_\varepsilon=\begin{bmatrix}I_{ns} & \\ & I_{nf}\end{bmatrix}$，$ns=\frac{n_1(n_1+1)}{2}$，$I_{nf}=n_1n_2+\frac{n_2(n_2+1)}{2}$。

将式（24）代入式（23）并转换为克罗内克积形式：

$$[\boldsymbol{x}^T(t+T)\otimes\boldsymbol{x}^T(t+T)-\boldsymbol{x}^T(t)\otimes\boldsymbol{x}^T(t)]\times\boldsymbol{\pi}_\varepsilon\mathrm{vecm}(\boldsymbol{P}^i)+2(\boldsymbol{I}\otimes\boldsymbol{R})\times\int_t^{t+T}x^T\otimes(\boldsymbol{u}^*+\boldsymbol{K}_i\boldsymbol{x})^T\mathrm{d}\tau\mathrm{vec}(\boldsymbol{K}_{i+1})=\int_t^{t+T}\boldsymbol{x}^T(\boldsymbol{Q}^i+\boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i^T)\boldsymbol{x}\mathrm{d}\tau+$$

$$a\int_{t+T}^{t+T}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)^T\boldsymbol{R}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)\mathrm{d}\tau \tag{34}$$

令：

$$\boldsymbol{Z}_i=\begin{bmatrix}\boldsymbol{Z}_1^1 & \boldsymbol{Z}_2^1\\ \vdots & \vdots\\ \boldsymbol{Z}_1^l & \boldsymbol{Z}_2^l\end{bmatrix},\boldsymbol{\vartheta}_i=\begin{bmatrix}\boldsymbol{\vartheta}_i^1\\ \vdots\\ \boldsymbol{\vartheta}_i^l\end{bmatrix} \tag{35}$$

其中：

$$\boldsymbol{Z}_1^l=\pi_\varepsilon\{\boldsymbol{x}^T[t+(l-1)\boldsymbol{T}]\otimes x^T[t+(l-1)\boldsymbol{T}]-x^T(t+l\boldsymbol{T})\otimes x^T(t+l\boldsymbol{T})\}$$

$$\boldsymbol{Z}_2^l=2(\boldsymbol{I}\otimes\boldsymbol{R})\int_{t+(l-1)T}^{t+lT}\boldsymbol{x}^T\otimes(\boldsymbol{u}^*+\boldsymbol{K}_i\boldsymbol{x})^T\mathrm{d}\tau$$

$$\vartheta_i^l=\int_{t+(l-1)T}^{t+lT}\boldsymbol{x}^T(\boldsymbol{Q}^i+\boldsymbol{K}_i^T\boldsymbol{R}\boldsymbol{K}_i^T)\boldsymbol{x}\mathrm{d}\tau+\alpha\int_{t+T}^{t+T}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)^T\boldsymbol{R}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)\mathrm{d}\tau$$

注 2：由于 $[\mathrm{vecm}(\boldsymbol{P}_\varepsilon^i)^T,\mathrm{vec}(\boldsymbol{K}_{i+1})^T]$ 是一个具有 $N=n(n+1)/2+nm$ 个独立元素的对称矩阵。因此，至少需要 $1\geqslant N$ 个数据集来求解以下的最小二乘：

$$[\mathrm{vecm}(\boldsymbol{P}_\varepsilon{}^i)^T\quad\mathrm{vec}(\boldsymbol{K}_{i+1})^T]=(\boldsymbol{Z}_i^T\boldsymbol{Z}_i)^{-1}\boldsymbol{Z}_i^T\vartheta \tag{36}$$

为了保证 $Z_i^TZ_i$ 可逆，数据集必须满足持续激励条件。在控制输入加入探测噪声 e，令 $\boldsymbol{u}^*+e$ 代替公式（35）中的 $\boldsymbol{u}^*$。

将式（17）两边同时乘以 $\boldsymbol{x}$，并对其从 t 到 $t+T$ 的积分，得：

$$\int_t^{t+T}x^T\boldsymbol{Q}^{i+1}\boldsymbol{x}\mathrm{d}\tau=\int_t^{t+T}x^T\boldsymbol{Q}^i\boldsymbol{x}\mathrm{d}\tau+\alpha\int_{t+T}^{t+T}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)^T\boldsymbol{R}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)\mathrm{d}\tau \tag{37}$$

将式（37）带转换为克罗内克积形式：

$$\int_t^{t+T}\boldsymbol{x}^T\otimes\boldsymbol{x}^T\mathrm{d}\tau\mathrm{vecm}(\boldsymbol{Q}^{i+1})=\int_t^{t+T}\boldsymbol{x}^T\boldsymbol{Q}^i\boldsymbol{x}\mathrm{d}\tau+\alpha\int_{t+T}^{t+T}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)^T\boldsymbol{R}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)\mathrm{d}\tau \tag{38}$$

令：

$$\boldsymbol{\varphi}_i=\begin{bmatrix}\int_t^{t+T}\boldsymbol{x}^T\otimes\boldsymbol{x}^T\mathrm{d}\tau\\ \vdots\\ \int_{t+(h-1)T}^{t+hT}\boldsymbol{x}^T\otimes\boldsymbol{x}^T\mathrm{d}\tau\end{bmatrix},\boldsymbol{\gamma}_i=\begin{bmatrix}\boldsymbol{\gamma}_i^1\\ \vdots\\ \boldsymbol{\gamma}_i^l\end{bmatrix} \tag{39}$$

其中：$\gamma_i^l=\int_{t+(l-1)T}^{t+lT}\boldsymbol{x}^T\boldsymbol{Q}^i\boldsymbol{x}\mathrm{d}\tau+\alpha\int_{t+T}^{t+T}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)^T\times\boldsymbol{R}(\boldsymbol{K}_i\boldsymbol{x}-\boldsymbol{u}^*)\mathrm{d}\tau$。

注 3：由于 $\mathrm{vecm}(\boldsymbol{Q}^{i+1})$ 是一个具有 $N_1=n(n+1)/2$ 个独立元素的对称矩阵。因此，至少需要 $h\geqslant N_1$ 个数据集来求解以下的最小二乘，由于 $N>N_1$，因此，这里仍收集 l 组数据：

$$\mathrm{vecm}(\boldsymbol{Q}^{i+1})=(\varphi_i^T\varphi_i)^{-1}\varphi_i^T\gamma \tag{40}$$

同理为了满足持续激励条件。在控制输入加入探测噪声 e，令 $\boldsymbol{u}^*+e$ 代替（38）中的 $\boldsymbol{u}^*$。

算法 2：无模型 off policy IRL 的奇异摄动系统最优控制算法

第一步：初始化。假设 $i=0$，其中 i 为迭代步长。给定初始的状态处罚矩阵 $\boldsymbol{Q}^0\in\boldsymbol{R}^{n\times n}\geqslant 0$，对于任意 $\boldsymbol{R}=\boldsymbol{R}^T\in\boldsymbol{R}^{m\times m}>0$，给定初始增益为，调节参数为。

第二步：数据采集。收集 $l\geqslant n$ 组数据$(\boldsymbol{x},\boldsymbol{u}^*)$并加入探测噪声 e，形成式(36)中的 $\boldsymbol{Z}_i$ 和 $\boldsymbol{\vartheta}_i$ 与式(40)中的 $\boldsymbol{\varphi}_i$ 和 $\boldsymbol{\gamma}_i$。

第三步：策略评估。过求解(36)得 $\boldsymbol{K}_{i+1}$。

第四步：逆最优控制。通过求解(37)得 $\boldsymbol{Q}^{i+1}$。

第五步：令 $i=i+1$，转到第三步，直到对任意小的正常数，时停止。

根据以上过程，算法 2 给出了 SPSs 的无模型 off policy IRL 的奇异摄动系统最优控制算法。

定理 3：给定与算法 1 中相同的 $\boldsymbol{R}$，算法 2 求出的 $\boldsymbol{Q}^{i+1}$ 收敛于与算法 1 中相同的权重 $\boldsymbol{Q}$，此时，控制策略增益 $\boldsymbol{K}_{i+1}$ 收敛于 $\boldsymbol{K}_\varepsilon^*$，即 $\lim\limits_{i\to\infty}\boldsymbol{K}_i=\boldsymbol{K}_\varepsilon^*$。因此，系统可以跟踪最优行为 $[\boldsymbol{x}(t),\boldsymbol{u}^*(t)]$。

证明：给定初始的状态处罚矩阵 $\boldsymbol{Q}^0\in\boldsymbol{R}^{n\times n}\geqslant 0$ 和初始化稳定的 $\boldsymbol{K}_0$，对（32）求导，并分别去掉方程中左右两边的 $\boldsymbol{x}$，再进行合并同类项就得到式（11）类似的式子。同理对式（37）求导分别去掉方程中左右两边的 $\boldsymbol{x}$，并让式（11）减去去掉左右两边 $\boldsymbol{x}$ 的方程，可得式（12）类似的式子。因此算法 2 与算法 1 等价。同时，由于算法 2 产生唯一解。因此，算法 2 的解与算法 1 的解相同。由于算法 1 收敛且稳定，因此算法 2 也收敛且稳定。因此，算法 2 求出的 $\boldsymbol{Q}^{i+1}$ 能增加到与算法 1 中相同等价权重 $\boldsymbol{Q}$，此时控制策略增益 $\boldsymbol{K}_{i+1}$ 收敛于 $\boldsymbol{K}_\varepsilon^*$。因此，系统可以跟踪最优行为 $[\boldsymbol{x}(t),\boldsymbol{u}^*(t)]$。

定理 4：在式（32）中加入的探测噪声对算法 2 的结果不会产生偏差。

证明：文献［25，27，29］已经证明 off policy IRL 算法能消除探测噪声的影响，因此探测噪声对算法 2 的结果不会产生偏差。

4　实验结果与分析

在本节中，将应用模数转换目标的 RC 梯形电路系统[18]来验证算法 2 的有效性。其系统矩阵如下所示：

$$\boldsymbol{A}_\varepsilon=\begin{bmatrix}-\frac{3}{2RC} & \frac{1}{RC} & 0 & 0\\ \frac{1}{RC} & -\frac{2}{RC} & 0 & 0\\ 0 & 0 & -\frac{2}{\varepsilon RC} & \frac{1}{\varepsilon RC}\\ 0 & 0 & \frac{1}{\varepsilon RC} & -\frac{3}{2\varepsilon RC}\end{bmatrix} \tag{41}$$

$$\boldsymbol{B}_\varepsilon=\begin{bmatrix}\frac{1}{2RC} & 0\\ 0 & 0\\ 0 & 0\\ 0 & \frac{1}{2\varepsilon RC}\end{bmatrix} \tag{42}$$

其中：$R=5\times10^{3}\ \Omega$，$C=1\times10^{-4}$ F，$\varepsilon=0.05$。使用文献［18］相同的代价函数参数来生成最优行为数据，其中矩阵 $\boldsymbol{R}^{*}=\boldsymbol{I}_{2}$，状态惩罚权重：

$$\boldsymbol{Q}^{*}=\begin{bmatrix}50 & 0 & 0.5 & 1\\ 0 & 100 & 0.5 & 1.5\\ 0.5 & 0.5 & 1 & 0\\ 1 & 1.5 & 0 & 1\end{bmatrix}$$

根据这些参数可以得到该奇异摄动系统的最优增益为：

$$\boldsymbol{K}_{\varepsilon}^{*}=\begin{bmatrix}5.3933 & 2.9167 & 0.0161 & 0.0234\\ 0.4687 & 0.6741 & 0.1145 & 0.2338\end{bmatrix}$$

假设在双时间尺度系统动力学、代价函数权重（$\boldsymbol{Q}$，$\boldsymbol{R}$）与最优增益 $\boldsymbol{K}^{\varepsilon*}$ 都完全未知的情况下，仿真通过使用算法 2 来解决问题 1。也就是选择任意权重 $\boldsymbol{R}=\boldsymbol{R}^{T}\in\boldsymbol{R}^{m\times m}>0$，仅利用已知的最优行为数据［$\boldsymbol{x}(t)$，$\boldsymbol{u}^{*}(t)$］来重构代价函数，即为系统（1）找到一个定义 1 中的等价权重 $\boldsymbol{Q}$，使其能产生与（5）中的最优控制策略增益 $\boldsymbol{K}_{\varepsilon}^{*}$，从而使系统跟踪学习最优行为［$\boldsymbol{x}(t)$，$\boldsymbol{u}^{*}(t)$］。

在仿真中，这里选取矩阵 $\boldsymbol{R}=\boldsymbol{I}_{2}$，设置状态惩罚权重初始值 $\boldsymbol{Q}^{0}=$zeros（4，4），状态初始值 $\boldsymbol{x}(0)=[1,1,-1,-1]^{T}$，初始增益 $K_{0}=0$，正常数 $\sigma_{1}=0.1$，调节参数 $a=0.3$，采样时间 $T=0.01$ s，从 $t=0$ s 到 $t=2$ s 之间加入探测噪声为：

$$e=\sum_{i=1}^{100}\sin(\omega_{i}t)/100 \tag{43}$$

其中：ω_{i}，$i=1,\cdots,100$ 是在［-500，500］中随机选择的。

通过算法 2 使用 Matlab 进行仿真实验。算法 2 $\boldsymbol{Q}^{i}$ 与 $\boldsymbol{K}_{i}$ 的收敛过程如图 1 和图 2 所示。算法 2 经过迭代后的状态惩罚权重和控制增益收敛为：

$$\boldsymbol{Q}^{i+1}=\begin{bmatrix}49.8477 & 0.1662 & 0.4800\\ 0.1662 & 99.4454 & 0.5701\\ 0.4800 & 0.5701 & 0.9188\\ 1.0061 & 1.4622 & 0.2429\end{bmatrix} \tag{44}$$

$$\boldsymbol{K}_{i+1}=\begin{bmatrix}5.3836 & 2.9141 & 0.0159 & 0.0234\\ 0.4685 & 0.6732 & 0.1169 & 0.2298\end{bmatrix} \tag{45}$$

与用来生成最优行为数据的状态惩罚权重 $\boldsymbol{Q}^{*}$ 和其求得的最优增益 K_{ε}^{*} 高度相似。由于所提出的算法是 off policy 的，可以消除以满足激励持久性条件而加入的探测噪声的影响[25,27,29]，因此使用其他类型的探测噪声，也能收敛到相似的状态惩罚权重和控制增益。图 3 则显示将学习完的 $\boldsymbol{K}_{i}$ 应用于系统中，作为慢状态的电容电压跟踪最优行为的性能。图 4 则显示将学习完的 $\boldsymbol{K}_{i}$ 应用于系统中，作为快状态的电阻电压跟踪最优行为的性能。从图 3 和图 4 可以看出算法 2 在双时间尺度系统动力学、代价函数权重（$\boldsymbol{Q}$，$\boldsymbol{R}$）与最优增益 $\boldsymbol{K}^{\varepsilon*}$ 完全未知的情况下，能跟踪已知的最优行为［$\boldsymbol{x}(t)$，$\boldsymbol{u}^{*}(t)$］，且具有很好的跟踪学习效果。

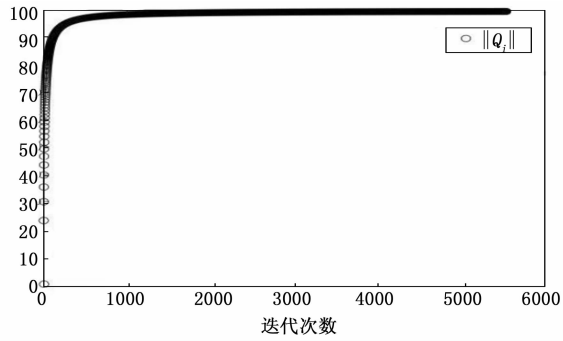

图 1 算法 2 Q^{i} 的收敛性

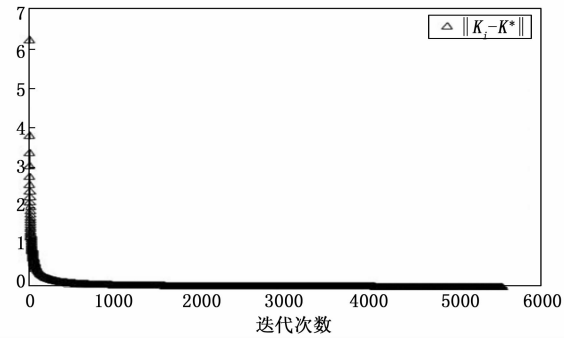

图 2 算法 2 K_{i} 的收敛性

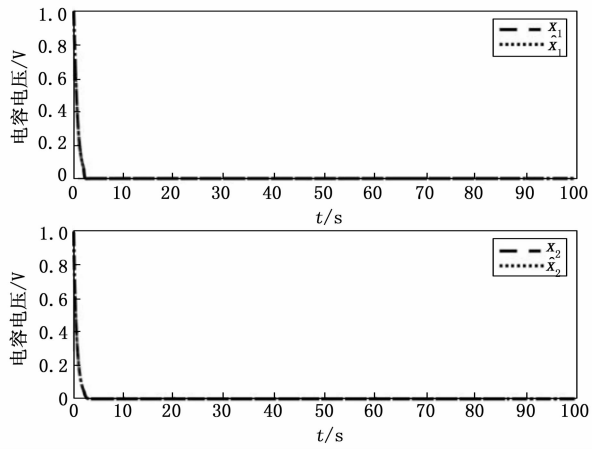

图 3 算法 2 慢状态的跟踪性能

在评估系统跟踪最优行为轨迹的效果时，采用了两个关键指标：绝对误差积分（IAE，integral absolute error）和误差均方差（MSE，mean square error）。为了更加直观地展现算法 2 的优越性，将算法 2 对算法 1 的跟踪性能进行对比，表 1 中显示了算法 1 执行结束后各状态变量的 *IAE* 和 *MSE* 值，表 2 中显示了算法 2 执

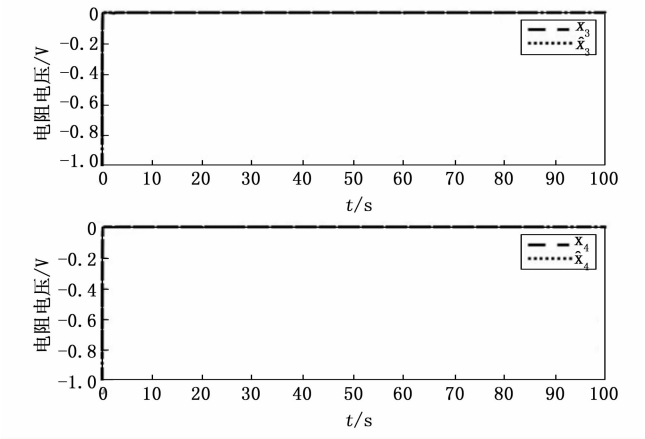

图 4　算法 2 快速状态的跟踪性能

行结束后各状态变量的 *IAE* 和 *MSE* 值：

$$IEA = \sum_{j=i}^{i+n} |\hat{x}(jT) - x(jT)| \tag{46}$$

$$MSE = \sqrt{\frac{\sum_{j=i}^{i+n} |\hat{x}(jT) - x(jT)|^2}{n}} \tag{47}$$

通过对比表 1 和表 2 中的数据，可以清晰地看到算法 2 在 *IAE* 和 *MSE* 两个指标上均优于算法 1，这充分证明了算法 2 在跟踪最优行为轨迹方面具有更高的精度。

表 1　算法 1 的评估指标

$i>7\ 000 \quad n=3\ 000$	*IEA*	*MSE*
$x_1(t)$	0.306 1	$1.036\ 6\times10^{-3}$
$x_2(t)$	0.224 3	$8.233\ 1\times10^{-4}$
$x_3(t)$	0.146 3	$5.172\ 5\times10^{-4}$
$x_4(t)$	0.330 5	$1.140\ 3\times10^{-3}$

表 2　算法 2 的评估指标

$i>7\ 000 \quad n=3\ 000$	*IEA*	*MSE*
$x_1(t)$	0.229 1	$7.817\ 5\times10^{-4}$
$x_2(t)$	0.183 4	$7.089\ 3\times10^{-4}$
$x_3(t)$	0.138 6	$4.897\ 2\times10^{-4}$
$x_4(t)$	0.312 5	$1.079\ 5\times10^{-3}$

5　结束语

本文针对未知连续 SPSs 的数据驱动 IRL 问题进行研究，提出了新的有效解决方案。结合现有的 IRL 技术和 SPSs 的两个时间尺度特征，与传统处理 SPSs 的两个时间尺度特使用的快慢时间尺度分离的方法不同，本文直接在全阶模型框架下求解 SPS 问题，不仅简化了问题处理的复杂度，还明显提升了求解的精度。首先，构建了一个基于基础模型的 IRL 框架，该框架利用系统动力学特性和最优控制增益重构未知系统的成本函数，从而模仿演示的最优行为。在此基础上，本文设计了完全基于演示最优行为数据的数据驱动无模型 off-policy IRL 学习算法，摆脱对系统精确模型的依赖，仅通过观测到的最优行为轨迹来推断成本函数。此算法在保证收敛性和无偏性的同时，还确保了闭环系统的稳定性，为处理实际中普遍存在的模型不确定性问题提供了工具。然而状态惩罚权重行或列全为零情况下，本文方法虽然也能模仿系统的最优行为，但重构的奖励函数未能达到预期效果。在未来的工作将侧重于状态惩罚权重行或列全为零情况下的逆强化学习问题以及考虑系统状态的完全测量可能是困难或昂贵的情况下输出反馈来控制系统。同时，也将解决实际工业环境中常见的随机扰动，并探索将该方法应用于更复杂的系统，以增强其在实际工业环境中的实用性和适应性。

参考文献：

[1] GOMEZEA，GOMEZQC，DZAFIC I. State estimation in two time scales for smart distribution systems [J]. IEEETransactions on Smart Grid，2015，6 (1)：421 - 430.

[2] WU F，TIAN T，RAWLINGS J B，et al. Approximate method for stochastic chemical kinetics with two-time scales by chemical langevin equations [J]. The Journal of Chemical Physics，2016，144 (17)：174112.

[3] XUE W，FAN J，LOPEZ V，et al. New methods for optimal operational control of industrial processes using reinforcement learning on two time scale [J]. IEEE Transactions on Industrial Informatics，2020，16 (5)：3085 - 3099.

[4] 孙凤琪，姜思汇，阚晓慧. 奇异摄动离散系统理论与应用综述 [J]. 吉林师范大学学报（自然科学版），2015，36 (3)：73 - 77.

[5] KOKOTOVIC P，KHALIL H，REILLY J. Singularly perturbation methods in control：analysis and design [M]. Philadelphia，PA：Society for Industrial and Applied Mathematics，1999.

[6] SU W，GAJIC Z，SHEN X. The exact slow-fast decomposition of the algebraic Ricatti equation of singularly perturbed systems [J]. IEEE Transactions on Automatic Control，1992，37：1456 - 1459.

[7] LI Y，WANG J，YANG G. Sub-optimal linear quadratic control for singularly perturbed systems [C] //the 40th IEEE Conference on Decision and Control，Orlando，FL，USA：IEEE，2001：3698 - 3703.

[8] LIU L，LIU Y，HAN C. Sub-optimal output tracking control for singularly perturbed systems [C] //Chinese Automation Congress (CAC)，Hangzhou，China：IEEE，2019：4521 - 4525.

[9] BARTO A，SUTTON R. Reinforcement learning：an introduction [M]. Cambridge，MA，USA：MIT Press，1998.

[10] PANG B，JIANG Z，MAREELS I. Reinforcement learning for adaptive optimal control of continuous-time linear periodic systems [J]. Automatica，2020，118：109035.

[11] 田奋铭，刘 飞. 带状态约束的事件触发积分强化学习控制 [J]. 计算机测量与控制，2023，31 (7)：143－149.

[12] 周盛世，单 梁，常 路，等. 基于改进 DDPG 算法的机器人路径规划算法研究 [J]. 南京理工大学学报，2021，45 (3)：265－270.

[13] DAI W，LI T，ZHANG L，JIA Y，YAN H. Multi-rate layered operational optimal control for large-scale industrial processes [J]. IEEE Transactions on Industrial Informatics，2022，18 (7)：4749－4761.

[14] ZHAO J，YANG C，DAI W，GAO W. Reinforcement learning-based composite optimal operational control of industrial systems with multiple unit devices [J]. IEEE Transactions on Industrial Informatics，2022，18 (2)：1091－1101.

[15] JIANG Y，JIANG Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics [J]. Automatica，2012，48 (10)：2699－2704.

[16] LI J，KIUMARSI B，CHAI T，LEWIS F L，FAN J. Off-policy reinforcement learning for tracking in continuous-time systems on two time scale [J]. IEEE Transactions on Neural Networks and Learning Systems，2021，32 (10)：4334－4346.

[17] SHEN M，LIU F. H∞ tracking control of tow time scale linear system based on off-policy reinforcement learning [C] //IEEE 13th Data Driven Control and Learning Systems Conference (DDCLS)，Kaifeng，China：IEEE，2024：1228－1233.

[18] ZHAO J，YANG C，GAO W. Reinforcement learning based optimal control of linear singularly perturbed systems [J]. IEEE Transactions on Circuits and Systems II：Express Briefs，2022，69 (3)：1362－1366.

[19] GAJIC Z，LIM M. Optimal control of singularly perturbed linear systems and applications：high-accuracy techniques [M]. New York，NY，USA：Marcel Dekker，2001.

[20] ABBEEL P，NG A. Apprenticeship learning via inverse reinforcement learning [C] //Proceedings of the Twenty-First International Conference on Machine Learning，Banff：Association for Computing Machinery，2004：1－8.

[21] ZHAN H，UMENBERGER J，HU X. Inverse optimal control for discrete-time finite-horizon linear quadratic regulators [J]. Automatica，2019，110：108593.

[22] MOLLOY T L，FORD J J，PEREZ T. Finite-horizon inverse optimal control for discrete-time nonlinear systems [J]. Automatica，2019，87：442－446.

[23] XUE W，KOLARIC P，FAN J，et al. Inverse reinforcement learning in tracking control based on inverse optimal control [J]. IEEE Transactions on Cybernetics，2022，52 (10)：10570－10581.

[24] XUE W，LIAN B，FAN J，et al. Inverse reinforcement Q-learning through expert imitation for discrete-time systems [J]. IEEE Transactions on Neural Networks and Learning Systems，2023，34 (5)：2386－2399.

[25] LIAN B，XUE W，XIE Y，et al. Off-policy inverse Q-learning for discrete-time antagonistic unknown systems [J]. Automatica，2023，155：111171.

[26] LIAN B，XUEW，LEWISFL，et al. Robust inverse Q-learning for continuous-time linear systems in adversarial environments [J]. IEEE Transactions on Cybernetics，2022，52 (12)：13083－13095.

[27] LIAN B，DONGE V S，LEWIS F L，et al. Data-driven inverse reinforcement learning control for linear multiplayer games [J]. IEEE Transactions on Neural Networks and Learning Systems，2024，35 (2)：2028－2041.

[28] 刘 文，范家璐，薛文倩. 基于输出反馈逆强化 Q 学习的线性二次型最优控制方法 [J]. 控制理论与应用，2024，41 (8)：1469－1479.

[29] LIAN B，XUE，W，LEWISFL，et al. Inverse value iteration and Q-learning：algorithms，stability，and robustness [J/OL]. IEEE Transactions on Neural Networks and Learning Systems，2024. https://doi.org/10.1109/TNNLS.2024.3409182.

[30] SONG J，NIU Y，LAM H，et al. Asynchronous slidingmode control of singularly perturbed semi-Markovian jump systems：application to an operational amplifier circuit [J]. Automatica，2020，118：109026.

[31] KLEINMAN D. On an iterative technique for Riccati equation computations [J]. IEEE Transactions on Automatic Control，1968，13：114－115.

[32] HORN R，JOHNSON C. Matrix analysis [M]. U.K.：Cambridge University，Press，1985.

[33] BITTANTI S，LAUB A J，WILLEMS J C. The Riccati equation [M]. Secaucus，USA：Springer Science & Business Media，2012.

[34] VRABIE D，PASTRAVANU O，ABU-KHALAF M，et al. Adaptive optimal control for continuous-time linear systems based on policy iteration [J]. Automatica，2009，45 (2)：477－484.

[35] LAUB A. Matrix analysis for scientists and engineers [M]. Philadelphia，PA，USA：SIAM，2005.