華東理工大學学报 自然科学版 Natural Science Edition
JOURNAL OF EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY

基于混合动作安全强化学习的汽油调合调度优化

石林林 朱志伟 赵云蒙

Gasoline Blending Scheduling Optimization Based on Hybrid-Action Safe Reinforcement Learning

SHI Linlin, ZHU Zhiwei, ZHAO Yunmeng

在线阅读 View online: http://doi.org/10.14135/j.cnki.1006-3080.20260222001

您可能感兴趣的其他文章

Articles you may be interested in

基于强化学习和角度惩罚距离的冰晶连续优化算法

Ice Crystal Continuous Optimization Algorithm Based on Reinforcement Learning and Angle Penalty Distance

华东理工大学学报（自然科学版）. 2021, 47(1): 89-100 http://doi.org/10.14135/j.cnki.1006-3080.20191125003

基于强化学习的正弦优化算法求解能耗分布式流水车间节能调度问题

Sine Optimization Algorithm Based on Reinforcement Learning for Energy-Efficient Distributed Flow Shop Scheduling Problem

华东理工大学学报（自然科学版）. 2025, 51(5): 623-632 http://doi.org/10.14135/j.cnki.1006-3080.20250117002

基于组稀疏优化的强化学习稀疏表征

Reinforcement Learning with Sparse Representation via Sparse Overlapping Group Lasso

华东理工大学学报（自然科学版）. 2024, 50(6): 913-919 http://doi.org/10.14135/j.cnki.1006-3080.20231223001

离散水波优化算法求解带批处理的混合流水车间批量流调度问题

Discrete Water Wave Optimization Algorithm for Hybrid Flowshop Lot-Streaming Scheduling Problem with Batch Processing

华东理工大学学报（自然科学版）. 2021, 47(5): 598-608 http://doi.org/10.14135/j.cnki.1006-3080.20200831001

基于多目标离散正弦优化算法的混合零空闲置换流水车间调度

Mixed No-Idle Permutation Flow Shop Scheduling Problem Based on Multi-Objective Discrete Sine Optimization Algorithm

华东理工大学学报（自然科学版）. 2022, 48(1): 76-86 http://doi.org/10.14135/j.cnki.1006-3080.20201201005

多种群混合进化算法求解带工序跳跃的分布式异构批量流混合流水车间调度问题

Multiple-Population Hybrid Evolutionary Algorithm for Solving Distributed Heterogeneous Lot-Streaming Hybrid Flowshop Scheduling Problem with Missing Operations

华东理工大学学报（自然科学版）. 2025, 51(2): 228-241 http://doi.org/10.14135/j.cnki.1006-3080.20240411001

扫码关注公众号，获取更多信息！

文章编号：1006-3080(2026)05-0735-13 DOI：10.14135/j.cnki.1006-3080.20260222001

基于混合动作安全强化学习的汽油调合调度优化

石林林， 朱志伟， 赵云蒙
（华东理工大学能源化工过程智能制造教育部重点实验室，上海 200237）

摘要：针对汽油调合调度中传统优化方法难以兼顾实时性、安全性与模型适配性的问题，提出了一种基于混合动作安全强化学习的调合调度优化方法，旨在实现随机扰动下的高效安全决策。该方法采用多通道前向传播(Multi-pass)结构消除混合动作评估中的虚假梯度，并结合比例-积分-微分(PID)调节的拉格朗日乘子机制处理质量与库存约束。仿真结果表明，所提方法在随机扰动下实现了产品质量与库存约束零违背，单步决策时间约5 ms。尽管平均收益较确定性最优解低2.5%，但该方法在复杂扰动环境下仍表现出优越的鲁棒性与实时应用潜力。

关键词：汽油调合调度；安全强化学习；混合动作空间；拉格朗日乘子；优化问题

中图分类号：TP391.9 **文献标志码：**A

调度是炼厂生产计划管理与装置生产运行之间的枢纽，将长周期计划方案转化为短周期执行指令，其核心目标是在满足质量与库存约束的前提下，通过合理组织生产以最大化企业利润[1-2]。汽油是炼厂经济价值最高的产品之一[3]，汽油调合调度旨在满足汽油调合产品质量指标的同时，通过精确控制调合组分油的配比与调度逻辑(即涉及储罐指派、流路切换及作业时序安排的离散决策)，尽可能降低高价值组分油的过量添加，降低质量过剩。从数学角度看，汽油调合调度是一个典型的混合整数非线性规划(Mixed-Integer Nonlinear Programming, MINLP)问题[4]，存在离散决策-连续控制-非线性约束的耦合：(1)牌号选择/切换与任务指派引入混合整数型离散变量；(2)在离散决策给定后，还需优化各组分油的流量/配比等连续决策变量以完成配方控制；(3)研究法辛烷值(RON)/马达法辛烷值(MON)等关键性质与配比之间存在非线性交互，需借助非线性调合模型刻画[5]。

在建模层面，为减少离散变量规模，调度模型多采用连续时间表示方法[1]，通过将规划周期划分为若干灵活的时间段来组织生产活动。然而，受混合整数变量与前述非线性属性约束的双重影响，此类调度问题在理论上呈现非确定性多项式时间困难(NP-hard)特征[6]，目前尚不存在能在多项式时间内求解的通用算法。尽管工程实践中常采用混合整数线性规划(Mixed-Integer Linear Programming, MILP)进行近似处理以降低计算复杂度，但这种简化往往难以兼顾求解精度与非线性特征[7]。当生产规模扩大或调度约束增多时，传统的数学规划方法往往面临“维数灾难”，难以在合理的计算时间内获得全局最优解，导致其在需要高频次响应的在线决策场景中表现受限[8-9]。为降低计算负担，现有研究多采用降维或分解策略：例如通过分层思想分离非线性部分[2]，构建适用的连续与离散通用模型[8]，或利用库存夹点分析减少决策变量的调整频率[10]。然而，这些策略本质上多遵循确定性优化范式，而实际生产工况中，上游装置(如催化裂化、连续重整等)的负荷调整或原料切换，会导致直调组分油的关键属性(辛烷值、密度等)发生非平稳的随机漂移。这种直调油扰动极易导致离线优化的静态模型与实时工况失配，进而引发调度方案失效或频繁重调度的计算负担。需要指出的是，汽

收稿日期：2026-02-22
基金项目：国家科技重大专项(2025ZD1608400)；国家自然科学基金重大项目(62394343)；中央高校基本科研业务费专项(222202617006)；高等学校学科创新引智计划(B17017)
作者简介：石林林(2001—)，男，湖北钟祥人，硕士生，主要研究方向为石油化工过程建模与优化、强化学习。E-mail：linmakeit@163.com
通信联系人：赵云蒙，E-mail：yunmeng.zhao@ecust.edu.cn
引用本文：石林林，朱志伟，赵云蒙. 基于混合动作安全强化学习的汽油调合调度优化[J]. 华东理工大学学报(自然科学版)，2026，52(5)：735-747.
Citation：SHI Linlin, ZHU Zhiwei, ZHAO Yunmeng. Gasoline Blending Scheduling Optimization Based on Hybrid-Action Safe Reinforcement Learning[J]. Journal of East China University of Science and Technology, 2026, 52(5): 735-747.

油调合调度在实际炼厂中并非一次性离线优化后静态执行,而是具有在线滚动决策特征的动态过程。随着直调组分油性质、流量的变化,需要依据最新生产信息及时修改后续生产安排,以避免质量越界、库存冲突等风险。因此,该场景下的调度方法不仅需要获得较优离线方案,还应具备快速在线重调度能力。然而,当调合模型存在明显非线性且系统状态频繁波动时,传统数学规划方法往往需反复重建并求解优化模型,在线响应速度受限。

针对实际生产中不可忽略的环境扰动,现有研究多通过引入鲁棒优化和随机规划等方法来提升调度方案的适应能力。代表性工作包括:结合启发式算法的鲁棒优化[11]、基于数据分布刻画的机会约束规划[12]、数据驱动的原油/汽油鲁棒调度[13],以及处理多重随机参数的两阶段鲁棒模型[14]。这些方法旨在提升方案对不确定性的稳健性,但鲁棒优化往往导致过度保守的调度方案,随机规划则依赖于精确的概率分布先验且计算量随场景数指数增长,难以满足炼厂敏捷调整的需求。总体而言,上述方法普遍存在计算成本较高、对生产环境模型精度依赖性强等问题。当调合模型存在明显非线性或发生模型失配时,这类方法的优化性能可能显著下降,从而限制了其在大规模、动态生产环境中的应用效果[15]。

近年来,深度强化学习(Deep Reinforcement Learning, DRL)作为一种数据驱动的智能优化方法,在流程工业的控制与决策领域展现出巨大的应用潜力[16-18]。与传统方法相比,DRL不依赖于精确的机理模型,能够通过与环境的交互自主学习最优策略,且在线决策速度快,极具实时优化的前景[5, 19]。然而,将DRL直接应用于汽油调合调度面临挑战:其一,汽油调合天然具有混合动作特性——既包含成品油牌号切换的离散决策,又包含各组分流量配比的连续控制,传统的单一动作空间强化学习(Reinforcement Learning, RL)算法难以实现二者的协同优化;其二,生产过程的安全性要求,即必须严格满足产品质量、库存限制等多种硬性约束,而常规RL算法通常以最大化长期收益为目标,难以保证探索过程中的约束满足。

针对上述挑战,本文从建模与算法两个维度提出一种基于混合动作安全强化学习的汽油调合调度优化方法。首先,针对混合动作空间问题,本文将调度过程建模为参数化动作马尔可夫决策过程,并提出一种最大熵多通道混合软演员-评论家算法(Multi-Pass Hybrid Soft Actor-Critic, MPH-SAC)。该算法通过多通道前向传播(Multi-pass)结构消除虚假梯度,实现了离散调度策略与连续配方参数的协同优化。其次,针对安全约束问题,本文将调度任务进一步形式化为约束马尔可夫决策过程(Constrained Markov Decision Process, CMDP),并引入基于比例-积分-微分(PID)调节的拉格朗日乘子法。通过构建奖励与代价的双重评估体系,动态调整约束惩罚强度,有效引导智能体在安全可行域内搜索高收益调度策略。本文旨在克服传统数学规划方法在计算效率与抗扰动能力及在线调度响应速度方面的不足,为不确定环境下的炼厂智能调度提供一种高效、鲁棒的在线决策新范式。

1 汽油调合调度优化模型

1.1 汽油调合调度优化难点分析

(1)复杂的离散逻辑与混合整数特性:调合过程涉及大量的离散决策,如调合头的启动与停止、成品油牌号的切换、以及储罐的指派与选择。在数学建模中,这些逻辑决策必须由混合整数变量来刻画。由于调度周期长、操作频繁,离散变量的数量随时间步呈爆炸式增长,极大地增加了求解的计算复杂度。(2)属性关联的强非线性:汽油调合的核心目标是使成品油质量达标。然而,各组分油在混合时,其质量指标并非组分体积的线性加权。为保证精度,通常需引入如Ethyl-RT70等复杂关联模型,这使得模型中的质量约束呈现显著的非线性。(3)环境扰动导致的模型失配:在实际生产中,环境并非静态。由于上游生产装置的波动,进入调合装置的直调组分油属性(如辛烷值、密度等)往往存在随机扰动,且各组分油的流率也受设备工况影响而具有不确定性。这种随机扰动会导致离线计算的最优方案在执行时发生“模型失配”,轻则造成产品质量过剩(浪费昂贵的组分油),重则导致产品不合格,迫使系统必须具备在线重调度的能力。

1.2 问题描述

如图1所示,典型调合系统由直调油(Direct-feed Component Oils)、罐式组分油(Tank Component Oils)、组分油罐(CT)、成品油罐(PT)与调合头(B)组成。汽油多周期调合调度可抽象为带多类约束的优化问题:约束集合通常覆盖产品质量指标、调合头加工能力、组分油/成品油库存上下限、订单交付与提货规则,以及批次切换等工艺逻辑。在满足可行性的前提下,调合头需要按照既定配方与加工顺序完成指定牌号的生产,从而兼顾生产运行稳定性与市场供给需求。基于上述过程,本文采用约束优化模型刻画多周期调合调度,并以生产利润最大化作

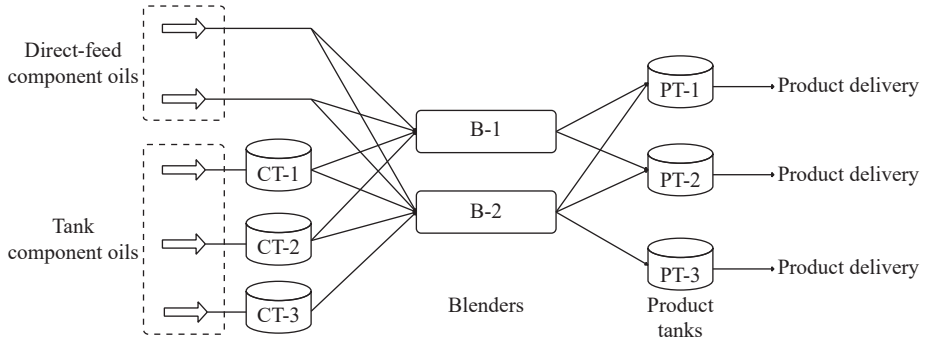


图 1　典型汽油调合系统

Fig. 1　Typical gasoline blending systems

为目标函数:

$$\text{obj}=\sum_{k}\left\{\sum_{g}\left\{P(g)\times\sum_{i}V_{\text{cout}}(i,g,k)\right\}-\sum_{i}\left\{C(i)\times\sum_{g}V_{\text{cout}}(i,g,k)\right\}\right\} \tag{1}$$

其中, i、g 和 k 分别表示组分油种类、成品油牌号和调度周期; $C(i)$ 和 $P(g)$ 分别为组分油和成品油的单价; $V_{\text{cout}}(i,g,k)$ 表示第 k 个调合周期内组分油 i 调合成品油 g 的体积量。

在理想物料平衡假设下, 组分油库存随周期的演化可由库存递推关系给出, 并要求期末库存落在允许区间内:

$$V_{\text{end}c}(i,k)=V_{\text{begin}c}(i,k)+\eta_{\text{cin}}(i)-\sum_{g}V_{\text{cout}}(i,g,k) \tag{2}$$

$$V_{c\min}(i)\leqslant V_{\text{close}c}(i,k)\leqslant V_{c\max}(i) \tag{3}$$

其中, $V_{\text{end}c}(i,k)$ 和 $V_{\text{begin}c}(i,k)$ 分别为组分油 i 对应储罐在第 k 个调合周期的期末库存和初始库存; $\eta_{\text{cin}}(i)$ 为组分油产率; $V_{\text{cout}}(i,g,k)$ 为第 k 个调合周期内组分油 i 调合成品油 g 的体积量; $V_{c\max}(i)$ 和 $V_{c\min}(i)$ 分别表示组分油 i 库存的上限和下限。

成品油库存同样满足递推关系, 并受罐容上下限的限制:

$$V_{\text{end}p}(g,k)=V_{\text{begin}p}(g,k)+V_{b}(g,k)-D_{p}(g,k) \tag{4}$$

$$V_{p\min}(g)\leqslant V_{\text{end}p}(g,k)\leqslant V_{p\max}(g) \tag{5}$$

其中, $V_{\text{end}p}(g,k)$ 和 $V_{\text{begin}p}(g,k)$ 分别为成品油 g 对应储罐在第 k 个调合周期的期末库存和初始库存; $V_b(g,k)$ 第 k 个调合周期内调合成品油 g 的体积; $D_p(g,k)$ 为第 k 个调合周期内成品油 g 的提货量; $V_{p\max}(g)$ 和 $V_{p\min}(g)$ 表示成品油 g 库存的上限和下限。

对任一周期内正在生产的牌号, 组分配方应满足归一性要求:

$$V_{\text{cout}}(i,g,k)=r(i,g,k)\times\sum_{bl}V_{b}(g,bl,k) \tag{6}$$

$$\sum_{i}r(i,g,k)=1 \tag{7}$$

其中, $V_b(g,bl,k)$ 表示第 k 个调合周期内调合头 bl 调合成品油 g 的体积, $r(i,g,k)$ 为第 k 个调合周期内组分油 i 调合成品油 g 的配方。满足对同一成品牌号的配方求和为 1。

每个周期内调合头的加工量需落在允许范围内, 并由二值变量刻画其是否运行:

$$F_{b\min}\times\delta(g,bl,k)\leqslant V_{b}(g,bl,k)\leqslant F_{b\max}\times\delta(g,bl,k) \tag{8}$$

$$n_{p\min}\leqslant\sum_{g}\delta(g,bl,k)\leqslant n_{p\max} \tag{9}$$

其中, $\delta(g,bl,k)$ 为第 k 个调合周期内调合头 bl 是否调合成品油 g 的二值变量; $F_{b\max}$ 和 $F_{b\min}$ 分别表示调合头加工能力上限和下限; $n_{p\max}$ 和 $n_{p\min}$ 分别表示调合头可调合成品油的种类范围。

对于和体积分数呈线性关系的油品属性, 调合头处存在以下约束:

$$Q_{pb}(g,s,k)\times V_{b}(g,k)=\sum_{i=1}^{n}Q_{c}(i,s,k)\times V_{\text{cout}}(i,g,k) \tag{10}$$

在一般情况下, 调合头产物并不需要立即进入分销环节, 因此可以选择不对调合头出口属性设置硬约束, 而仅要求最终进入成品油罐后的罐内属性满足指标。这样的优势是: 当罐内属性偏低时, 模型会倾向于提高当期调合头处属性以“拉回达标”; 反之亦然, 从而在调度时域内实现更灵活的“卡边生产”。对应的成品油罐属性约束为:

$$Q_{p}(g,s,k)\times\left(\sum_{g}\sum_{bl}V_{b}(g,bl,k)+V_{p\text{heel}}(g,k)\right)=Q_{pb}(g,s,k)\times\sum_{g}\sum_{bl}V_{b}(g,bl,k)+Q_{p\text{heel}}(g,s,k)\times V_{p\text{heel}}(g,k) \tag{11}$$

$$Q_{p\min}(g,s)\leqslant Q_{p}(g,s,k)\leqslant Q_{p\max}(g,s) \tag{12}$$

其中, $Q_p(g,s,k)$ 和 $Q_{pb}(g,s,k)$ 分别为第 k 个调合周

期成品油 g 和调合头出口成品油 g 的属性 s；$Q_c(i,s,k)$ 表示第 k 个调合周期组分油 i 的属性 s；$V_{p\text{heel}}(g,k)$ 和 $Q_{p\text{heel}}(g,s,k)$ 分别表示第 k 个调合周期成品油 g 的罐底油体积和对应的罐底油属性 s；$Q_{p\max}(g,s)$ 和 $Q_{p\min}(g,s)$ 分别为成品油 g 属性 s 的上限和下限。在以上的成品油属性约束中，只考虑让成品油罐内的属性满足指标要求，而不对调合头处的汽油属性做约束。当成品油储罐内油品属性低于指标要求时，约束会促使调合头处油品属性高于指标以保证最终成品油属性达标，反之亦然，以此来实现卡边生产。

为简化提货操作，本文采用阈值触发的提货规则：当罐内库存未超过阈值时不提货；一旦超过阈值，则在当天进行提货，提货量按规则计算：

$$D_p(g,k)=V_{p\text{heel}}(g,k)-V_{p\min}(g) \tag{13}$$

其中，$V_{p\min}(g)$ 表示成品油的最小库存。为将该类规则融入调度模型，可用“最小产量约束”近似替代调度时域内的最小提货要求：

$$\sum_{k}\sum_{bl}V_b(g,bl,k)\geqslant Y_{\min}(g) \tag{14}$$

其中，$Y_{\min}(g)$ 表示各个牌号汽油最小产量。传统离线调度通常建立在“参调组分油性质恒定、装置无扰动”的理想前提下；但在实际生产中，环境变化会使初始方案偏离最优，甚至引发质量/产量风险，从而需要频繁的人工干预与调整，带来额外成本。考虑到管道调合广泛应用、直调油为节省罐容常从上游装置直接进入调合头，其属性会受装置波动影响而随机变化。本文聚焦直调油属性不确定性，并用一阶无截距自回归过程描述其随时间的演化：

$$Q_c(d,s,k+1)=Q_c(d,s,k)+\epsilon(d,s,k) \tag{15}$$

其中，$Q_c(d,s,k)$ 表示在第 k 个调合周期直调油 d 的属性 s，ϵ 表示正态分布的噪声。该模型在缺乏精细机理模型或大规模样本数据时具有合理性：它既能体现属性演变的时间相关性与连续性，又能通过高斯噪声覆盖足够扰动情形，便于在训练阶段提升策略的泛化能力与鲁棒性。

2 基于约束 MPH-SAC 算法的汽油调合调度优化

2.1 混合动作特性与建模需求

从决策变量的组织形式来看，汽油调合调度属于典型的离散-连续混合决策问题。对任一调度周期而言，策略需先完成离散层选择：确定当期投入生产的成品油牌号；随后在该选择约束下，再给出与之匹配的连续层控制量，即各罐式组分油在该周期内的配比。

在现有强化学习研究中，针对同时包含离散动作及其连续参数的问题，Masson 等[20] 提出了参数化动作（Parameterized Actions, PA）建模框架，并将该类问题形式化为参数化动作马尔可夫决策过程（Parameterized Action Markov Decision Process, PAMDP）。设状态空间为连续空间 $S\subseteq\mathbb{R}^n$，离散动作集合 $\mathcal{A}_d$ 对应调度层的离散决策（如选择生产成品油牌号，记为 $k\in\mathcal{A}_d$），且每一离散动作 k 对应一组连续参数空间 $x_k\in\mathcal{X}_k\subseteq\mathbb{R}^{m_k}$，表示在该牌号选择约束下可行域中的连续参数向量，例如各罐式组分油的配方 x_k，其中，m_k 为连续参数向量的维数，整体动作可表示为：

$$\mathcal{A}=\bigcup_{k\in[K]}\{a_k=(k,x_k)|x_k\in\mathcal{X}_k\} \tag{16}$$

整体的策略表示为：

$$\pi(k,x|s)=\pi^{\mathrm{d}}(k|s)\pi^{\mathrm{c}}(x|s) \tag{17}$$

然而，将问题置于 PAMDP 形式，直接套用常见的混合动作强化学习范式仍存在明显困难。其一，若把离散选择强行“连续化”，或把连续控制量粗粒度离散化，往往会模糊离散操作与连续调节之间的差异，削弱策略对两类决策的刻画能力，进而影响表示效果与收敛表现[21]；其二，在采用共享函数逼近结构的混合动作方法中，即使某一离散动作未被执行，与其绑定的连续参数分支仍可能在训练中产生梯度传播，带来与当前决策无关的更新信号，扰乱价值估计并使策略性能受损。对于工艺约束明确、操作逻辑严格的工业调度问题，上述负效应通常会被进一步放大。

2.2 Multi-pass 建模思想

在以 Q 函数为核心的混合动作强化学习框架中[22]，动作价值函数需要对“离散选择 + 连续参数”构成的联合动作进行统一评估，其对应的贝尔曼递推形式可写为：

$$Q(s_t,k_t,x_{k_t})=\mathbb{E}_{r_t,s_{t+1}}\left[r_{\mathrm{t}}+\gamma\max_{k\in[K]}\sup_{x_k\in\mathcal{X}_k}Q(s_{t+1},k,x_k)\middle|s_t=\mathrm{s},a_t=(k_t,x_{k_t})\right] \tag{18}$$

在此框架中，一般会引入一个关键假设：Q 函数只依赖于当前被选取的离散动作所绑定的连续参数，而与其余离散动作对应的参数无关。该“理想独立性”可形式化表示为：

$$\frac{\partial Q_k}{\partial x_j}=0,\forall k,j\in[K],j\neq k \tag{19}$$

但在实际实现中, 动作价值函数通常由神经网络近似。由于不同离散动作的连续参数往往在网络内部共享特征表示, 上述独立性假设很难严格成立, 进而使策略更新时产生与当前所选离散动作无关的连续参数梯度, 并在训练中累积为对价值估计不利的“虚假梯度”。其推导可写为:

$$\nabla_{\theta_x} L_x(\theta_x) = -\sum_{k=1}^{K} \nabla_x Q(s,k,x(s;\theta_x);\theta_Q) \nabla_{\theta_x} x(s;\theta_x) \tag{20}$$

$$\nabla_x Q = \left(\frac{\partial Q_1}{\partial x_1} + \frac{\partial Q_2}{\partial x_1} + \cdots + \frac{\partial Q_K}{\partial x_1}, \cdots, \frac{\partial Q_1}{\partial x_K} + \cdots + \frac{\partial Q_K}{\partial x_K} \right) \tag{21}$$

结合汽油调合调度的工艺过程, 上述“虚假梯度”具有直观的物理含义: 在任一调度周期内, 系统真正执行的只可能是被选中成品油牌号对应的组分油用量/配方; 而对未被选中的牌号而言, 其连续参数在该周期内并不会落地执行, 因而不应参与当前联合动作的价值评估。如果价值网络在学习时无法有效区分“当前生效的参数”与“未生效的参数”, 就可能带来两方面后果: 一是离散层的牌号选择出现偏置, 二是连续层的配方搜索被无关梯度干扰, 从而整体性能下降。为此, Multi-pass 结构在每次对动作价值网络进行前向计算时, 仅保留当前离散动作所绑定的连续参数, 并将其余离散动作对应的连续参数统一置为零, 以此切断无关梯度传播路径, 从机制上抑制虚假梯度[23], 其数学形式如下:

$$\begin{pmatrix} Q(s,a_1,xe_1;\theta_Q) & \cdots & Q(s,a_k,xe_1;\theta_Q) \\ \vdots & \ddots & \vdots \\ Q(s,a_1,xe_k;\theta_Q) & \cdots & Q(s,a_k,xe_k;\theta_Q) \end{pmatrix} = \begin{pmatrix} Q_{11} & \cdots & Q_{1K} \\ \vdots & \ddots & \vdots \\ Q_{K1} & \cdots & Q_{KK} \end{pmatrix}$$

$$Q_{ij} = \begin{cases} Q(s,a_i,xe_i;\theta_Q), i = j \\ 0, i \neq j \end{cases} \tag{22}$$

其中, Q_{ij} 表示第 i 个离散动作与第 j 组连续参数组合所对应的价值项, Multi-pass 结构仅保留 $i = j$ 的有效组合, 而对 $i \neq j$ 的无效组合统一置为零, 从而将价值函数更新限制在当前真实可执行的联合动作上, 避免无关参数参与学习。在汽油调合调度问题中, Multi-pass 结构可理解为在算法层面对价值评估施加了一种“操作逻辑一致性”约束: 价值计算严格遵循“先确定成品油牌号, 再分配组分油配方/用量”的工艺流程。基于这一稳定的价值评估机制, 后续才能更可靠地在混合动作空间中引入最大熵策略优化与安全约束处理。

2.3　基于最大化熵的 MPH-SAC 算法

在借助 Multi-pass 机制抑制混合动作价值评估中的虚假梯度之后, 汽油调合调度仍面临更深层的学习难题: 连续控制量维度较高、成品属性约束具有非线性耦合, 同时直调油扰动会引起环境动态变化, 进而导致训练过程易振荡、探索效率不足。为应对上述问题, 本文以 MP-DQN[23] 为基础, 融合最大熵强化学习的随机策略思想, 并配合更稳健的价值估计方式, 提出约束 MPH-SAC 算法。具体而言, 在离散动作价值评估部分, MPH-SAC 引入对偶 (Dueling) 结构[24], 以增强价值函数表示能力并减轻估计偏差带来的训练不稳定。

针对传统双网络(Twin)结构在计算目标值时常采用“取最小”操作而可能引发的价值低估问题, 本文引入 Zhou 等[25] 提出的双均值(Double Average)思路: 以目标 Twin 网络两路输出的均值作为更新目标, 从而在抑制过估计的同时, 减轻由低估带来的过度保守探索。相应地, 离散动作价值网络的更新目标可定义为:

$$y = r + \gamma \mathop{\mathbb{E}}_{s\sim\mathcal{D},x'\sim\pi^{\mathrm{c}}} \mathrm{avg}_{i=1,2} \left[\pi^{\mathrm{d}}(k'|s',x') Q\left(s',k',x';\theta^{-}_{Q,i}\right) - \alpha \log \pi^{\mathrm{c}}(x'|s') \right] \tag{23}$$

在策略表达上, MPH-SAC 以随机策略取代 MP-DQN 中针对连续参数的确定性输出[26]。具体而言, 连续参数由高斯分布建模并通过采样获得: 一方面, 这种参数化方式不再依赖反转梯度等启发式手段来处理参数越界问题[27]; 另一方面, 它能够与最大熵强化学习的目标形式自然兼容, 从而在存在扰动与不确定性的场景中提供更充分、更稳定的探索。价值网络与策略网络的交互关系如图 2 所示: 离散动作分支基于动作价值函数构造玻尔兹曼分布, 并与贪心选择相结合形成混合探索机制, 其对应的策略分布可写为:

$$\pi^{\mathrm{d}}(k_a|s,x) = (1-p) \frac{\mathrm{e}^{Q(s,k_a,x)/T}}{\sum_{k\in\mathcal{A}_{\mathrm{d}}} \mathrm{e}^{Q(s,k,x)/T}} + p \frac{1}{|\mathcal{A}_{\mathrm{d}}|} \tag{24}$$

在最大熵强化学习框架中, 策略优化不仅最大化期望回报, 还会在目标中加入显式的熵正则项, 以在探索与利用之间取得权衡[26]。进一步利用条件概率的分解关系, 可将连续参数策略网络的优化目标整理为如下损失函数形式:

$$L_{\mathrm{x}}(\theta_x) = \mathop{-\mathbb{E}}_{s\sim\mathcal{D},x\sim\pi_{\theta_x}} \left[\sum_k \pi^{\mathrm{d}}(k|s,x) Q\left(s,k,x(s;\theta_x);\theta^{-}_{Q,1}\right) \right] - \alpha \mathop{\mathbb{E}}_{s\sim\mathcal{D}} [\mathrm{H}(\pi^{\mathrm{c}}(\cdot|s))] \tag{25}$$

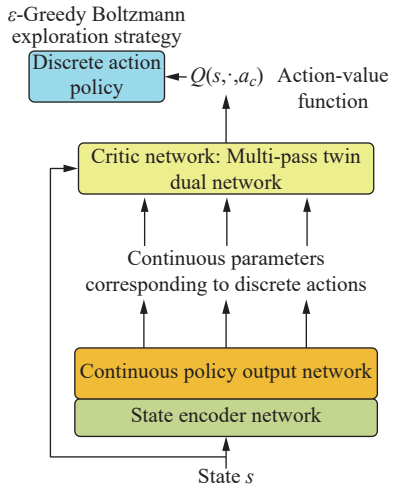


图 2 MPH-SAC 框架

Fig. 2 Framework of MPH-SAC

其中，目标中的期望项采用蒙特卡洛采样进行近似估计。需要说明的是，从严格定义出发，混合动作策略的熵应同时覆盖离散选择与连续参数两部分；但在本文实现中，仅对连续参数分支显式加入熵正则项，而离散动作的探索性则由式 (24) 所给出的随机化选择机制来提供与维持。图 3 展示了 MPH-SAC 算法的网络更新与信息流示意。

综上，MPH-SAC 在保留 Multi-pass 机制优势的基础上，进一步结合最大熵随机策略与更稳健的价值估计策略，使得混合动作空间下的探索效率与训练稳定性得到同步提升。该特性对汽油调合调度尤为重要：面对直调油属性扰动与多重非线性质量约束，策略既要满足安全与可行性要求，又需避免因探索不足而过早陷入局部最优。

为直观验证 MPH-SAC 的有效性，我们在混合动作空间基准环境 Platform Domain 下对 MPH-SAC 与 MP-DQN 算法进行了对比测试。实验结果表明，在相同的网络结构与超参数设置下，MPH-SAC 的平均成功率达到 99.67%，显著优于 MP-DQN 的 93.60%。从学习曲线(图 4)来看，MPH-SAC 表现出更快的收敛速度与更小的训练波动，具体而言，其收敛耗时 29 min，而 MP-DQN 的收敛耗时为 42 min；其奖励标准差仅为 0.015 6，较 MP-DQN(0.017 0)降低了约 8.20%，这有力地证明了双网络-对偶(Twin Dueling)结构对价值过估计的抑制作用。此外，得益于最大熵机制对策略空间的充分探索，MPH-SAC 的平均奖励达到 0.999 1，高于 MP-DQN 的 0.995 6，避免了基础算法易陷入次优局部极值的问题。

2.4 软约束安全强化学习与约束 MPH-SAC 算法

汽油调合调度在追求经济效益的同时，还必须满足成品油质量、组分油库存、调合头能力以及调度期总产量等多类工艺约束。上述约束既包括需要在每个调度周期即时满足的单步约束，也包括只能在调度期结束后通过累积结果判定的跨周期累积约束。在直调油属性存在扰动与不确定性时，若将这些约束全部以硬约束方式直接施加，策略在训练早期往往会因频繁触发不可行而难以获得有效学习信号，甚至导致训练失败[28]。基于此，本文将汽油调合调度建模为约束马尔可夫决策过程。

在 CMDP 框架下，本文采用拉格朗日乘子法对应的原始-对偶优化思路[29]，将多约束策略优化转化为等价的无约束形式：针对每一条工艺约束引入相应乘子，并将各成本项以加权方式并入策略目标，使策略迭代过程中能够在“收益提升”与“约束满足”之间进行动态权衡。鉴于混合动作空间与随机策略设定下的动作价值难以解析求解，基于采样轨迹，本文在乘子更新环节采用蒙特卡洛估计，对各约束对应的成本效益量进行近似计算。进一步考虑多约束并存时乘子更新容易出现振荡或响应滞后的问题，本文引入 PID 拉格朗日乘子更新机制。具体而言，比

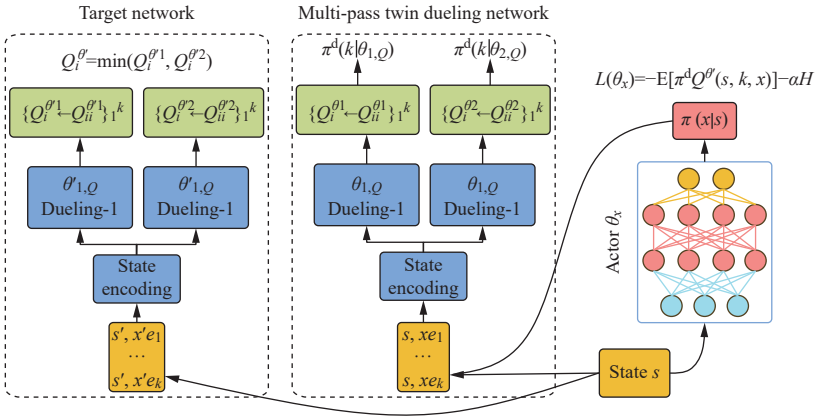


图 3 MPH-SAC 网络更新示意图

Fig. 3 Network update diagram of MPH-SAC

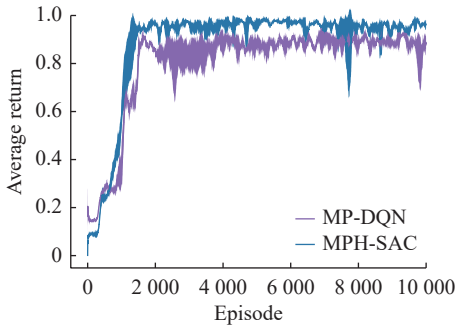


图 4　Platform 测试环境学习曲线

Fig. 4　Learning curves on Platform

例项用于对当前违约进行快速反馈, 积分项用于累积长期偏差以降低稳态误差, 微分项用于抑制更新过程中的剧烈波动, 从而减少策略在约束边界附近反复“贴边-反弹”的振荡。借助该机制, 训练初期可对高风险调度方案施加更强约束压力, 引导策略尽快远离严重违约区域; 当约束逐步被满足后, 又能自动减弱惩罚强度, 避免策略过度保守[30]。综合 Multi-pass 价值评估机制、最大熵随机策略以及基于 PID 的乘子更新方法, 本文构建了 MPH-SAC 算法, 其完整流程见表 1。

2.5　约束 MPH-SAC 在汽油调合调度中的应用

在约束 MPH-SAC 框架下, 汽油调合调度被刻画为随调度周期滚动推进的序列决策过程。每一周期对应一次“智能体-环境”交互。二者共同定义当期的实际调合操作。当周期执行结束后, 环境返回当期收益以及各约束对应的成本信息。成本函数用于量化当前调度方案在成品油属性、库存、产量等方面的违约风险, 而拉格朗日乘子则表征长期约束压力, 并通过在线更新持续反馈到后续的策略优化中。随着上述交互-更新过程在各调度周期内反复迭代, 策略逐步学会在收益提升与风险控制之间形成稳定权衡。由此, 汽油调合调度被完整转化为一个可由约束 MPH-SAC 求解的 CMDP 问题。值得注意的是, 与传统数学规划方法在每一个决策步均需实时求解高维非线性优化问题不同, 本文方法采用“离线学习, 在线决策”的范式。在部署阶段, 策略输出仅依赖于神经网络的一次前向推理, 单步决策的时间复杂度降至 $O(1)$ 常数级, 这为应对复杂工况下的高频在线

表 1　带约束的 MPH-SAC 算法优化流程

Table 1　Optimization process of MPH-SAC with constraints

Algorithm: Optimization procedure of the MPH-SAC safe reinforcement learning algorithm based on the PID Lagrange multiplier method

Initialize learning rates: lr_{θ_Q}, lr_{θ_x}, lr_{λ}, lr_{θ_α}; Network parameters: θ_Q, θ_x; Mini-batch size: B; Soft update rate of target networks: τ; Total training steps: K_1; Number of critic updates per batch: K_2; Number of value network updates: $n_Q = 0$; Number of policy network updates: $n_x = 0$; Initial discrete action temperature coefficient: T; Exponential moving average of discrete action entropy: $\overline{H}$; Minimum discrete action entropy: H_0; Scaling factor: k; Counter: $i = 0$; Constraint Threshold: K_3

for $t = 1$ to K_1 do
  Sample discrete actions and continuous parameters from the parameter network output distribution. $x_k, k \leftarrow x_k(s;\theta_x)$;
  Execute action: $a_t = (k_t, x_{k_t})$, Receive Rewards: r_t, Cost: $c_{1,t}, c_{2,t}, \ldots, c_{N_c,t}$, Observe the next state s_{t+1};
  $\{s_t, a_t, r_t, c_{1,t}, c_{2,t}, \ldots, c_{N_c,t}, s_{t+1}\}$ stored in the experience pool $\mathcal{D}$
  if $t \bmod K_1 = 0$
    Randomly sample state transition data from the experience pool $\{s_b, a_b, r_b, c_{1,b}, c_{2,b}, \ldots, c_{N_c,b}, s_{b+1}\}_{b\in[B]}$;
    $n_x = 0$;
    for $n_Q = 1$ to K_2 do
      Calculate the loss function for the Critic network: L_{θ_Q}, update θ_Q, θ_{Q_j};
      Calculate the estimated value of the cost-benefit function, update the multiplier using the PID multiplier method;
      Compute the moving average of the critic-network loss, denoted by $\overline{L}_{n_Q}$;
      if $n_x < n_Q / \left(2 - \exp\left(-\overline{L}_{n_Q}\right)\right)$
        Compute the Q-weighted values and use them to calculate the discrete action distribution of the Boltzmann distribution;
        Calculate the entropy coefficient loss function, update θ_α;
        Compute the parameter network loss function, update θ_x;
        if $\overline{H} < H_0$, then $i = i + 1$;
        if $i \geqslant K_3$, then $T = T * k$;
        $n_x \leftarrow n_x + 1$
    end for
  end for

重调度提供了理论上的实时性保障。

3 仿真实验结果与分析

3.1 仿真环境与参数设置

考虑调度周期为7 d的单调合头调度优化问题，系统生产两类成品油，成品油1与成品油2，分别记为oil1与oil2。参调原料包含5种组分油，为便于表述，统一记为组分油A~E，其与实际油品的对应关系为：A为OCTMD，B为SZORBD，C为REFOR，D为NOARO，E为SZORBG。其中，组分油A、B来自上游装置的直调物流，受工况波动影响，其进料流量与辛烷值等关键属性呈随机扰动特征，从而增加oil1、oil2的RON与MON达标不确定性。模型中将A、B的流量视为不可控输入；C~E作为属性稳定的罐式组分油，其用量(或配比)由决策变量给出，用于在扰动存在时对成品属性进行补偿与平衡。本文重点考虑与体积分数呈非线性关系的质量指标RON与MON，并采用Ethyl-RT70调合模型计算产品性质，具体属性边界见表2。进一步地，假设直调组分油A、B的流量与辛烷值按式(15)引入随机扰动：噪声服从正态分布，均值取0.05、方差取0.2；为避免波动幅度过大，对辛烷值扰动项施加小于0.5的缩放系数，对流量扰动项施加小于10的缩放系数。产量目标要求oil 1不低于35000 m^3，oil 2不低于28000 m^3。约束集合包括：调度期总产量约束、两类成品油RON/MON非线性质量约束、组分油库存约束、调合头能力约束，以及每个调合周期仅允许生产一种成品的工艺逻辑约束；合并后共计12条约束。提货阈值分别设为40000 m^3与15000 m^3；罐式组分油C、D、E的日进料量分别设为1350、850、450 m^3；oil 1与oil 2的价格分别为8500元/m^3与8700元/m^3。

表2 成品汽油属性

Table 2 Properties of product oils

Type	RON			MON			Reid vapor pressure/kPa			φ(Olefins)/%			φ(Aromatics)/%			Inventory/m^3		
	Initial	Min	Max	Initial	Min	Max	Initial	Min	Max	Initial	Min	Max	Initial	Min	Max	Initial	Min	Max
Oil 1	92.2	92.2	93.0	82.3	82.0	100	72.0	0	72.0	26.0	0	26.0	35.1	0	38.0	4000	2100	59500
Oil 2	94.3	94.2	95.0	82.1	82.0	100	72.0	0	72.0	16.6	0	26.0	36.8	0	38.0	4000	900	17000

为保证算法对比的公正性与计算效率评价的准确性，本文所有实验均在同一计算平台上运行。硬件配置如下：处理器(CPU)为AMD Ryzen 9 7945HX(主频2.50 GHz，16核心/32线程)，运行内存(RAM)为16.0 GB。软件环境如下：操作系统为Windows 11(64位)，编程语言采用Python 3.9，深度学习框架为PyTorch 2.0。此外，数学规划基线方法采用商用求解器Gurobi 12.0.0，并在求解配置中开启了16线程并行计算及针对非凸混合整数二次约束规划(Mixed-Integer Quadratically Constrained Programming，MIQCP)问题的全局寻优模式(设置NonConvex=2)。在算法设置上，MPH-SAC的各类网络均采用256×256两层全连接结构。单步决策由“离散选择+连续控制”两部分组成：离散动作维度为2，分别对应当日生产oil1或oil2；在离散动作确定后，再输出与之绑定的3维连续参数，用于表征可控罐式组分油在该日的用量分配。状态输入为包含时间信息在内的18维观测向量。为保证对比的可比性，本文引入MINLP数学规划作为基线方法，并采用Gurobi求解器，最优间隙设为2.5%。考虑到模型包含多项非线性属性约束，Gurobi在缺乏初始可行解时通常收敛较慢。为此，本文将训练收敛后的MPH-SAC策略输出作为求解器初始解，并在此基础上进行迭代求优，以提升规划基线的求解效率与解质量。奖励构造以单步生产收益为主体，同时叠加多类惩罚与软约束项以促进可行解的形成与收敛：对于组分油库存约束与调合能力约束，将违背量汇总后施加统一惩罚；对成品油的非线性属性约束，采用“边界激励+超限惩罚”的分段形式；在调度期末，若总产量未满足要求则追加惩罚项；此外，对关键约束进一步通过拉格朗日乘子构造相应成本函数，以软约束方式融入策略学习过程。MPH-SAC的主要训练超参数列于表3。

3.2 实验结果与分析

图5给出了算法训练阶段平均探索回报的收敛曲线：智能体每与环境交互224步，统计并记录该区间内回报的均值。由图可见，基于MPH-SAC的混合动作强化学习方法在所构建的CMDP环境中能够实现稳定收敛，表明其训练过程具有较好的稳定性与可收敛性。

图6给出了在PID乘子更新机制下，成品油非线性属性约束(以RON上下限为代表)对应拉格朗日乘子的演化轨迹。可以观察到，针对oil1与oil2的

表 3　MPH-SAC 训练超参数表

Table 3　Training hyperparameter list of MPH-SAC

Total training step	B	Learning rate of Q-Actor	Learning rate of Param-Actor	Discount factor	τ	Target entropy of continuous policy
3×10^5	560	2×10^{-5}	2×10^{-5}	0.98	5×10^{-6}	−6.0

Minimum entropy of discrete policy	T	Multiplier update batch size	Cost discount factor	PID coefficient		
				k_P	k_I	k_D
0.173	0.5	896	0.98	1×10^{-6}	1×10^{-7}	1×10^{-7}

B—Training batch size; τ—Q-actor coefficient; T—Initial temperature coefficient

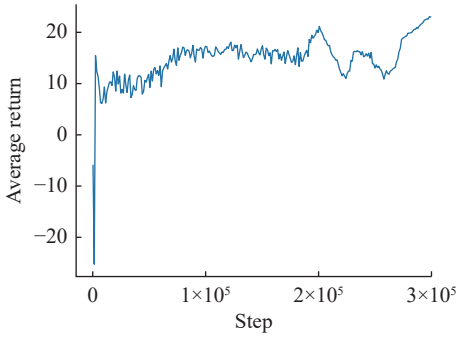


图 5　MPH-SAC 学习曲线

Fig. 5　Learning curve of MPH-SAC

RON 上限/下限约束, 乘子整体能够在训练过程中逐渐达到稳定区间; 随着约束违背程度被持续压缩, 乘子的调整幅度明显减小并呈现收敛趋势, 且在 PID 调节作用下, 个别乘子会在阶段性超调后出现一定回落。需要说明的是, 累积折扣成本通常基于批量采样轨迹进行估计, 同一批次内不同回合可能分别触发“上限超限”或“下限不足”等不同类型的违约情形, 因此在统计意义上可能出现上下限对应乘子同步增大的现象。另外, 部分约束在奖励设计中已显式引入惩罚项进行引导, 使得其对应乘子在训练早期即被推至较稳定水平, 从而表现为更快的收敛特征。

使用训练完成的智能体开展调度测试, 图 7 和图 8 给出了单次实验中 MPH-SAC 与 Gurobi 得到的调度方案。可以看出, 两种方法对 oil1 与 oil2 的生产天数分配一致; 同时, 在可控罐式组分油的配置结构上也表现出相近倾向, 即以组分油 C 与 D 作为主要调合来源, 并尽量压低组分油 E 的用量。其原因在于组分油 E 具备“供给规模较小且单位成本较高”的特征, 在库存约束与成品油性质约束共同作用下, 方案更倾向于通过用量更充裕的组分油 C 进行补偿性替代, 从而降低对高成本组分油 E 的依赖并维持属性可行。

在对比调度方案的分布特征后, 本文进一步评估了两种算法在实际执行阶段的计算效率。由于传统的 MINLP 方法在每一决策步均需要通过迭代求解非线性规划问题, 其计算耗时受系统规模与非线性强度的影响巨大。实测显示, Gurobi 求解器在处理包含非线性属性约束的初始调度问题时, 求解至 2.89% 最优间隙所需时间约为 394 s。而 MPH-SAC 采用离线训练、在线推理的模式, 决策过程仅涉及神经网络的一次前向计算, 单步平均响应时间稳定在 5 ms 左右, 相比于传统方法实现了 5 个数量级的效率提升。

为进一步验证所提方法在真实扰动环境下的在线动态调度能力, 本文构建了一个 7 d 的滚动调度场景。在初始时刻, 系统基于组分油库存、标称属性和调度任务生成初始状态; 随后, 在每个决策时刻 t, 先获取直调组分油的实时信息并更新系统状态, 再对剩余时域 $t,\cdots,T$ 的计划进行在线重调度(在每一决策时刻开始时, 根据最新观测到的系统状态, 对剩余时域的调度方案重新生成), 并执行当前调度时域的调度方案, 随后滚动进入下一调度时域, 直至整个调度周期结束。由此形成实时观测-状态更新-重调度-执行当前动作-滚动推进的在线闭环流程。滚动调度测试沿用第 3.1 节所述的直调组分油 A、B 随机扰动设置。在每次独立测试中, 首先生成一条长度为 7 的扰动轨迹, 并将其逐日注入仿真环境; 两种算法在相同的初始条件和相同的扰动轨迹下分别进行测试, 以保证比较公平。

对于 Gurobi 方法, 重调度表现为重新建立并求解剩余时域的 MINLP 模型, 再提取当前调度时域方案执行; 对于 MPH-SAC, 重调度则表现为将更新后的状态输入已完成训练的策略网络, 直接输出当前时域的离散动作和连续参数, 无需重新训练或进行全局优化求解。

图 9 给出了两种方法在 7 d 滚动调度实验中的在线决策耗时。可以看出, 随着剩余时域逐渐缩短, Gurobi 的重算时间虽有所下降, 但在整个滚动过程中始终处于数十秒级至数百秒级范围: 第 1 天约为

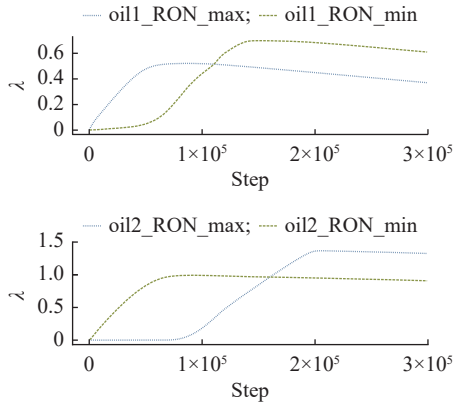


图 6 MPH-SAC 乘子更新曲线

Fig. 6 Multipliers update curve of MPH-SAC

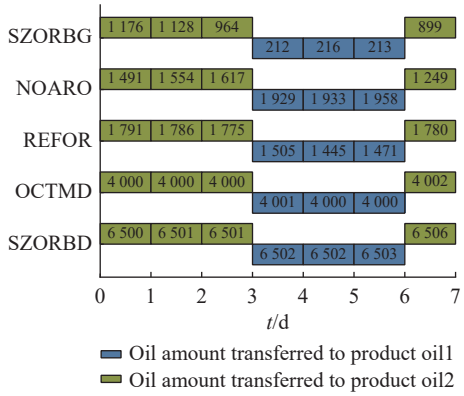


图 7 基于 MPH-SAC 的各组分油调合调度计划(单位: m^3)

Fig. 7 Scheduling plan of components based on MPH-SAC (unit: m^3)

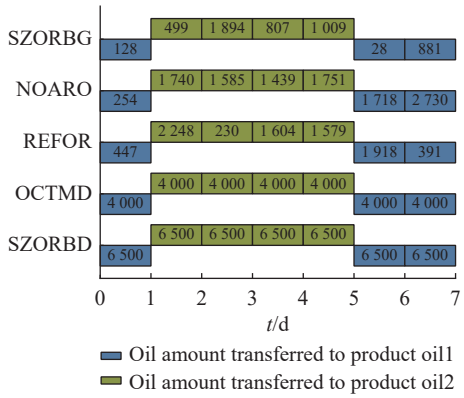


图 8 基于 Gurobi 的各组分油调合调度计划(单位: m^3)

Fig. 8 Scheduling plan of components based on Gurobi (unit: m^3)

394.0 s, 至第 7 天仍需约 26.0 s, 单步平均耗时为 187.9 s。相比之下, MPH-SAC 在全部决策时域均稳定在约 5 ms, 几乎不随调合时域变化而波动。二者平均在线决策时间相差约 10^4 倍。这表明, 在扰动环境下, 当系统需要根据最新状态频繁修正后续调度计划时, 传统数学规划方法由于反复求解高维非线性优化模型, 难以满足实时在线应用需求; 而 MPH-SAC 仅通过一次神经网络的前向推理即可快速生成当前动作(该调度时域的调度方案), 因而更适用于实际的在线动态调度场景。

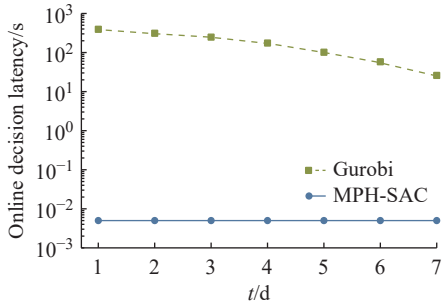


图 9 不同调度算法在线决策耗时对比

Fig. 9 Comparison of online decision latency for different scheduling algorithms

为评估两种方法在扰动条件下的约束满足能力与经济表现, 本文在随机扰动环境中进行 200 次独立测试, 并统计调度结果的约束满足率与收益指标。表 4 汇总了两种方法在测试集上的约束满足概率: 若在整个调度时域内, 任一调度周期的成品油 RON 均未超出其边界值 ±0.05 的容差范围, 则判定该次实验满足 RON 约束; MON 约束采用同样判定规则。对于组分油库存约束, 若所有组分油库存全程均落在允许范围内, 则认为该次实验库存约束得到满足。

表 4 不同算法下的约束满足率统计

Table 4 Statistics of constraint satisfaction rate under different algorithms

Algorithm	RON constraint/%	MON constraint/%	Component oil inventory constraint/%
MPH-SAC	100	100	100
Gurobi	49.5	100	100

由表 4 可见, 在直调油属性与流量存在随机扰动时, 即便扰动幅度较小, 基于“组分油属性恒定”假设求解的数学规划方案仍可能出现较高概率的约束失效。这一结果对确定性规划方法而言具有合理性: 求解器在精确模型下倾向于追求最优目标值, 而汽油调合调度的高收益方案往往伴随对质量约束的“贴边”配置; 当不确定性介入后, 原本依赖贴边可行性的解更容易因扰动而越界。理论上可以通过对规划模型中的属性约束进行收缩/保守化来提升鲁棒性(降低失效概率), 但这通常会带来两方面代价: 一是策略更保守导致收益下降, 二是可行域收缩与非线性约束叠加可能提高求解难度。相比之下, MPH-SAC 学到的是基于状态观测的反馈策略, 能够随扰动引起的系统状态变化在线调整决策, 因此即使缺乏对扰动的先验信息, 也能以较高概率维持约束可

行性。同时, 由于回报设计以经济收益为主导, 策略在保证可行的前提下仍会主动偏向更高回报的调度行为。

图 10 给出了 200 次实验的收益统计, 其中 MPH-SAC 取 200 次回合的均值与 Gurobi 方案进行对比。结果表明, MPH-SAC 产生的产品总价值与总成本均较高(对应更积极的生产与配比选择), 其平均收益约为 7.94×10^7 元; Gurobi 方案收益约为 8.13×10^7 元。总体而言, MPH-SAC 的收益水平与 Gurobi 相当, 仅低约 2.5% 左右, 同时在扰动环境下表现出更稳定的约束满足能力。

在代表性单次实验中, 两种方法在质量约束满足方式上呈现出明显差异。由图 11 可见, MPH-SAC 在调度执行阶段表现出更强的安全裕度意识: oil1 与 oil2 的 RON/MON 轨迹整体与约束边界保持一定距离。以 oil2 的 RON 为例, 其控制水平显著高于 94.2 的下限要求, 通常维持在 94.5 以上, 从而为扰动留出缓冲空间。相对地, 由 Gurobi 求解 MINLP 模型所得出的确定性方案更倾向于贴近约束线运行, 即典型的“贴边可行”策略; 在随机扰动作用下, oil2 的 RON 多次出现轻微跌破下限的现象, 导致可行性被破坏。

将结论扩展到 200 次独立实验的统计结果后, 上述差异更加直观。图 12 中实线表示 RON 的样本均值, 阴影表示标准差范围。对于 Gurobi, 其均值几乎与 94.2 的约束线重合, 而波动带中相当一部分落

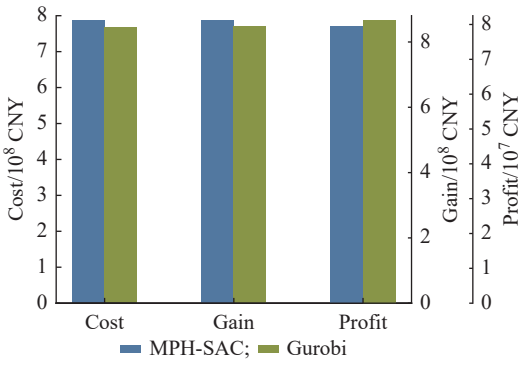


图 10　不同算法调度收益对比

Fig. 10　Revenue comparison of scheduling under different algorithms

在约束线以下, 这与表 4 中 RON 约束满足率仅为 49.5% 的结果一致, 说明在不确定环境下, 依赖“贴边”获得高收益的确定性方案对扰动高度敏感, 任意小幅偏移都可能触发违约。相比之下, MPH-SAC 由于采用软约束安全机制, 在训练过程中会对“靠近边界导致的违约风险”形成显式反馈: 属性贴近边界时更易产生软约束超标, 从而抬升成本项并推动拉格朗日乘子增大, 进而削弱对应动作的价值评估并抑制该类策略被反复选择。因此, 智能体最终收敛到更保守的配方调节策略, 使 oil2 的 RON 均值稳定在约 94.6 附近; 即便考虑波动下界(约 94.4), 仍明显高于 94.2 的指标要求。

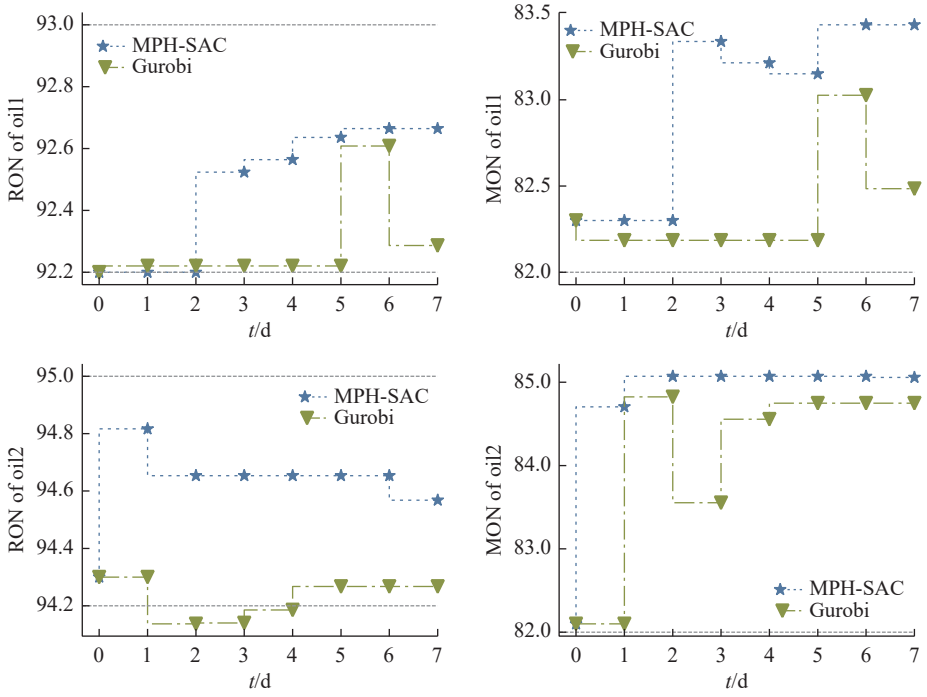


图 11　调度时域内成品油属性变化情况

Fig. 11　Changes in product properties within the scheduling horizon

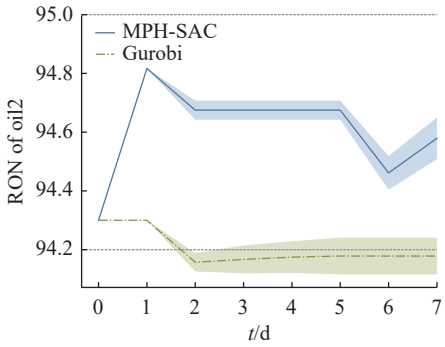


图 12 成品油 2 属性值平均变化情况

Fig. 12 Average changes in attribute values of product oil 2

4 结 论

本文提出了一种基于混合动作空间的安全强化学习调度方法，该方法通过引入 Multi-pass 结构消除混合动作评估中的虚假梯度，并结合 PID 调节的拉格朗日乘子机制，协同优化离散牌号切换与连续配方变量。仿真结果表明，所提方法在扰动环境下具备良好的在线动态重调度能力：滚动调度实验中，Gurobi 平均在线决策耗时为 187.9 s，而 MPH-SAC 的仅为 5 ms。进一步地，所提算法在复杂不确定环境下实现了质量与库存约束的完全满足，并与 Gurobi 保持近似的收益水平。上述结果表明，本文方法能够在不确定工况下兼顾实时性、安全性与经济性，为炼油企业汽油调合调度提供了一种有效的智能决策新范式。

参考文献：

[1] CASTILLO-CASTILLO P A, MAHALEC V. Inventory pinch gasoline blend scheduling algorithm combining discrete-and continuous-time models[J]. Computers & Chemical Engineering, 2016, 84: 611-626.

[2] GLISMANN K, GRUHN G. Short-term scheduling and recipe optimization of blending processes[J]. Computers & Chemical Engineering, 2001, 25(4/6): 627-634.

[3] LI J, KARIMI I A. Scheduling gasoline blending operations from recipe determination to shipping using unit slots[J]. Industrial & Engineering Chemistry Research, 2011, 50(15): 9156-9174.

[4] CERDÁ J, PAUTASSO P C, CAFARO D C. Optimizing gasoline recipes and blending operations using nonlinear blend models[J]. Industrial & Engineering Chemistry Research, 2016, 55(28): 7782-7800.

[5] ZHU Z, YANG M, HE W, *et al*. A deep reinforcement learning approach to gasoline blending real-time optimization under uncertainty[J]. Chinese Journal of Chemical Engineering, 2024, 71: 183-192.

[6] TERRAZAS-MORENO S, GROSSMANN I E. A multiscale decomposition method for the optimal planning and scheduling of multi-site continuous multiproduct plants[J]. Chemical Engineering Science, 2011, 66(19): 4307-4318.

[7] CERDÁ J, PAUTASSO P C, CAFARO D C. Optimization approaches for efficient crude blending in large oil refineries[J]. Industrial & Engineering Chemistry Research, 2018, 57(25): 8484-8501.

[8] JOLY M, MORO L F L, PINTO J M. Planning and scheduling for petroleum refineries using mathematical programming[J]. Brazilian Journal of Chemical Engineering, 2002, 19: 207-228.

[9] FRANZOI R E, MENEZES B C, KELLY J D, *et al*. Large-scale optimization of nonconvex MINLP refinery scheduling[J]. Computers & Chemical Engineering, 2024, 186: 108678.

[10] CASTILLO P A C, MAHALEC V, KELLY J D. Inventory pinch algorithm for gasoline blend planning[J]. AIChE Journal, 2013, 59(10): 3748-3766.

[11] ZHAO X, WANG Y. Gasoline blending scheduling based on uncertainty[C]//2009 International Conference on Computational Intelligence and Natural Computing. Wuhan, China: IEEE, 2009: 84-87.

[12] YANG Y. Optimal blending under general uncertainties: A chance-constrained programming approach[J]. Computers & Chemical Engineering, 2023, 171: 108170.

[13] DAI X, ZHAO L, LI Z, *et al*. A data-driven approach for crude oil scheduling optimization under product yield uncertainty[J]. Chemical Engineering Science, 2021, 246: 116971.

[14] 罗春鹏, 荣冈. 不确定条件下汽油调和调度的鲁棒优化模型 [J]. 石油学报 (石油加工), 2009, 25(3): 391-400.

[15] HANEVELD K, VAN DER VLERK M H. Stochastic Programming [M]. New York: Springer New York, 2020.

[16] SKORDILIS E, MOGHADDASS R. A deep reinforcement learning approach for real-time sensor-driven decision making and predictive analytics[J]. Computers & Industrial Engineering, 2020, 147: 106600.

[17] 郎峻, 殷晓明, 顾幸生. 基于强化学习的正弦优化算法求解能耗分布式流水车间节能调度问题 [J]. 华东理工大学学报 (自然科学版), 2025, 51(5): 623-632.

[18] 蔡林逸, 冯翔, 虞慧群. 基于组稀疏优化的强化学习稀疏表征 [J]. 华东理工大学学报 (自然科学版), 2024, 50(6): 913-919.

[19] 许毅, 冯翔, 虞慧群. 基于强化学习和角度惩罚距离的冰晶连续优化算法 [J]. 华东理工大学学报 (自然科学版), 2021, 47(1): 89-100.

[20] MASSON W, RANCHOD P, KONIDARIS G. Reinforcement learning with parameterized actions[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence.

Arizona, USA: AAAI Press, 2016: 1934-1940.

[21] KANERVISTO A, SCHELLER C, HAUTAMÄKI V. Action space shaping in deep reinforcement learning[C]//2020 IEEE conference on games (CoG). Osaka, Japan: IEEE, 2020: 479-486.

[22] XIONG J, WANG Q, YANG Z, *et al*. Parametrized deep q-networks learning: Reinforcement learning with discrete-continuous hybrid action space[EB/OL]. (2018-10-10). https://arxiv.org/abs/1810.06394.

[23] BESTER C J, JAMES S D, KONIDARIS G D. Multi-pass q-networks for deep reinforcement learning with parameterised action spaces[EB/OL]. (2019-5-10). https://arxiv.org/abs/1905.04388.

[24] WANG Z, SCHAUL T, HESSEL M, *et al*. Dueling network architectures for deep reinforcement learning[C]// International conference on machine learning. New York, USA: PMLR, 2016: 1995-2003.

[25] ZHOU H, LIN Z, LI J, *et al*. Revisiting discrete soft actor-critic[EB/OL]. (2022-9-22). https://arxiv.org/abs/2209.10081.

[26] HAARNOJA T, TANG H, ABBEEL P, *et al*. Reinforcement learning with deep energy-based policies[C]//International Conference on Machine Learning. Sydney, Australia: PMLR, 2017: 1352-1361.

[27] HAUSKNECHT M, STONE P. Deep reinforcement learning in parameterized action space[EB/OL]. (2015-12-10). https://arxiv.org/abs/1511.04143.

[28] 王雪松, 王荣荣, 程玉虎. 安全强化学习综述 [J]. 自动化学报, 2023, 49(9): 1813-1835.

[29] STOOKE A, ACHIAM J, ABBEEL P. Responsive safety in reinforcement learning by pid lagrangian methods[C]//International Conference on Machine Learning. Virtual Event: PMLR, 2020: 9133-9143.

[30] XU Y, HU D, LIANG L, *et al*. Target entropy annealing for discrete soft actor-critic[EB/OL]. (2021-12-06). https://arxiv.org/abs/2112.02852.

Gasoline Blending Scheduling Optimization Based on Hybrid-Action Safe Reinforcement Learning

SHI Linlin, ZHU Zhiwei, ZHAO Yunmeng

(Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China)

Abstract: A safe reinforcement learning method is proposed for gasoline blending scheduling under random disturbances in the properties and flow rates of direct-feed component oils. The scheduling task requires the coordinated optimization of discrete scheduling decisions and continuous blending recipes under strict product quality and inventory constraints. To solve this problem, the scheduling process is formulated in a hybrid action space so that discrete product switching decisions and continuous recipe regulation can be handled in a unified framework. On this basis, a Multi-pass hybrid soft actor-critic algorithm (MPH-SAC) is developed for gasoline blending scheduling. The Multi-pass structure removes spurious gradients in hybrid action evaluation and improves the coordination between discrete scheduling decisions and continuous blending recipes. In addition, a proportional-integral-derivative (PID) regulated Lagrangian multiplier mechanism is introduced to dynamically enforce safety constraints while maintaining economic performance during learning. The proposed method is tested in a gasoline blending scheduling environment with uncertain direct-feed component oil conditions. Simulation results show that the proposed method achieves zero violations of product quality and inventory constraints under random disturbances. The average online decision time is about 5 ms, whereas the deterministic optimization model solved by Gurobi requires about 394 s to obtain a near-optimal solution. Although the average profit of the proposed method is about 2.5% lower than the deterministic optimum, it maintains strong robustness against disturbances in the properties and flow rates of direct-feed component oils and avoids the model mismatch caused by deterministic assumptions. These results demonstrate that the proposed method can effectively balance profitability, operational safety, and real-time decision requirements, and it provides a practical solution for real-time gasoline blending scheduling in complex and uncertain operating environments.

Key words: gasoline blending scheduling; safe reinforcement learning; hybrid action space; Lagrange multiplier; optimization problem

（责任编辑：李娟）