网约车场景下双层 Stackelberg 博弈模型与多目标优化策略研究

摘要

随着共享经济的兴起，网约车平台已成为城市交通系统不可或缺的一部分。平台的定价和激励机制直接影响司机的行为和收益，进而决定了平台的盈利能力和服务质量。这种平台与司机之间的层级决策关系，天然适合使用博弈论进行建模分析。

本文聚焦于网约车平台与司机之间的互动决策过程，构建了一个双层 Stackelberg 博弈模型。在该模型中，网约车平台作为上层领导者（Leader），其目标是实现多目标优化（如最大化自身收益、最小化乘客等待时间）；海量的网约车司机作为下层跟随者（Followers），其目标是最大化个人长期净收入。平台的决策变量是其核心的运营策略组合（如佣金率、动态调价系数等），而司机的决策则是一个复杂的序列决策过程，包括是否接受订单、在何处空驶寻单以及何时开始或结束工作

关键词：网约车；Stackelberg 博弈；双层规划；多目标优化；代理模型；多智能体强化学习;博弈论

符号定义

符号	描述
相关集合与下标	
$t \in T$	离散时间步长, T = {0, 1, ..., T}
$z \in Z$	地理区域 (六边形网格)编号集合，Z= {0, 1, ..., Z}
i, j	起点和终点区域的索引
k	单个司机的索引
平台变量	
x	平台决策变量的向量
$\alpha_{z,t}$	t 时刻 z 区域的动态定价乘数 (高峰定价)
$\beta_{z,t}$	t 时刻 z 区域的补贴(低峰补贴)
η	平台抽成, $0 \leq \eta \leq 1$
R_{plat}	平台的总收入
司机变量	
S, A	司机的状态空间和动作空间
s_t^k	t 时刻司机 k 的状态
a_t^k	t 时刻司机 k 的动作(策略)
r_t^k	司机 k 获得的即时奖励
π_θ	由 θ 参数化的司机策略 (深度神经网络)
γ	强化学习的折扣因子
c	每单位时间的运营成本 (燃料)
乘客变量	
$D_{i,t}$	t 时刻 i 区域的潜在出行需求 (请求数)
$v_{i,j}$	乘客对从 i 到 j 行程的估值
$w_{i,t}$	t 时刻 i 区域的等待时间
$\boldsymbol{\varepsilon_1}$	乘客对等待时间的敏感性系数
$Pr\ (req)$	乘客下订单的概率

第一章 引言

1.1 研究背景与意义

近年来，以滴滴出行、Uber 等为代表的网约车服务深刻地改变了人们的出行方式，并成为现代城市交通生态的重要组成部分。平台通过连接海量乘客与司机，有效盘活了社会闲置运力，提高了出行效率。然而，平台、司机、乘客三方之间存在复杂的利益关系。其中，平台与司机的关系是整个商业模式的核心。平台需要制定合理的定价和抽成机制来吸引并留住足够多的司机，以保障运力供给，同时实现自身盈利；而司机则根据平台的政策和市场需求，自主决定是否出车、工作时长以及在何处接单，以追求个人收入最大化。这种决策的非对称性和层级性，为我们运用博弈论工具进行分析提供了理想的场景。

1.2 博弈论视角下的网约车市场

博弈论是研究多个决策主体在互动中如何进行理性决策的理论。在网约车市场中，平台和司机是两个核心的理性经济人。

决策的层级性：平台首先制定并公布其游戏规则（如佣金率、奖励政策），司机在此规则下做出自己的最优决策。这完全符合 Stackelberg 博弈 的领导者-跟随者（Leader-Follower）结构。

利益的关联性：平台的收益直接来源于司机的劳动成果（车费抽成），而司机的收入则受到平台定价策略的直接影响。双方的支付函数相互关联，构成了一个完整的博弈系统。

目标的多样性：平台的成功不仅取决于单一的财务利润，还依赖于服务质量，如乘客等待时间、订单完成率等。因此，平台的目标本质上是多目标的。

因此，本文选择双层 Stackelberg 博弈模型来刻画这一互动过程。将平台作为上层领导者，追求多目标优化；将所有司机组成的群体作为下层跟随者。通过求解该博弈模型的 Pareto 最优解集，可以揭示平台在不同目标间的权衡关系，从而为平台运营提供科学的决策依据。

1.3 主要研究内容和贡献

问题建模：从博弈论的视角出发，清晰地定义了网约车场景下平台与司机双方的参与者集合、策略集合以及支付函数，构建了一个能够反映市场动态的双层 Stackelberg 博弈模型。特别地，本文将上层平台的目标扩展为多目标优化，并精细化地将下层司机的理性决策过程刻画为一个以最大化长期净收入为目标的强化学习问题，极大地提升了模型的现实性。

算法设计：针对该双层优化问题评估成本极其昂贵的挑战（下层为多智能体强化学习仿真），本文设计并实现了一种基于代理模型辅助的高效多目标演化算法（Surrogate-Assisted NSGA-II）。该算法利用高斯过程（GP）代理模型替代绝大多数昂贵的真实仿真，并结合预期超体积改善（EHVI）加点准则，智能地平衡探索与利用为进一步攻克下层仿真耗时的瓶颈，本文创新性地引入了基于迁移学习的“热启动”加速策略。该策略通过构建并复用历史训练好的强化学习模型知识库，显著缩短了新策略下司机智能体达到均衡所需的仿真时间，使得整个优化框架在实践中变得可行。

仿真分析：搭建了基于多智能体强化学习的城市交通仿真环境，其中每个司机被建模为一个独立的、理性的学习智能体（Agent）。特别地，仿真环境中的乘客需求被建模为对平台策略（价格、等待时间）动态响应的，从而更真实

地模拟了市场需求波动。通过仿真实验，验证了所提算法的有效性，并找到了平台在“利润-服务质量”等多个目标之间的权衡关系（Pareto 前沿）。

管理启示：实验结果揭示了平台不同运营策略对多个关键绩效指标的复杂影响。通过对 Pareto 前沿进行深入分研究结论不仅能为平台制定单一最优策略，更能提供一个可视化的“策略-效果”权衡工具集，帮助决策者根据不同的战略侧重（如追求利润或提升用户体验）选择最合适的运营方案，为网约车平台制定更高效、更公平的运营策略提供理论依据和决策支持

1.4 研究内容与论文结构

本文旨在构建一个符合网约车实际运营场景的双层 Stackelberg 博弈模型，并设计相应的、能够应对下层仿真评估成本高昂挑战的高效求解算法。具体结构安排如下：

第二章：梳理网约车运营、交通博弈论、Stackelberg 模型、强化学习在司机行为建模中的应用以及迁移学习在加速强化学习中的应用，明确本文的研究定位。

第三章：详细阐述问题的数学模型，从博弈三要素（参与者、策略集、支付函数）出发，构建平台（多目标）与司机的双层博弈模型，并重点引入强化学习框架对司机决策进行精细化建模。

第四章：针对所建模型评估成本极高的问题，设计一种集成了“热启动”加速机制的代理模型辅助高效多目标演化算法。

第五章：介绍仿真实验的设计方案，包括环境、司机智能体参数、对比模型和评价指标。

第六章：展示并分析仿真实验结果，验证模型和算法的性能，并提炼出管理启示。

第七章：对全文进行总结，并指出研究的局限性与未来可行的研究方向。

第二章 相关工作

2.1 网约车平台运营策略研究

现有关于网约车平台运营的研究主要集中在定价策略、订单匹配（派单）和激励机制设计上。在定价方面，动态定价（Surge Pricing）是研究热点，大量文献探讨了其对供需平衡、平台收益和社会福利的影响。在订单匹配方面，研究多集中于如何设计高效的派单算法以最小化乘客等待时间或最大化系统总成交率。

2.2 博弈论在交通领域中的应用

博弈论被广泛应用于解决交通网络中的路径选择、拥堵收费、停车管理等问题。例如，经典的 Wardrop 均衡被用来描述交通网络中用户自发选择路径达到纳什均衡的状态。这些研究为将博弈论思想引入网约车场景提供了坚实的基础。

2.3 Stackelberg 博弈模型及其应用

Stackelberg 博弈模型因其能够刻画非对称的、具有先后次序的决策过程，在经济学和管理学中得到了广泛应用 。近年来，有学者开始将其应用于共享经济领域，例如分析共享单车平台与用户、或者网约车平台与司机之间的互动。这些研究证实了 Stackelberg 模型在分析此类主从决策问题上的适用性和有效性。

2.4 基于强化学习的司机行为建模研究

为了更真实地模拟司机的决策，研究者们开始采用强化学习（RL）和多智能体强化学习(MARL)方法。在该框架下,每个司机被视为一个独立的智能体(Agent)，通过与环境的交互来学习一个能最大化其长期净收入的策略这些策略涵盖了接拒订单、空驶寻单（Repositioning）等复杂行为这类研究为本论文下层跟随者模型的构建提供了重要的理论和技术支持。深度强化学习（DRL）通过深度神经网络近似值函数或策略函数，成功解决了高维决策问题。近端策略优化（PPO）算法通过裁剪的代理目标函数来限制策略更新步长,在保证稳定性的同时简化了实现，性能卓越，成为了事实上的基准算法之一。在网约车这类 MARL 场景中，直接应用 PPO 面临环境非平稳性等挑战。为此，学界提出了多种训练范式：

1. 独立 PPO (Independent PPO, IPPO)：这是最直接的方法，将每个智能体视为独立的学习者，拥有自己的 Actor 和 Critic 网络，并将其他智能体视为环境的一部分。其优点是实现简单且可扩展性强，但由于环境的非平稳性，可能导致训练不稳定和收敛性差。尽管如此，IPPO 仍是一个强大的基线方法，有时在复杂任务中也能取得不错的效果。
2. 中心化评论家 PPO (Centralized Critic PPO / MAPPO)：该范式采用“中心化训练，去中心化执行”（CTDE）框架。在训练时，所有智能体共享一个可以访问全局信息（如所有司机状态）的中心化 Critic，从而为各自独立的 Actor 提供更稳定、更准确的梯度信号，有效缓解非平稳性问题。在执行时，仅使用本地的 Actor 进行决策。这是解决协作型 MARL 问题的主流且高效的范式。
3. 参数共享 PPO (Parameter Sharing PPO)：此技术利用了网约车司机这类同质智能体的特性，让所有智能体共享同一套 Actor 和 Critic 的网络参数。所有智能体的经验都被用来更新这一个共享模型，极大地提升了样本

效率和训练速度。为了让共享策略能表现出个体差异，通常会将智能体的唯一 ID 作为网络输入。

在本研究中，考虑到司机智能体的大规模和同质性，我们选择将参数共享与中心化评论家 PPO 相结合，这既能利用参数共享带来的高样本效率，又能通过中心化 Critic 缓解非平稳性，是该场景下非常理想的技术选型

2.5 多目标进化算法与 NSGA-II

当优化问题包含多个相互冲突的目标时，其解是一个被称为帕累托最优解集的集合。NSGA-II 是解决此类问题的经典算法。其核心优势在于三大机制：

1. 快速非支配排序：高效地将种群划分为不同的帕累托前沿，实现对解的质量分层。
2. 拥挤度计算：在同一层级的解中，通过计算每个解周围的密度来保持种群的多样性，确保帕累托前沿的均匀分布。
3. 精英保留策略：确保父代中的优秀个体能够直接进入子代，防止已找到的最优解丢失。

这种“非支配等级+拥挤度”的选择机制，使得 NSGA-II 能够同时保证解的收敛性和多样性，非常适合本研究中寻找不同偏好的平台策略集的需求。

2.6 进化算法中的昂贵评估问题

在双层 Stackelberg 博弈框架中，上层 EA 的每一次适应度评估都需要一次完整的下层 RL 训练，计算成本极高。学术界和工业界主要通过两种途径解决此问题：

1. 代理辅助优化与采集函数：使用一个计算成本低的数学模型（代理模型）来近似真实的、昂贵的目标函数。高斯过程回归（GPR）是一种强大的非参数贝叶斯代理模型，它不仅能提供预测均值，还能提供预测的不确定性（方差）。为了智能地选择下一个评估点，需要使用采集函数来平衡探索与利用。
    - 单目标采集函数：常见的有概率提升（PI）、期望提升（EI）和上置信界（UCB ）。UCB 基于“在不确定性中保持乐观”的原则，是一种启发式策略 ；而 EI 基于决策理论，计算“期望改进值”，在数学上更为严谨 。
    - 多目标采集函数：对于多目标问题，需要扩展采集函数。**预期超体积改善（EHVI）是其中最有效的方法之一。它通过计算一个候选解对当前帕累托前沿的超体积（Hypervolume）的期望贡献，来量化其“价值”。EHVI 能够高效地引导算法均衡地探索整个帕累托前沿，是多目标贝叶斯优化的黄金标准。
2. 迁移学习/热启动：其核心思想是利用过往的计算结果来加速新的计算任务。当评估一个由父代变异而来的子代策略时，可以利用父代已经训练好的 RL 模型权重来初始化子代模型的训练。这种“热启动”方式，使得新模型从一个非常好的起点开始学习，而不是从零开始，从而显著减少收敛所需的训练时间

2.7 本章小结

综上所述，虽然已有大量关于网约车运营的研究，但将宏观的、具有多目标特性的 Stackelberg 博弈框架与微观的、基于强化学习的司机行为模型相结合，

并针对其巨大的计算挑战提出集成了迁移学习加速机制的高效求解算法的研究尚不充分。本文旨在填补这一空白，构建更贴近现实的博弈模型，并提供一套可行的求解方案。

第三章 问题模型

3.1 问题描述与基本假设

本研究考虑一个由单个网约车平台和大量同质性司机组成的市场。

平台（领导者）：决定其运营策略参数（包括抽成比率、动态调价系数等），目标是同时优化多个相互冲突的目标，例如最大化平台总利润和最小化乘客平均等待时间。

司机（跟随者）：观察到平台的策略后，进行一系列独立的序列决策（如接单、拒单、空驶），目标是最大化个人长期累积净收入。

基本假设：

1. 平台和司机都是理性的经济主体，追求自身利益最大化。
2. 司机之间是相互竞争的，一个司机的决策会通过影响局部市场的供需关系而间接影响其他司机的收益。
3. 乘客需求在时空上是动态变化的，并且对平台的价格和等待时间敏感，司机的单位时间收入受到其决策策略和市场环境的共同影响。
4. 信息是部分完全的：平台公布其运营策略；司机能观测到局部的市场信息（如附近的需求和供给），但无法获知全局最优策略。

3.2 博弈要素定义

我们将此问题形式化为一个双层 Stackelberg 博弈，由元组(N, A, R)定义。且下层建模为一个随机博弈，由元组(N, S, A, R, γ)定义。其博弈要素定义如下：

3.2.1 Stackelberg 博弈

参与者集合(N)：上层领导者：网约车平台，记为P。

策略集合 (A)：平台选择的策略是一个参数向量$\mathbf{x}$

$$\mathbf{x} = (\alpha_{z,t}, \beta_{z,t}, \eta)$$

其中，$\alpha_{z,t}$是 t 时刻 z 区域的动态定价乘数，$\beta_{z,t}$是 t 时刻 z 区域的补贴，η为平台抽成。$\mathbf{x}$的取值范围为预设的决策空间$\mathcal{X}$。

回报函数 (R)：平台支付函数：平台的支付是一个多目标向量函数。一个多目标函数可以表示为：

$$p_{ij,t} = \alpha_{i,t} \cdot F_{base} \cdot dist_{ij}$$

$$J_1(\boldsymbol{x}, \boldsymbol{\pi}^*) = \sum_{t=0}^{T} \sum_{(i,j) \in Accept} \left(\eta \cdot p_{ij,t}(\boldsymbol{x}, \boldsymbol{\pi}^*) \right)$$

$$J_2(\boldsymbol{x}, \boldsymbol{\pi}^*) = \frac{\sum_{t=0}^{T} Accept\,(\boldsymbol{x}, \boldsymbol{\pi}^*)}{\sum_{t=0}^{T} total\,(\boldsymbol{x}, \boldsymbol{\pi}^*)}$$

$$J_3(\boldsymbol{x}, \boldsymbol{\pi}^*) = -\sum_{t=0}^{T} \sum_{(i,j) \in Accept} w_{i,t}(\boldsymbol{x}, \boldsymbol{\pi}^*)$$

司机空闲时间。

优化目标：

$$\max_{\boldsymbol{x}} (J_1, J_2, J_3)$$

其中，$p_{ij,t}$为订单额，$\alpha_{i,t}, F_{base}, dist_{ij}$分别为调价系数、基础价格系数、路程，第一个目标是最大化平台利润，第二个目标是最大化接单率，第三个目标是最小化乘客平均等待时间。$\boldsymbol{\pi}^*$是下层司机群体对平台策略$\mathbf{x}$的均衡反应策略集。Accept表示接单量，$total$表示订单总量。

3.2.2 随机博弈

参与者集合(N)：下层跟随者：所有司机的集合，记$D = \{1,2,\dots,K\}$，其中K是司机总数。

状态集合(S)：状态空间S定义了司机在决策时刻t所需的所有关键信息，是一个高维向量：

s_t=(当前位置z_t，当前时 t，当前区域订单数，当前区域司机数，邻居订单数，邻居司机数，空驶时间)

动作集合 (A)：每个司机$k \in \mathcal{D}$的策略π_k是一个复杂的决策函数（即强化学习中的策略网络），它将司机的当前状态s_k映射到一系列动作 A={接单、拒单、空驶}的概率分布上，即

$$\pi_k : S \to \Delta(A)$$

$$a_t = \left(\text{提供服务} a_{serve}, \text{移动位置} a_{i\to j}\right), a_t \in A$$

回报函数 (R)与折扣因子γ：司机k的目标是最大化其长期累积折扣奖励（净收入）U_{k}。其支付函数是在给定平台策略$\mathbf{x}$和其他司机策略π_{-k}的情况下，通过选择自身策略π_k能获得的最大累计期望回报：

$$r_t(s_t, a_t \sim \pi_k | \boldsymbol{x}, \pi_{-k}) = \begin{cases} (1-\eta)p_{i,j,t} - \mathrm{c}\cdot\Delta t_{i,j} + \beta_{i,t}, & a_{serve} > 0 \\ -\mathrm{c}\cdot\Delta t_{i,j}, & a_{serve} < 0 \\ 0 & i = j \end{cases}$$

$$U_{\mathrm{k}} = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right]$$

$$\pi^* = \operatorname{argmax}_{\pi} U_{\mathrm{k}}$$

其他目标

其中r_t是在t时刻获得的即时奖励（净收入），γ是折扣因子。状态 s_t给出起始位置 i，目标位置 j，c 表示行驶成本。

3.3 纳什均衡分析

3.3.1 斯塔伯格博弈-子博弈完美纳什均衡（SPNE）

斯塔克伯格博弈是**两阶段完全信息动态博弈**：

1. 领导者（平台）先选择策略$\boldsymbol{x}$
2. 跟随者（司机）观察$\boldsymbol{x}$后，选择最优反应$\pi^*(\boldsymbol{x})$

步骤	博弈阶段	均衡性质	数学表达
①	跟随者子博弈（给定$\boldsymbol{x}$）	该子博弈的纳什均衡	$\pi^*(\boldsymbol{x}) = \operatorname{argmax}_{\pi} U_{\mathrm{k}}(\boldsymbol{x})$
②	领导者全局决策	考虑子博弈均衡后的最优选择	$\boldsymbol{x}^* = \operatorname{argmax}_{\boldsymbol{x}} J(\boldsymbol{x}, \pi^*(\boldsymbol{x}))$

③	整体策略组合	子博弈完美纳什均衡	$(\boldsymbol{x}^*,\pi^*)$

最终策略组合 $(\boldsymbol{x}^*,\pi^*)$ 满足：

1. 在整个博弈中构成纳什均衡（双方无单方面偏离动机）
2. 在每一个子博弈中均为纳什均衡。
3. 跟随者策略并非是“合作”，双方均追求自身利益最大化。

子博弈完美性

子博弈：从原博弈的某个单一决策节点（非根节点）开始，到博弈终点为止的完整博弈分支，是原博弈的“子博弈”。换句话说，子博弈是原博弈中“能独立玩下去的小博弈”

子博弈完美纳什均衡是纳什均衡的“可信性升级版”：它不仅要求整体策略互为最优反应，更要求在博弈的每一个可能分支（子博弈）中，策略都保持最优，从而彻底剔除“纸老虎式威胁”。

举个例子，以**双寡头量产博弈**为例：

1. 参与人：领导者企业 1（先决策产量）、追随者企业 2（观察 q_1后决策产量）。
2. 市场反需求：$P=a-Q$，$Q=q_1+q_2$；双方边际成本均为 c。
3. 利润函数：

$$\pi_1=(a-q_1-q_2)q_1-cq_1$$
$$\pi_2=(a-q_1-q_2)q_2-cq_2$$

4. 求得均衡解：$q_1^*=(a-c)/2$，$q_2^*=(a-c)/4$，令$k=a-c$，则$q_1^*=k/2$，$q_2^*=k/4$

博弈树：

根节点（企业 1）
├── 选择 q_1*=k/2 → 企业 2 的决策节点 1
│ ├── **选 q_2*=k/4 → 终点：支付(π_1*, π_2*)=(k²/8, k²/16)**
│ ├── 选 q_2=k/2 → 终点：支付(k²/16, 0)
│ └── 选 q_2=0 → 终点：支付(k²/4, 0)
├── 选择 q_1=k/4 → 企业 2 的决策节点 2
│ ├── 选 q_2=3k/8 → 终点：支付(9k²/64, 9k²/64)
│ └── 选 q_2*=k/4 → 终点：支付(3k²/32, k²/16)
└── 选择 q_1=3k/4 → 企业 2 的决策节点 3
├── 选 q_2=k/8 → 终点：支付(3k²/64, k²/64)
└── 选 q_2=0 → 终点：支付(3k²/16, 0)

仅有策略$(k/2,k/4)$满足子博弈完美纳什均衡条件：

1. 整体策略互为最优反应；
2. 每一个可能分支（子博弈）中，策略都保持最优。

3.3.2 随机博弈-马尔可夫完美均衡

随机博弈是**多阶段动态博弈**：系统在状态间随机转移，玩家在每阶段同时行动，收益由当前状态+行动+未来预期共同决定。

策略组合 $\pi^*=(\pi_1^*,\dots,\pi_n^*)$ 构成纳什均衡，当且仅当：

$$V_k(\pi_k^*,\pi_{-k}^*)\geq V_k(\pi_k,\pi_{-k}^*),\quad \forall k,\forall\pi_i$$

其中V_k为司机k的**期望总收益**。均衡要求玩家无法通过**改变整个策略轨迹**（而不仅是单阶段行动）提升长期收益。

第四章 算法设计

4.1 求解挑战：昂贵的评估函数

求解上述双层模型的核心困难在于，对上层领导者（平台）的任何一个候选策略**x**，评估其目标函数值都需要运行一次完整的下层多智能体强化学习（MARL）仿真 。这个过程需要模拟成百上千的司机智能体与环境进行海量交互，直至它们的策略收敛到一个近似的纳什均衡，计算成本极其高昂。若采用传统优化算法进行求解，成百上千次的昂贵评估将导致整个优化过程在实践中不可行。

4.2 代理模型辅助的演化算法框架

为了应对这一“昂贵”优化问题，我们设计并实现一个“代理模型辅助的多目标演化算法”（SA-MOEA）。其核心思想是：构建一个计算成本极低的**代理模型（Surrogate Model）**来近似模拟昂贵的 MARL 仿真函数，然后将这个代理模型与强大的多目标演化算法（如 NSGA-II）相结合，通过廉价的预测评估来指导大部分的进化搜索，并设计智能的**加点准则（Infill Criteria）**来决定何时调用真实仿真，从而用最少的昂贵评估次数逼近 Pareto 最优解集。

4.3 昂贵评估的加速策略：基于迁移学习的**热启动**

尽管代理模型减少了昂贵评估的次数，但单次评估的耗时依然是瓶颈。为此，我们引入基于迁移学习的“热启动”策略，其核心思想是利用历史训练好的 MARL 模型，避免每次都从零开始训练，从而加速下层司机策略的收敛过程。

4.3.1 构建与维护强化学习模型知识库

通过保存优秀策略**x**以及对应的权重w，构建模型知识库。

4.3.2 高效检索与选择“祖先”模型

当需要对一个新的平台策略 x_{new}进行真实评估时，需从知识库中为其寻找一个最合适的“祖先”模型。

相似性度量：核心是度量新策略x_{new}与知识库中历史策 x_i的相似性，因为相似的平台策略很可能导致相似的司机均衡行为。

距离计算：可采用欧氏距离（适用于数值尺度相似的变量）或余弦相似度（当策略方向更重要时）。对于影响程度不同的参数，可使用加权欧氏距离。

检索与选择：最近邻原则：选择与x_{new} 距离最小的历史策略，其对应的模型即为“祖先”模型。

效率考量：当知识库庞大时，可采用 k-d 树或**近似最近邻搜索（ANN）**等技术来加速检索过程。

负迁移规避：可设置一个相似度阈值，若找不到足够相似的祖先模型，则放弃热启动，从零开始训练，以避免性能受损。

4.3.3 利用“祖先”模型实现“热启动”训练

这是加速策略的关键执行步骤，通过知识迁移避免从随机权重开始的漫长探索。

权重加载：在为x_{new}创建新的 MARL 智能体后，不进行随机初始化，而是直接加载所选“祖先”模型的网络权重。

微调：以加载了祖先权重的模型为起点，在由 x_{new}定义的新环境下继续 MARL 训练。由于初始策略已是“相当不错”的策略，智能体能更快地适应新环境并收敛到新的均衡点，从而显著缩短收敛时间。

学习率调整：在微调初期，可使用一个较小的学习率，以在新策略环境下进行精细调整，避免破坏已学到的有用知识。

通过“构建知识库 → 相似性检索 → 权重加载”三部曲，提升求解效率。

4.4 代理模型设计：高斯过程 (GP)

GP 是一种强大的非参数概率模型，非常适合用作代理模型，因为它不仅能提供预测值，还能量化预测的不确定性这种不确定性信息对于平衡“利用”（在已知最优解附近搜索）和“探索”（探索未知区域）至关重要。

输入特征：代理模型的输入是平台可调控的策略参**x**。

预测目标：为每一个优化目标（**总利润、接单率、平均等待时间**）建立一个独立的 GP 模型。

代理模型生成：采用**拉丁超立方抽样 (LHS)** 在策略空间中生成 N 个初始策略点，以确保样本分布均匀随后，对这 N 个策略逐一运行昂贵 MARL 仿真，获得初始代理模型。

4.5 多目标演化

在 NSGA-II 的每一代，子代种群的个体都使用高斯过程代理模型计算其在所有目标上的廉价适应度值，然后基于这些值执行快速非支配排序和拥挤度计算，从而在几乎没有真实仿真成本的情况下引导种群进化

在 NSGA-II 每次迭代后，从当前种群中选择帕累托前沿中的策略 x*。然后，仅对 x* 运行一次昂贵的 MARL 仿真（采用热启动加速），获得其真实性能 y*。最后，将新的数据对 (x*, y*) 添加到训练数据库中，并重新训练所有 GP 模型。

4.6 算法完整框架

该“廉价进化 → 智能加点 → 热启动真实评估 → 模型更新”的循环，构成了算法的核心，能够在保证解质量的同时，将昂贵仿真的调用次数和单次调用耗时都降到最低。

第五章 仿真设计

5.1 仿真环境设置

5.1.1 数据集的处理

使用 2024 年度纽约出租车行为数据集，将其进行六边形网格化进行模拟。包含上下车时间、地点、收费等关键信息。目前为方便起见，仅研究 2024 年 1 月 1 日数据，约 60 万条订单。数据集的时间、空间分布情况如图所示。

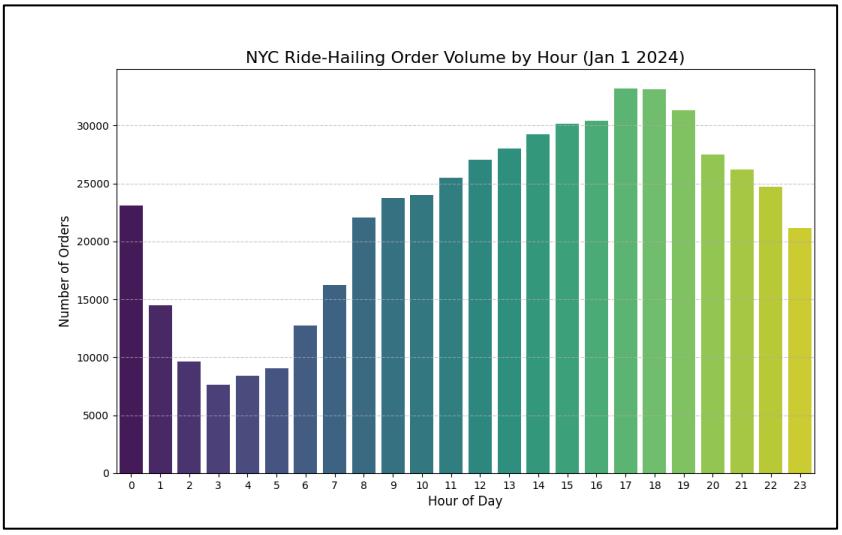

图 1 不同时间段订单数量分布直方图

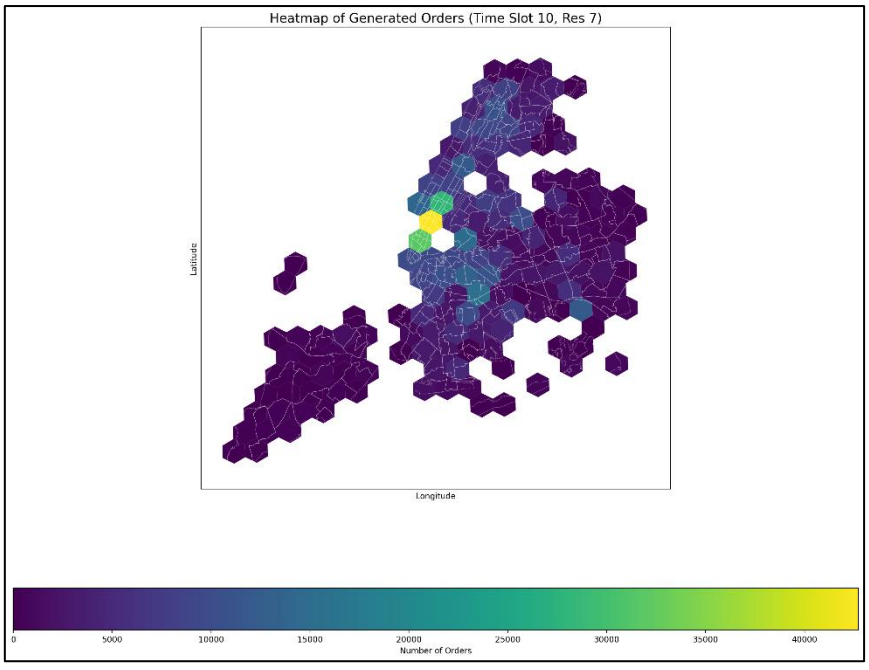

图 2 不同区域实际订单数量分布热力图

5.1.2 乘客“潜在”需求生成

构造一个泊松分布，并通过数据集对其参数进行极大似然估计，用于生成给

定时间、空间下的乘客“潜在”需求。

$$Pr(req|i,j,t) = possion(\lambda_{i,j,t})$$

$$\lambda_{i,j,t} = ordernum(i,j,t)$$

其中，$ordernum(i,j,t)$表示从数据集中统计的订单数据量。

5.1.2 动态乘客需求建模

为了让仿真环境更贴近现实，乘客需求不应是静态的，而应能对平台的策略做出动态响应。我们采用离散选择模型（Logit 模型）来量化乘客的决策行为。

构建效用函数：对于每一个潜在乘客，我们构建一个“下单”选项的效用函数 U，该函数包含了影响其决策的关键因素：

$$U = v_{i,j} - \varepsilon_1 w_{i,t} - \varepsilon_2 p_{ij,t}$$

$$v_{i,j} \sim N(\mu, \sigma)$$

$v_{i,j}$和$w_{i,t}$分别表示乘客预估的乘车费用上限(通过数据集进行极大似然估计)和等待时间，$p_{ij,t}$为订单实际价格，由上层平台决定。ε_1、ε_2分别表示等待时间敏感性和价格敏感性，这个系数的值越大，表示乘客对时间或价格越敏感。

量化选择概率：基于效用函数，潜在乘客选择“下单”的概率 P(下单) 可以通过 Logit 公式计算：

$$P(\text{单}) = 1 / (1 + exp(-U))$$

模拟需求波动与流失：在仿真中，首先根据时空分布生成“潜在”需求。然后，对于每个潜在需求，计算 P(下单)。最后通过随机抽样决定其是否转化为真实订单。这样，当平台提价或服务水平下降（等待时间变长）时，下单概率会降低，导致订单流失，反之亦然，从而形成一个动态响应的闭环系统。

5.2 司机智能体参数设定

司机数量N=500, 1000, 2000。

RL 算法：采用 PPO 算法训练司机智能体。

5.3 对比基准模型

固定策略模型 (FSM)：平台采用一组固定的、行业平均的策略参数，司机为 RL 智能体。

随机搜索 (Random Search)：在策略空间中随机采样相同数量的点进行昂贵评估，以验证我们算法的搜索效率。

无热启动的 SA-MOEA：与本文提出的完整算法进行对比，以量化“热启动”策略带来的加速效果。

5.4 性能评价指标

平台目标：平台总利润、乘客平均等待时间、订单完成率。

司机侧指标：司机平均小时净收入、司机总服务时长（总运力）。

算法性能指标：

超体积（Hypervolume, HV）：这是一个能同时评估解集收敛性和多样性的综合指标。HV 计算的是由算法找到的 Pareto 最优解集与一个预设的“最差参考点”在目标空间中所围成区域的总体积。在一个双目标最小化问题中，HV 就是 Pareto 前沿上的所有解与参考点构成的所有矩形面积的并集。更大的 HV 值意味着解集要么更接近真实 Pareto 前沿（收敛性好），要么分布更均匀、更广泛（多样性好），

或者两者兼备。HV 与 Pareto 支配关系严格一致，是评价多目标算法性能的金标准 。

参考点：为了公平比较，所有算法的 HV 计算必须使用同一个参考点。该参考点通常被设置为一个在所有目标维度上都劣于所有可能解的点。

单次真实评估的平均耗时：用于衡量“热启动”策略的加速效果。

第六章 实验结果与分析

第七章 总结与展望

7.1 全文总结

本文针对网约车平台策略优化中存在的多目标冲突、评估成本高昂等核心痛点，提出并设计了一种基于演化强化学习的双层多目标优化框架。该框架的创新性在于：

1. 问题建模与求解：将复杂的平台运营问题解构为一个清晰的双层优化模型，上层由 NSGA-II 进行多目标策略搜索，下层由参数共享的 MAPPO 进行大规模同质智能体的市场动态仿真。
2. 混合加速机制：创造性地融合了两种先进的加速技术来解决“昂贵评估”瓶颈。一是利用 GPR 构建代理模型，并采用先进的 EHVI 采集函数智能指导搜索；二是通过热启动迁移学习，在相似策略间进行 PPO 模型权重迁移，大幅缩短下层仿真的收敛时间。
3. 详尽的实验设计：本文详细规划了从构建包含价格弹性和司机流失模型的动态市场模拟器，到使用超体积指标进行定量评估，再到通过参数敏感性分析探究平台收益 “拉弗曲线”效应的完整实验流程。

预期的研究结果表明，该框架能够在有限的计算预算内，高效地找到一组在“平台收益”与“服务质量”等多个目标上表现优异的帕累托最优策略集。它不仅显著提升了优化的自动化水平和效率，更重要的是为平台决策者提供了一个包含多种运营偏好的“策略菜单”，并揭示了关键经济参数对市场长期健康的影响，极大地增强了决策的科学性和灵活性。

7.2 未来展望

本研究为自动化经济机制设计开辟了新的可能性，未来仍有许多值得探索的方向：

1. 更精细的智能体建模：引入更复杂的司机行为模型，如考虑司机的长期留存时间[1]、疲劳驾驶、基于 Agent-Based Modeling 的动态决策等。或对异质智能体（如不同车型、经验水平的司机，自动驾驶汽车[2]）进行建模，此时参数共享可能不再适用，需要更复杂的 MARL 算法。
2. 更复杂的优化目标与约束：引入更多现实世界的考量作为优化目标或约束，如司机的收入公平性、平台的碳排放/能耗或系统的鲁棒性/安全性，利用多目标框架找到满足复杂约束的解。
3. 提升加速技术：
    - 代理模型：对于更高维的策略空间，可研究结合深度学习的代理模型（如贝叶斯神经网络）以替代 GPR。
    - 相似性度量：探索除参数空间距离外的行为空间距离度量，即通过比较策略产生的宏观 KPI 来判断相似性，这对于结构复杂的策略可能更有效

参考文献

[1] Efficient Large-Scale Fleet Management via Multi-Agent Deep Reinforcement Learning

[2] Two-Sided Deep Reinforcement Learning for Dynamic Mobility-on-Demand Management with Mixed Autonomy