西安电子科技大学研究生学位论文
撰写要求（2015年修订版）

作者姓名 张三

指导教师姓名、职称 李四 教授

申请学位类别 工学硕士

学校代码 10701　　学　号 1101110071
分 类 号 TN82　　密　级 秘密

西安电子科技大学

硕士学位论文

西安电子科技大学研究生学位论文撰写要求（2015年修订版）

作者姓名：张三

一级学科：电子科学与技术

二级学科：电磁场与微波技术

学位类别：工学硕士

指导教师姓名、职称：李四 教授

学　　院：电子工程学院

提交日期：20xx年x月

Thesis/Dissertation Guide for Postgraduates of XIDIAN UNIVERSITY

A Thesis submitted to

XIDIAN UNIVERSITY

in partial fulfillment of the requirements

for the degree of Master

in Electromagnetic Field and Microwave Technology

By

Zhang San

Supervisor: Li Si Professor

February 2015

西安电子科技大学
学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文若有不实之处，本人承担一切法律责任。

本人签名：_____　　日　　期：_____

西安电子科技大学
关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属于西安电子科技大学。学校有权保留送交论文的复印件，允许查阅、借阅论文；学校可以公布论文的全部或部分内容，允许采用影印、缩印或其它复制手段保存论文。同时本人保证，获得学位后结合学位论文研究成果撰写的文章，署名单位为西安电子科技大学。

保密的学位论文在_____年解密后适用本授权书。

本人签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘要

网约车动态定价技术在缓解流量负载、控制交通拥塞、平衡供需、提高平台收入方面有重要作用。有效的动态定价策略不仅能够提高订单响应率，从而提高车辆利用率和打车服务提供者（平台方和司机）的收入，而且能增强乘客和司机的满意度。网约车供需不平衡问题在时间和空间上都普遍存在，时间上的供需不平衡体现在，需求随着时间变化，当前可能处于打车平峰期，下一时段可能就进入了打车高峰期，或者由打车高峰期过渡到平峰期；空间上的供需不平衡体现在，人流量大的地区的车辆供应比人流量少的地区的车辆供应更紧张。网约车动态定价基于需求的变化而相应地改变订单价格去达到平衡供需和提高平台订单响应率和收入的目的。由于当前的价格决策不仅会影响当前的收入也会对未来的收入造成影响，因此价格的决策应该致力于去优化累积的订单响应率和平台收入。然而，现存的大多数网约车动态定价算法旨在优化当前收入，或者前后两个时段的短期总收入，面临着优化长期收入时的灵活性和扩展性差以及复杂度高的问题。

强化学习从代理与环境的交互经验中学习优化长期收入，因此本文提出使用基于目标导向的强化学习技术解决网约车动态定价中面临的问题。本文首先将动态定价问题形式化为马尔科夫决策过程。考虑到空间上供需不平衡问题的存在，订单的价格由出发地和目的地不同而不同，同时价格设为作为连续值，避免动作空间离散化。其次，虽然动态定价旨在提高平台的收入，但是仅将收入作为奖励函数并不适合网约车平台的动态定价问题，因此本文针对设计了更适合的奖励函数。奖励函数通过计算即时收入、订单响应率和下一时段的供需分布的KL散度三者的和得到，相比于用即时收入作为奖励函数，其收敛更快更好。由于动作空间是连续值，在学习过程中需要充分探索，因此本文使用探索能力较为优秀且适合于优化连续动作空间问题的SAC算法优化本文中的动态定价模型。

使用基于单代理的强化学习来解决网约车动态定价问题会面临动作空间随着划分的区域数增加而增长的问题，导致模型难以收敛，因此为了降低动作空间的维度，本文同时提出基于多代理的强化学习方法去解决网约车动态定价问题，每个区域使用一个代理去决策其前往其它各区域的订单价格，多个代理共同优化平台的累积订单响应率和累积收入。多代理相比于单代理的方法面临的主要挑战是代理协同，该方法通过参数共享实现代理间的协同，参数共享使得每个代理的策略可以使用所有代理的经验训练，训练更有效。

本文使用真实数据集验证了本文方法的有效性，实验表明，基于强化学习的方法能大大提高订单响应率和平台收入，并且本文所设计的奖励函数能让模型收敛得

更快更好；实验表明基于多代理的方法相比代理的方法效果要差，但是基于多代理的方法很好地实现了代理间的协同，相比单代理的方法其适用范围更广。

关键词：动态定价， 供需平衡， 强化学习， SAC， 多代理

ABSTRACT

The Abstract is a brief description of a thesis or dissertation without notes or comments. It represents concisely the research purpose, content, method, result and conclusion of the thesis or dissertation with emphasis on its innovative findings and perspectives. The Abstract Part consists of both the Chinese abstract and the English abstract. The Chinese abstract should have the length of approximately 1000 Chinese characters for a master thesis and 1500 for a Ph.D. dissertation. The English abstract should be consistent with the Chinese one in content. The keywords of a thesis or dissertation should be listed below the main body of the abstract, separated by commas and a space. The number of the keywords is typically 3 to 5.

The format of the Chinese Abstract is what follows: Song Ti, Small 4, justified, 2 characters indented in the first line, line spacing at a fixed value of 20 pounds, and paragraph spacing section at 0 pound.

The format of the English Abstract is what follows: Times New Roman, Small 4, justified, not indented in the first line, line spacing at a fixed value of 20 pounds, and paragraph spacing section at 0 pound with a blank line between paragraphs.

Keywords: XXX, XXX, XXX, XXX, XXX

插图索引

表格索引

符号对照表

符号	符号名称
XXX	XXX
XXX	XXX
XXX	XXX
…	

缩略语对照表

缩略语	英文全称	中文对照
XXX	XXX	XXX
XXX	XXX	XXX
XXX	XXX	XXX

...

目录

第一章　绪论

1.1　研究背景与意义

网约车平台如滴滴出行、优步等平台自上线及不断发展以来，不仅满足了人们的日常出行需求，而且给乘客提供了方便快捷的出行服务和舒适优质的乘车环境。根据中国最大的移动出行平台滴滴出行所发布的《滴滴出行企业公民报告2017》[1] 显示，2017年滴滴平台的网约车服务在400多个城市覆盖，日订单量达到2500万，在滴滴上获得收入的司机超2100万，为社会创建了巨大的经济效益和社会效益。不仅如此，由于出行市场品质化和专业化发展，网约车基本实现了连续增长。由CNNIC发布的第46次《中国互联网发展状况报告》[2]显示，截止2020年6月，中国网民规模达到9.40亿，而网约车用户规模达3.40亿，占网民整体的36.2%，这充分说明网约车已成为人们日常出行当中的不二选择。

网约车平台作为出行乘客和司机的中间媒介，负责将制定合适的订单价格，吸引有出行需求的乘客打车，激励司机加入平台去服务乘客，将尽可能多的订单与司机进行匹配，从已完成的订单提取部分收入作为其收入。因此，平台需要设计合适的乘车价格来服务更多乘客，提高平台的订单响应率和总收入。

在提高订单响应率和平台总收入方面面临的挑战之一就是如何更好地平衡供需。一方面，乘客和司机间的时空不匹配问题在实际生活中是普遍存在的。例如，平峰期会出现供过于求的现象而在高峰期则是供不应求的情况。特别是在供不应求情况下，可用车辆被系统调度去服务距离其当前位置较远的乘客。结果，导致出现WGC现象（”Wild Goose Chase”）,司机会花费较长的时间去接乘客，一直处于忙碌状态，这将加剧空闲司机的稀缺性和交通的低效率[3]。另一方面，平台当前的决策会显式地影响未来司机的地理分布。如果平台使用短视的策略来制定当前的订单价格和车辆调度策略来平衡当前供需，忽略未来的供需状况，将导致未来可用司机的分布则会越来越分散，从而会造成未来需要用车的地区面临供应紧张的问题。因此，平台需要设计具有前瞻性的策略，引导更多的司机未来出现在高需求的地区，提高未来可服务的订单数。

目前有很多的方法能够平衡供需，动态定价便是其中一项关键技术。不管是乘客还是司机对订单价格都比较敏感，比如，乘客更希望价格越低越好，如果价格超过其心理价位，那么乘客就会放弃打车；相反，司机则希望价格越高越好，这样就能提高其收入，价格越高，司机愿意加入平台并服务乘客的积极性就越高。动态定价正是利用乘客和司机对价格的敏感性，基于实时的供需状况调整订单的价格，影

响乘客和司机的行为。动态定价对于缓解高峰负载均衡问题、提高订单调度率和最大化收入上有重要的作用。

本文研究如何根据实时的供需状况设计具有前瞻性的动态定价策略来最大化平台长期的累积订单响应率和累积收入。考虑到乘车的不稳定性和价格的后影响性，所设计的动态定价策略不仅要考虑平衡当前的供需，同时要尽可能平衡未来的供需。比如，当预知到未来某一地区会出现需求激增的情况时，当前的动态定价策略应该引导更多乘客前往该地区，从而使得未来能有更多的可用司机出现在该地区，避免未来该地区由于无车可用而造成订单流失或者需要远距离调车而导致WGC问题。前瞻性定价对于提高车辆的利用率和司机的收入，缩短乘客的等待时间，改善乘客的出行体验以及提高平台总收入具有重要意义。

大多数现有的网约车动态定价算法不具有前瞻性。优步和Lyft网约车平台采用的峰时定价作为其定价策略，当乘车需求高于车辆供应时，订单的基础价格就会乘上一个大于1的溢价系数，chen等人验证了峰时定价能激励司机在峰时工作更长时间[4]，Castillo等人则证明峰时定价能一定程度上缓解WGC问题[3]。但是许多乘客在峰时会被收取昂贵的乘车费时，大部分司机也只有在峰时才愿意出现。Qian等人针对该问题提出了TOD定价策略，与峰时定价只有溢价系数不同，TOD还引入了折扣系数，虽然乘客在高峰时期的乘车价格会溢价，但是处于平峰时期时，平台会给出相应的折扣来补偿乘客[5]。Bimpikis等人提出空间定价策略，即把城市划分成多个地区，订单的价格不仅与出发地区的供需有关，而且与目的地的供需有关。这些方法都致力于最大化平台当前的收入，并未兼顾平台未来的收入。Battifarano等人提出了一个峰时定价预测方法，通过构建机器学习模型建模交通流特征与峰时系数之间的关系，从而预测未来几分钟或几个小时内的峰时系数的变化，并且向司机和乘客共享预测结果，从而帮助平台有效地分配车辆，节省用户的金钱和时间[6]，不过该方法并没有设计新的动态定价策略。Asghari等人则考虑了当前和未来两个时段的短期收入最大化，通过估计未来的需求和司机分布，在当前决策时则考虑引导更多司机出现在未来出现供需激增的地区[7]，但该方法只是兼顾了短期收入，如果想扩展模型兼顾到更长期（几个小时，一天）的收入，则面临计算复杂度高的问题。

强化学习通过研究环境中的代理如何选择动作以实现给定的目标[8], 代理当前选择的动作考虑了即时的收益也考虑到了对未来的影响，以优化问题的长期目标。目前已经有不少使用强化学习来解决动态定价问题的相关研究[9–13],代理通过利用与环境互动收集来的经验学习提高定价策略。但是，这些方法把动态定价问题的价格（动作）离散化，之后使用传统的Q-learning方法求解，导致选择一个合适的离散动作数比较困难。如果离散动作数过于少，一个较大区间范围内的很多价格对代理来说都是一样的，代理收到关于价格动作的反馈也是不准确的，导致得到次优的策

略解。同样地，如果离散动作数过大，模型的优化过程中也会面临维度灾难和计算负担过大的问题。并且，已有的基于强化学习的动态定价算法不是为网约车场景所设计的，其目标函数与网约车场景下的目标函数也不同，因此不能直接扩展这些方法用在网约车场景下。城市的交通环境是复杂多变的，而强化学习不需要事先了解城市的交通情况，仅通过与城市环境的不断交互就可以学习如何制定合适的价格策略，同时可以根据环境的变化适当的做出响应。所以，强化学习可以帮助智能化定价，而智能化定价能大大促进智慧交通的发展。

1.2　国内外研究现状

动态定价对于解决智能交通系统中的拥塞控制、流量负载均衡和网约车的车辆调度问题起着重要作用，被广泛研究并应用于各种交通问题，比如车费定价、停车费定价、拥塞定价等，来最大化乘车服务或者停车服务的提供者的收入、缓解高峰负载。网约车动态定价则属于车费定价，网约车平台使用不同的动态定价技术提高平台收入和平衡供需。随着各类网约车平台的发展和移动出行的普及，近年来出现了不少关于网约车动态定价的研究，如峰时定价、需求定价、分时定价等，这些方法或使用传统的基于数学优化的解来求解最佳价格，或基于深度学习和强化学习来优化订单价格。本文中涉及网约车动态定价和强化学习的内容，因此本节将简要介绍网约车动态定价技术和强化学习技术的国内外研究现状。

1.2.1　动态定价

现存的网约车动态定价研究有很多，一般的优化目标是最小化平台运营成本[14]、最大化社会福利[15]或者最大化平台收入[16]等。Banerjee等人结合经济模型和排队模型来研究能使得平台收入最大化的最优价格，并对比了动态定价和静态定价两种定价策略下的吞吐量，发现动态定价能带来更多的吞吐量[17]。zha等人提出了打车市场不同劳动力供给行为假设下的均衡模型，并考察了峰时定价的表现，结果发现与静态定价相比，动态定价能为平台和司机带来更高的收入[18]。峰时定价是动态定价较为流行的一种定价模型，已被应用到优步[19]和Lyft[20]等网约车平台上，峰时定价能激励司机在道路上工作更长时间[4]，解决WGC问题[3]。以上的定价策略都旨在最大化平台收入，Fang等人发现基于收入最大化的定价模型会由于司机供应短缺而限制其收入，从而提出引入补贴来鼓励司机提供乘车服务。胡天宇[21]等人基于排队论模拟司机在系统中的流动，并根据动态定价构建社会福利最大化模型以及平台收入最大化模型，对比两个模型发现平台利润最大化模型获得的平台定价更高。。上述文献均考虑优化平台当前收入为研究目标，是一种短视的定价策略，为考虑当前的定价策略对平台未来的收入的影响。随着需求预测准确度的提高，未来

几分钟内或1个小时内的乘车需求可以较为准确地估计[22, 23]。Guda等人通过共享平台对乘车市场的需求预测结果给司机，从而引导司机从供应过剩的区域转移到供应短缺的区域[24]。Asghari等人结合下一时段的需求预测结果，通过降低当前时段某些地区的乘车价格，刺激更多前往未来供应短缺地区的乘车需求，从而能够使得更多司机出现在未来需求激增的地区，平衡未来的供需。但是这些模型仅考虑了当前和下一时段的短期收入优化问题，若需要将这些模型扩展到优化更长期的收入，则面临建模困难以及模型优化困难的问题。

强化学习能够优化长期目标，并且基于无模型的强化学习模型不需要建模环境的动态性，通过代理与环境的交互来地响应环境的变化，因此也被广泛应用于解决不同场景的动态定价问题。Kim等人[25]和Lu等人[10]利用强化学习来优化电力市场的价格平衡供需以降低能源消耗，强化学习无需关于电网系统环境的先验信息，解决动态定价中缺乏消费者信息和电网系统存在多种不确定性的问题。针对需求不平稳的收入管理问题，Rana等人[9]提出一个无模型方法来自适应地响应需求的变化，最大化收入。Maestre等人[26]在利用强化学习指定合适的价格以权衡系统收入和差异性定价造成的不公平性。以上的这些方法使用Q-learning方法来优化定价策略，因此需要将价格按照一定的粒度进行离散化。但是，现实中大部分动态定价问题的都需要价格是连续的数值，将问题建模在离散动作空间会影响结果的准确度。Turan等人[27]则在连续的动作空间中建模电车调度和电价的定价问题，其目标是平台收入最大化，故而将即时收入作为奖励函数，之后使用TRPO优化一段时间内的累积收入。然而，使用即时收入作为奖励函数可能对于其它问题很适用，但是对于网约车市场，打车需求是不平衡的，收入不仅受到订单价格的影响，而且与乘车需求有关，因此使用即时收入作为奖励函数在网约车动态定价问题上不一定适合。

1.2.2 强化学习

近年来，强化学习已成为机器学习领域的一个研究热点，被广泛应用来解决金融[28]、交通[29]、机器人控制[30]等领域问题强化学习是通过代理与动态环境交互来进行行为学习[31]，学习如何把状态映射到动作，以最大化数值奖励信号[8]。与监督学习不同，强化学习不需要训练数据，而是通过环境所反馈的奖励信号判断其所选择动作的好坏，从而做出相应的调整。强化学习其中一个挑战便是平衡探索和利用，一方面，代理需要充分地探索环境，获取更多信息避免局部优化；另一方面，代理也需要利用目前已经学习到的知识做出决策。常见的平衡探索和利用的方法有$\epsilon-$贪婪算法[8]、置信空间上限（Upper Confidence Bound，简称UCB）[32]、添加噪声项[33]、添加熵正则项[34]等。按照环境中代理的数目，强化学习模型可以分为单代理模型和多代理模型。基于单代理的强化学习方法比较简单，适合动作空间维

度不是特别大的问题，在平衡探索和利用上采用常用的方法即可。而当动作空间维度过高，基于单代理的方法即使结合深度学习技术，也需要花费大量的时间进行充分的探索，并且单个模型参数过多会导致收敛困难。而解决动作维度过高的一个解决方法是采用多代理的强化学习方法，将问题分解成一个个子问题，每个子问题由一个代理去处理，从而降低了问题的动作空间维度。基于多代理的强化学习需要各个代理间形成协作，共同去实现全局目标。常见的可以让代理们达成协同关系的方法有参数共享[35]、中心化训练[36]等。中心化训练会随着代理数目的增长而导致状态空间和动作空间的维度增长，从而导致训练和收敛困难。参数共享则所有代理共享单个策略的参数，状态空间和动作空间的维度不受代理数目的影响。

1.3　研究内容与创新点

本文为网约车平台提出一种新的动态定价方法，以最大化长期的APR（Accumulative Platform Revenue，简称APR）和ORR（Order Response Rate，简称ORR）。订单响应率能够体现平台的服务水平，一般来说，订单响应率越高，收入也越高。为了找到前瞻性的定价策略，在决策时有必要考虑未来的供需关系。强化学习在做决策时权衡了当前收益与未来的收益，天然适合于解决最大化长期奖励的问题。之前大部分动态定价算法是基于数学优化，需要预先建模环境的动态性，强化学习不需要事先建模复杂的城市交通环境，仅通过代理与环境的交互来学习。强化学习与深度学习结合来近似动作值函数, 能避免动作空间离散化。因此，本文应用深度强化学习来求解网约车平台的动态定价问题。

基于深度强化学习的网约车动态定价算法需要解决以下的挑战，首先，需要设计适合的奖励函数，不合适的奖励函数会导致不收敛的策略。其次，不同的调度策略会得到不同的司机地理分布，有可能会使得目标函数陷入次优，因此控制可用司机的未来的地理分布是重要的。最后，由于本文将问题建模在连续的动作空间中，因此需要让代理充分地探索环境，避免局部最优。

为了进一步解决上述的第一个挑战，本文设计了一个新的奖励函数。基于其它场景的动态定价问题一般将即时收入作为奖励函数，在需求是平稳的场景下，直接将即时收入作为奖励函数是合适的。但是对于网约车场景下，需求是不平稳的，在时间和空间上均存在波动，这时，平台收入也受到需求的影响，即时收入作为奖励函数将导致错误的优化方向。所以需要另外设计合适的奖励函数。考虑到目标函数是最大化APR和ORR，本文的奖励函数包含了即时收入和订单响应率, 除了这两项以外，奖励函数还加上了下一时段各地区的司机供应分布和打车需求分布的KL散度，这一项能够让代理在做决策时兼顾到未来的供需状况。针对第二个问题，Chen等人[37]提出只把订单价格作为唯一市场调节手段，价格会较高和不稳定。并且只优

化价格并不能做到全局优化，且系统不能控制司机的未来分布。因此，本文将订单价格和司机调度联合优化。对于第三个问题，为了提高代理的探索能力，本文应用SAC（Soft Actor-Critic，简称SAC）算法来优化定价策略。SAC算法通过最大化奖励函数的期望和策略熵来进行充分的探索，并且能保证模型的稳定性[38]。本文将基于深度强化学习的动态定价方法命名为DRLDP（Deep Reinforcement Learning for Dynamic Pricing）。DRLDP是基于单代理的方法，其代理的策略包含两部分，一部分是价格策略，另一部分是调度策略。本文将城市分成N个区域，订单价格根据其起点区域和终点区域的供需状况进行差异化定价，因此需要优化的订单价格有N^2个，同时，代理亦需要优化从一个区域调度去服务前往其它区域的订单的司机比例，所以调度比例也有N^2个需要优化。总的动作数为$2N^2$，该动作数会随着区域划分数的增长而增长，会给模型的优化和收敛带来困难。为了解决动作维度过高而导致的优化和收敛问题，本文另提出了基于多代理的强化学习方法MRLDP（Multi-agent Reinforcement Learning for Dynamic Pricing)，每个区域前往其它区域的订单策略和调度策略有一个代理负责优化，这样就能把动作维度从$2N^2$降到$2N$，极大降低了动作维度，方便模型的收敛。MRLDP需要在多个代理间形成协同，共同优化全局的APR和ORR，为此，本文使用共享参数策略让所有代理形成协作。在共享参数下，每个代理输入自身的局部观测状态和索引，便能得到其对应的策略，同时所有代理与环境交互获得的经验被共同用来优化策略网络，因此能极大加速模型的收敛。

本文的研究内容和创新点总结如下：

（1）针对之前的动态定价模型为优化长期收入来得到前瞻性的定价策略的问题，本文提出了一个深度强化学习框架DRLDP来解决网约车平台的动态定价问题。DRLDP将同时优化订单价格和车辆调度以谋求全局优化，并将动态定价问题建模在连续的动作空间上，并使用SAC算法找到最优策略。

（2）对在乘车需求不稳定情况下直接将即时收入作为网约车动态定价并不适合的问题，本文设计了一个新的奖励函数。奖励函数包括即时收入、订单响应率和下一时间步的车辆供应分布和乘车需求分布之间的KL散度之和。

（3）针对基于单代理的DRLDP会由于动作空间维度增长而导致优化和收敛困难的问题，本文另提出基于多代理的MRLDP方法来解决该问题，每个区域由一个代理负责其定价策略和调度策略的优化，使用共享参数让所有代理达成协同，来优化全局的APR和ORR。

（4）本文在真实的数据集上与基准算法进行比较，验证了本文所提出的两个方法的性能。

1.4　论文章节安排

本论文的章节安排如下：

第一章：绪论，介绍了本文的研究背景与意义，所研究问题所涉及的动态定价领域和强化学习领域的国内外研究现状，以及本文的主要研究内容和创新点。

第二章：问题形式化和相关理论技术，将动态定价问题进行形式化，并介绍本文所涉及和应用到的相关理论和技术。

第三章：基于单代理的动态定价方法。详细介绍基于单代理的DRLDP方法，包括马尔科夫决策过程中的基本元素如状态、动作、奖励函数和状态转移函数的设计，并且展示在真实数据集上的实验结果。

第四章：基于多代理的动态定价方法。介绍为了解决动作维度增长而导致优化和收敛困难问题，提出的基于多代理的MRLDP方法，介绍各个代理的马尔科夫决策过程，以及促进代理间协作的方法。最后进行对比实验，验证该方法的性能。

第五章：总结与展望。对本文的研究工作做一个总结，发现其优点与不足，并对其不足和未来可改进之处做出展望。

第二章　问题形式化描述与相关理论技术

本章首先对问题做一个形式化描述，介绍网约车场景下的动态定价问题，描述动态定价的一般过程以及本文实验对比时使用到的基准算法。然后介绍强化学习的相关理论，包括马尔科夫决策过程以及基于多代理强化学习的常见的代理协同方法。

2.1　问题形式化描述

2.1.1　问题描述

（1）交通网络和乘车需求模型

本文将使用包括N 个节点的全连通图来表示交通网络，图中的每个节点表示一个区域，可以看做订单的起点或者终点，并且每个节点都包含着一定数量的潜在乘客和司机，所有节点的集合使用$G = \{v_i | i \in 1, 2, ..., N\}$ 表示。本文假设任意两个节点之间的距离是相等的，事实上，本文所提出的方法很容易被扩展去考虑节点间距离不相等的情况，该假设是为了与已有基准算法的假设保持一致，方便实验的对比，本文拟最大化平台一天的收入，两个连续的时间步之间的间隔时长为Δ_t, 一天被划分成T个时间片。在每个时间步，潜在的乘客会出现在各个网络节点，并且会有一定数量的乘客会提交订单，所提交的订单会同一节点的可用的并愿意提供服务的司机所服务，本文假设所有的订单会在一个时间步之后到达目的地。

当一个用户打开手机上的打车应用软件，选择其起点和目的地进行相应的乘车价格、通行时间、通行距离或者附近可用车辆的查询时，则表明这名乘客有打车的意向，不管该用户最终是否提交打车订单，平台会将其标记为一名潜在的乘客。在每个时间片内，平台可以统计得到各个节点的潜在乘车需求数，本文将从节点i到节点j的潜在乘车需求数使用Λ_{ij}^t表示。对于每个潜在乘客用户来说，用户会有一个对于本次行程的预留价格，预留价格影响着用户的行为。当平台给出的订单的价格低于其预留价格时，用户会愿意提交打车订单，反之，用户会放弃乘车退出打车平台。所以，在这个过程中可以看出，最终愿意提交订单的乘车需求与乘车价格有着一定的关系。动态定价问题大都需求设计一个模型来描述需求随价格变换之间的响应情况。需求模型的研究比较复杂且超过本文的范围，所以本文使用在常见的网约车动态定价策略研究中[7, 17, 24, 37, 39]的需求模型来建模乘车需求随乘车价格变化的情况。

假设在乘客愿意付钱乘车的意愿上，所有乘客都是同质的。本文使用函

数$f^r(p)$表示当乘车订单价格设置为p时，乘客愿意提交订单的概率, 其对应的累积分布函数为$F^r(p)$。本文将乘车价格的范围设置为$p \in [0, p_{max}]$，且价格按照起点和终点进行差异化定价。在时间步为t时，从起点i到终点j的乘车价格设置为p_{ij}^t，最终愿意提交订单的乘车需求数表示为$A_{ij}^t(p_{ij}^t)$, 形式化为：

$$A_{ij}^t(p_{ij}^t) = \Lambda_{ij}^t(1 - F^r(p_{ij}^t)) \tag{2-1}$$

（2）司机供应模型

司机与乘客一样，对订单价格也是敏感的。与乘客不同，乘客希望乘车价格越低越好，节省其乘车成本，而司机则希望乘车价格越高越好，这样便能提高其收入。司机对于每次的行程也有一个预留价格，当订单价格超过司机的预留价格，司机才会积极地加入平台，为平台广播给司机的订单提供服务。假设在t时，节点i的可用司机供应数为V_i^t。平台将指派一定比例的可用司机服务从节点i前往节点j的订单，假设指派比例为b_{ij}^t，则会有$b_{ij}^t V_i^t$数量的可用司机被平台调度来服务从节点i前往节点j的订单。$b_{ij}^t \in [0, 1]$，并且满足$\sum_{i=1}^N b_{ij}^t = 1$的等式约束。在$b_{ij}^t V_i^t$数量的可用司机中，最终只有预留价格不高于订单价格的司机才会愿意接单。本文使用函数$f^v(p)$表示当订单价格为p时，司机愿意接单的概率，其对应的累积分布函数为$F^v(p)$。所以，在t时，当交通网络中从地区i前往地区j的订单价格为p_{ij}^t时，最终的司机供应用$U_{ij}^t(p_{ij}^t)$表示，形式化为：

$$U_{ij}^t(p_{ij}^t) = V_i^t b_{ij}^t F^v(p_{ij}^t) \tag{2-2}$$

（3）平台模型

平台通过调整订单价格来影响司机和乘客之间的行为，并充当司机和乘客之间的媒介，将已经提交的订单指派给一定数量的司机，并将乘客和愿意提供服务的司机进行匹配。在t时，假设从节点i前往节点j的乘车需求数为$A_{ij}^t(p_{ij}^t)$, 对应的愿意提供服务的司机供应数为$U_{ij}^t(p_{ij}^t)$, 当$U_{ij}^t(p_{ij}^t) < A_{ij}^t(p_{ij}^t)$, 供不应求下，平台按照一定的规则（如先来先服务）选择$U_{ij}^t(p_{ij}^t)$数量的乘客，为其匹配对应的司机，而剩下未被匹配的乘客会被平台放弃。同理，当$U_{ij}^t(p_{ij}^t) > A_{ij}^t(p_{ij}^t)$, 供过于求时，所有提交的订单都会被匹配到对应的司机。接到订单的司机通过服务乘客而转移到订单的目的地，完成其订单后，在目的地重新变得可用，未接单的司机则留在原处等待接单。因此，不管是供不应求还是供过于求，最终平台在t时，设置从节点i前往节点j的订单价格为p_{ij}^t后，能服务的订单数为：

$$O_{ij}^t(p_{ij}^t) = \min\{A_{ij}^t(p_{ij}^t), U_{ij}^t(p_{ij}^t)\} \tag{2-3}$$

平台会从已完成的订单所收到的钱中，提取λ比例的金钱作为其收入，剩下的$1 - \lambda$比例的金钱由司机获得。在t时，平台在整个交通网络获得的收入表示

为Rev^t，形式化为：

$$Rev^t = \lambda \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}^t O_{ij}^t(p_{ij}^t) \tag{2-4}$$

（4）收入最大化问题

平台不能直接控制司机和乘客的行为，只能通过订单价格间接影响。因此网约车动态定价问题是通过设置合适的订单价格，来最大化平台的收入，提高订单响应率。之前大部分关于网约车动态定价的研究只把订单价格p作为优化的参数，而把车辆调度比例b作为常数项，对每个时间步的价格进行单独的优化。然而，由于司机在网络中的移动，会显著地影响未来可用司机在各地区的分布。仅优化价格而不考虑未来的供需状态会得到短视的定价策略，优化的是短期收入，限制了收入的提高。平台应该考虑优化长期收入，来将更多司机转移到未来需求激增的地区，服务即将到达的订单。所以平台需要同时优化每个订单的价格和车辆调度比例，制定前瞻性的定价策略和调度策略，从而平台的优化长期收入。本文中拟打算最大化平台一天的总收入，所以本文需要解决下面的收入最大化问题：

$$\max_{p_{ij}^t, b_{ij}^t} \sum_{t=1}^{T} Rev^t = \max_{p_{ij}^t, b_{ij}^t} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij}^t O_{ij}^t(p_{ij}^t) \tag{2-5}$$

上式把λ去掉了，λ不是需要优化的参数，所以去掉并不影响最终的优化结果。

2.1.2　已有基准算法

已有的基准算法一般是基于数学优化的算法，通过求解当前时段的收入最大化问题，或者短期内的收入最大化问题。下面分别介绍属于这两个类型的经典算法。

（1）DPCRM算法

网约车平台（如优步、Lyft）一般根据网络的供需状况来实时地决定订单的价格。DPCRM（Dynamic Pricing for Current Revenue Maximization，简称DPCRM）考虑网络当前的需求情况，当乘车需求高于可用司机供应时，订单的价格会提高以鼓励更多的司机加入平台。DPCRM算法的订单价格根据起点进行差异化定价，且仅优化订单价格，车辆调度比例与从一个地区前往各个地区的订单比例保持一致。DPCRM通过优化每个时段的收入，得到每个时段各地区的最优价格p_i^{t*}，形式化表示如下：

$$p_i^{t*} = \max_{p_i^t} \sum_{i=1}^{N} p_i^t O_i^t(p_i^t) \tag{2-6}$$

平台将每个时段的最优收入进行相加便可得到一天的总收入，即：

$$TotalRev = \sum_{t=1}^{T} \sum_{i=1}^{N} p_i^{t*} O_i^t(p_i^{t*}) \tag{2-7}$$

可以看出，DPCRM是短视的动态定价算法，没有同时优化司机的调度，有未将下一时段的供需情况纳入考虑。

（2）POD算法

POD（Predicting demand at Origin & Destination，简称POD）算法是文献[7]所提出的定价算法，其结合了下一时段的需求预测结果，通过降低当前时间步往未来会出现需求激增的地区的订单价格，来鼓励更多潜在乘客提交订单，从而使得未来能有更多可用司机出现在需求激增的地区。该算法依据订单的起点和终点的目的地进行差异化定价，且同时优化车辆的调度比例。

POD优化连续两个时段的总收入，当前时段的收入可以通过公式2-4计算得到。接单的司机通过接送乘客到目的地从而发生响应的转移，所以可以估计得到未来可用司机的在节点i的数量为：

$$V_i^{t+1} = V_i^t - \sum_{j=1}^{N} O_{ij}^t + \sum_{j=1}^{N} O_{ij}^t + \delta_i^{t+1} \tag{2-8}$$

其中，$\sum_{j=1}^{N} O_{ij}^t$是从地区i载客离开的司机数量，$\sum_{j=1}^{N} O_{ij}^t$ 是从其它各个地区载客到达地区i的司机数，δ_i^{t+1}是在$t+1$时段加入平台的司机数量和离开平台的司机数量的差。结合下一时段的潜在需求预测结果Λ^{t+1}，便可以通过DPCRM算法求得下一时段的收入Rev^{t+1}

POD算法通过优化以下目标函数求得t时的最优价格${p_{ij}^t}^*$和最优调度比例${b_{ij}^t}^*$：

$${p_{ij}^t}^*, {b_{ij}^t}^* = \max_{p_{ij}^t, b_{ij}^t} (Rev^t + Rev^{t+1}) \tag{2-9}$$

POD最后的总收入通过把各时段的收入相加得到。

POD算法虽然在制定订单价格和车辆调度策略时具备一定的前瞻性，能有效缓解下一时间步由于某些地区需求激增所造成的的供需紧张的问题。但是，对于未来乘车需求较少的地区并且当前没有足够的车辆服务所有的订单时，POD会从价格和调度比例上抑制生成前往这些地区的订单。且由于司机只能通过载客发生转移，未载客的司机会继续原地等候订单，造成某些地区存在大量的可用车辆，而有些地区则供不应求。对于这种情况，仅考虑前后优化前后两个时段的短期收入是不够的，需要从更长远的视角来进行价格和调度比例的优化。POD优化前后两个时段的收入需要$O(N^3)$的时间复杂度，继续扩展优化更长期的收入则面临更高的计算复杂度。并且在继续建模$t+2, t+3, ...$的收入也存在困难，缺乏扩展性和灵活性。

2.2　相关理论技术

2.2.1　马尔科夫决策过程

强化学习问题可以通过建模成马尔科夫决策过程（Markov Decision Process，简称MDP）来形式化表示，MDP的基本思想是通过代理与环境进行交互来达到实际问题的目标。MDP可以用五元组$< \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma >$表示, 其中$\mathcal{S}$是状态空间，$\mathcal{A}$是动作空间。$\mathcal{R}$是奖励函数，将状态-动作对$(s_t, a_t)$映射到奖励值，$\mathcal{R} = \mathcal{R}(s_t, a_t) : \mathcal{R} \times \mathcal{A} \to \mathbb{R}$。$\mathcal{P} = \mathcal{P}(s_{t+1|s_t,a_t}) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$，是状态转移函数，描述了在给定当前状态$s_t$和动作$a_t$时，环境从$s_t$转移到$s_{t+1}$的概率，$\gamma \in [0, 1]$是延迟因子，用于权衡即时奖励和未来收益。

代理与环境的交互发生在离散的连续时间步中，$t = 0, 1, 2, 3, ...$，在每个时间步t，代理通过观测所处环境的状态$s_t \in \mathcal{S}$, 基于此状态并按照所遵循的策略$\pi(a_t|s_t) : \mathcal{S} \times \mathcal{A} \to [0, 1]$，选择一个动作$a_t \in \mathcal{A}$作用于环境。环境转移到下一个状态$s_{t+1} \sim \mathcal{P}(S_{t+1}|s_t, a_t)$，并给代理反馈回去一个奖励值$r_t \sim \mathcal{R}(s_t, u_t) \in \mathbb{R}$。代理与环境在每个时间步的交互过程如图2.1所示。

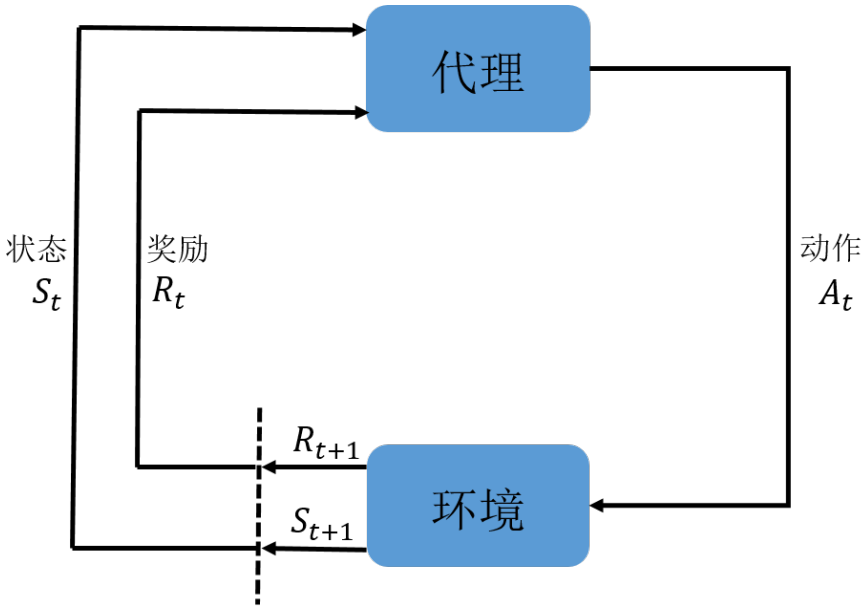

图 2.1　代理与环境的交互过程

环境中的代理根据所收到的奖励信号更新其策略π，即更新给定状态s下，各个动作的概率分布，其目标是学习如何最大化每个情节的奖励值的累积折扣和的期望，累积折扣和称之为回报，形式化为：

$$G_t = \sum_{t=1}^{T} \gamma^{t-1} r_t \tag{2-10}$$

代理通过最大化期望回报来找到最佳策略π^*，可形式化为，

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}[\sum_{t=1}^{T} \gamma^{t-1} r_t] \tag{2-11}$$

目前存在很多强化学习算法来求解最佳策略，基于Q-表的方法如Q-learning，通过查询存储着各个状态-动作对的Q值来获得最优策略。由于要存储所有状态-动作对的Q值，所以这类方法适合状态和动作空间是离散的且维度较低的问题。对于状态或者动作空间是连续的问题，则可以结合深度神经网络来去近似状态值函数或者状态-动作值函数，这类方法属于深度强化学习方法。常见的比如A2C、TRPO、DDPG、SAC等，这类方法在动作空间维度不是很高的情况下，能很好地帮助代理找到最佳策略。

2.2.2 深度多代理强化学习

深度强化学习方法如A2C、TRPO、DDPG、SAC等大多用于处理单代理设置的问题，即环境中仅存在一个代理来与环境交互学习全局的决策。但是，现实生活中的很多场景都包含着多个代理，如自动驾驶、包裹派送、车辆调度等问题。多代理设置下，代理们处在同一个环境中，每个代理根据其观测状态，独立地做出自己的决策，最大化全局目标或者个体目标。

多代理强化学习通过将问题建模为SC（Stochastic game，简称SG）[40]来形式化，SG可以用$< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, Z, \mathcal{O}, n, \gamma >$定义，其中，$n$表示环境中代理的数量，用$u$来标识环境中的一个代理，$u \in \mathcal{U} \equiv \{1, .., n\}$。$\mathcal{S}$表示环境的全局状态，$A$是动作空间。在每个时间步，每个代理$u$从联合动作$a \in A \equiv A^n$选择一个动作$a^u \in A$。环境根据其状态转移函数$\mathcal{P}(s'|s, a^u) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$进行相应的状态转移，每个代理所收到的奖励由奖励函数给出，$R = R(s, a, u) : \mathcal{S} \times \mathcal{A} \times \mathcal{U} \to \mathbb{R}$。与单代设置一样，$\gamma \in [0, 1]$是折扣因子。在局部可观测设置下，每个代理$u$会对应一个局部观测状态$o^u \in O$，该局部状态可以根据观测函数$Z(s, u) : \mathcal{S} \times \mathcal{U} \to O$得到。本文默认是采用局部可观测设置，即每个代理仅可以感知环境的局部状态。

根据代理是否能完全感知全局状态，多代理强化学习的环境设置可以分为完全可观测设置和部分可观测设置，相应地，学习的策略也分为中心化策略和去中心化策略。中心化策略下一般对应于在完全可观测设置，其策略$\pi(a|s_t)$通过将全局状态s_t映射为联合动作a的概率分布，即：

$$\pi(a|s_t) : \mathcal{A} \times \mathcal{S} \to [0, 1] \tag{2-12}$$

但是，基于中心化策略的方法会面临两个挑战，首先，联合动作空间的维度会随着代理的数目呈指数形增长。其次，在真实世界的问题场景下，代理仅能观测到

局部状态，故而学习中心化策略是不实际的。基于去中心化策略的方法则没有这个问题，去中心化策略下，每个代理自身对应一个策略$\pi^u(a^u|s_t)$，将状态映射到该代理所对应动作的概率分布上。联合动作上的概率分布可由各代理的概率分布结合得到，形式化为：

$$P(a|s_t) = \prod_u \pi^u(a^u|s_t), u = 1...n \tag{2-13}$$

当基于多代理强化学习来最大化问题的全局目标时，代理之间需要达成协作，来最大化全局奖励。环境可以反馈给所有代理相同的奖励值来让代理达成完全合作的关系，即每个代理的奖励函数为：

$$r(s, a, u) = r(s, a, u'), \forall u, u' \tag{2-14}$$

与单代理一样，每个代理需要最大化自身的期望回报：

$$J^u(\pi) = \mathbb{E}_\pi[\sum_{t=1}^{T} \gamma^{t-1} r_t^u] \tag{2-15}$$

目前有很多方法可以处理多代理的强化学习问题，本文主要集中于深度多代理强化学习（Deep Learning for Multi Agent Reinforcement Learning，简称DMARL），DMARL结合了深度神经网络来表示各个代理的值函数和策略。在合作设置下，常见的让代理达成协作关系的方法是参数共享，即不同的代理使用相同的参数θ来参数化其值函数和策略。参数共享极大节省了计算成本，所有代理的策略可以并行计算。同时，不同代理的经验都可以用来更新网络参数，提高了样本利用率。

Tan等人[41]通过结合IQL（Independent Q- leanrning，简称IQL）[42]和DQN[43]提出了比较经典的IDQN算法（Independent DQN，简称IDQN）。IDQN通过共享网络参数和使用索引来区别代理来扩展IQL算法，使其能应用到大规模多代理场景中。网络参数通过最小化以下的损失函数来更新：

$$\mathbb{E}[Q(s_t^u, a_t^u; \theta) - (r_{t+1}^u + \gamma \max_{a_{t+1}^u} Q(s_{t+1}^u, a_{t+1}^u; \theta'))] \tag{2-16}$$

其中θ'是目标Q网络的参数，$Q(s_t^u, a_t^u; \theta)$则表示代理u的状态-动作对(s_t^u, a_t^u)所对应的状态动作值，该方法一般用于动作空间是离散的问题，但很容易能扩展应用于连续的动作空间上。

2.3　本章小结

本章将动态定价问题进行了形式化建模并介绍其优化的目标函数，之后介绍了两个常用的动态定价方法，作为本文实验对比的基准算法之一。由于本文所提出的

算法涉及强化学习的相关理论技术，所以本章同时介绍基于单代理的马尔科夫决策过程和深度多代理强化学习，并介绍局部观测和合作设置下，能让多代理达成合作的方法。为后续研究基于单代理和动态定价算法以及扩展到基于多代理的动态定价算法打下基础。

第三章　研究生学位论文的编辑、打印、装订要求

3.1　学位论文封面的编辑和打印要求

学位论文的封面由研究生院按国家规定统一制定印刷，封面内容必须打印，不得手写。

3.2　学位论文的版面设置要求

（1）行间距：固定值 20 磅（题名页除外）。

（2）字符间距：标准。

（3）页眉设置：单面页码页眉标题为章节题目，每一章节的起始页必须在单面页码，双面页码页眉标题统一为“西安电子科技大学博/硕士学位论文”，页眉标题居中排列，字体为宋体，字号为五号。页眉文字下添加双横线，双横线宽度为 0.5 磅，距正文距离为：上下各 1 磅，左右各 4 磅。

（4）页码设置：学位论文的前置部分和主体部分分开设置页码，前置部分的页码用罗马数字标识，字体为 Times New Roman ，字号为小五号；主体部分的页码用阿拉伯数字标识，字体为宋体，字号为小五号。页码统一居于页面底端中部，不加任何修饰。

（5）页面设置：为了便于装订，要求每页纸的四周留有足够的空白边缘，其中页边距为上 3 厘米、下 2 厘米；内侧 2.5 厘米、外侧 2.5 厘米；装订线为 0.5 厘米；页眉 2 厘米，页脚 1.75 厘米。

3.3　学位论文的打印、装订要求

（1）打印：学位论文必须用 A4 纸页面排版，双面打印；

（2）装订：依次按照中文题名页、英文题名页、声明、摘要、插图索引、表格索引、符号对照表、缩略语对照表、目录、正文、附录（可选）、参考文献、致谢、作者简介的顺序，用学校统一印制的学位论文封面装订成册。盲审论文必须删除致谢部分的文字内容（致谢标题须保留）以及封面和研究成果中的作者和指导教师姓名，研究成果列表中应体现作者的排序，如第一作者、第一发明人等。

3.4　其他说明

本规定由研究生院负责解释，从申请 2015 年 9 月毕业和授位的研究生开始执

行，其它有关规定同时废止。研究生毕业论文撰写要求参照学位论文撰写要求执行。

第四章　图、表、公式示例

图：包括曲线图、示意图、流程图、框图等。图序号一律用阿拉伯数字分章依序编码，如：图 1.3 、图 2.11 。

每一个图应有简短确切的图名，连同图序号置于图的正下方。图名称、图中的内容字号为五号，中文字体为宋体，英文字体为 Times New Roman ，行距一般为单倍行距。图中坐标上标注的符号和缩略词必须与正文保持一致。引用图应在图题右上角标出文献来源；曲线图的纵横坐标必须标注“量、标准规定符号、单位”，这三者只有在不必要标明（如无量纲等）的情况下方可省略。

图与正文之间一般应空一行。

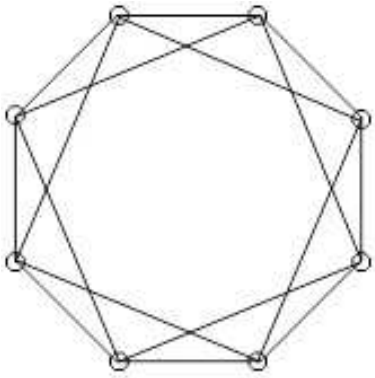

图 4.1　插图示例

公式：正文中的公式、算式、方程式等必须编排序号，序号一律用阿拉伯数字分章依序编码，如：(3-32)、(6-21)。

对于较长的公式，另起行居中横排，只可在符号处（如：+、-、*、/、<>等）转行。公式序号标注于该式所在行（当有续行时，应标注于最后一行）的最右边。连续性的公式在“=”处排列整齐。大于 999 的整数或多于三位的小数，一律用半个阿拉伯数字符的小间隔分开；小于 1 的数应将 0 置于小数点之前。公式的行距一般为单倍行距。

公式与正文之间一般应空一行。

$$\begin{aligned} X_{e1}\left(s,n_1,k_1\right) = & \left(\begin{array}{c} k_1 \\ s \end{array} \right) \frac{n_1!}{(n_1-s)!} \sum_{v=0}^{\min(n_1-s,k_1-s)} (-1)^v \left(\begin{array}{c} k_1-s \\ v \end{array} \right) \\ & \times \frac{(n_1-s)!}{(n_1-s-v)!} (n_1-s-v)^{k_1-s-v} \end{aligned} \tag{4-1}$$

表：包括分类项目和数据，一般要求分类项目由左至右横排，数据从上到下竖列。

分类项目横排中必须标明符号或单位，竖列的数据栏中不要出现“同上”、“同左”等词语，一律要填写具体的数字或文字。表序号一律用阿拉伯数字分章依序编码，如：表 2.5 、表 10.3 。

每一个表格应有简短确切的题名，连同表序号置于表的正上方。表名称、表中的内容居中排列，字号为五号，中文字体为宋体，英文字体为 Times New Roman ，行距一般与正文保持一致。表格线统一用单线条，磅值为 0.5 磅。

表格与正文之间一般应空一行。

表 4.1 表格示例

馈电方式 / 电性能参数	探针	环形缝隙	探针和缝隙		缝隙和CPW	
			探针	缝隙	缝隙	CPW
谐振频率	9.5 GHz	8.8 GHz	9.4 GHz	9.8 GHz	9.2 GHz	9.3 GHz
带宽 $\lvert S_{11}\rvert$ <-10 dB）	7.3%	4.5%	6.9%	6.8%	4.9%	5.3%
隔离度（带内最差）	-16.5 dB	-17 dB	-31 dB		-22 dB	
方向图	不对称	对称	不对称	对称	对称	对称
交叉极化电平	高	低	高	低	低	低

计量单位：学位论文中出现的计量单位一律采用国务院 1984 年 2 月 27 日发布的《中华人民共和国法定计量单位》标准。

参考文献

[1] 滴滴出行. [R]. .

[2] 中国互联网络信息中心. [R]. .

[3] CASTILLO J C, KNOEPFLE D, WEYL G. Surge pricing solves the wild goose chase[C] // Proceedings of the 2017 ACM Conference on Economics and Computation. 2017 : 241 – 242.

[4] CHEN M K, SHELDON M. Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform.[C] // Ec. 2016 : 455.

[5] QIAN X, UKKUSURI S. Time-of-Day Pricing in Taxi Markets[J/OL]. IEEE Transactions on Intelligent Transportation Systems, 2017, PP. http://dx.doi.org/10.1109/TITS.2016.2614621.

[6] BATTIFARANO M, QIAN Z S. Predicting real-time surge pricing of ride-sourcing companies[J]. Transportation Research Part C: Emerging Technologies, 2019, 107 : 444 – 462.

[7] ASGHARI M, SHAHABI C. ADAPT-pricing: a dynamic and predictive technique for pricing to maximize revenue in ridesharing platforms[C] // Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2018 : 189 – 198.

[8] SUTTON R S, BARTO A G, OTHERS. Introduction to reinforcement learning : Vol 135[M]. [S.l.] : MIT press Cambridge, 1998.

[9] RANA R, OLIVEIRA F S. Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning[J]. Omega, 2014, 47 : 116 – 126.

[10] LU R, HONG S H, ZHANG X. A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach[J]. Applied Energy, 2018, 220 : 220 – 230.

[11] PETERS M, KETTER W, SAAR-TSECHANSKY M, et al. A reinforcement learning approach to autonomous decision-making in smart electricity markets[J]. Machine learning, 2013, 92(1) : 5 – 39.

[12] SHEN W, PENG B, LIU H, et al. Reinforcement Mechanism Design: With Applications to Dynamic Pricing in Sponsored Search Auctions[C] // Proceedings of the AAAI Conference on Artificial Intelligence : Vol 34. 2020 : 2236 – 2243.

[13] WANG S, BI S, ZHANG Y J A. Reinforcement learning for real-time pricing and scheduling control in ev charging stations[J]. IEEE Transactions on Industrial Informatics, 2019.

[14] LONG J, TAN W, SZETO W, et al. Ride-sharing with travel time uncertainty[J]. Transportation Research Part B: Methodological, 2018, 118 : 143 – 171.

[15] FANG Z, HUANG L, WIERMAN A. Prices and subsidies in the sharing economy[J]. Performance Evaluation, 2019, 136 : 102037.

[16] HU L, LIU Y. Joint design of parking capacities and fleet size for one-way station-based carsharing

systems with road congestion constraints[J]. Transportation Research Part B: Methodological, 2016, 93 : 268 – 299.

[17] BANERJEE S, JOHARI R, RIQUELME C. Pricing in ride-sharing platforms: A queueing-theoretic approach[C] // Proceedings of the Sixteenth ACM Conference on Economics and Computation. 2015 : 639 – 639.

[18] ZHA L, YIN Y, DU Y. Surge pricing and labor supply in the ride-sourcing market[J]. Transportation Research Part B: Methodological, 2018, 117 : 708 – 722.

[19] REMPEL J. A review of uber, the growing alternative to traditional taxi service[J]. AFB AccessWorld® Magazine, 2014, 51(6).

[20] YAN C, ZHU H, KOROLKO N, et al. Dynamic pricing and matching in ride-hailing platforms[J]. Naval Research Logistics (NRL), 2019.

[21] 胡天宇, 张勇. 网约车平台的动态定价策略[J]. 山东科学, 2020, 33(2) : 79 – 90.

[22] YAO H, WU F, KE J, et al. Deep multi-view spatial-temporal network for taxi demand prediction[C] // Proceedings of the AAAI Conference on Artificial Intelligence : Vol 32. 2018.

[23] ZHANG X, HUANG C, XU Y, et al. Spatial-Temporal Convolutional Graph Attention Networks for Citywide Traffic Flow Forecasting[C] // Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020 : 1853 – 1862.

[24] GUDA H, SUBRAMANIAN U. Your Uber Is Arriving: Managing On-Demand Workers Through Surge Pricing, Forecast Communication, and Worker Incentives[J]. Management Science, 2019.

[25] KIM B-G, ZHANG Y, VAN DER SCHAAR M, et al. Dynamic pricing and energy consumption scheduling with reinforcement learning[J]. IEEE Transactions on Smart Grid, 2015, 7(5) : 2187 – 2198.

[26] MAESTRE R, DUQUE J, RUBIO A, et al. Reinforcement learning for fair dynamic pricing[C] // Proceedings of SAI Intelligent Systems Conference. 2018 : 120 – 135.

[27] TURAN B, PEDARSANI R, ALIZADEH M. Dynamic pricing and management for electric autonomous mobility on demand systems using reinforcement learning[J]. arXiv preprint arXiv:1909.06962, 2019.

[28] JIANG Z, XU D, LIANG J. A deep reinforcement learning framework for the financial portfolio management problem[J]. arXiv preprint arXiv:1706.10059, 2017.

[29] WEI H, ZHENG G, YAO H, et al. Intellilight: A reinforcement learning approach for intelligent traffic light control[C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018 : 2496 – 2505.

[30] NGUYEN H, LA H. Review of deep reinforcement learning for robot manipulation[C] // 2019 Third IEEE International Conference on Robotic Computing (IRC). 2019 : 590 – 595.

[31] KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: A survey[J]. Journal of artificial intelligence research, 1996, 4 : 237 – 285.

[32] AUER P, CESA-BIANCHI N, FISCHER P. Finite-time analysis of the multiarmed bandit problem[J]. Machine learning, 2002, 47(2-3) : 235 – 256.

[33] FORTUNATO M, AZAR M G, PIOT B, et al. Noisy networks for exploration[J]. arXiv preprint arXiv:1706.10295, 2017.

[34] AHMED Z, LE ROUX N, NOROUZI M, et al. Understanding the impact of entropy on policy optimization[C] // International Conference on Machine Learning. 2019 : 151 – 160.

[35] GUPTA J K, EGOROV M, KOCHENDERFER M. Cooperative multi-agent control using deep reinforcement learning[C] // International Conference on Autonomous Agents and Multiagent Systems. 2017 : 66 – 83.

[36] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C] // Advances in neural information processing systems. 2017 : 6379 – 6390.

[37] CHEN Y, HU M. Pricing and matching with forward-looking buyers and sellers[J]. Rotman School of Management Working Paper, 2018(2859864).

[38] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[J]. arXiv preprint arXiv:1801.01290, 2018.

[39] BIMPIKIS K, CANDOGAN O, SABAN D. Spatial pricing in ride-sharing networks[J]. Operations Research, 2019.

[40] L.S.SHAPLEY. ”Stochastic Games”[J]. Proceedings of the National Academy of Sciences of the United States of America, 1953, 39 : 1095 – 1100.

[41] TAMPUU A, MATIISEN T, KODELJA D, et al. Multiagent cooperation and competition with deep reinforcement learning[J]. PloS one, 2017, 12(4) : e0172395.

[42] TAN M. Multi-agent reinforcement learning: Independent vs. cooperative agents[C] // Proceedings of the tenth international conference on machine learning. 1993 : 330 – 337.

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

致谢

本论文是在导师的悉心指导下完成的，从论文的选题到论文的撰写，无不渗透着导师的心血，……值此论文完稿之际，谨对导师的辛勤培育以及谆谆教诲表示最衷心的感谢!

作者简介

1. 基本情况

张三，男，陕西西安人， 1982 年 8 月出生，西安电子科技大学 XX 学院 XX 专业 2008 级硕士研究生。

2. 教育背景

2001.08～ 2005.07，西安电子科技大学，本科，专业：电子信息工程

2008.08～ ，西安电子科技大学，硕士研究生，专业：电磁场与微波技术

3. 攻读硕士学位期间的研究成果

3.1 发表学术论文

[1] XXX, XXX, XXX. Rapid development technique for drip irrigation emitters[J].RP Journal,UK.,2003,9(2): 104-110.(SCI: 672CZ, EI: 03187452127)

[2] XXX, XXX, XXX. 基于快速成型制造的滴管快速制造技术研究[J]. 西安交通大学学报, 2001, 15(9): 935-939. (EI: 02226959521)

[3] . . .

3.2 申请（授权）专利

[1] XXX, XXX, XXX等. 专利名称: 国别,专利号[P]. 出版日期.

[2] . . .

3.3 参与科研项目及获奖

[1] XXX项目, 项目名称, 起止时间, 完成情况, 作者贡献。

[2] XXX, XXX, XXX等. 科研项目名称. 陕西省科技进步三等奖, 获奖日期.

[3] . . .