强化学习性能最优控制框架及其在高压给水加热器运行优化中的应用

周东阳[1,2]，曹军[2]，毕胜山[1]，邵壮[3]，司风琪[3]
（1.西安交通大学热流科学与工程教育部重点实验室，710049，西安；2.西安热工研究院有限公司，710054，西安；3.东南大学能源热转换及其过程测控教育部重点实验室，210096，南京）

摘要：针对现阶段火电机组运行工况频繁波动的情况，为了解决复杂动态过程难以辨识、控制器设定点无法确定的问题，提出了一种基于历史运行数据与强化学习算法的性能最优控制框架。在现有控制器的输出上叠加少量随机噪声，采用均匀化网格算法构建并维护包含典型工况的数据缓冲区，采用基于粒子群优化的连续批量Q学习算法离线求解性能最优控制策略函数。以高压给水加热器控制任务为研究对象，得到了一种无需系统辨识也无需确定设定点即可保持变工况控制品质与换热性能的控制器求解方法。为了验证所提框架的通用性，利用某600 MW机组高压加热器的仿真模型对水位控制过程进行了分析。结果表明，基于强化学习的性能最优控制框架不需要建立系统模型，可以直接利用历史运行数据求解以累积性能最优为目标的控制策略函数，不仅在动态过程中可以达到较好的控制品质，稳态下也能使系统维持在性能较优的状态，相当于同时实现了设定值优化与设定点跟踪控制。

关键词：给水加热器；强化学习；最优控制；运行优化
中图分类号：TK26 **文献标志码**：A
DOI：10.7652/xjtuxb202208004 **文章编号**：0253-987X(2022)08-0032-11

OSID码

Reinforcement-Learning-Based Performance Optimal Control Framework and Its Application in Operation Optimization of High Pressure Feedwater Heaters

ZHOU Dongyang[1,2], CAO Jun[2], BI Shengshan[1], SHAO Zhuang[3], SI Fengqi[3]
(1. MOE Key Laboratory of Thermo-Fluid Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. Xi'an Thermal Engineering Research Institute Co., Ltd., Xi'an 710054, China; 3. Key Laboratory of Energy Thermal Conversion and Control, Ministry of Education, Southeast University, Nanjing 210096, China)

Abstract: In view of the frequent fluctuations in the operating conditions of thermal power units, to solve the problems that the complex dynamic process is difficult to be identified and that the controller setpoint cannot be determined, this paper proposes a performance optimal control framework based on historical operating data and reinforcement learning algorithms. Firstly, a small amount of random noise is superimposed on the output of the existing controller; then a data buffer containing typical operating conditions is built and maintained with the homogenization grid algorithm; and finally the performance optimal control policy function is solved offline with batch Q-learning algorithm based on particle swarm optimization. Focusing on

收稿日期：2021-11-02。 作者简介：周东阳(1979—)，男，在职博士生；毕胜山(通信作者)，男，教授。 基金项目：国家自然科学基金资助项目(51776171)；陕西省重点研发计划资助项目(2020KW-021)。
网络出版时间：2022-04-26 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20220425.1243.002.html

the research into the control task of high-pressure feedwater heaters, this paper proposes a controller solving method that can maintain the control quality and heat transfer performance under variable operating conditions without system identification or setpoint determination. In order to verify the versatility of the performance optimal control framework, the water level control process was analyzed by using the simulation model of a high-pressure heater of a 600 MW unit. The results show that the reinforcement-learning-based performance optimal control framework works without the need of establishing a system model, and can directly solve the control policy function using historical data for the optimal cumulative performance, which not only realizes better control in the dynamic process, but also keeps the system in its optimal performance in a steady state. This method achieves setpoint optimization and setpoint tracking control at the same time.

Keywords: feedwater heater; reinforcement learning; optimal control; operation optimization

电力行业在持续发展过程中对火电机组的生产技术提出更高的要求，除了最基本的运行稳定之外，还要求过程的高效与智能[1]。传统控制理论中较为成熟的控制器算法，通常将大多数控制任务都简化为设定点跟踪问题，从而通过将输出调节到设定值来确保闭环过程的稳定性。在大多数情况下，这些设定值是根据经验手动设置的，但是现代复杂的火电机组发电过程还需要对设备的性能指标进行优化，使其在运行工况不断变化的过程中保持最优，此时单一不变的设定值往往难以满足需求。现代控制理论在自动控制技术的发展中起着积极的作用，并衍生了最优控制[2]、自适应控制[3]、鲁棒控制[4]和模型预测控制(MPC)[5]等先进控制算法，可以在控制过程中同时实现系统性能优化，但是它们通常依赖对象的动态特性模型，因此对于某些动态特性难以辨识或存在时变的对象[6]，这些基于模型的方法往往难以达到预期的效果。

凝结式给水加热器水位控制是火电机组的一项经典控制任务，目前使用最广泛的是比例-积分-微分(PID)控制器。控制器根据实时水位与目标水位的偏差，通过调节疏水阀，保持水位在目标水位附近。然而，机组负荷在运行过程中持续变化，会改变加热器的边界参数和系统的动态特性。作为设定点跟踪问题，控制器的目标是在不断变化的边界参数下保持水位稳定，但是无法考虑诸如加热器端差和给水温升之类的性能指标。为此，学者们围绕水位与加热器性能的关系开展了研究。Hossienalipour 等[7]建立了一个数学模型来评估加热器的性能，定量分析表明，在某些工况下水位对加热器的换热性能影响很大，给定的水位设定值在大多数情况下都会使加热器偏离其最佳运行状态。Xu 等[8]通过建立一个变工况特性模型，针对性能指标分析了变工况下的最佳水位设定曲线。但是，建立一个精确的数学模型来描述高压蒸汽在复杂物理结构中的凝结过程是非常困难的，其中有大量的换热特性参数需要通过实验手段获取。此外，文献[9]中还分析了的热交换器表面存在的劣化现象，这进一步阻碍了基于模型的优化控制方法的应用。

近些年，学者们提出了许多方法来满足运行优化控制的需求[10]，包括基于模型的方法(如 MPC[5]和实时优化(RTO)[11])和无模型的方法(如数据驱动的优化(DDO)[12-13]和强化学习[12,14-15])。基于强化学习的无模型最优控制方法可以直接利用观测数据求解控制器，而无需建立描述系统的解析表达式。由于计算能力的飞速发展，强化学习近年来受到了极大的关注，并显示出其在许多领域中解决广义控制问题的能力，如国际象棋[16]、跳棋[17]、网络资源分配[18]、视频游戏[19]、围棋[20]等。强化学习结合了动态规划和机器学习两种理论，用于求解序列决策问题的最优策略，在面对维度诅咒和模型不确定性的问题时具有一定优势[21]。通过观察控制器与对象交互的状态转移和相应的奖励信号，强化学习在累积奖励最大化的方向上更新状态、状态-动作组合的价值估计或直接更新控制器参数[22-23]，以逐步改进控制器作用于对象的控制品质。目前，强化学习已在诸如电力系统控制[24]、飞行控制[25]、动态功率管理[26]、无人机[27]和机器人控制[28]等领域实现了应用。在过程控制领域，Jiang 等[29]以浮选工艺为例，设计了基于强化学习的最优控制方法，使用过程生产效率的性能指标取代原有的设定点跟踪目标，证明了强化学习有助于提高过程控制品质和生产效益。

本文以高压给水加热器的水位控制为研究对

象，首先介绍高压给水加热器的物理系统，并对最优控制的数学问题进行形式化，然后介绍基于强化学习的性能最优控制框架，最后利用某 600 MW 机组高压加热器的仿真模型对本文提出的方法进行验证。

1 高压给水加热器水位性能最优控制

目前，火电机组给水加热器主要使用 PID 控制器将水位控制在一个固定设定值附近[1]。然而，持续波动的机组负荷导致加热器的运行工况也在不断变化，而不同工况下最佳水位设定值却是不同的[8]。因此，如果水位设定值固定，则会使加热器偏离其最佳运行状态[7]。考虑到加热器凝结过程较为复杂，难以利用模型来确定不同工况的最佳水位设定值，本文采用基于强化学习的性能最优控制框架来解决高加水位控制问题。

图 1 给出了加热器的物理结构，其中给水从右下侧流入底部水室，平行地流经 U 型管，同时从管壁吸收热量，最终进入顶部水室并流向下一级的加热器。蒸汽侧分为蒸汽冷却段、凝结段和疏水冷却段共 3 个区域。过热蒸汽首先进入蒸汽冷却段，与管壁进行交叉对流换热，冷却至饱和状态后进入冷凝区，在 U 型管表面冷凝成水滴，并流入加热器底部形成疏水，随后通过水封进入疏水冷却区，与管壁进行交叉对流换热。过冷的疏水最终从加热器排出，通过控制阀，流入下一级加热器，控制阀通过调节疏水流量，使加热器水位达到给定的目标值水位。

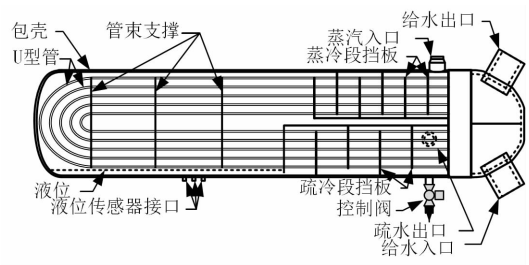


图 1 典型高压加热器的物理结构

Fig. 1 Physical structure of a typical high-pressure feedwater heater

疏水水位随液滴凝结量的增加而升高，随疏水流量的增加而降低，其中液滴凝结量主要取决于冷凝区的换热量，它同时与管侧的给水流量、温度以及壳侧的蒸汽压力、温度等有关，疏水流量则取决于疏水调节阀的开度、当前水位及加热器压力。可见，给水的流量和温度、蒸汽的压力和温度是加热器的 4 个边界条件，影响加热器的动态平衡。水位的动态变化与上述边界条件之间的关系可描述为

$$\begin{cases}\dfrac{\partial l}{\partial t}=\dfrac{Q_{\mathrm{s}}(P_{\mathrm{s}},T_{\mathrm{s}},Q_{\mathrm{w}},T_{\mathrm{w}})-Q_{\mathrm{d}}(P_{\mathrm{s}},l,V)}{A(l)}\\ \dfrac{\partial V}{\partial t}=G(V,a)\end{cases}\tag{1}$$

式中：$A(l)$为水位为 l 时的横截面积；Q_{s} 为凝结液滴的总质量流量；P_{s} 为蒸汽入口压力；T_{s} 为蒸汽入口温度；Q_{w} 为给水入口质量流量；T_{w} 为给水入口温度；Q_{d} 为疏水质量流量；P_{s} 为当前抽汽级压力；l 为水位；V 为疏水阀开度；$G(V,a)$为执行器的动态特性；a 为阀门开度变化率的控制信号。

为了保持加热器的热交换过程稳定且高效，在运行过程中应始终保持合适的水位。当水位太高时，疏水会浸没 U 型管，从而减少凝结段的传热面积；当水位太低时，疏水管中会混有蒸汽，降低蒸汽的利用率，还影响下一级加热器的换热过程。加热器的运行性能指标包括给水温升 ΔT_{w}、给水端差 ΔT_{ttd}和疏水端差 ΔT_{dtd}。ΔT_{w} 是给水出口温度与给水入口温度之差，ΔT_{w} 越高则热力系统效率越高；ΔT_{ttd}是蒸汽入口压力对应的饱和温度与给水出口温度之差，ΔT_{ttd} 越小则说明凝结段的传热性能越好；ΔT_{dtd}是疏水温度和给水入口温度之差，ΔT_{dtd} 越小则说明疏水冷却段的传热性能越好。因此，给定蒸汽和给水的入口参数，好的加热器的运行状态所对应的 ΔT_{w} 大、ΔT_{ttd}小、ΔT_{dtd}小。边界条件和水位都会对 ΔT_{w}、ΔT_{ttd} 和 ΔT_{dtd} 造成影响[30]，因此水位控制需要考虑在满足安全性与稳定性的同时优化上述性能指标。

2 性能最优控制问题的数学形式

将式(1)连续时间状态空间方程转化为离散形式

$$\begin{bmatrix}l_{t+1}\\ V_{t+1}\end{bmatrix}=\begin{bmatrix}F(P_{\mathrm{s},t},T_{\mathrm{s},t},Q_{\mathrm{w},t},T_{\mathrm{w},t},l_t,V_t)\\ G(V_t,a_t)\end{bmatrix}\tag{2}$$

式中：$F(\cdot)$为水位动态特性的差分方程；$G(\cdot)$为执行器动态特性的差分方程。

高压给水加热器的离散时间性能最优控制问题的优化目标为

$$\begin{cases}\underset{a_t=\pi(\cdot)}{\arg\max}\displaystyle\sum_{t=t_0}^{\infty}\gamma^{t-t_0}\Big[\omega_1\Delta T_{\mathrm{w},t}-\omega_2\Delta T_{\mathrm{ttd},t}-\\ \qquad\omega_3\Delta T_{\mathrm{dtd},t}-\omega_4\Big(\dfrac{\partial l_t}{\partial t}\Big)^2-\lambda_{\mathrm{l}}P_{\mathrm{l},t}-\lambda_{\mathrm{a}}P_{\mathrm{a},t}\Big]\\ \text{s.t.}\quad -1\leqslant a_t\leqslant 1\end{cases}\tag{3}$$

式中：$\gamma\in(0,1]$为折扣因子；$[\omega_1,\omega_2,\omega_3,\omega_4]^{\mathrm{T}}\in\mathbb{R}^4$为 3个性能指标和水位变化率的平方的权重向量；$P_{\mathrm{l},t}$为水位超限惩罚函数；$P_{\mathrm{a},t}$为避免疏水阀全开或

全关的软约束函数；λ_l 和 λ_a 分别为惩罚的权重；$\pi(\cdot)$为控制策略函数；在限值之外的二次项形式是为了保证优化目标的一阶导数连续；$P_{l,t}$和 $P_{a,t}$均为不等式约束，利用拉格朗日乘子将其引入到目标函数中，公式为

$$\begin{cases} P_{l,t}=\begin{cases}0, & \text{if } l_{\min}\leqslant l_t\leqslant l_{\max}\\ (l_t-l_{\min})^2, & \text{if } l_t<l_{\min}\\ (l_t-l_{\max})^2, & \text{if } l_t>l_{\max}\end{cases}\\ P_{a,t}=\begin{cases}0, & \text{if } V_{\min}\leqslant V_t\leqslant V_{\max}\\ (V_t-V_{\min})^2, & \text{if } V_t<V_{\min}\\ (V_t-V_{\max})^2, & \text{if } V_t>V_{\max}\end{cases}\end{cases} \tag{4}$$

其中，$l_{\max}$和 $l_{\min}$分别为水位上下限，$V_{\max}$和 $V_{\min}$分别为阀位的上下限。

3　基于强化学习的性能最优控制框架

为了使用异策略连续动作强化学习算法解决式(3)所示的优化问题，同时避免性能较差的初始策略函数参与真实物理系统的运行。本文首先提出了基于强化学习的性能最优控制框架，然后重点介绍其中数据缓冲区的数据处理算法和用于求解策略函数的强化学习算法，最后利用两个算例对框架的性能进行验证。

图 2 给出了基于强化学习的性能最优控制框架。由图可知，基于强化学习的性能最优控制框架包括在线控制、数据预处理和策略函数求解共 3 个主要环节。首先通过在线控制环节生成大量历史运行数据，然后在数据预处理环节，利用均匀化网格算法(homogenization grid algorithm，HGA)算法对训练样本进行整理，最后在策略函数求解环节，利用基于粒子群优化的连续批量 Q 学习算法(particle swarm optimization continues batch Q-learning algorithm，PSO-CBQ)算法训练控制策略函数。最终得到的控制策略函数在通过性能测试之后，可以替代现有控制器，以改善系统的运行水平。

3.1　在线控制环节

图 2 中的在线控制环节描述了真实物理系统受外部扰动和控制动作的共同影响而持续地进行状态转移的过程。真实物理系统在时刻 t 的内部状态为 $s_{i,t}$，它在外部扰动 $s_{d,t}$和控制动作 a_t 的影响下，于 $t+1$ 时刻转变为 $s_{i,t+1}$，由 t 到 $t+1$ 的状态变化称为一组状态转移样本。

为了提高状态转移样本的多样性，本文在现有控制器的输出上叠加了少量随机噪声，最终作用在真实物理系统的控制动作 a_t 满足以现有控制器实际输出为均值的正态分布

$$a_t \sim N(a_{\text{online},t},\sigma) \tag{5}$$

式中：$a_{\text{online},t}$为现有控制器的输出；σ 为扰动的方差。

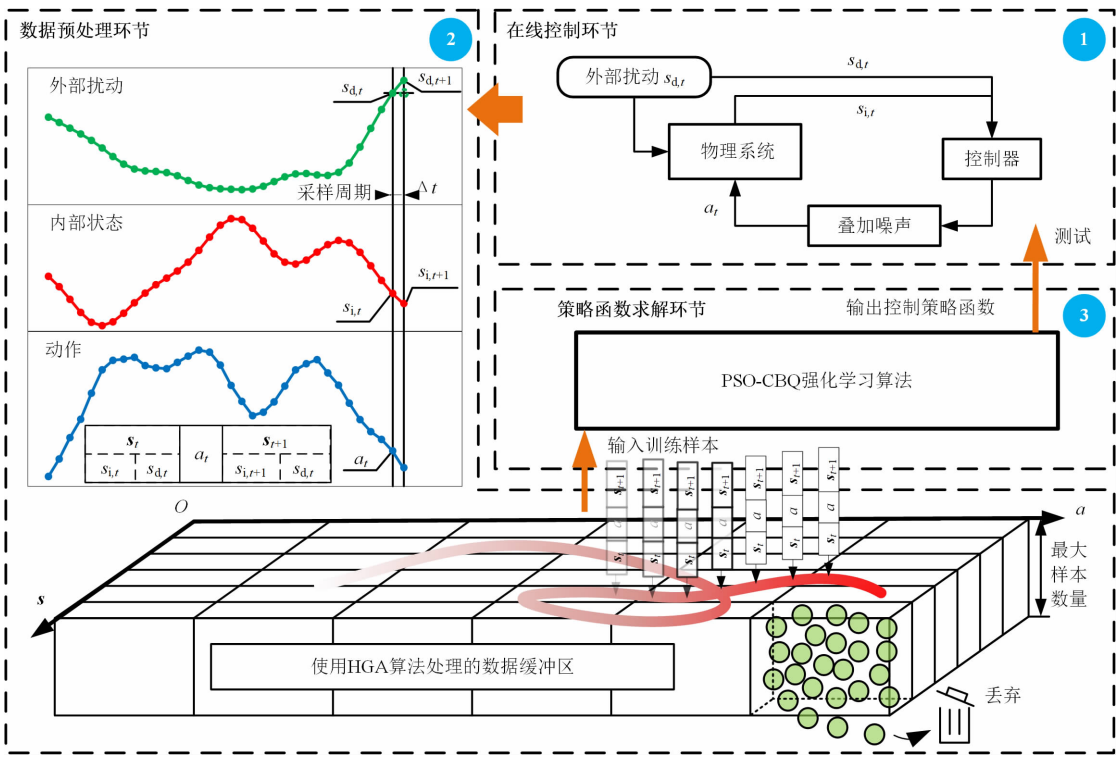


图 2　基于强化学习的性能最优控制框架

Fig. 2　Performance optimal control framework based on reinforcement learning

3.2 数据预处理环节

图 2 中的数据预处理环节从时间序列中采集状态转移样本并生成用于训练策略函数的数据集。在样本采集过程中，需要将时间序列数据构造成 $\boldsymbol{s}_t$，a_t，$\boldsymbol{s}_{t+1}$元组的形式，其中 $\boldsymbol{s}_t=[s_{\mathrm{i},t},s_{\mathrm{d},t}]^{\mathrm{T}}$ 而 $\boldsymbol{s}_{t+1}=[s_{\mathrm{i},t+1},s_{\mathrm{d},t}]^{\mathrm{T}}$。这里需要注意的是，$\boldsymbol{s}_{t+1}$所包含的外部扰动是 $s_{\mathrm{d},t}$而不是 $s_{\mathrm{d},t+1}$，这是因为 $t+1$ 时刻的外部扰动 $s_{\mathrm{d},t+1}$与 t 时刻的外部扰动 $s_{\mathrm{d},t}$无关且不由 a_t 决定，这样的设置隐含着外部扰动不会变化的假设，从而使策略函数倾向于将系统调节至稳定状态。

考虑到训练数据集中样本分布不均匀容易导致策略函数训练发散，本文提出了均匀化网格算法，其伪代码如下。

算法 均匀化网格算法

初始化：结构为 $d_{\mathrm{s}}\times d_{\mathrm{a}}\times N_{\max}$ 的数组 $\mathcal{B}$，其中 $N_{\max}$是单个网格中的最大样本数量，下边界向量 $\boldsymbol{B}_{\mathrm{l}}=[s_{\min},a_{\min}]^{\mathrm{T}}$，上边界向量 $\boldsymbol{B}_{\mathrm{u}}=[s_{\max},a_{\max}]^{\mathrm{T}}$，网格数量向量 $\boldsymbol{n}_{\mathrm{c}}=[n_{\mathrm{c,s}},n_{\mathrm{c,a}}]^{\mathrm{T}}$
重复：
  对于每一个从时间序列数据库采集的新元组 $\boldsymbol{s}_t,a_t,\boldsymbol{s}_{t+1}$：
    定义工况向量 $\boldsymbol{v}_{\mathrm{c}}=[\boldsymbol{s}_t,a_t]^{\mathrm{T}}$
    如果 $\boldsymbol{v}_{\mathrm{c}}\in[\boldsymbol{B}_{\mathrm{l}},\boldsymbol{B}_{\mathrm{u}}]$：
      计算网格索引向量 $\boldsymbol{v}_{\mathrm{i}}=\mathrm{floor}(((\boldsymbol{v}_{\mathrm{c}}-\boldsymbol{B}_{\mathrm{l}})./[(\boldsymbol{B}_{\mathrm{u}}-\boldsymbol{B}_{\mathrm{l}})./\boldsymbol{n}_{\mathrm{c}}]))$，floor(·)为逐元素向下取整
      将 $\boldsymbol{s}_t,a_t,\boldsymbol{s}_{t+1}$插入网格 $\mathcal{B}[\boldsymbol{v}_{\mathrm{i}}]$尾部
      如果网格 $\mathcal{B}[\boldsymbol{v}_{\mathrm{i}}]$长度大于 $N_{\max}$：
        将网格 $\mathcal{B}[\boldsymbol{v}_{\mathrm{i}}]$的头部元素删除

HGA 首先在 $\boldsymbol{s}_t$-a_t 空间划分网格，建立数组结构的数据缓冲区 $\mathcal{B}$，处理时间序列数据时把 $\boldsymbol{s}_t$，a_t，$\boldsymbol{s}_{t+1}$元组依次插入网格中，通过平衡各网格的元组数量，保证历史数据集在 $\boldsymbol{s}_t$-a_t 空间分布均匀。HGA 中的数据缓冲区 $\mathcal{B}$ 具有 3 个特性：①$\mathcal{B}$ 中数据总量有限，可以避免冗余数据无限积累；②$\mathcal{B}$ 中所有网格的数据量均处于同一数量级；③存储在每个网格中的数据会不断更新，更新速度取决于时间序列中该状态动作对出现的频率。相较于其他均匀化算法，HGA 算法尽管不适合处理状态空间维度过高的问题，但是其计算量更小，因此在面对采样周期较短的连续动态优化问题时，可以有效地降低计算负载。

3.3 策略函数求解环节

图 2 中的策略函数求解环节利用从数据缓冲区采集的样本，使用异策略连续动作强化学习算法离线地求解控制策略函数。需要注意的是，当状态 $\boldsymbol{s}_t$ 为连续变量时，要使用参数化 Q 值函数 $Q(\boldsymbol{s},a|\omega^Q)\in\mathbb{R}$ 来近似 Q 值函数，其中 ω^Q 为 Q 值函数的参数。当状态 $\boldsymbol{s}_t$ 和动作 a_t 同时为连续变量时，还要使用参数化策略函数 $\pi(\boldsymbol{s}|\theta)$，其中 θ 为策略函数的参数，此时关于 $Q(\boldsymbol{s},a|\omega^Q)$的最大化运算 $\max_a$、$\arg\max_a$ 求解效率较低，因此有必要对算法进行改进。

考虑到火电厂连续动态优化问题通常具有动作空间维度低的特点，本文使用粒子群优化[31-32]算法来求解最大化运算，即对于给定状态 $\boldsymbol{s}_t$ 和 Q 值函数，以 $Q(\boldsymbol{s}_t,\cdot|\omega^Q)$为粒子群优化的适应度函数，通过在动作空间随机搜索，找到使适应度最大的动作，即为 $\arg\max_a$ 的解，对应的适应度为 $\max_a$ 的解。

结合粒子群优化算法和 Q 学习算法，本文提出了基于粒子群优化的连续批量 Q 学习算法算法，其伪代码如下。

算法 基于粒子群优化的连续批量 Q 学习算法

已知：数据缓冲区 $\mathcal{B}$
初始化：Q 值函数 $Q(\boldsymbol{s},a|\omega^Q)$，$\forall(\boldsymbol{s},a)$，$Q(\boldsymbol{s},a|\omega^Q)=0$，目标 Q 值函数 $Q_{\mathrm{d}}(\boldsymbol{s},a|\omega^{Q_{\mathrm{d}}})$，其中，$\omega^{Q_{\mathrm{d}}}$为目标 Q 值函数的参数，$\omega^{Q_{\mathrm{d}}}\leftarrow\omega^Q$，随机参数初始化的策略函数 $\pi(\boldsymbol{s}|\theta)$，给定 Q 值函数学习率 α、策略函数学习率 α_θ
重复 直到 ω^Q 稳定：
  从 $\mathcal{B}$ 中采集 N 个状态转移样本
  对于第 j 个样本 $\boldsymbol{s}_{t,j},a_{t,j},\boldsymbol{s}_{t+1,j}$：
    $q_{\max}=Q_{\mathrm{d}}(\boldsymbol{s}_{t+1,j},\arg\max_a Q(\boldsymbol{s}_{t+1,j},a|\omega^Q)|\omega^{Q_{\mathrm{d}}})$，其中使用 PSO 求解 $\arg\max_a Q(\boldsymbol{s}_{t+1,j},a|\omega^Q)$
    $r_j=R(\boldsymbol{s}_{t,j},a_{t,j},\boldsymbol{s}_{t+1,j})$
    $q_j=Q(\boldsymbol{s}_{t,j},a_{t,j}|\omega^Q)+\alpha[r_j+\gamma q_{\max}-Q(\boldsymbol{s}_{t,j},a_{t,j}|\omega^Q)]$
  以 $L=\frac{1}{N}\sum_{j=1}^{N}(q_j-Q(\boldsymbol{s}_j,a_j\mid\omega^Q))^2$ 为损失函数，采用梯度下降法更新 ω^Q
  更新 $\omega^{Q_{\mathrm{d}}}\leftarrow(\omega^Q+\omega^{Q_{\mathrm{d}}})/2$
重复 直到 θ 稳定：
  从状态空间均匀采集 M 个样本，利用 Q 值函数和采样策略梯度对 θ 进行一步更新：

$$\theta\leftarrow\theta+\alpha_\theta\left[\frac{1}{M}\sum_{k=1}^{M}\nabla_a Q(\boldsymbol{s}_k,a_k\mid\omega^Q)\nabla_\theta\pi(\boldsymbol{s}_k\mid\theta)\right]$$

首先从数据缓冲区 $\mathcal{B}$ 中收集单步状态转移样本，然后根据初始 Q 值函数 $Q(\boldsymbol{s},a|\omega^Q)$，结合粒子

群优化算法计算样本集中所有状态-动作对的损失函数，并使用梯度下降更新 Q 值函数的参数 ω^{Q}[33]，随后重复该过程直到 ω^{Q} 收敛，最后计算策略函数 $\pi(\boldsymbol{s}|\theta)$ 的策略梯度 $\nabla_{\theta}J$，并使用梯度上升更新策略函数 $\hat{\pi}$ 的参数 θ 直至收敛。

4　算例分析

本文在 Python-3.7 环境下，在 TensorFlow-2.0.0 深度学习库的基础上实现了强化学习性能最优控制框架，并基于仿真运行数据来求解式(3)所示优化问题，以获得水位控制策略函数。首先，使用 APROS-6.04 仿真软件[34]，以某 600 MW 机组#1 高压加热器为研究对象建立了仿真模型，将其作为真实物理系统生成运行数据。该高压加热器模型的结构与参数如图 3 所示，该模型稳定状态在机组 THA 工况附近。

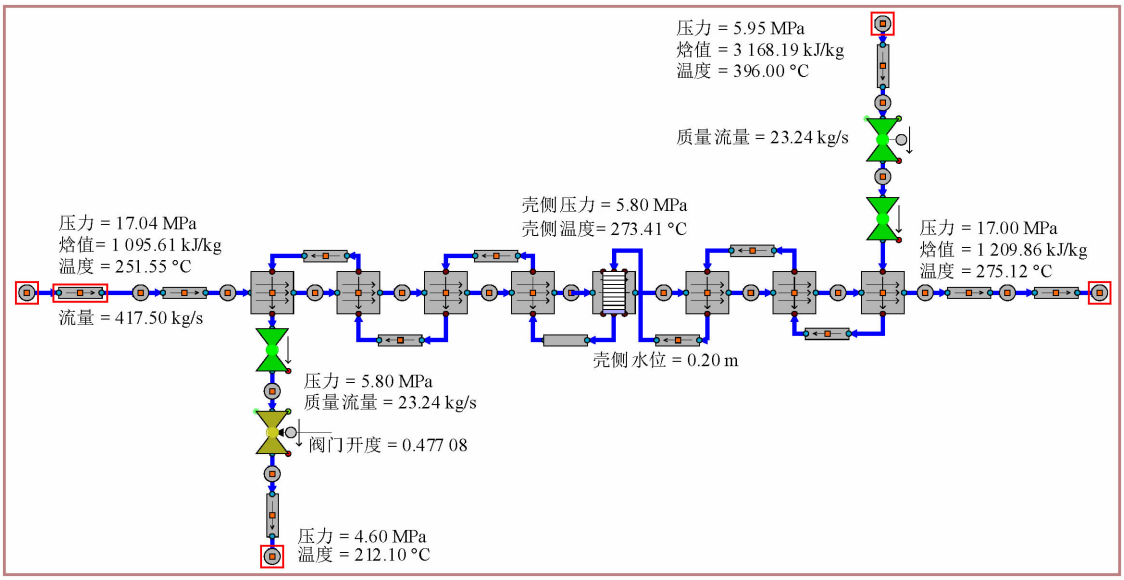


图 3　APROS 高压加热器仿真模型

Fig. 3　APROS simulation model of high pressure heater

采用基于强化学习的性能最优控制框架求解策略函数的超参数如表 1 所示，其中加热器运行工况范围等效于在 500～600 MW 之间。

实验发现，策略函数采用浅层网络效果不佳，Q 值网络采用深层网络的效果不佳。可能的原因在于式(3)所示的优化目标较为复杂，策略函数需要具备足够的特征变换能力，才能具备较好的控制效果，但又为了避免训练过程不稳定，因此适合选择多层少节点的网络结构。Q 值网络需要具备较强的泛化能力，以防止对价值估计的过拟合，综合考虑适合选择少层而多节点的网络结构。奖励函数中的性能指标 $\Delta T_{\mathrm{w},t}$、$\Delta T_{\mathrm{ttd},t}$、$\Delta T_{\mathrm{dtd},t}$ 为

$$\begin{bmatrix}\Delta T_{\mathrm{w},t}\\ \Delta T_{\mathrm{ttd},t}\\ \Delta T_{\mathrm{dtd},t}\end{bmatrix}=\begin{bmatrix}0 & 1 & 0 & -1\\ 0 & -1 & T_{\mathrm{sat}}(\cdot) & 1\\ 1 & 0 & 0 & -1\end{bmatrix}\begin{bmatrix}T_{\mathrm{do},t}\\ T_{\mathrm{wo},t}\\ P_{\mathrm{s},t}\\ T_{\mathrm{w},t}\end{bmatrix}\tag{6}$$

式中 $T_{\mathrm{sat}}(\cdot)$ 为饱和蒸汽温度关于蒸汽压力的函数，根据 IAPWS-IF97 标准公式计算。

图 4 给出了 Q 值神经网络及其策略神经网络参数的平均绝对变化率的变化趋势。图中，蓝色阴影为 10 轮不同训练过程中 95%置信水平对应的置信区间。可以看出，学习过程是不稳定的，网络参数的平均绝对变化率没有单调下降，且在 300 次迭代之前均存在较大的方差，不过在 300 次迭代之后逐渐稳定收敛。由此可知，使用神经网络逼近器的性能最优控制框架可以使学习过程收敛至局部最优解。

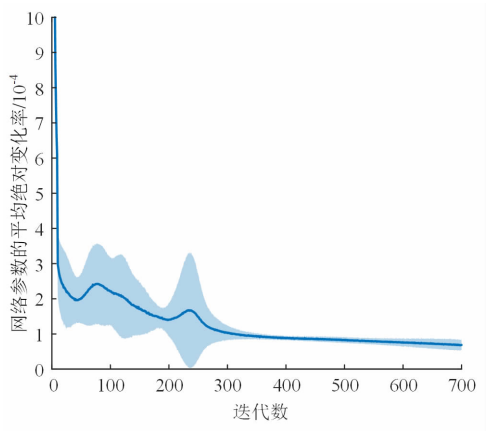
为了验证基于强化学习的性能最优控制框架得到的策略函数的性能，使用某一个收敛的策略神经网络在设计工况附近进行阶跃实验，表 2 给出了设计工况稳定状态下系统的过程参数。

在稳定状态下，将加热器水位设置为 1，观察水位 l_t、阀位 V_t、疏水出口温度 $T_{\mathrm{do},t}$ 和给水出口温度 $T_{\mathrm{wo},t}$ 的响应曲线，如图 5 所示。可以看出，水位可以快速地被调节至初始状态，且过程中阀门开度被限制在的软约束之内。在稳定状态下，分别将蒸汽压力从 5.95 MPa 升至 6.45 MPa、将蒸汽温度从 396 ℃升至 400 ℃、将给水质量流量从 417.5 kg/s 升至 422.5 kg/s、将给水入口温度从 249.5 ℃升至 251.5

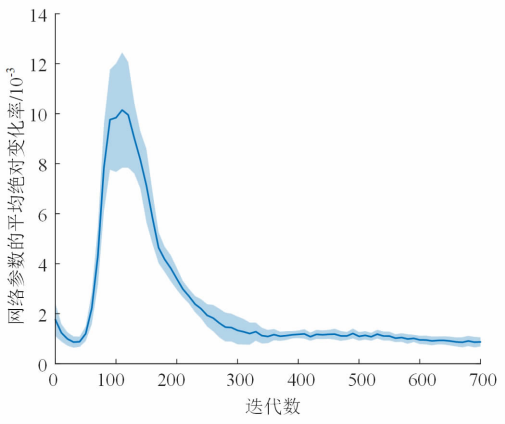
表 1 高压加热器水位性能最优控制算例的超参数

Table 1 Hyperparameter of optimal water level performance control for high pressure heater

参数	数值
σ	0.5
Δt	0.1
$(d_s, d_a, N_{\max})$	(6,1,10)
$\boldsymbol{B}_{\mathrm{l}}$	$[0,-0.003,5.4,393,413,247,0]^{\mathrm{T}}$
$\boldsymbol{B}_{\mathrm{u}}$	$[1,0.003,6.5,399,422,252,1]^{\mathrm{T}}$
$\boldsymbol{n}_{\mathrm{c}}$	$[20,1,3,3,3,3,20]^{\mathrm{T}}$
α	0.01
α_θ	0.001
γ	0.9
N	10 000
M	10 000
现有控制器	$a_{\mathrm{online},t}=-10(l_t-0.5)-100\partial l/\partial t$
$\boldsymbol{s}_t$	$[l_t,\partial l/\partial t_t,P_{\mathrm{s},t},T_{\mathrm{s},t},Q_{\mathrm{w},t},T_{\mathrm{w},t}]^{\mathrm{T}}$
$\boldsymbol{s}_{t+1}$	$[l_{t+1},\partial l/\partial t_{t+1},P_{\mathrm{s},t},T_{\mathrm{s},t},Q_{\mathrm{w},t},T_{\mathrm{w},t}]^{\mathrm{T}}$
奖励函数	$r_t=R^2_{\mathrm{nor}}(\Delta T_{\mathrm{w},t})-R^2_{\mathrm{nor}}(\Delta T_{\mathrm{ttd},t})-R^2_{\mathrm{nor}}(\Delta T_{\mathrm{dtd},t})-\left(\frac{\partial l_t}{\partial t}\right)^2-\frac{P_{\mathrm{l},t}}{2}-\frac{P_{\mathrm{a},t}}{2}$ 其中，$R_{\mathrm{nor}}(\cdot)$是 0－1 归一化函数，$P_{\mathrm{l},t}$和$P_{\mathrm{a},t}$均为不等式约束
$(l_{\min}, l_{\max})$	(0.1,0.9)
$(V_{\min}, V_{\max})$	(0.1,0.9)
策略函数	输入为 $\boldsymbol{s}_t$ 输出为 a_t 的神经网络，3 个隐层各 6 个节点，ReLU 激活函数，输出层激活函数为 tanh
Q 值函数	输入为 $\boldsymbol{s}_t$、a_t 的神经网络，1 个隐层有 128 个节点，tanh 激活函数，输出层无激活函数



(a)Q 值神经网络



(b)策略神经网络

图 4 Q 值神经网络及其策略神经网络参数平均绝对变化率的变化趋势

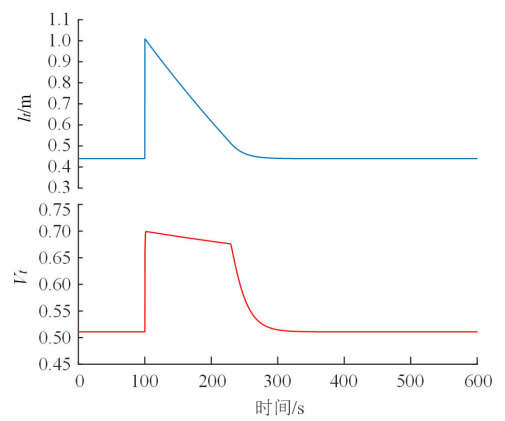
Fig. 4 Trend of Q-value neural network and its policy neural network parameter average absolute rate of change

表 2 设计工况稳定状态下的过程参数

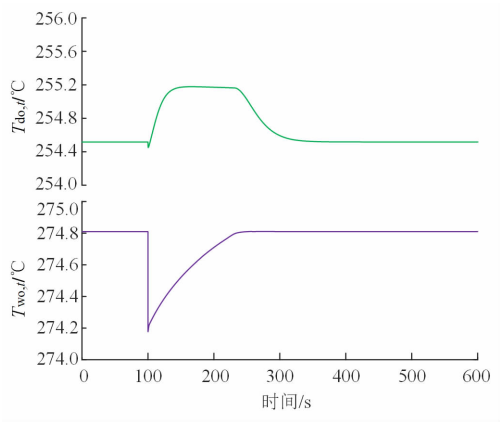
Table 2 Design parameters in steady state of working condition

边界条件	数值	参数	数值
蒸汽入口压力 $P_{\mathrm{s},t}$/MPa	5.95	阀位 V_t	0.51
蒸汽入口温度 $T_{\mathrm{s},t}$/℃	396	水位 l_t/m	0.44
给水质量流量 $Q_{\mathrm{w},t}$/(kg·s^{-1})	417.5	疏水出口温度 $T_{\mathrm{do},t}$/℃	254.5
给水入口温度 $T_{\mathrm{w},t}$/℃	249.5	给水出口温度 $T_{\mathrm{wo},t}$/℃	274.8

℃，观察各参数的响应曲线，如图 6 所示。可以看出，各边界条件不仅影响疏水和给水的出口温度，还改变了新稳态下的水位值，变化过程较快且没有出现超调。



(a)水位和阀位的响应曲线



(b)疏水出口温度和给水出口温度的响应曲线

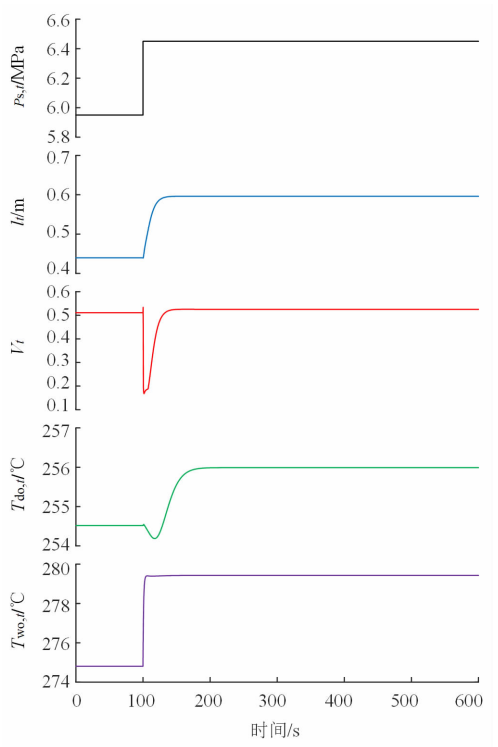
图 5　水位扰动后各参数的响应曲线

Fig. 5　Response curves of each parameter after water level disturbance

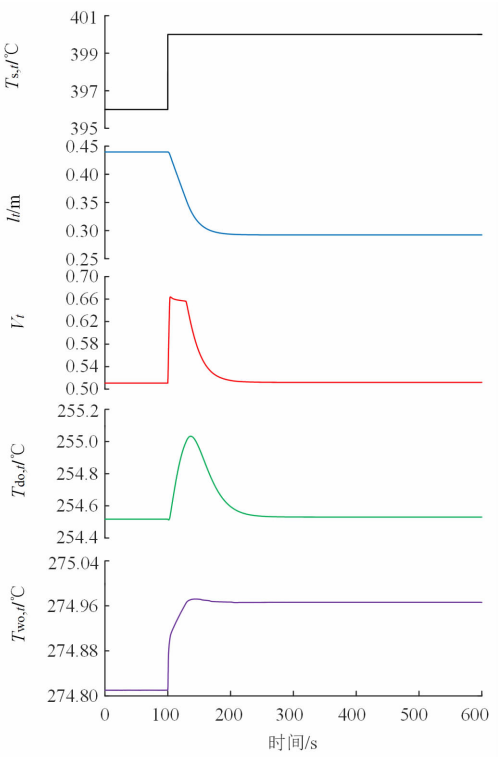
从 249.5 ℃阶跃至 251.5 ℃为了进一步说明基于强化学习的性能最优控制框架得到的策略函数的合理性，将策略函数在对应工况下的稳定水位与试验最优水位进行对比。试验最优水位来自于 APROS 仿真模型的设定值试验优化[34]，它是一种经典的工程优化方法，通过在试验中手动调整控制系统的设定值，以确定各边界条件下以性能指标为目标的最优设定值，并拟合最优设定值与各边界条件的关系曲线以参与控制。

图 7 给出了强化学习策略函数稳定水位与试验最优水位关于每个边界条件的特性对比情况。可以看出，基于强化学习的性能最优控制框架得到的策略函数在各工况下的稳定水位与设定值试验优化得到的最优水位比较接近，相对于边界条件的趋势也相似。在变蒸汽压力条件下，策略函数稳定水位和试验最优水位趋势存在差异。可以看出，试验最优水位的趋势比较平滑，而策略函数稳定水位的曲线在 5.9 MPa 附近存在一个拐点，可能的原因是策略函数采用了 ReLU 的隐层激活函数，导致其函数曲面不连续。可能的改进方法是减少策略函数层数，并使用平滑连续的激活函数。考虑到相关控制策略学习算法的特性，一般训练控制策略所使用数据的工况范围与其适用的工况范围是相近的，因此不建议在训练数据所在范围之外使用得到的控制策略。本文选择在 THA 稳定工况附近对模型的有效性进行了验证，而在范围外的实验效果不佳。

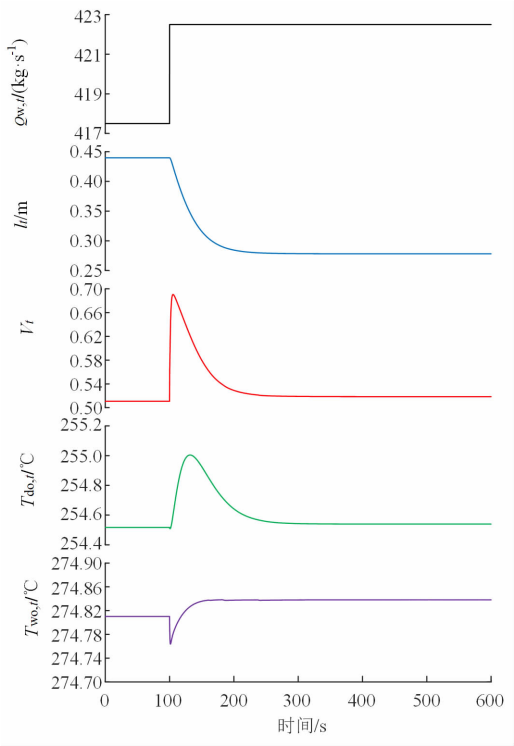
在实际应用中，设定值试验优化在面对多边界条件的场景时，需要进行大量的组合试验以确定各工况下边界条件与最优设定值的关系，而采用基于强化学习的性能最优控制框架可以直接利用历史运行数据求解控制策略函数，不仅在动态过程中可以达到较好的控制品质，稳态下也能使系统维持在性能较优的状态，相当于同时实现了设定值优化与设



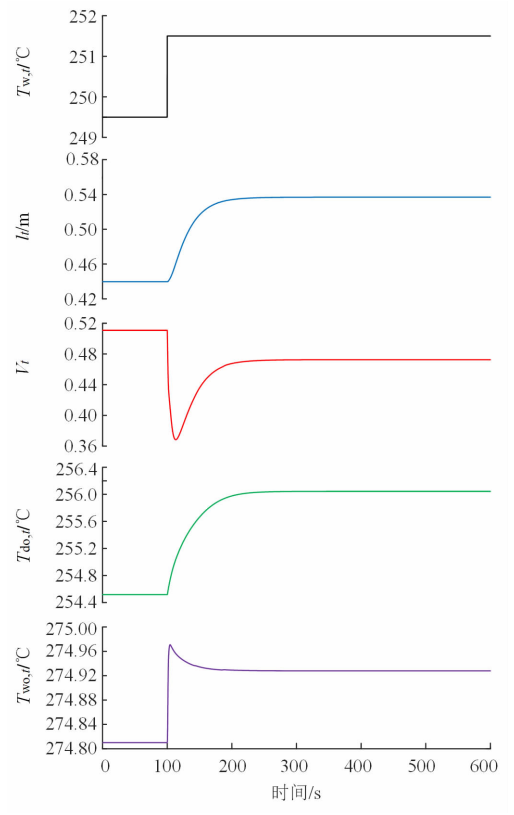
(a)$P_{s,t}$从 5.95 MPa 阶跃至 6.45 MPa



(b)$T_{s,t}$从 396 ℃阶跃至 400 ℃



(c)$Q_{w,t}$从 417.5 kg/s 阶跃至 422.5 kg/s

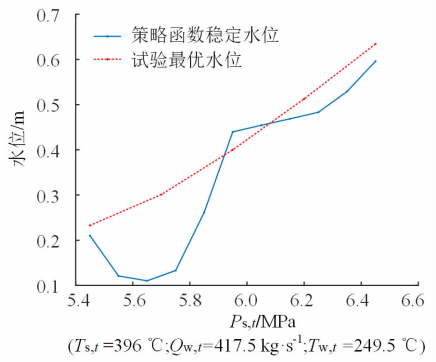


(d)$T_{w,t}$从 249.5 ℃阶跃至 251.5 ℃

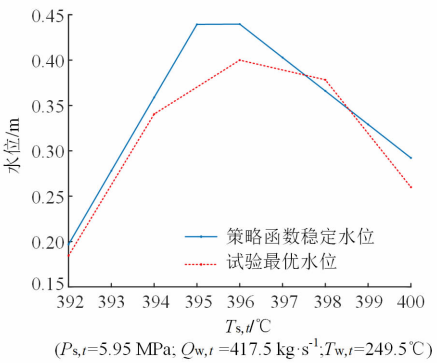
图 6 各边界条件阶跃后各参数的响应曲线

Fig. 6 Step response curves of each parameter with respect to the boundary conditions

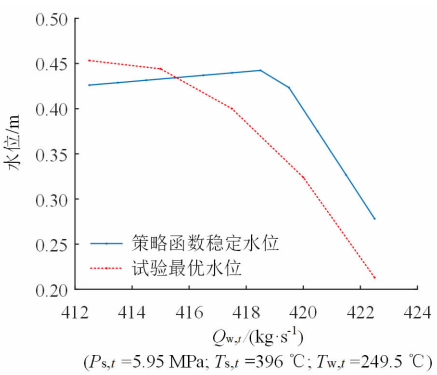
定点跟踪控制。然而,Q 值函数存在近似误差,框架得到的策略函数尚达不到理论最优的控制品质。这是由于 Q 值本身是对单步状态转移的奖励估计,而优化目标是最大化多步累积奖励。采用机器学习算法拟合单步奖励必然会存在误差,在常规的监督学习任务中,这种误差的影响不大,而在强化学习任务中,单步误差的多步累积,可能导致多步优化目标存在较为明显的差异,因此得到的控制策略与解析法相比往往是次优的,但是其优势在于可以处理解析法无法解决的问题,对于解决包含复杂目标的过程控制任务具有较大的潜力。



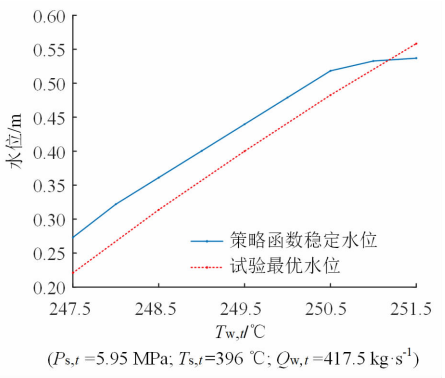
(a)不同 $P_{s,t}$ 下策略函数稳定水位与试验最优水位对比



(b)不同 $T_{s,t}$ 下策略函数稳定水位与试验最优水位对比



(c)不同 $Q_{w,t}$ 下策略函数稳定水位与试验最优水位对比



(d)不同 $T_{w,t}$ 下策略函数稳定水位与试验最优水位对比

图 7　强化学习策略函数稳定水位与试验最优水位关于各边界条件的对比

Fig. 7　Comparison of the stable water level between the experimental optimal water level and reinforcement learning policy function with respect to various boundary conditions

5　结　论

考虑到火电厂对象特性复杂、动作空间维度较低、策略函数在训练期间无法与物理系统交互等特点，本文提出了基于强化学习的性能最优控制框架。在框架的数据预处理环节提出了 HGA 算法，以较低的计算负载解决了数据不平衡问题。在策略函数求解环节提出了 PSO-CBQ 算法，使用粒子群优化准确快速地实现了动作值迭代计算中的最大化运算，解决了连续动作强化学习求解效率低的问题。在高压给水加热器性能最优控制算例中，将基于强化学习的性能最优控制框架训练得到的策略函数与试验最优水位设定值控制器进行了对比。结果表明，基于强化学习的性能最优控制框架不需要建立系统模型，可以直接利用历史运行数据求解以累积性能最优为目标的控制策略函数，不仅在动态过程中可以达到较好的控制品质，稳态下也能使系统维持在性能较优的状态，相当于同时实现了设定值优化与设定点跟踪控制。

参考文献：

[1] LEWIS F L, VRABIE D L, SYRMOS V L. Optimal control [M]. Hoboken, NJ, USA: Wiley, 2012.

[2] WHITE D J. Dynamic programming, Markov chains, and the method of successive approximations [J]. Journal of Mathematical Analysis and Applications, 1963, 6(3): 373-376.

[3] ÅSTRÖM K J. Adaptive control [M]. [S. l.]: Courier Corporation, 2013.

[4] DOYLE J. Robust and optimal control [C]// Proceedings of 35th IEEE Conference on Decision and Control. Piscataway, NJ, USA: IEEE, 1996: 1595-1598.

[5] 李书臣，徐心和，李平. 预测控制最新算法综述 [J]. 系统仿真学报，2004，16(6)：1314-1319，1349.
LI Shuchen, XU Xinhe, LI Ping. An overview of new algorithms in predictive control [J]. Journal of System Simulation, 2004, 16(6): 1314-1319, 1349.

[6] LEWIS F L, VRABIE D. Reinforcement learning and adaptive dynamic programming for feedback control [J]. IEEE Circuits and Systems Magazine, 2009, 9(3): 32-50.

[7] HOSSIENALIPOUR S M, KARBALAEE M S, FATHIANNASAB H. Development of a model to evaluate the water level impact on drain cooling in horizontal high pressure feedwater heaters [J]. Applied Thermal Engineering, 2017, 110: 590-600.

[8] XU Jianqun, YANG Tao, SUN Youyuan, et al. Research on varying condition characteristic of feedwater heater considering liquid level [J]. Applied Thermal Engineering, 2014, 67(1/2): 179-189.

[9] ANTAR M A, ZUBAIR S M. The impact of fouling on performance evaluation of multi-zone feedwater heaters [J]. Applied Thermal Engineering, 2007, 27(14/15): 2505-2513.

[10] CHAI Tianyou, QIN S J, WANG Hong. Optimal operational control for complex industrial processes [J]. Annual Reviews in Control, 2014, 38(1): 81-92.

[11] YIN Shen, LUO Hao, DING S X. Real-time implementation of fault-tolerant control systems with performance optimization [J]. IEEE Transactions on Industrial Electronics, 2014, 61(5): 2402-2411.

[12] LU Xinglong, KIUMARSI B, CHAI Tianyou, et al. Data-driven optimal control of operational indices for a class of industrial processes [J]. IET Control Theory & Applications, 2016, 10(12): 1348-1356.

[13] DAI Wei, CHAI Tianyou, YANG S X. Data-driven optimization control for safety operation of hematite grinding process [J]. IEEE Transactions on Industrial Electronics, 2015, 62(5): 2930-2941.

[14] WANG Ding, HE Haibo, MU Chaoxu, et al. Intelligent critic control with disturbance attenuation for affine dynamics including an application to a microgrid system [J]. IEEE Transactions on Industrial Electronics, 2017, 64(6): 4935-4944.

[15] WANG Ding, HE Haibo, LIU Derong. Improving the critic learning for event-based nonlinear H_∞ control de-

sign [J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3417-3428.

[16] THRUN S. Learning to play the game of chess [M]// Advances in Neural Information Processing Systems. Cambridge, MA, USA: The MIT Press, 1995: 1069-1076.

[17] SAMUEL A L. Some studies in machine learning using the game of checkers [J]. IBM Journal of Research and Development, 2000, 44(1/2): 206-226.

[18] 吴晓军，张成，原盛，等. 基于强化学习的云资源混合式弹性伸缩算法 [J]. 西安交通大学学报，2022，56(1)：142-150.
WU Xiaojun, ZHANG Cheng, YUAN Sheng, et al. Blended elastic scaling method for cloud resources following reinforcement learning [J]. Journal of Xi'an Jiaotong University, 2022, 56(1): 142-150.

[19] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533.

[20] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of Go without human knowledge [J]. Nature, 2017, 550(7676): 354-359.

[21] KHAN S G, HERRMANN G, LEWIS F L, et al. Reinforcement learning and optimal adaptive control: an overview and implementation examples [J]. Annual Reviews in Control, 2012, 36(1): 42-59.

[22] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. 2nd ed. Cambridge, MA, USA: The MIT Press, 2018.

[23] 张铁，廖才磊，邹焱飚，等. 采用强化学习的多轴运动系统时间最优轨迹优化 [J]. 西安交通大学学报，2021，55(7)：33-40.
ZHANG Tie, LIAO Cailei, ZOU Yanbiao, et al. Time-optimal trajectory optimization of multi-axis motion system by reinforcement learning [J]. Journal of Xi'an Jiaotong University, 2021, 55(7): 33-40.

[24] ERNST D, GEURTS P, WEHENKEL L. Tree-based batch mode reinforcement learning [J]. The Journal of Machine Learning Research, 2005, 6: 503-556.

[25] KAMPEN E V, CHU Q P, MULDER J A. Online adaptive critic flight control using approximated plant dynamics [C]// 2006 International Conference on Machine Learning and Cybernetics. Piscataway, NJ, USA: IEEE, 2006: 256-261.

[26] LIU Wei, TAN Ying, QIU Qinru. Enhanced *Q*-learning algorithm for dynamic power management with performance constraint [C]// 2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010). Piscataway, NJ, USA: IEEE, 2010: 602-605.

[27] STINGU P E, LEWIS F L. Adaptive dynamic programming applied to a 6DoF quadrotor [M]// Computational Modeling and Simulation of Intellect: Current State and Future Perspectives. Hershey, PA, USA: IGI Global, 2011: 102-130.

[28] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: a survey [J]. The International Journal of Robotics Research, 2013, 32(11): 1238-1274.

[29] JIANG Yi, FAN Jialu, CHAI Tianyou, et al. Data-driven flotation industrial process operational optimal control based on reinforcement learning [J]. IEEE Transactions on Industrial Informatics, 2018, 14(5): 1974-1989.

[30] 赵则飞，赵轶韬，赵四海，等. 600 MW 汽轮机高压加热器水位运行优化 [J]. 内蒙古电力技术，2017，35(3)：54-56，60.
ZHAO Zefei, ZHAO Yitao, ZHAO Sihai, et al. Water level operation optimization of high-pressure heater in 600 MW unit [J]. Inner Mongolia Electric Power, 2017, 35(3): 54-56, 60.

[31] 杨维，李歧强. 粒子群优化算法综述 [J]. 中国工程科学，2004，6(5)：87-94.
YANG Wei, LI Qiqiang. Survey on particle swarm optimization algorithm [J]. Engineering Science, 2004, 6(5): 87-94.

[32] 王珊，刘明，严俊杰. 采用粒子群算法的热电厂热电负荷分配优化 [J]. 西安交通大学学报，2019，53(9)：159-166.
WANG Shan, LIU Ming, YAN Junjie. Optimizing heat-power load distribution of thermal power plants based on particle swarm algorithm [J]. Journal of Xi'an Jiaotong University, 2019, 53(9): 159-166.

[33] RIEDMILLER M. Neural fitted *Q* iteration: first experiences with a data efficient neural reinforcement learning method [C]// 16th European Conference on Machine Learning. Cham, Germany: Springer, 2005: 317-328.

[34] STARKLOFF R, ALOBAID F, KARNER K, et al. Development and validation of a dynamic simulation model for a large coal-fired power plant [J]. Applied Thermal Engineering, 2015, 91: 496-506.

（编辑 陶晴 亢列梅）