信息与控制　2023 年　第 52 卷　第 3 期：343 ~ 351

DOI：10.13976/j.cnki.xk.2023.2263　　文章编号：1002 - 0411(2023) - 03 - 0343 - 09

基于近端策略优化和广义状态相关探索算法的双连续搅拌反应釜系统跟踪控制

史洪岩，付国城，潘多涛

沈阳化工大学信息工程学院，辽宁省化工过程控制技术重点实验室，辽宁 沈阳　110142

基金项目：国家重点研发计划(2018YFB1700200)；辽宁省自然科学基金(2019 - ZD - 0069)；辽宁省教育厅科研面上项目(LJKZ0433)

通信作者：付国城，fuguochengsh@163.com　收稿/录用/修回：2022 - 06 - 15/2022 - 09 - 20/2022 - 10 - 28

摘要

连续搅拌反应釜(continuous stirring tank reactor，CSTR)是经典的化工设备，被广泛应用于化工过程。由于其具有较强的非线性和时滞性，传统的控制方法无法满足其跟踪控制的精度要求。针对连续搅拌反应釜提出一种基于广义状态相关探索(generalized state-dependent exploration，gSDE)的近端策略优化(proximal policy optimization，PPO)算法的跟踪控制方法。首先使用机理模型模拟真实环境与 PPO 智能体进行交互；其次利用 gSDE 使每个回合的探索更稳定且方差更小，同时保证了探索的效果；最后通过增加反馈奖励的方式，解决环境稀疏奖励的问题，使得智能体学会如何对 CSTR 进行跟踪控制。将该算法应用于双 CSTR 系统进行测试。仿真结果表明，该算法对复杂非线性系统的跟踪控制具有训练过程平稳、控制误差小、对干扰的反应迅速等优势。

关键词

双连续搅拌反应釜
深度强化学习
非线性系统
近端策略优化算法
广义状态相关探索

中图法分类号：TP273

文献标识码：A

Two-CSTR System Tracking Control Based on PPO-gSDE Algorithm

SHI Hongyan，FU Guocheng，PAN Duotao

Liaoning Key Laboratory of Chemical Process Control Technology，*School of Information Engineering*，*Shenyang University of Chemical Technology*，*Shenyang* 110142，*China*

Abstract

A continuous stirring tank reactor (CSTR) is classic chemical equipment widely used in chemical processes. The traditional control methods fail to meet the precision requirements of tracking control because of their strong nonlinearity and associated time delay. Thus, in this study, we propose a tracking control method of proximal policy optimization (PPO) based on generalized state-dependent exploration (gSDE) for continuous stirred reactors. First, the mechanism model simulates the real environment and interacts with the PPO agent. Second, gSDE is used to make the exploration of each round more stable and with less variance, ensuring the effect of exploration. Finally, a feedback reward is added to resolve sparse reward issues in the environment, such that the agent can learn how to track and control the CSTR. The algorithm is applied to a double CSTR system to examine its effectiveness. Our simulation results show that the algorithm exhibits a stable training process, small control error, and rapid response to disturbance.

Keywords

two continuous stirring tank reactors;
deep reinforcement learning;
nonlinear system;
proximal policy optimization (PPO) algorithm;
generalized state-dependent exploration

0 引言

双 CSTR 广泛应用于制药、生物和化学加工行业。它通常在规定的温度下通过一系列物理和化学反应将原料转化为产品[1]。由于其具有高度复杂性，且双 CSTR 在动力学本质上是易受影响和开环不稳定的[2]。当出现外部干扰和噪声时，系统可能会在不久后偏离中间稳态。因此，迫切需要为双 CSTR 系统设计一个可靠的控制器，这将有助于提高生产率，并通过减少化学废物来减轻环境污染。

目前双 CSTR 受到国内外广泛的关注，出现了大量与跟踪控制相关的研究成果。例如，Al SEYAB[3]使用了一种新的高效非线性模型预测控制算法(nonlinear model predictive control，NMPC)和连续时间递归神经网络(continuous-time recurrent neural network，CTRNN)模型的非线性系统辨识算法对双 CSTR 系统进行跟踪控制，实验表明此算法能够解决在线计算问题，并且可以有效实现系统的跟踪控制。Al SEYAB 等[4]提出了一种训练广义微分递归神经网络(differential recurrent neural network，DRNN)的有效算法。并将其与 NMPC 算法结合，通过双 CSTR 案例研究，证明了 DRNN 训练算法和 NMPC 方法的有效性，具有良好的控制性能。WANG 等[5]提出了一种基于深度学习的模型预测控制(deep learning-based model predictive control，DeepMPC)来建模和控制双 CSTR 系统，该算法由一个增长的深度信念网络(growing deep belief network，GDBN)和一个最优控制器组成。实验表明，DeepMPC 在建模、跟踪和抗干扰方面可以表现出更好的性能。周红标等[6]提出了一种基于自适应模糊神经网络(adaptive fuzzy neural network，AFNN)的模型预测控制(model predictive control，MPC)方法，AFNN-MPC 控制器具有较高的跟踪精度和较强的自适应能力，能够满足复杂非线性系统的智能控制需求。此类型方法需要大量的系统数据和复杂的神经网络模型以保证模型辨识的精确度，并且有许多超参数要调节，同时需要平衡在线计算量和控制效果，在大规模随机多输入多输出(multiple-input multiple-output，MIMO)问题中的工业应用仍然受到限制[7]。XIN 等[8]提出了基于动态面控制(dynamic surface control，DSC)的双 CSTR 系统的自适应模糊反推控制，仿真分析证明了控制器的有效性和鲁棒性，同时减少了在线计算量，但未考虑干扰问题，且温度控制上存在超调。

深度强化学习(deep reinforcement learning，DRL)是建立在随机最优控制的基础上，与现代控制技术相比，随机最优控制可能在某些系统中更具优势[9]。DRL 通过离线预计算最优解，克服了在线计算时间长的问题。DRL 已被广泛应用于机器人[10-11]、路径规划[12]、实体博弈对抗[13]等领域，但并未应用于双 CSTR 系统。

因此本文基于强化学习 PPO-gSDE 算法对双 CSTR 系统进行跟踪控制，gSDE 可以保证 PPO 算法探索的稳定性且方差更小[14]，而通过添加多个反馈奖励的方式，则可以解决智能体训练中稀疏奖励导致策略网络不收敛或收敛速度慢等问题，使得 PPO 智能体可以学会如何高效地控制双 CSTR 系统。使用双 CSTR 系统的机理模型模拟真实化工过程进行实验，实验表明，本文方法在双 CSTR 系统的跟踪控制上，具有训练过程平稳、调参简单、控制误差小、对干扰的反应迅速等优势。

1 双 CSTR 系统的数学模型

双 CSTR 系统是经典的多输入多输出系统，其在实际生产过程中具有较强的非线性、时滞性和输入信号约束等特征，它常被用作控制方案的测试模型。典型双 CSTR 系统如图 1 所示，假设系统是完全混合的，其反应过程的非线性微分方程为[15]

$$\frac{\mathrm{d}C_{O1}}{\mathrm{d}t}=-K_1x_1+Q_{I1}\frac{(C_{I1}-x_1)}{V_1} \tag{1}$$

$$\frac{\mathrm{d}T_{O1}}{\mathrm{d}t}=\Delta HK_1x_1+Q_{I1}\frac{(T_{I1}-x_2)}{V_1}-U_{a1}\frac{(x_2-x_3)}{V_1} \tag{2}$$

$$\frac{\mathrm{d}T_{CWO1}}{\mathrm{d}t}=\frac{1}{V_{J1}}(Q_{CW1}(T_{CW1}-x_3)+U_{a1}(x_2-x_3)) \tag{3}$$

$$\frac{\mathrm{d}C_{O2}}{\mathrm{d}t}=-K_2x_4+K_{V1}\sqrt{V_1}\frac{(x_1-x_4)}{V_2}+Q_{I2}\frac{(C_{I2}-x_4)}{V_2} \tag{4}$$

$$\frac{\mathrm{d}T_{O2}}{\mathrm{d}t}=\Delta HK_2x_4+K_{V1}\sqrt{V_1}\frac{(x_2-x_5)}{V_2}+Q_{I2}\frac{(T_{I2}-x_5)}{V_2}-U_{a2}\frac{(x_5-x_6)}{V_2} \tag{5}$$

$$\frac{\mathrm{d}T_{CWO2}}{\mathrm{d}t}=\frac{1}{V_{J2}}(Q_{CW2}(T_{CW2}-x_6)+U_{a2}(x_5-x_6)) \tag{6}$$

其中，

$$K_1=K_0\mathrm{e}^{-\frac{E}{Rx_2}} \tag{7}$$

$$K_2=K_0\mathrm{e}^{-\frac{E}{Rx_5}} \tag{8}$$

式中 6 个过程状态变量分别为 CSTR1 的出口浓度 C_{O1}、出口温度 T_{O1}、冷却剂温度 T_{CWO1}、CSTR2 的出口浓度 C_{O2}、出口温度 T_{O2}、冷却剂温度 T_{CWO2}。模型中的其余参数如表 1 所示[16]。

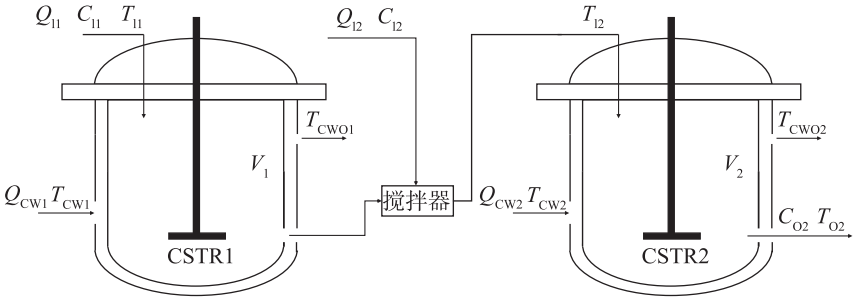


图 1　双 CSTR 结构图

Fig.1　Double CSTR structure diagram

表 1　双 CSTR 系统的模型参数

Tab.1　Model parameters of a dual CSTR system

变量名及符号	数值	单位
CSTR1 的体积 V_1	4.489	m^3
CSTR2 的体积 V_2	5.493	m^3
CSTR1 的出口阀常数 K_{V1}	0.16	$m^{3/2}/s$
CSTR2 的出口阀常数 K_{V2}	0.256	$m^{3/2}/s$
传热系数乘 CSTR1 的传热面积 U_{a1}	0.35	m^3/s
传热系数乘 CSTR2 的传热面积 U_{a2}	0.35	m^3/s
活化能系数 E/R	6 000	K
反应热系数 ΔH	5	$m^3 \cdot K/mol$
阿伦尼乌斯常数 K_0	2.7×10^8	s^{-1}
CSTR1 冷却夹套体积 V_{J1}	1	m^3
CSTR2 冷却夹套体积 V_{J2}	1	m^3
CSTR1 的入口流量 Q_{I1}	0.339	m^3/s
CSTR2 的入口流量 Q_{I2}	0.261	m^3/s
CSTR1 的冷却水流量 Q_{CW1}	0.45	m^3/s
CSTR2 的冷却水流量 Q_{CW2}	0.272	m^3/s
CSTR1 的入口温度 T_{I1}	300	K
CSTR2 的入口温度 T_{I2}	300	K
CSTR1 冷却剂的入口温度 T_{CW1}	300	K
CSTR2 冷却剂的入口温度 T_{CW2}	300	K
CSTR1 的入口浓度 C_{I1}	20	mol/m^3
CSTR2 的入口浓度 C_{I2}	20	mol/m^3

2　强化学习

2.1　马尔可夫决策过程(Markov decision process, MDP)

强化学习研究智能体(agent)与环境(environment)的相互作用，通过不断学习最优策略(policy)，做出序列决策并获得最大奖励(reward)。强化学习问题通常被建模为马尔可夫决策过程($\boldsymbol{S}$, $\boldsymbol{A}$, $\boldsymbol{p}$, r)，其中 $\boldsymbol{S}$ 是状态空间、$\boldsymbol{A}$ 是动作空间、$\boldsymbol{p}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$为状态转移概率、$r(\boldsymbol{s}_t, \boldsymbol{a}_t)$为奖励函数。在每一个时间步 t 中，智能体都会在当前状态 $\boldsymbol{s}_t$ 中按照其策略 $\boldsymbol{\pi}$ 选取并执行动作 $\boldsymbol{a}_t$，依据奖励函数 $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$和状态转移概率 $\boldsymbol{p}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)$返回奖励 r_t 和状态 $\boldsymbol{s}_{t+1}$。智能体将重复以上操作直到环境终止，如图 2 所示。

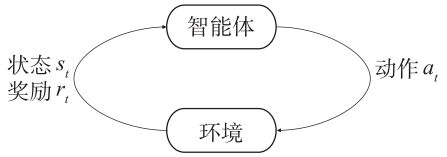


图 2　马尔可夫决策过程

Fig.2　Markov decision process

智能体的目标是使得长期奖励回报最大化，即在其策略 $\boldsymbol{\pi}$ 生成的动作轨迹 $\boldsymbol{\rho}_\pi$ 上，最大化折扣奖励总和的期望：

$$\sum_t E(s_t, \boldsymbol{a}_t) \sim \boldsymbol{\rho}_\pi(\gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t)) \tag{9}$$

其中，$\gamma \in [0, 1)$为折扣因子，为了权衡最大化短期回报和长期回报。

2.2　PPO 算法

SCHULMAN 等[17]提出了近端策略优化算法，首先采用当前策略下的行动概率 $\pi_\theta(\boldsymbol{a}|\boldsymbol{s})$能体行动的效果，即新旧策略的比值记为

$$r_t(\boldsymbol{\theta}) = \frac{\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\theta\text{old}}(\boldsymbol{a}_t|\boldsymbol{s}_t)} \tag{10}$$

其次 PPO 算法通过截断的方式将策略更新限制于一个小范围内，以此来提高智能体训练时的稳定性。即 PPO 的目标函数为

$$L^{\text{CLIP}}(\boldsymbol{\theta}) = \hat{E}_t(\min(\boldsymbol{r}_t(\boldsymbol{\theta})\hat{A}_t, \text{clip}(r_t(\boldsymbol{\theta}), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)) \tag{11}$$

其中，ε 为截断常数，通常取值为 0.1 或 0.2；clip 函数为截断函数，将新旧策略参数 $r_t(\boldsymbol{\theta})$ 的值限定在$[1-\varepsilon, 1+\varepsilon]$区间，如图 3 所示。目标函数使用 min 函数表示选取新旧策略概率比与截断函数中较小的值。

当优势函数 $\hat{A}$ 为正时，说明当前时刻的动作对优化目标有积极效果，所以应该增加其出现的概率，同时要限制其更新范围在 $1+\varepsilon$ 以下。当 $\hat{A}$ 为负时，说明当前行为是消极的，所以应当被阻止，同时降低其概率到 $1-\varepsilon$。图 4 显示了 PPO 算法训练的整体框架[18]。

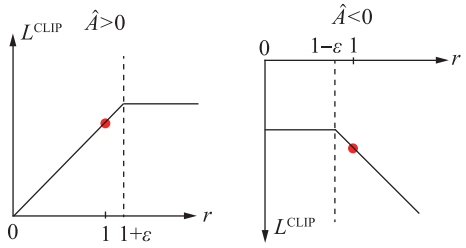


图 3 clip 函数示意图

Fig.3 Schematic diagram of clip function

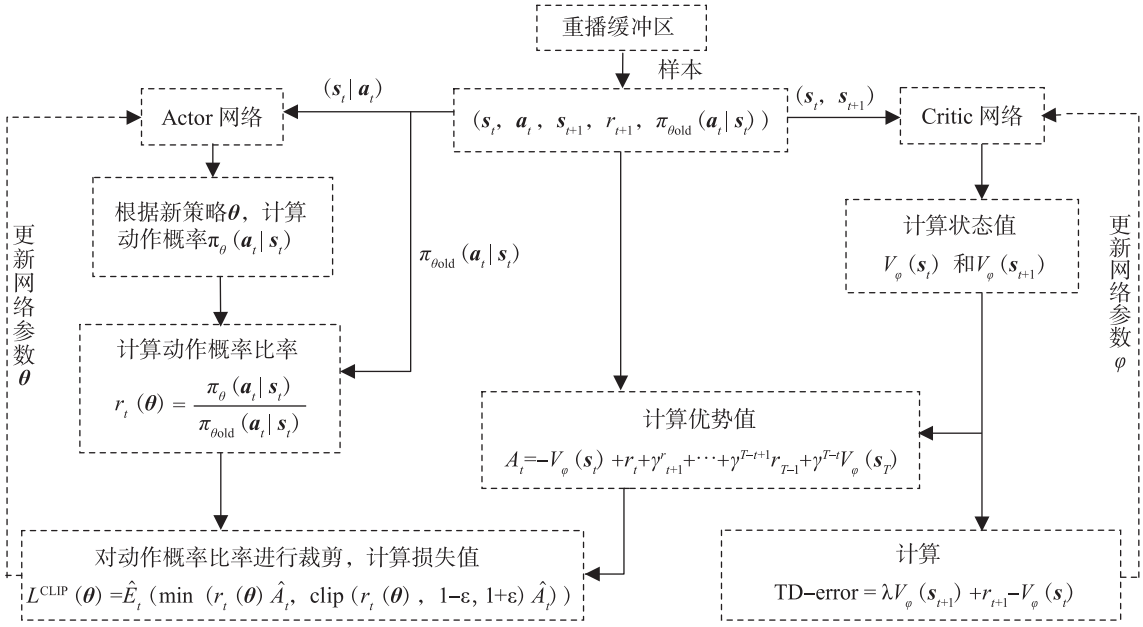


图 4 PPO 算法框图

Fig.4 PPO algorithm block diagram

2.3 动作空间的探索

2.3.1 状态相关探索

状态相关探索（state-dependent exploration，SDE）依赖状态探索是一个中间的解决方案[19]，它将噪声作为状态 $\boldsymbol{s}_t$ 的函数添加到确定性动作 $\boldsymbol{\mu}(\boldsymbol{s}_t)$ 中，θ_μ 是确定性策略的参数。在每一轮探索开始时，探索函数的参数 θ_ε 由高斯分布产生，由此动作 $\boldsymbol{a}_t$ 可通过式(12)得出：

$$\boldsymbol{a}_t=\boldsymbol{\mu}(\boldsymbol{s}_t;\theta_\mu)+\boldsymbol{\varepsilon}(\boldsymbol{s}_t;\theta_\varepsilon),\ \theta_\varepsilon\sim N(0,\boldsymbol{\sigma}^2) \tag{12}$$

在本文剩余部分中为了避免符号重载，将去掉时间下边 t，即符号 $\boldsymbol{s}$ 代表 $\boldsymbol{s}_t$。而 $\boldsymbol{s}_j$ 或 $\boldsymbol{a}_j$ 现在代表状态向量或者动作向量中的一个元素。

对于利用高斯分布进行运算的线性探索函数 $\boldsymbol{\varepsilon}(\boldsymbol{s};\theta_\varepsilon)=\theta_\varepsilon\boldsymbol{s}$，RÜCKSTIEβ 等[20]表明，动作元素 $\boldsymbol{a}_j$ 是正态分布的

$$\boldsymbol{\pi}_j(a_j|s)\sim N(\boldsymbol{\mu}_j(s),\hat{\sigma}_j^2) \tag{13}$$

其中，$\hat{\boldsymbol{\sigma}}$ 是元素$\hat{\boldsymbol{\sigma}}_j=\sqrt{\sum_i(\boldsymbol{\sigma}_{ij}\boldsymbol{s}_i)^2}$ 的对角矩阵。

因为知道策略的轨迹，所以可以得到其对数似然 $\ln\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s})$ 对于方差 $\boldsymbol{\sigma}$ 导数

$$\frac{\partial\ln\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s})}{\partial\boldsymbol{\sigma}_{ij}}=\frac{(\boldsymbol{a}_j-\mu_j)^2}{\hat{\boldsymbol{\sigma}}_j^3}\cdot\frac{-\hat{\boldsymbol{\sigma}}_j^2\boldsymbol{s}_i^2\boldsymbol{\sigma}_{ij}}{\hat{\boldsymbol{\sigma}}_j} \tag{14}$$

对于非线性探索函数，策略 $\boldsymbol{\pi}(\boldsymbol{a}|\boldsymbol{s})$ 的分布大部分时间是未知的。因此，计算精确导数是非常困难的，可能需要近似推理。

2.3.2 广义状态相关探索

由于式(13)和式(14)具有局限性[21]：

1）噪声在一次探索回合中不会改变，如果回合过长，整个回合的探索会受到限制。

2）策略的方差 $\hat{\boldsymbol{\sigma}}_j=\sqrt{\sum_i(\hat{\boldsymbol{\sigma}}_{ij}\boldsymbol{s}_i)^2}$ 的维度大小取决于状态空间的维度，这就意味着不同的问题要调整初始的 $\boldsymbol{\sigma}$。

3）状态和探测噪声之间只有线性相关性，这限制了可能性。

4）状态必须标准化，因为梯度和噪声大小取决于状态大小。

为了解决上述问题并使其适应深度强化学习算法，提出了两项改进：

a）参数 θ_ε 每 n 步采样一次，而不是一次探索循环采样一次。

b）可以使用任何特征作为噪声函数的输入，所以选择策略特征 $\boldsymbol{Z}_\mu(\boldsymbol{s};\theta_{Z_\mu})$（确定性输出前的最后一层 $\boldsymbol{\mu}(\boldsymbol{s})=\boldsymbol{\theta}_\mu\boldsymbol{Z}_\mu(s;\theta_{Z_\mu})$）作为噪声函数的输入 $\boldsymbol{\varepsilon}(\boldsymbol{s};\theta_\varepsilon)=\theta_\varepsilon\boldsymbol{Z}_\mu(\boldsymbol{s})$。

参数 θ_ε 每 n 步采样一次可以解决上述局限性问题 1），并产生一个统一的框架，其中包括非结构化探索（$n=1$）和原始 SDE（$n=$ 循环长度）。使用策略特征可以解决上述局限性问题 2）、3）、4），状态 s 和噪声 $\boldsymbol{\varepsilon}$ 的关系是非线性的，策略的变化只取决于网络的结构。这使得 gSDE 更加独立于任务（因为网络架构通常保持不变），并且在处理大型状态空间（如图像）时节省了大量参数和计算。参数和操作的数量只是最后一层大小和操作维度的函数，而不再是状态空间大小的函数。因此，当使用状态作为噪声函数的输入或策略为线性时，该式（12）更为通用，并包含原始 SDE 描述。

3　基于强化学习的跟踪控制

3.1　跟踪控制目标

由于双 CSTR 系统为多输入多输出系统，本文的控制目标为：在双 CSTR 系统冷却水温度具有扰动 d_1 和 d_2 的情况下，通过控制冷却水流量[Q_{CW1}，Q_{CW2}]，使出口温度[T_{O1}，T_{O2}]达到期望[q_1，q_2]的输出效果，即控制输入 $\boldsymbol{u}=[Q_{\mathrm{CW1}},\ Q_{\mathrm{CW2}}]$，被控输出 $\boldsymbol{y}=[T_{\mathrm{O1}},\ T_{\mathrm{O2}}]$。设定 6 个过程状态变量的初始值分别为[0.084，363，327.57，0.053，363，335.447]。其中干扰、控制输入和期望输出的约束条件如下：

$$\begin{cases}-10\leqslant d_1\leqslant 10,\ -10\leqslant d_2\leqslant 10\\ 0.05\leqslant u_1\leqslant 0.8,\ 0.01\leqslant u_2\leqslant 0.8\end{cases}\tag{15}$$

$$\begin{cases}q_1(t)=q_2(t)=364,\ 0<t\leqslant 30\\ q_1(t)=q_2(t)=365,\ 30<t\leqslant 60\\ q_1(t)=q_2(t)=366,\ 60<t\leqslant 90\\ q_1(t)=q_2(t)=364,\ 90<t\leqslant 130\end{cases}\tag{16}$$

其中，d_1，d_2，q_1，q_2 的单位是热力学温度 K，u_1，u_2 的单位是 $\mathrm{m^3/s}$，t 的单位为 s。

3.2　状态空间和动作空间

图 5 为双 CSTR 系统跟踪控制算法框架图，为了能达到跟踪控制的目的，选取的状态变量为反应器浓度和温度，其均可通过传感器得到数据，便于在实际系统上实现。双 CSTR 系统环境的状态空间为 6 个过程状态变量和出口温度的期望[q_1，q_2]，并且均为连续空间。状态可定义为

$$\boldsymbol{s}_t=[C_{\mathrm{O1}t},\ T_{\mathrm{O1}t},\ T_{\mathrm{CWO1}t},\ C_{\mathrm{O2}t},\ T_{\mathrm{O2}t},\ T_{\mathrm{CWO2}t},\ q_{1t},\ q_{2t}]\tag{17}$$

动作空间则是由冷却水流量[Q_{CW1}，Q_{CW2}]组成的连续空间，即动作可定义为

$$\boldsymbol{a}_t=[Q_{\mathrm{CW1}t},\ Q_{\mathrm{CW2}t}]\tag{18}$$

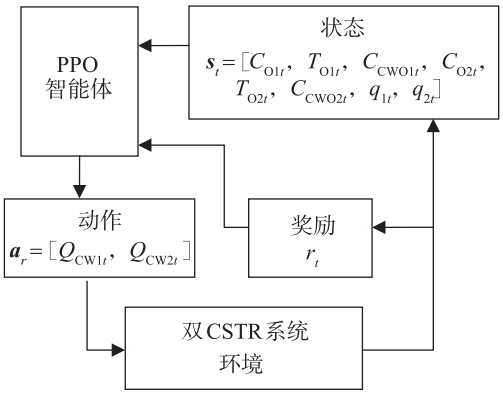


图 5　双 CSTR 系统跟踪控制算法框架图

Fig.5　Block diagram of tracking control algorithm for two-CSTR system

3.3　奖励函数

深度强化学习算法与环境交互得到相应的奖励，通过将预期累积奖励最大化来实现学习到正确的策略。在双 CSTR 系统环境中，每次实验固定为 1300 回合，每回合表示 0.1 s 的抽样，总计为 130 s。为了避免稀疏奖励导致学习效率低而无法收敛的情况，在主线奖励的基础上增加 2 种误差反馈奖励。

奖励函数是决定强化学习算法能否成功收敛的关键[22]，为了保证跟踪控制的有效性和精确性，选取主线奖励为当误差 $e_1(t)=\sqrt{(r_1(t)-y_1(t))^2}$ 和 $e_2(t)=\sqrt{(r_2(t)-y_2(t))^2}$ 同时小于 0.002 时奖励加 1，由于主线奖励在探索过程中很难实现，在初始训练阶段在奖励上无法产生差异，会导致 PPO 智能体难以收敛。为了保证全局范围的探索，将误差 $e_1(t)$ 和 $e_2(t)$ 的负数作为奖励函数的一部分，使得智能体能够向误差更小的方向学习，减小智能体训练的难度。但当误差减小到 0.01 附近时，仅依靠 $e_1(t)$ 和 $e_2(t)$ 的负数所得到的奖励提升太小，已经

无法使智能体向误差更小的方向学习了。因此当 $0.002 \leqslant e_1(t) \leqslant 0.01$ 或 $0.002 \leqslant e_2(t) \leqslant 0.01$ 时，将 $\frac{1}{1\,000 \times e(t)}$ 加入奖励函数之中，能够加速智能体在目标附近的收敛，并满足主线奖励。式(19)为全部的奖励函数。

$$r_t = \begin{cases} -(e_1(t) + e_2(t)), & \\ +1, & e_1(t) \leqslant 0.002,\ e_2(t) \leqslant 0.002 \\ +\dfrac{1}{1\,000 \times e_1(t)}, & 0.002 \leqslant e_1(t) \leqslant 0.01 \\ +\dfrac{1}{1\,000 \times e_2(t)}, & 0.002 \leqslant e_2(t) \leqslant 0.01 \end{cases} \tag{19}$$

4 仿真实验

4.1 环境与训练参数

本文采用双 CSTR 系统作为强化学习的环境，系统参数如表 1 所示。实验设置连续状态空间，智能体与环境交互得到一组状态向量来表示观察结果。通过 PPO-gSDE 算法对智能体进行训练学习，在训练过程中，遵循奖励函数式(19)对动作进行评价。每回合固定步长为 1 300 步，当达到最大步长时，对环境进行初始化，使得环境恢复到初始状态。本实验策略网络和价值网络均为 3 隐含层结构，隐含层大小分别为 64、128、64，激活函数为 Tanh。训练的样本数为 1.5×10^6，gSDE 的采样步数 $n=8$，PPO 参数设置如表 2 所示。

表 2 PPO 算法参数设置

Tab.2 Parameter settings of the PPO algorithm

参数	数值
学习率	5×10^{-5}
折扣因子 λ	0.99
截断常数 ε	0.2
批量大小	512
泛化优势估计(GAE λ)	0.95
优化代理损失时的回合数(n epochs)	40
每回合最大步数	1 300

4.2 仿真结果与对比分析

经过训练后得到的策略网络就可以对双 CSTR 系统进行控制，图 6 为训练过程中平均奖励的变化曲线，图 7 为在有干扰的情况下，智能体对双 CSTR 系统跟踪控制的效果。

由图 6 所示，随着迭代次数的增加，平均奖励也在不断的上升，奖励最大达到约 287，最终时刻的奖励为 265。说明网络已经基本收敛并在其终值附近波动，根据设定的奖励函数可知，奖励大于 0 说明大部分时刻的控制误差应小于 0.01，在有干扰存在的情况下，其控制效果必然是存在波动的，所以奖励达到 260 左右已经可以保证良好的跟踪控制效果。

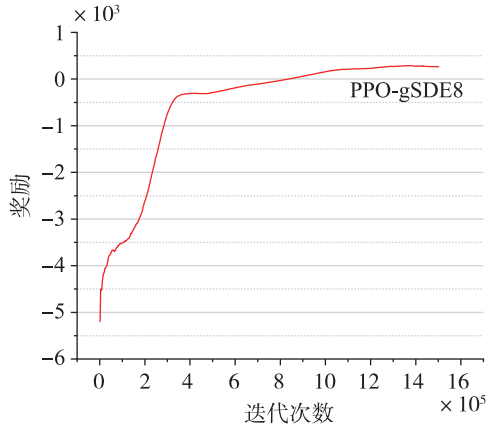
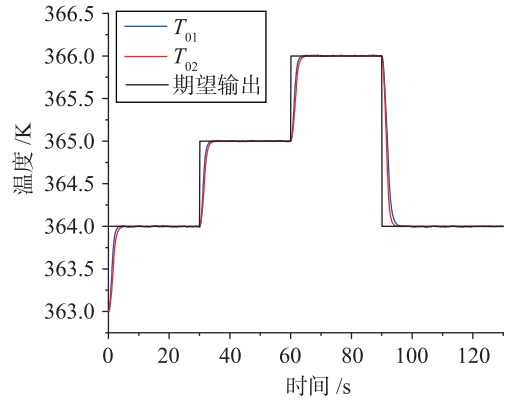
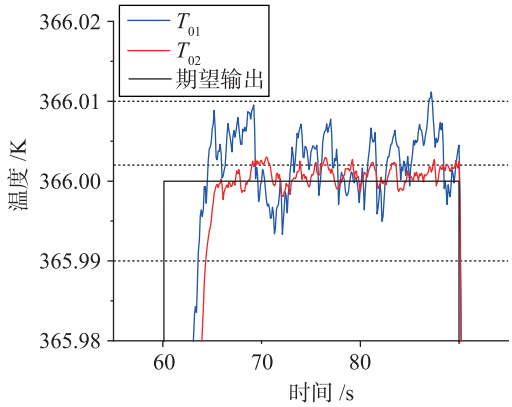


图 6 PPO-gSDE8 奖励变化曲线

Fig.6 PPO-gSDE8 reward change curve



(a) 跟踪控制



(b) 图 (a) 细节放大（55 s~95 s）

图 7 PPO-gSDE8 对双 CSTR 跟踪控制

Fig.7 PPO-gSDE8 for two-CSTR tracking control

如图 7(a)和 7(b)所示，2 个被控输出均成功地跟

踪上期望输出，其在期望输出附近波动是由于干扰造成的。图 7(b)是图 7(a)在 55 s～95 s 之间的细节图，可以看出 T_{O1} 在有干扰的情况下，误差在 0.01 以下，而 T_{O2} 误差小于 0.003。说明 PPO-gSDE 算法对干扰能够迅速地反应，并做出正确的决策以消除干扰带来的影响，保证跟踪控制的精确性。由图 7 得出结论，除了干扰引起的一些波动外，PPO-gSDE 通常以更高的精度跟踪 T_{O1} 和 T_{O2} 的参考输出。

为了充分证明 PPO-gSDE 算法和奖励函数的优越性，本文将与其他几种方法进行比较。探索方式包括高斯分布噪声和 gSDE 探索，其中 gSDE 探索的采样步数分别取 2、8、32、64 进行对比，对比结果如图 8 所示。

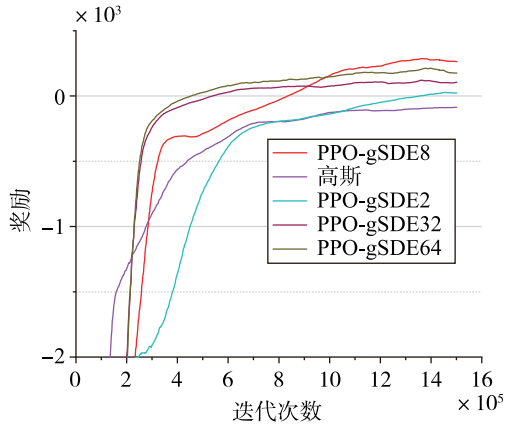


图 8　奖励曲线对比

Fig.8　Reward curve comparison

由图 8 和表 3 可以看出，基于 gSDE 的 PPO 算法在采样步数 $n=8$ 时，既保证了训练的速度，又训练出了更加优秀的智能体，得到的最终奖励远远大于高斯噪声探索得到的最大奖励。虽然在前一半全局探索时期收敛速度相比于 $n=32$ 和 $n=64$ 时的 gSDE 探索较为缓慢，但在后半段的局部探索时期表现优秀，可以使智能体获得更大的奖励。

表 3　算法对比

Tab.3　Algorithm contrast

探索方式	最大奖励	最终奖励	训练时间
gSDE8	286.80	246.06	33 min 47 s
gSDE2	27.165 2	23.68	35 min 24 s
gSDE32	120.53	105.82	34 min 16 s
gSDE64	212.76	175.89	32 min 37 s
高斯(Gaussian)	−84.28	−84.28	33 min 41 s

为了验证 2 种误差反馈奖励的效果，本文做了 2 组对照实验。一组为只有主线奖励，即当误差 $e_1(t)$ 和 $e_2(t)$ 同时小于 0.002 时奖励加 1，另一组在主线奖励的基础上加入误差反馈奖励，即误差 $e_1(t)$ 和 $e_2(t)$ 的负数作为奖励函数的一部分。由于奖励函数的变动，无法从平均奖励来判断其效果的好坏，这里本文直接对比使用不同奖励时的跟踪效果。

由图 9 可以得出，只有主线奖励时由于稀疏奖励，使得无法学习到有用的信息，智能体奖励无法收敛，达不到任何控制效果。而图 10 显示，在添加

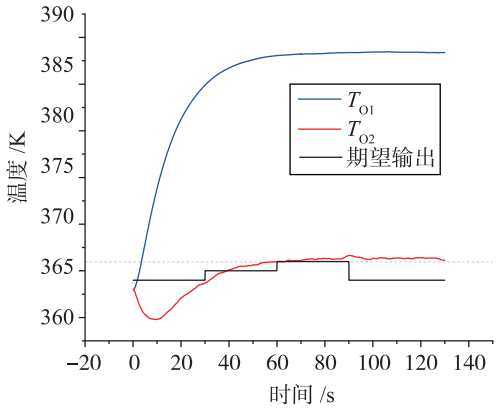
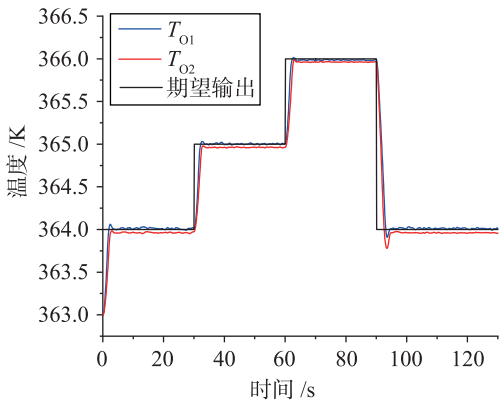
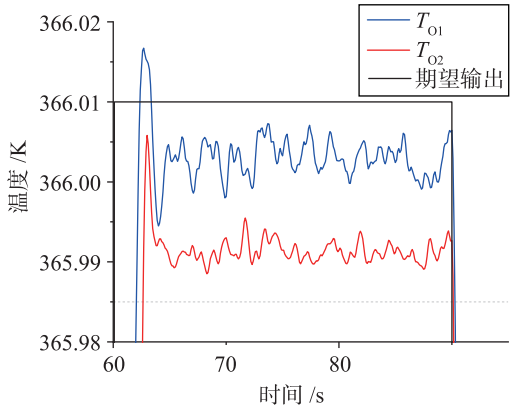


图 9　PPO-gSDE8 只有主线奖励的控制效果

Fig.9　PPO-gSDE8 only has the control effect of mainline reward



(a) 跟踪控制



(b) 细节放大（55 s~95 s）

图 10　PPO-gSDE8 引入一种误差反馈奖励的控制效果

Fig.10　PPO-gSDE8 introduces a control effect of error feedback reward

了 1 种误差反馈奖励以后，稀疏奖励现象得到了有效的改善，使智能体能够学习到相应的控制策略，达到了较为不错的控制效果，但与引入 2 种误差反馈奖励相比，其误差较大。由图 10(b)和图 7(b)对比可知，引入 1 种误差反馈奖励的 T_{01} 误差是引入 2 种误差反馈奖励的 T_{01} 误差的 2 倍，引入 1 种误差反馈奖励的 T_{02} 误差是引入 2 种误差反馈奖励的 T_{02} 误差的 13.3 倍，由此可得，误差反馈奖励在本实验中发挥重要作用。

为了验证本文所提方法的优越性，表 4 和表 5 分别给出了 PPO-gSDE8 与强化学习 SAC、PPO、PID、非线性模型预测控制(NMPC)等算法对于 T_{01} 和 T_{02} 的控制效果。控制指标采用绝对误差积分(integral of absolute error, IAE)，平方误差积分(integral of square error, ISE)和最大绝对误差(maximal deviation from setpoint, DEV^{max})，具体表达式见文[23]。

从表 4 和表 5 中可以看出，相比于其他强化学习算法如 PPO 算法和 SAC(soft actor critic)算法，PPO-gSDE 算法在性能指标上有较大提升，尤其是 PPO-gSDE8 的效果最好。与传统控制算法 PID(proportional-integral-differential)相比，跟踪控制误差更小，大大提高了控制的精确度。而对比预测控制器 NMPC，在 T_{01} 上仅有很小的提升，但对于 T_{02} 的控制却有很大程度的改善。说明本算法在多输入多输出系统的跟踪控制中，能够平衡每个被控变量的误差，实现对双 CSTR 系统的精确控制。

表 4 不同算法在 T_{01} 的控制效果

Tab.4 Control results of different algorithms for T_{01}

控制算法	IAE	ISE	DEV^{max}
PPO-gSDE8	7.751	7.735	1.984
PPO-gSDE64	8.197	7.780	2.004
PPO-gSDE32	8.331	7.820	1.991
PPO-gSDE2	9.136	7.822	2.000
PPO	10.601	8.160	2.028
NMPC	7.798	7.761	1.995
PID	10.088	9.011	2.035
SAC	23.542	12.746	2.223

表 5 不同算法在 T_{02} 的控制效果对比

Tab.5 Control results of different algorithms for T_{02}

控制算法	IAE	ISE	DEV^{max}
PPO-gSDE8	8.072	7.738	1.993
PPO-gSDE64	8.482	7.912	1.995
PPO-gSDE32	8.945	8.116	2.013
PPO-gSDE2	12.048	7.865	2.012
PPO	9.220	8.356	2.029
NMPC	9.221	8.460	2.000
PID	10.665	8.623	2.035
SAC	26.864	14.614	2.073

5 结论

本文提出了基于深度强化学习的多输入多输出系统的跟踪控制，训练得到的智能体能够更加迅速且稳定的对双 CSTR 系统进行控制并达到期望的效果。近端策略优化算法依靠广义状态探索可以在保证探索效果的同时尽量减小状态的变化，以减小在探索过程中对系统的伤害，智能体通过与环境不断交互，经过反复试错得到较优的策略。奖励函数的选择非常重要，本文通过添加反馈奖励使得智能体的控制误差大幅度减小，获得更加明显的控制效果。仿真结果表明，PPO-gSDE 算法与传统的控制算法对比控制误差更小，在有干扰的情况下，能够有效地应用于多输入多输出的双 CSTR 系统，使得系统输出可以迅速且稳定的跟随期望输出。

参考文献

[1] LUYBEN W L. Chemical reactor design and control[M]. New York, USA: John Wiley & Sons, 2007.

[2] BEQUETTE B W. Process control: Modeling, design, and simulation[M]. New Jersey, USA: Prentice Hall Professional, 2003.

[3] AL SEYAB R K S. Nonlinear model predictive control using automatic differentiation[D]. England, UK: Cranfield University, 2006.

[4] AL SEYAB R K, CAO Y. Differential recurrent neural network based predictive control[J]. Computers & Chemical Engineering, 2008, 32(7): 1533 – 1545.

[5] WANG G, JIA Q S, QIAO J, et al. Deep learning-based model predictive control for continuous stirred-tank reactor system[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(8): 3643 – 3652.

[6] 周红标，张钰，柏小颖，等. 基于自适应模糊神经网络的非线性系统模型预测控制[J]. 化工学报，2020，71(7)：

3201 – 3212.
ZHOU H B, ZHANG Y, BAI X Y, et al. Model predictive control of nonlinear system based on adaptive fuzzy neural network [J]. CIESC Journal, 2020, 71(7): 3201 – 3212.

[7] MARAVELIAS C T, SUNG C. Integration of production planning and scheduling: Overview, challenges and opportunities[J]. Computers & Chemical Engineering, 2009, 33(12): 1919 – 1930.

[8] XIN L P, YU B, ZHAO L, et al. Adaptive fuzzy backstepping control for a two continuous stirred tank reactors process based on dynamic surface control approach[J/OL]. Applied Mathematics and Computation, 2020[2022 – 04 – 02]. https://www.sciencedirect.com/science/article/pii/s0096300320301077. https://doi.org/10.1016/j.amc.2020.125138.

[9] RAWLIK K, TOUSSAINT M, VIJAYAKUMAR S. On stochastic optimal control and reinforcement learning by approximate inference[C]//International Joint Conference on Artificial Intelligence. Keystone, USA: AAAI Press, 2013: 3053 – 3060.

[10] LEVINE S, PASTOR P, KRIZHEVSKY A, et al. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection[J]. The International Journal of Robotics Research, 2018, 37(4/5): 421 – 436.

[11] HUA J, ZENG L, LI G, et al. Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning[J/OL]. Sensors, 2021[2022 – 05 – 16]. https://www.medscil.cn/sci/show.paper.app? id = b61d01239698e369. DOI: //doi.org/10.3390/s21041278.

[12] 黄东晋，蒋晨凤，韩凯丽. 基于深度强化学习的三维路径规划算法[J]. 计算机工程与应用，2020，56(15)：30 – 36.
HUANG D J, JIANG C F, HAN K L. 3D path planning algorithm based on deep reinforcement learning[J]. Computer Engineering and Applications, 2020, 56(15): 30 – 36.

[13] 张振，黄炎焱，张永亮，等. 基于近端策略优化的作战实体博弈对抗算法[J]. 南京理工大学学报，2021，45(1)：77 – 83.
ZHANG Z, HUANG Y Y, ZHANG Y L, et al. Battle entity confrontation algorithm based on proximal policy optimization[J]. Journal of Nanjing University of Science and Technology, 2021, 45(1): 77 – 83.

[14] RAFFIN A, KOBER J, STULP F. Smooth exploration for robotic reinforcement learning[C]//Conference on Robot Learning. New York, USA: PMLR, 2022: 1634 – 1644.

[15] CAO Y, BISS D. An extension of singular value analysis for assessing manipulated variable constraints[J]. Journal of Process Control, 1996, 6(1): 37 – 48.

[16] CAO Y, YANG Z. Multiobjective process controllability analysis[J]. Computers & chemical engineering, 2004, 28(1/2): 83 – 90.

[17] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. (2017 – 08 – 28)[2022 – 09 – 10]. https://doi.org/10.48550/arXiv.1707.06347.

[18] 施伟，冯旸赫，程光权，等. 基于深度强化学习的多机协同空战方法研究[J]. 自动化学报，2021，47(7)：1610 – 1623.
SHI W, FENG Y H, CHENG G Q, et al. Research on multi-aircraft cooperative air combat method based on deep reinforcement learning [J]. Acta Automatica Sinica, 2021, 47(7): 1610 – 1623.

[19] RÜCKSTIESS T, SEHNKE F, SCHAUL T, et al. Exploring parameter space in reinforcement learning[J]. Paladyn, 2010, 1(1): 14 – 24.

[20] RÜCKSTIE T, FELDER M, SCHMIDHUBER J. State-dependent exploration for policy gradient methods[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer, 2008: 234 – 249.

[21] VAN H H, TANNEBER G D, PETERS J. Generalized exploration in policy search[J]. Machine Learning, 2017, 106(9): 1705 – 1724.

[22] 杨惟轶，白辰甲，蔡超，等. 深度强化学习中稀疏奖励问题研究综述[J]. 计算机科学，2020，47(3)：182 – 191.
YANG W Y, BAI C J, CAI C, et al. Survey on sparse reward in deep reinforcement learning[J]. Computer Science, 2020 , 47(3): 182 – 191.

[23] 周红标. 基于自组织模糊神经网络的污水处理过程溶解氧控制[J]. 化工学报，2017，68(4)：1516 – 1524.
ZHOU H B. Dissolved oxygen control of the wastewater treatment process using self-organizing fuzzy neural network[J]. CIESC Journal, 2017, 68(4): 1516 – 1524.

作者简介

史洪岩(1977 –)，女，博士，副教授。研究领域为工业过程优化和控制，复杂系统分析设计。

付国城(1999 –)，男，硕士生。研究领域为工业过程优化和控制。

潘多涛(1979 –)，男，博士，副教授。研究领域为系统生物学，复杂流程工业过程建模与优化控制。