DOI:10.3724/j.issn.1004-3918.2026.02.021

# 自适应强化学习驱动的电商动态定价与消费者行为博弈模型

李红新

（西安欧亚学院 职业教育学院，西安 710065）

**摘 要：** 针对电商动态定价中需求与竞争环境高度非平稳、消费者异质偏好难以显式建模以及收益、稳定性与合规约束并存的问题，提出一种自适应强化学习博弈框架ARL-GT。该框架以斯塔克尔伯格双层结构刻画“平台定价—消费者响应”的最优反应关系，并采用最大熵逆强化学习从历史交互日志中学习消费者效用函数，生成可解释的期望需求以稳定奖励信号并缓解策略训练的稀疏反馈问题。在策略层将MAML与PPO算法结合，学习可迁移初始化，通过少量梯度更新实现跨时段、跨品类和不同竞争状态下的快速恢复与自适应调整。同时，在目标函数中引入CVaR尾部风险度量、调价摩擦与缺货罚金，以抑制过度调价与激进探索，并将价格区间、频率与幅度等业务规则作为可审计约束进行统一评估。仿真测试包含促销冲击、供给扰动与竞品策略切换以模拟真实市场。基于RetailRocket日志与高保真多智能体仿真评估的结果表明，ARL-GT相较强基线方法，累计收益提升约9.3%，市场份额达到41.3%且价格波动标准差降至0.48。在学习型竞品突入场景下，中位恢复时间约21个回合。风险敏感配置进一步将95%置信水平下CVaR提升至1.51并降低了收益方差。消融结果显示，逆强化学习的需求校准与元学习初始化对收益提升与恢复加速贡献最关键。总体上，该方法兼顾收益、稳定性与可审计性，为动态定价在线部署提供了可解释的博弈决策方案。

**关键词：** 电商动态定价；强化学习；元学习；偏好学习；博弈论

**中图分类号：** TP18；O225 **文献标志码：** A **文章编号：**1004-3918(2026)02-0439-11

# Adaptive Reinforcement Learning-Driven Dynamic Pricing and Consumer Behavior Game Model in E-Commerce

LI Hongxin

（School of Vocational Education, Xi'an Eurasia University, Xi'an 710065, China）

**Abstract:** To address the challenges in e-commerce dynamic pricing arising from highly non-stationary demand and competitive environments, the difficulty of explicitly modeling heterogeneous consumer preferences, and the coexistence of profitability, stability, and compliance constraints, this paper proposes an adaptive reinforcement learning game-theoretic framework, ARL-GT. The problem setting considered in this study emphasizes not only revenue maximization, but also the practical need to maintain stable pricing behavior and rule-compliant operation under continuously changing market conditions. The framework employs a Stackelberg bi-level structure to characterize the optimal-response relationship between platform pricing and consumer response, so that the sequential interaction of platform decisions and consumer behavioral feedback can be represented in a structured, explicit, and interpretable way. Within this structure, the pricing side and the response side are linked through an

**收稿日期：** 2025-10-27 **修回日期：** 2026-01-31 **接受日期：** 2026-02-25

**基金项目：** 陕西省“十四五”教育科学规划2024年度课题(SGH24Y2570)

**作者简介：** 李红新(1973-)，男，副教授，硕士，主要研究方向为网络营销与电商运营，E-mail:hongxin202509@163.com

**引用格式：** 李红新. 自适应强化学习驱动的电商动态定价与消费者行为博弈模型[J]. 河南科学，2026，44(2)：439-449.

Li H X. Adaptive reinforcement learning-driven dynamic pricing and consumer behavior game model in e-commerce[J]. Henan Science，2026，44(2)：439-449.

optimal-reaction formulation, which is used to describe the decision process of "platform pricing-consumer response" in a unified game-theoretic form. The method further adopts maximum-entropy inverse reinforcement learning to learn a consumer utility function from historical interaction logs, and then uses the learned utility to generate interpretable expected demand. This demand representation is introduced to stabilize reward signals during training and to alleviate the sparse-feedback problem that commonly affects policy optimization in dynamic pricing tasks. At the policy layer, MAML is combined with PPO to learn a transferable initialization, and this initialization can be adapted with only a small number of gradient updates. As a result, the pricing policy is able to achieve rapid recovery and adaptive adjustment across different time periods, product categories, and competitive states, which directly supports cross-scenario adaptation in non-stationary environments. At the same time, the objective function incorporates a CVaR tail-risk measure, price-adjustment friction, and stockout penalties to suppress excessive repricing and aggressive exploration, thereby reducing unstable policy behavior while retaining profitability-oriented optimization. In addition, business rules including price range, adjustment frequency, and adjustment magnitude are incorporated as auditable constraints for unified evaluation, so that performance, stability, and compliance can be assessed within the same framework. The simulation tests include promotional shocks, supply disturbances, and competitor strategy switching to mimic real markets, and these settings are used to evaluate the robustness of the proposed method under representative market disturbances. Results based on RetailRocket logs and high-fidelity multi-agent simulation evaluation show that, compared with strong baseline methods, ARL-GT improves cumulative revenue by approximately 9.3%, achieves a market share of 41.3%, and reduces the standard deviation of price fluctuations to 0.48. In scenarios involving the entry of learning-based competitors, the median recovery time is about 21 episodes, indicating faster recovery of pricing performance after competitive regime changes. A risk-sensitive configuration further improves CVaR at a 95% confidence level to 1.51 and reduces revenue variance, showing the effectiveness of the introduced risk-control design under the same framework. Ablation results show that inverse-reinforcement-learning-based demand calibration and meta-learning initialization make the most critical contributions to revenue improvement and recovery acceleration. Overall, the proposed method jointly considers profitability, stability, and auditability, and provides an interpretable game-theoretic decision-making scheme for the online deployment of dynamic pricing.

**Key words**: dynamic pricing in e-commerce; reinforcement learning; meta-learning; preference learning; game theory

## 0 引言

动态定价是电子商务平台在竞争压力、库存约束与运营规则共同作用下，提升收益与资源配置效率的关键手段。随着流量结构与促销机制频繁变化、竞品策略快速迭代以及供给侧扰动日益增多，电商市场呈现显著的非平稳特征，平台需要在信息不完备条件下持续做出序贯定价决策[1]。强化学习（Reinforcement Learning，RL）通过"交互—试错—长期回报优化"的机制学习策略，可在不依赖显式需求函数的前提下处理高维状态与动态反馈，为动态定价提供了可行范式[2]。

现有研究主要从"需求刻画—策略优化—竞争建模"三个维度推进。其一，在消费者行为与需求刻画方面，部分工作从交互数据与心理因素出发，分析购买意愿与偏好差异，以支撑定价与营销策略联动[3]；也有研究从单智能体到多智能体视角讨论学习主体与激励机制的设计，为在复杂环境中刻画用户与对手反应提供方法论参考[4-6]。其二，在策略优化方面，已有研究将强化学习引入动态定价，以直接优化长期收益并取得一定效果[7]；其中，Q学习（Q-Learning）奠定了序贯决策的价值迭代求解基础[8]，深度Q网络（Deep Q-Network，DQN）进一步提升了高维状态下的表示与逼近能力[9]；在此基础上，近端策略优化（Proximal Policy Optimization，PPO）、深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）与软行动者—评论家（Soft Actor-Critic，SAC）等策略梯度与连续控制算法进一步改善了样本效率

与训练稳定性[10-12]。其三，在竞争建模方面，多智能体强化学习（Multi-Agent Reinforcement Learning，MARL）将对手显式作为独立学习体，有助于表达策略相互影响与竞争反馈[13]；但相关研究也指出，多个基于强化学习的定价体在同一市场中可能在无通信条件下形成“算法默示合谋”，带来价格偏离竞争均衡的合规风险[14-15]。

为提升可解释性与可审计性，斯塔克尔伯格博弈等序贯博弈结构常用于刻画平台与消费者或竞争对手的层级互动关系[16-17]。在消费者偏好学习方面，离散选择模型可用于刻画异质偏好并支撑效用函数估计[18]；逆强化学习（Inverse Reinforcement Learning，IRL）可从行为轨迹反推潜在奖励，以支持偏好学习的奖励函数估计[19-20]。在跨场景快速适应方面，模型无关元学习（Model-Agnostic Meta-Learning，MAML）可在任务分布层面学习策略先验，从而提升跨时段、跨品类迁移时的再适应效率[21]。

尽管上述研究取得一定进展，面向真实电商场景的部署仍存在结构性缺口：一是环境非平稳导致经验分布与状态转移持续漂移，致使训练稳定性与泛化能力容易退化；二是消费者反应与偏好刻画若依赖强假设或固定结构，易引入系统性偏差并传导至定价策略；三是竞争互动常被简化为外生扰动，缺乏可复核的博弈结构与统一的风险控制口径，难以同时满足收益最优化与合规审计的工程需求。尤其在存在市场冲击与分布错配的场景下，如何在保持盈利能力的同时实现快速适应与风险可控，仍是强化学习定价落地的关键挑战[22]。

为此，本文提出一种自适应强化学习博弈（Adaptive Reinforcement Learning and Game-Theoretic，ARL-GT）框架。该框架的创新之处在于其一体化的设计，贯通了“可解释需求、结构化博弈、元学习适应、风险与合规约束”四大核心环节。其主要优势在于：①结构化博弈建模：采用斯塔克尔伯格双层结构显式刻画“平台—消费者”的最优响应关系，使策略优化对竞争互动具备可解释性、可约束性和可优化性；②数据驱动的偏好与需求建模：基于最大熵逆强化学习（Max-Ent IRL）从历史交互日志中学习消费者效用函数，生成可解释的期望需求，缓解因预设需求函数带来的奖励错配问题；③快速跨场景自适应：在策略层结合MAML与PPO方法，以少量数据和少步更新实现对不同时段、品类和竞争状态的快速恢复与迁移；④面向部署的稳健性约束：在目标中引入风险敏感与运营摩擦，以降低尾部风险与价格波动，同时支持合规性监测。

## 1　方法与建模

本文提出的ARL-GT框架如图1所示，由四个模块构成：元学习器、基于PPO的Actor-Critic基础策略网络、消费者效用网络（由IRL训练）以及包含斯塔克尔伯格博弈层约束的多智能体仿真环境。在该框架中，平台作为领导者给出价格，消费者与竞争对手作为跟随者产生响应。为便于求解，下层最优响应由黑盒近似器实现。

框架的流程主要包含以下三个核心环节：首先，在内层循环中，以外层学习得到的可迁移初始化为起点，针对单个市场任务运行仿真交互并使用PPO进行少步策略更新，实现快速适应；其次，在外层循环中，汇总多个任务的内层更新结果与对应回报，更新可迁移初始化以提升跨任务泛化能力，并将该初始化回传用于下一轮内层适应；最后，消费者效用网

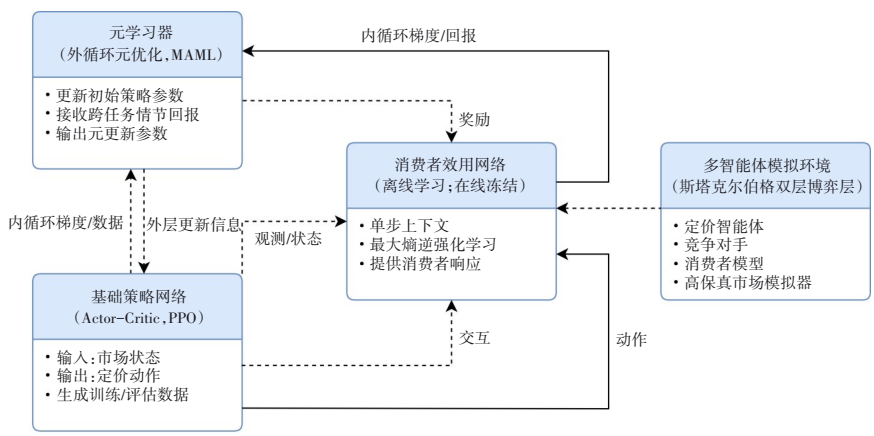


**图1　自适应强化学习博弈（ARL-GT）框架**

Fig.1　Adaptive Reinforcement Learning and Game-Theoretic（ARL-GT）framework

络在离线数据上进行拟合，为在线训练提供可解释的需求信号。为贴近真实部署场景，采用风险敏感目标函数（以CVaR度量尾部风险），并在利润中加入价格调整摩擦与缺货罚金，同时在环境中嵌入促销冲击、供给扰动和竞争对手价格漂移等非平稳因素。

### 1.1 问题假设与变量定义

假设电子商务市场环境由定价智能体$A_p$、竞争对手智能体$A_t$和消费者群体响应模型$M_u$构成的交互系统。为增强模型的严谨性，对非平稳马尔可夫决策过程（Non-stationary Markov Decision Process，Non-stationary MDP）的核心组件进行显式建模。任务$T$被定义为一种特定的市场环境，其非平稳性主要体现在以下两个方面：①对奖励函数$R$的影响。市场需求的变化与供给成本的波动会直接改变同一价格决策所能带来的预期利润，导致奖励函数$R$随任务$T$变化，记为$R_T$。②对状态转移函数$P$的影响。竞争环境的变化，尤其是竞争对手调整其定价策略时，会改变市场状态的演化路径[22]。由于对手的未来价格是下一状态的关键组成部分，其策略的改变直接影响了从当前状态转移到未来各种可能状态的概率。因此，竞争格局的变化使得状态转移函数$P$也与任务$T$相关，记为$P_T$。

为刻画电商环境的非平稳性，本文作如下关键假设：每一种相对稳定的市场状况定义为一个独立任务$T$，且每个任务对应一个近似平稳的马尔可夫决策过程（Markov Decision Process，MDP）。在任务分布$p(T)$下，本文旨在学习能够快速适应新任务的策略先验（见第1.3节）。在此假设下，将任务$T$对应的平稳MDP表示为$M_T=(S,A,P_T,R_T,\gamma)$，其中奖励函数与状态转移函数均取决于当前任务所处的具体市场条件。状态$\boldsymbol{s}_t\in\boldsymbol{S}$为包含市场相关信息的向量，如我方产品的历史价格与库存、竞争对手价格、市场宏观信息等；动作$a_t\in A$表示定价智能体的决策，对应设定价格$p_t$。在单个任务$T$内，假设市场环境相对稳定，因此状态转移函数$P_T$与奖励函数$R_T$可视为近似平稳，$\gamma$为折扣因子。本文将“期望需求”统一记为：

$$D_\varphi(\boldsymbol{s},p)=\mathbb{E}\left[q_t|\boldsymbol{s},p\right], \tag{1}$$

式中：$\boldsymbol{s}$为市场状态向量；$p$为平台定价；$q_t$为随机销量。单步利润统一记为：

$$\Pi(\boldsymbol{s},p)=(p-c(s))D_\varphi(\boldsymbol{s},p)-\Omega_{\text{inv}}(\boldsymbol{s},p), \tag{2}$$

式中：$c(s)$为单位边际成本；$\Omega_{\text{inv}}$包含缺货惩罚与价格调整成本等。

### 1.2 斯塔克尔伯格博弈

为刻画商家（领导者）与消费者（跟随者）之间的序贯互动，本文将定价问题建模为双层优化问题[20]。该斯塔克尔伯格博弈结构适应于平台定价执行力较强、消费者相对分散且单个消费者对价格的反向影响可忽略的市场情形。虽然现实中消费者可能存在跨平台比价、参考价格效应和观望等待等策略性行为，但通过将竞品价格、历史价格与时间等变量纳入市场状态（或消费者效用函数），仍可用聚合最优响应来近似刻画整体需求。设市场状态为$s$，领导者定价策略为$p=\pi_\theta(s)$，跟随者基于效用最大化的聚合最优响应（即需求）为$q^*(s,p)$。对应的基本双层优化问题为：

$$\max_\theta \sum\nolimits_{s\in S} D(s)\cdot\Pi(s,\pi_\theta(s)), \tag{3}$$

$$\text{s.t. } q^*(s,p)=\operatorname{argmax}_{q\in Q}\left[U_\varphi(s,p,q)-\lambda C(q)\right], \tag{4}$$

式中：$D(s)$是状态$s$的概率质量函数；$q$为消费者（跟随者）的单步聚合需求（即购买量决策变量）；$Q$为可行需求集合；$U_\varphi$为通过IRL估计的效用函数；$C(q)$为引入的选择正则项（$\lambda\geqslant 0$）。对任一具体任务$T$，式（3）~（4）刻画了斯塔克尔伯格双层博弈结构：下层给出跟随者的最优响应$q^*(s,p)$，上层在该响应下选择领导者最优策略。本文采用参数化策略$\pi_\theta$表示领导者定价决策，因此实际求解的是在策略函数形式受限于$\pi_\theta$时的近似斯塔克尔伯格均衡解（通常为局部最优解）。为确保以上问题可近似求解，引入三项最小假设：①给定市场状态$s$与价格$p$时，下层问题关于需求变量$q$强凹并具有唯一最优响应；②利润函数$\Pi(\cdot)$与效用函数$U_\varphi(\cdot)$在自变量上一阶可微且梯度有界；③任务分布$p(T)$具有稳定性，使得跨任务期望具有统计意义。由于直接显式求解最优响应$q^*$及其对价格$p$的敏感度通常计算代价较高，因此本文借助高保真多智能体仿真环境，构造式（5）所示的“最优响应”黑盒算子，以近似模拟下层跟随者行为。

$$B(s,p)\approx q^*(s,p)。 \tag{5}$$

进一步地，为避免显式求解下层最优响应$q^*(s,p)$，本文用仿真环境生成的黑盒算子$B(s,p)$近似刻画消费者响应，并将其并入交互环境。由此，上层仅需在任务分布$p(T)$上优化定价策略$\pi_\theta$的期望回报目标，从而将双层问题转化为式（6）~（7）的单层跨任务目标：

$$\max_\theta \mathbb{E}_{T\sim p(T)}[J(\pi_\theta;B,T)], \tag{6}$$

$$J(\pi_\theta;B,T)=\mathbb{E}\left[\sum_{t=0}^{H-1}\gamma^t\Pi\left(s_t,\pi_\theta(s_t)\right)\right], \quad (7)$$

式中：$J(\pi_\theta;B,T)$表示在任务$T$下、消费者响应由$B$生成时，上层策略$\pi_\theta$的期望折扣累计收益；$H$为单条交互轨迹的长度。将下层最优响应$q^*(s,p)$转化为由模拟环境生成的黑盒算子$B(s,p)$，是求解该双层优化问题的关键步骤，但同时也引入了近似误差。其优势在于：当$B$作为奖励由消费者响应生成时，上层梯度可用策略梯度估计，无需显式对下层函数求导。分析该误差的来源及其对上层策略梯度估计稳定性的影响，对于厘清本方法的理论边界至关重要。近似误差$B(s,p)-q^*(s,p)$主要源于两个方面：

1）系统性偏差（Systematic Bias）：源于消费者效用网络$U_\varphi$对真实消费者偏好的拟合偏差。由于$U_\varphi$是从有限且可能带有噪声的历史数据中学习得到的，模拟环境中的消费者群体响应$B(s,p)$只是对真实最优响应$q^*(s,p)$的一个有偏估计，从而导致下层响应产生偏离。

2）随机性方差（Stochastic Variance）：即使效用模型完全准确，单次模拟交互生成的销量本质上仍是一个随机变量，而非其期望值。因此，单步观测到的奖励也带有随机噪声。这些误差会直接影响上层策略梯度估计的质量。随机性方差是强化学习算法的固有问题，以PPO为代表的基于蒙特卡洛采样的策略梯度方法，通过在大量轨迹上进行平均，能够有效地平滑这种噪声，获得对梯度方向的稳定估计。

### 1.3 基于元学习的自适应策略优化

针对任务分布的多样性与环境的非平稳性，本文采用MAML与PPO算法相结合的元强化学习框架[18]。整个学习过程分为内循环和外循环两个层面。

内循环（任务适应）的目标是使策略快速适应特定的市场任务。该过程首先从任务分布$p(T)$中采样一个任务$T_i$，并将当前的元策略参数$\theta$复制为该任务的初始策略参数$\theta_i$；随后，使用该初始策略与仿真环境交互，收集经验轨迹；最后，基于收集到的数据计算PPO的损失函数$\mathcal{L}_{T_i}(\theta_i)$，并执行一步或多步梯度下降来更新任务特定的策略参数。梯度更新过程如下：

$$\theta_i'=\theta_i-\alpha\nabla_{\theta_i}\mathcal{L}_{T_i}(\theta_i)。 \quad (8)$$

外循环（元优化）的目标是更新元策略参数$\theta$，使其具备更好的跨任务适应性。该循环以采样一批任务为起点，在每个任务上执行内循环，得到一组任务的适应后策略参数$\theta_i'$；随后，通过评估这些适应策略的性能，并据此构建元损失函数——即所有任务上适应后策略损失的总和$J(\theta)$；最后，根据元损失函数对元策略参数$\theta$求梯度并进行更新。该优化目标可表示为最大化所有任务上适应后策略的性能期望：

$$\theta\leftarrow\theta-\beta\nabla_\theta J(\theta), \quad (9)$$

$$J(\theta)=\frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{T_i}(\theta_i')。 \quad (10)$$

在式（8）~（10）中，$\alpha$和$\beta$分别为内循环和外循环的学习率。内循环在单个任务$T_i$上进行少步更新得到$\theta_i'$，使策略$\pi_{\theta_i'}$在给定$q^*(s,p)$下达到该任务的近似稳定点，从而获得任务内的斯塔克尔伯格近似解。外循环学习一个“易于适应”的初始化策略参数$\theta$，使多数任务以少量样本或迭代进入各自均衡解的邻域，并通过最大化跨任务期望，对梯度起到平滑作用以提升收敛速度与训练稳定性。值得注意的是，分布外任务可能削弱该优势。

### 1.4 基于强化学习的消费者偏好

本文通过离线训练一个消费者效用网络$U_\varphi$（$\varphi$为该效用函数的参数），为RL智能体提供稳定且可泛化的奖励生成机制，并使用多层感知机（Multi-layer Perceptron，MLP）对效用函数$U_\varphi(s,p)$进行参数化：

$$P(\text{buy}|s,p;\varphi)=\sigma(U_\varphi(s,p))。 \quad (11)$$

基于上述效用函数构建购买概率模型，定义状态条件期望需求（需求强度）为：

$$D_\varphi(s,p)=\Lambda(s)P(\text{buy}|s,p;\varphi), \quad (12)$$

式中：$\Lambda(s)$近似表示状态$s$下给定价格$p$时单位决策周期内的期望成交量。IRL训练使用历史单步数据$D=\{(s,p,y)\}$（其中$y$表示在状态$s$与报价$p$下观测到的购买行为标签，购买为1，否则为0），通过最大化对数似然函数进行参数估计：

$$\mathcal{L}(\varphi)=\sum_{(s,p,y)\in D}\log P(y|s,p;\varphi)。 \quad (13)$$

其梯度满足期望差分形式。最大化式（13）等价于最小化负对数似然损失，用于估计消费者响应模型参数$\varphi$，并输出对行为$y$的概率预测。需要指出的是，将消费者效用函数建模为$U_\varphi(s,p)$的形式，是一个重要的简化假设：该模型假定消费者的购买决策主要由可观测的市场状态$s$和平台报价$p$驱动。然而，消费者的行为还受到许多未被模型直接观测的混杂变量影响，例如用户接收到的具体促销信息强度、用户的实时购物意图或来自特定社交媒体渠道

的引流效应等。若忽略这些潜在的混杂变量，可能导致效用估计产生遗漏变量偏差（Omitted-Variable Bias，OVB）。例如，模型可能将由高度促销曝光引发的购买行为错误地归因于消费者对某一价格的偏好。在当前框架下，本文通过在状态$s$中尽可能包含丰富的上下文特征（如促销和优惠标识、时间戳特征、流量来源、商品与用户侧可观测属性等）来部分缓解这一问题，从而使模型间接捕捉部分潜在的用户意图。效用网络训练收敛后，冻结$\varphi$并嵌入仿真环境，用于生成期望需求$D_\varphi$。

在线部署过程中，消费者偏好可能随季节变化、营销活动与外部冲击而发生缓慢漂移，导致离线学习的效用网络与真实响应之间产生偏差，从而导致任务内的斯塔克尔伯格最优解随时间移动。本文假设在一个相对较短的决策窗口内，消费者偏好近似稳定。在此前提下，当漂移较缓慢且仍在训练任务族的覆盖范围内时，元学习提供的可迁移策略初始化，可使定价智能体通过少量交互与少步更新，在策略层跟踪新的最优定价。然而，当漂移幅度较大或出现结构性变化时，需要对效用网络进行再校准（如基于近期日志的周期性重新训练或轻量微调），以维持下层响应刻画的有效性。

### 1.5 风险敏感目标与非平稳扰动

实际市场中存在尾部亏损风险与环境动态漂移。为提升策略在此类不确定性情况下的稳健性，本文在优化目标中采用条件风险价值（Conditional Value at Risk，CVaR）作为风险敏感目标；同时，将促销冲击、供给扰动与竞争对手价格漂移三类非平稳因素参数化，并嵌入到需求与成本的生成机制中[15-18]。记平台在时刻$t$的报价为$p_t=\pi_\theta(s_t)$，折扣因子为$\gamma$，轨迹长度为$H$。

首先，用CVaR度量累积折扣利润的下分位数期望，得到风险敏感目标：

$$J_\eta(\pi_\theta)=\mathrm{CVaR}_\eta\left(\sum_{t=0}^{H-1}\gamma^t\Pi_t\right), \tag{14}$$

式中：$\eta\in(0,1)$为置信水平，$\alpha=1-\eta$。令$Z$表示随机的累积折扣利润，即$Z=\sum_{t=0}^{H-1}\gamma^t\Pi_t$。CVaR的严格数学定义为：在累积利润$Z$小于其$\alpha$分位数的条件下，$Z$的期望值。具体而言，它是最差$\alpha$概率情况下利润的平均值：

$$\mathrm{CVaR}_\eta(Z)=\mathbb{E}\left[Z\middle|Z\le F_Z^{-1}(1-\eta)\right], \tag{15}$$

式中：$F_Z^{-1}(1-\eta)$是利润分布$F_Z$的$\alpha$分位数。该风险敏感目标旨在提升策略在最差情况下的平均性能，从而增强其对尾部风险的鲁棒性。例如，当$\eta=0.95$时，该目标最大化的是“最差5%情况下的平均利润”。其定义为：

$$\tilde{\Pi}_t=\left(p_t-c(s_t)\right)D_\varphi(s_t,p_t)-\Omega_{\mathrm{inv}}(s_t,p_t), \tag{16}$$

式中：$c(s_t)$是状态相关的单位边际成本；$D_\varphi(s_t,p_t)$是由偏好模型给出的期望需求（见第1.4节），反映价格$p_t$在状态$s_t$下的平均销量；$\Omega_{\mathrm{inv}}$汇总了调价摩擦与缺货罚金等运营成本。

为进一步将环境的非平稳性显式嵌入模型，本文在需求通道与成本通道中分别引入外生扰动，并对竞争对手价格采用自回归漂移过程建模。具体而言，促销活动或流量冲击$\xi_t\left(\xi_t\sim\mathcal{N}\left(0,\sigma_\kappa^2\right)\right)$通过影响用户到达强度$\Lambda(s)$，从而调节期望需求：

$$D_\varphi(s_t,p_t)=(1+\xi_t)\Lambda(s_t)P\left(\mathrm{buy}\middle|s_t,p_t;\varphi\right)。 \tag{17}$$

供给侧成本扰动$\kappa_t\left(\kappa_t\sim\mathcal{N}\left(0,\sigma_\kappa^2\right)\right)$直接作用于成本基准$\bar{c}(s_t)$，即：

$$c(s_t)=\bar{c}(s_t)(1+\kappa_t)。 \tag{18}$$

竞争对手价格的动态漂移采用一阶自回归过程刻画：

$$p_t^{(c)}=\phi p_{t-1}^{(c)}+\varepsilon_t。 \tag{19}$$

随后，将竞争对手价格$p_t^{(c)}$纳入市场状态$s_t$，参与$\pi_\theta$与转移核$p_t$的计算。在仿真过程中，近似器$B(s,p)$结合上述扰动生成销量$q_t$，其期望与式（17）一致。需要说明的是，本文将竞争对手价格漂移建模为外生非平稳过程（式（19）），主要用于生成可复现且强度可控的扰动场景，以便稳定评估策略在竞争非平稳性下的鲁棒性。若需逼近策略性竞争行为，亦可将竞争对手建模为具备学习能力的智能体，使其定价随市场状态自适应演化。本文方法在训练与决策中仅使用可观测竞争对手价格作为状态输入，不依赖对手的内部策略或参数更新机制。

## 2 实验结果与分析

### 2.1 实验环境与数据

本文所有实验均在软件环境为Python 3.9和PyTorch 2.0，硬件配置为Intel（R）Xeon（R）Platinum 8358P CPU与NVIDIA RTX A6000（48 GB）GPU的服务器上进行。为保证结果的可靠性，所有报告的指标均取自5个不同随机种子运行结果的平均值。在元学习阶段，外循环批大小设为16个任务，每个任

务在内循环中采样20条轨迹，总训练迭代次数为5000轮。基础学习率设为$1\times10^{-4}$，优化器采用Adam。

为全面评估ARL-GT的性能，本文分别基于公开数据集和自建模拟环境进行实验。

1）公开数据集：在离线消费者建模阶段，为了确保效用模型的准确性与可复核性，本文使用Retail-Rocket数据集。该数据集来自真实电商网站，涵盖超4.5个月的交互日志，约275万条事件（浏览、加购和交易），涉及140万个用户和23万件商品，并提供随时间变化的商品属性。这些丰富的上下文特征为基于最大熵逆强化学习的消费者效用网络$U_\varphi$训练提供了高质量的数据基础。

2）自建模拟环境：在定价策略的在线元训练和评估阶段，本文构建了一个高保真的多智能体电子商务模拟平台。该平台的核心是消费者群体模型，该模型由在RetailRocket数据集上预训练好的效用网络驱动，从而确保消费者响应行为基于真实用户行为模式。为了测试模型在不同竞争格局下的鲁棒性，模拟环境中设置了多类竞争对手智能体，包括静态策略（如成本加成定价、匹配定价）和学习型策略（如独立的PPO智能体）。此外，本文构造了多类市场冲击情景，包括竞争对手策略变化、消费者偏好漂移以及新竞争者进入市场，用于定量评估模型的自适应能力。针对偏好漂移情景，模型令消费者效用函数中的价格敏感参数随时间缓慢变化，以模拟用户偏好在部署期的渐进漂移，并据此考察策略的在线恢复能力与再适应性能。

### 2.2 评价指标与基线模型

本文采用一系列指标来评价模型性能：累计收益是衡量长期盈利能力的核心指标；市场份额按整个评估周期内的总销量计算，反映模型的市场竞争力；价格波动性定义为整个评估周期内价格序列的标准差，用于衡量定价策略的稳定性；适应时间指市场发生冲击后策略恢复至稳定状态的时间，用于评估模型的动态适应性能。

本文将ARL-GT与多个基线模型进行了比较：Static-Best是表现最优的静态策略；PPO（Single-Agent）和SAC（Single-Agent）代表当前主流的、不使用元学习的RL定价方法；PPO-FT（Fine-Tuning）是一种传统的迁移学习方法，在市场冲击发生后对预训练模型进行微调；此外，本文还引入TD3（Twin Delayed Deep Deterministic Policy Gradient）和MADDPG（Multi-Agent Deep Deterministic Policy Gradient）这两种先进的强化学习算法作为更强的基线模型。

### 2.3 实验结果与分析

各模型的结果如表1所示，可以看出ARL-GT在所有指标上均优于各基线模型。其累计收益比次优的MADDPG模型与SAC模型分别提升约9.3%与13.9%，展现了更强的长期盈利能力。同时，该模型获得了最高市场份额，并保持相对较低的价格波动性，表明其策略兼具进取性与稳定性。

**表1 不同模型的评价指标对比**

Tab.1 Comparison of main performance metrics across models

| 模型 | 累计收益（相对值） | 市场份额/% | 价格波动性（标准差） |
|---|---|---|---|
| Static-Best | 1.00±0.00 | 28.5±0.4 | 0.00±0.00 |
| PPO | 1.43±0.15 | 33.1±2.1 | 0.87±0.11 |
| SAC | 1.65±0.10 | 37.1±1.6 | 0.61±0.06 |
| PPO-FT | 1.51±0.12 | 34.5±1.9 | 0.75±0.09 |
| TD3 | 1.58±0.11 | 35.8±1.7 | 0.69±0.08 |
| MADDPG | 1.72±0.09 | 38.5±1.4 | 0.57±0.05 |
| **ARL-GT** | **1.88±0.08** | **41.3±1.2** | **0.48±0.05** |

如图2所示，不同模型在训练过程中的累计收益（相对值）随回合数演化呈现显著差异，反映出各方法在盈利能力、收敛性与训练稳定性方面的不同。从学习曲线可以看出，本文提出的ARL-GT在早期回合即实现快速增益并较快进入平台期，收敛速度最快，且最终累计收益显著高于其他对比模型，说明其在动态环境下具备更强的长期收益优化能力；MADDPG与SAC的曲线形态与最终收益较为接近，整体表现次之，但仍明显低于ARL-GT；相比之下，PPO的收益增长更为缓慢、收敛阶段滞后，最终收益最低。

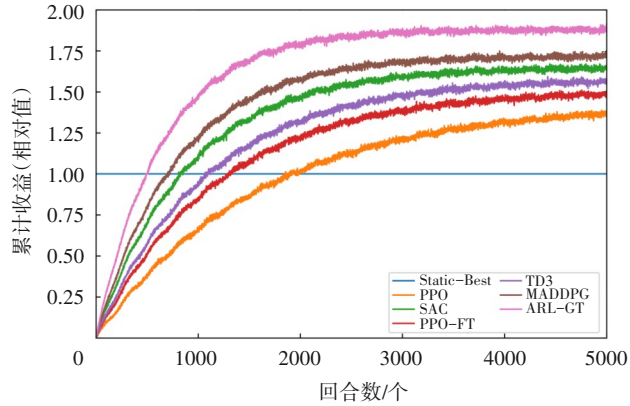


**图2 累计收益学习曲线**

Fig.2 Learning curve of cumulative return

图3展示了不同模型在市场冲击情境下的适应性表现。在第500个回合引入学习型PPO智能体竞争对手后，所有模型的收益均出现断崖式下降。然而，与基线模型相比，ARL-GT能快速恢复，并在较短时间内收敛甚至超过冲击前的水平，表明元学习机制显著增强了其定价策略的自适应能力。

为了更精细地评估模型的快速适应能力，本文对比了不同模型在“新竞争者进入”冲击场景下的适应时间。如图4所示，ARL-GT的中位恢复时间最短、分布最集中，表明其适应过程既迅速又稳定。相比之下，PPO、SAC和TD3的恢复时间中位数均超过150个回合，且离散程度较大，表明其适应过程缓慢且具有较高的不确定性；PPO-FT的表现虽优于标准RL，但仍显著逊于ARL-GT。整体结果进一步证明了元学习框架有助于提升模型快速且稳健的适应能力。

最后，为了直观地理解元学习过程如何区分不同的市场环境，本文采用基于任务统计向量的可视化，从元训练过程中采样多个任务（每个任务代表一种市场条件），对其对应的一段训练或评估轨迹进行汇总，构建包含回报水平与波动、价格水平与波动、需求水平与波动、价格对需求的近似弹性以及策略适应时间等摘要指标的任务统计向量。随后，对这些向量进行标准化，并使用主成分分析（Principal Component Analysis，PCA）法将其降至二维进行展示。如图5所示，不同颜色的点代表不同的任务类别（高需求-弱竞争、高需求-强竞争、中需求-中竞争、低需求-弱竞争、低需求-强竞争、季节性高峰）。同类市场条件在平面上呈现相对集中的分布，同时不同类别之间仍存在合理的重叠，反映了不同类别任务之间既有差异也有共性。

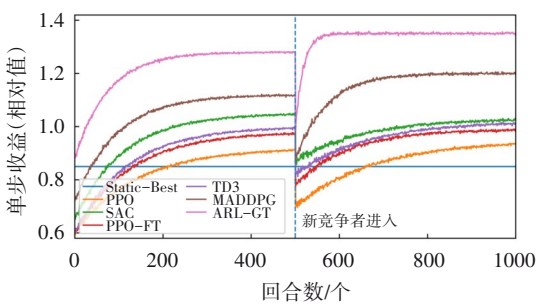


**图3　不同模型的市场冲击适应性对比**

Fig.3　Comparison of adaptability under market shocks

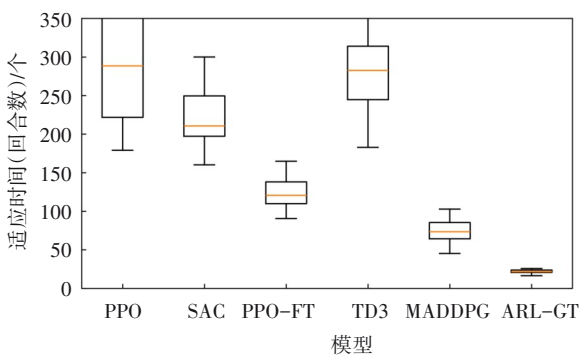


**图4　不同模型在市场冲击后的适应时间分布**

Fig.4　Distribution of adaptation time after market shocks for different models

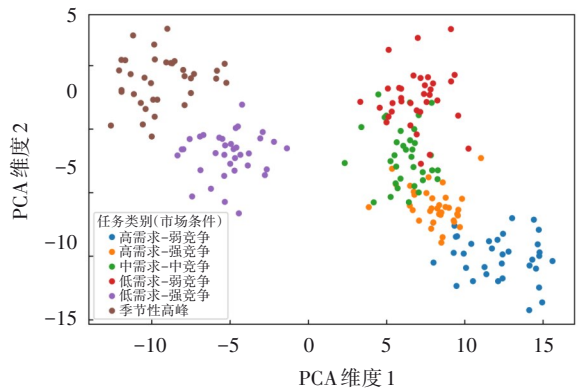


**图5　基于任务统计向量的不同市场任务在二维PCA空间中的分布**

Fig.5　Distribution of different market tasks in a 2D PCA space based on task statistic vectors

## 2.4　消融实验

为了科学验证ARL-GT框架中核心组件（MAML和IRL）的独立贡献，本文进行了消融研究。所有指标均以完整ARL-GT模型的性能为基准（设为100%），其余变体结果按比例进行归一化。

消融结果揭示了两个组件的不可或缺性。移除MAML（元学习）会导致模型的适应时间急剧增加，累计收益也随之下滑，这表明MAML的主要贡献在于提升快速跨情境适配能力。若没有元学习赋予的“学习如何学习”的能力，智能体在面对环境变化时缺乏适应能力。移除IRL（消费者建模）则会导致累计收益显著下降，说明如果优化的目标（奖励函数）本身仅是对真实世界的粗略近似，那么无论优化算法多么先进，智能体最终也只能收敛到一个次优的策略。

**表2　ARL-GT模型组件消融结果**

Tab.2　Results of ablation study on ARL-GT model components

| 模型 | 累计收益（相对值）/% | 适应时间（回合数）/个 |
|---|---|---|
| MAML+IRL | 100 | 21±3 |
| 移除MAML | 89.4±2.1 | 98±15 |
| 移除IRL | 86.2±2.5 | 23±4 |

此外，为验证第2.5节引入的风险敏感目标（CVaR）以及调价摩擦与缺货罚金机制的有效性，本文设计了两种消融配置进行对比：①Expected-only，仅最大化期望回报，不使用CVaR风险度量；②Risk-aware（本文方法），采用式（14）定义的CVaR目标，并引入调价摩擦与缺货罚金。如表3所示，Risk-aware在95%置信水平下的CVaR与两项波动性指标上均显著优于Expected-only，且其期望收益未出现统计意义上的显著下降。表明在不显著牺牲期望收益的前提下，风险项与摩擦项能够有效抑制尾部亏损和价格剧烈波动，从而验证了第2.5节提出的稳健性假设。

### 2.5 合谋与合规风险评估

为响应近年来关于算法定价可能引发隐性合谋的风险警示，本文在多智能体模拟环境中引入一组监管预警指标，包括价格相关性、同步定价比率、价格方差骤降筛查、价格离散度与超竞争加价率。实验环境设置、竞争对手族与冲击情景与主实验一致，消费者反应由离线预训练的Max-Ent IRL效用网络驱动，将每次价格曝光视为单步决策生成购买或选择概率，并沿用5个随机种子的统计评估原则。

从博弈结构看，ARL-GT采用斯塔克尔伯格序贯框架：平台在消费者显式响应$q^*(s,p)$、风险约束以及调价摩擦等限制下定价，其优化动因主要来自需求与风险之间的权衡，而非与竞争对手进行对称的共同学习。相比之下，MARL的共同学习更可能形成“相互抬价—相互确认”的正反馈，从而推高价格相关性与同步定价等合谋预警信号。而在本文框架中，高价偏离会通过需求响应与调价摩擦惩罚快速反馈为份额或收益损失，从而削弱长期同步高价的稳定性。

表4的结果进一步验证了上述分析：在相同信息与训练设定下，ARL-GT的价格相关性与同步定价比率低于部分MARL基线。在市场冲击后，其价格方差未出现系统性收敛，同时价格离散度与超竞争加价率保持在较低水平。整体上，ARL-GT未呈现易被识别的合谋模式，且与较低价格波动性结果一致。

## 3 局限性分析

本方法虽在多场景下优于基线模型，但其有效性仍依赖若干前提与边界条件。首先，消费者效用学习依赖历史日志的覆盖度与代表性，若数据存在覆盖不足、样本偏斜或关键混杂因素不可观测，则价格弹性与效用估计可能出现系统性偏差，进而污染奖励信号并影响上层策略优化方向。其次，本文采用的序贯博弈结构与下层最优响应近似虽然提升了模型的可解性，但对消费者侧的跨期策略性行为（如等待、比价、预期促销）以及更复杂的信息到达过程进行了简化，因此在强策略性消费场景下的适用性仍需进一步验证。再次，ARL-GT的跨任务自适应依赖任务划分的合理性与任务分布的相对稳定：当评估期出现显著分布漂移或偏好变化时，元先验可能失配，离线训练的效用网络可能过时，从而导致收益下滑、波动上升与冲击恢复延迟。最后，风险敏感与合规约束具有明显的情境依赖性，不同品类、监管

**表3 风险敏感目标与调价摩擦的消融对比**

Tab.3 Ablation comparison of risk-sensitive objective and price-adjustment friction

| 配置 | 期望收益 | 95%置信水平下CVaR | 收益标准差 | 价格波动标准差 |
|---|---|---|---|---|
| Expected-only | 1.90±0.07 | 1.34±0.05 | 0.52±0.06 | 0.82±0.09 |
| Risk-aware | 1.88±0.08 | 1.51±0.04 | 0.39±0.05 | 0.48±0.05 |

**表4 合谋与合规指标对比**

Tab.4 Comparison of collusion and compliance metrics

| 模型 | 价格相关性 | 同步定价比率 | 价格方差骤降 | 价格离散度 | 超竞争加价率/% |
|---|---|---|---|---|---|
| Static-Best | 0.62±0.03 | 0.41±0.02 | 0.00±0.00 | 0.00±0.00 | |
| PPO | 0.71±0.06 | 0.47±0.05 | 0.08±0.03 | 0.61±0.11 | +7.2±2.1 |
| SAC | 0.66±0.05 | 0.44±0.04 | 0.05±0.02 | 0.49±0.06 | +5.1±1.7 |
| PPO-FT | 0.69±0.05 | 0.46±0.04 | 0.06±0.03 | 0.55±0.09 | +6.1±1.8 |
| TD3 | 0.68±0.04 | 0.45±0.04 | 0.05±0.02 | 0.52±0.08 | +5.6±1.6 |
| MADDPG | 0.74±0.05 | 0.50±0.05 | 0.09±0.03 | 0.57±0.05 | +8.0±1.9 |
| **ARL-GT** | **0.59±0.04** | **0.39±0.03** | **0.01±0.01** | **0.48±0.05** | **+4.3±1.5** |

口径与运营目标下需要重新标定阈值与权重，否则可能出现过度保守（抑制收益）或约束不足（增加风险）的问题。

## 4 总结

本文面向现代电商动态定价中的非平稳性、消费者行为复杂性与策略性竞争三大挑战，设计并验证了一种新型动态定价框架ARL-GT。该方法的核心贡献包括：①构建斯塔克尔伯格序贯博弈结构，使竞争互动具备可解释、可约束与可优化的特点；②基于最大熵逆强化学习，从历史数据中学习消费者效用函数，并生成可解释需求以稳定奖励信号；③结合MAML与PPO实现少量数据和少步更新的快速跨场景适应；④通过CVaR风险度量尾部风险，引入调价摩擦和缺货罚金，将业务约束前置到优化目标；在模拟环境中嵌入关键非平稳因素，并设置合规预警机制监测潜在合谋信号。实验结果表明，ARL-GT在累计收益、市场份额、冲击恢复速度与价格稳定性等方面系统性优于强基线方法，且设置风险敏感目标后，尾部风险指标得到显著改善。

未来工作可以从三方面推进：一是提升计算效率与训练稳定性（如采用一阶MAML近似与更稳健的超参或约束自适应标定）；二是增强消费者建模与偏好漂移鲁棒性（如将专家先验与数据驱动IRL融合，并结合在线监测触发的周期性校准）；三是扩展博弈结构与评估体系（如拓展至动态斯塔克尔伯格或多领导者与更一般的多智能体博弈，并补充与MPC、BO等跨范式方法的对比，以界定适用边界）。

**参考文献：**


[1] 康琦，韦壹，王恺，等. 大型央国企数据资产管理运营创新体制机制研究[J]. 计算机技术与发展，2025，35(12)：1-9.
Kang Q，Wei Y，Wang K，et al. Research on data asset management and operation innovation mechanism of large central and state-owned enterprises[J]. Computer Technology and Development，2025，35(12)：1-9.

[2] Wang X，Wang S，Liang X X，et al. Deep reinforcement learning：A survey[J]. IEEE Transactions on Neural Networks and Learning Systems，2024，35(4)：5064-5078.

[3] 杨文贞. 人工智能主播特性对消费者购买意愿的影响研究[J]. 商业经济研究，2025(15)：87-90.
Yang W Z. Research on the influence of artificial intelligence anchor characteristics on consumers' purchase intention[J]. Journal of Commercial Economics，2025(15)：87-90.

[4] 李艺春，刘泽娇，洪艺天，等. 基于多智能体强化学习的博弈综述[J]. 自动化学报，2025，51(3)：540-558.
Li Y C，Liu Z J，Hong Y T，et al. Multi-agent reinforcement learning based game：A survey[J]. Acta Automatica Sinica，2025，51(3)：540-558.

[5] 刘晓雲，刘鸿雁，李劲松，等. 基于多元线性回归的学生成绩预测研究[J]. 计算机技术与发展，2022，32(3)：203-208.
Liu X Y，Liu H Y，Li J S，et al. Research on student achievement prediction based on multiple linear regression[J]. Computer Technology and Development，2022，32(3)：203-208.

[6] 黄昌勤，钟益华，王希哲，等. 从单智能体到多智能体：大模型智能体支持下的激励型学习活动设计与实证研究[J]. 华东师范大学学报（教育科学版），2025，43(5)：44-56.
Huang C Q，Zhong Y H，Wang X Z，et al. From single-agent to multi-agent：Motivational learning activities design and empirical study supported by LLM-based agents[J]. Journal of East China Normal University（Educational Sciences），2025，43(5)：44-56.

[7] 朱星辉，赵谦，陈欣，等. 基于强化学习的航班动态定价研究[J]. 航空计算技术，2025，55(2)：1-5.
Zhu X H，Zhao Q，Chen X，et al. Research on flight dynamic pricing based on reinforcement learning[J]. Aeronautical Computing Technique，2025，55(2)：1-5.

[8] Watkins C J C H，Dayan P. Q-learning[J]. Machine Learning，1992，8(3)：279-292.

[9] Mnih V，Kavukcuoglu K，Silver D，et al. Human-level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529-533.

[10] Schulman J，Wolski F，Dhariwal P，et al. Proximal policy optimization algorithms[PP/OL]. V2. arXiv（2017-08-28）[2025-09-10]. https://doi.org/10.48550/arXiv.1707.06347.

[11] Lillicrap T P，Hunt J J，Pritzel A，et al. Continuous control with deep reinforcement learning[PP/OL]. V6. arXiv（2019-07-05）[2025-09-10]. https://doi.org/10.48550/arXiv.1509.02971.

[12] Haarnoja T，Zhou A，Abbeel P，et al. Soft actor-critic：Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//International Conference on Machine Learning. PMLR，2018：1861-1870.

[13] Hernandez-Leal P，Kartal B，Taylor M E. A survey and

critique of multiagent deep reinforcement learning[J]. Autonomous Agents and Multi-Agent Systems, 2019, 33(6): 750-797.

[14] Calvano E, Calzolari G, Denicolò V, et al. Algorithmic collusion: Genuine or spurious?[J]. International Journal of Industrial Organization, 2023, 90: 102973.

[15] Klein T. Autonomous algorithmic collusion: Q-learning under sequential pricing[J]. The RAND Journal of Economics, 2021, 52(3): 538-558.

[16] Zhou Q, Yang Y F, Ma F F. A Stackelberg-based deep reinforcement learning approach for dynamic cooperative advertising in a two-echelon supply chain[J]. Computers & Chemical Engineering, 2025, 196: 109048.

[17] Kornowski G, Padmanabhan S, Sra S, et al. First-order methods for linearly constrained bilevel optimization[C]// Advances in Neural Information Processing Systems 37. Red Hook: Curran Associates, Inc., 2024, 37: 141417-141460.

[18] Haj-Yahia S, Mansour O, Toledo T. Incorporating domain knowledge in deep neural networks for discrete choice models[J]. Transportation Research Part C: Emerging Technologies, 2025, 171: 105014.

[19] Ziebart B D, Maas A L, Bagnell J A, et al. Maximum entropy inverse reinforcement learning[C]//20th Conference on Innovative Applications of Artificial Intelligence. Menlo Park: AAAI Press, 2008: 1433-1438.

[20] Abbeel P, Ng A Y. Apprenticeship learning via inverse reinforcement learning[C]//Proceedings of the 21st International Conference on Machine Learning. New York: ACM, 2004: 1-8.

[21] Beck J, Vuorio R, Liu Z E, et al. A tutorial on meta-reinforcement learning[J]. Foundations and Trends in Machine Learning, 2025, 18(2/3): 224-384.

[22] Lange F, Schlosser R. Dynamic pricing with waiting and price-anticipating customers[J]. Operations Research Perspectives, 2025, 14: 100337.

(编辑　关欣佳)