电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

专业学位硕士学位论文

MASTER THESIS FOR PROFESSIONAL DEGREE

论文题目 基于多智能体博弈的电动汽车充电市场多方策略研究

专业学位类别	**工程硕士**
学　　号	**201922080638**
作 者 姓 名	**张瑞昌**
指 导 老 师	**张彦如　教授**
学　　院	**计算机科学与工程学院 (网络空间安全学院)**

分类号 TP309.2 密级 公开

UDC[注1] 004.8

学 位 论 文

基于多智能体博弈的电动汽车充电市场多方策略研究

（题名和副题名）

张瑞昌

（作者姓名）

指导老师 **张彦如 教授**

电子科技大学 成都

（姓名、职称、单位名称）

申请学位级别 **硕士** 专业学位类别 **工程硕士**

专业学位领域 **计算机技术**

提交论文日期 **2022 年 5 月 20 日** 论文答辩日期 **2022 年 5 月 31 日**

学位授予单位和日期 **电子科技大学 2022 年 6 月**

答辩委员会主席 **李建平**

评阅人

注 1：注明《国际十进分类法 UDC》的类号。

Multilateral Decision Strategy of Electric Vehicle Charging Market based on Multi-Agent Game

A Master Thesis for Professional Degree Submitted to
University of Electronic Science and Technology of China

Discipline:	**Master of Engineering**
Student ID:	**201922080638**
Author:	**Ruichang Zhang**
Supervisor:	**Prof. Yanru Zhang**
School:	**School of Computer Science and Engineering(School of Cyberspace Security)**

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，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谢意。

作者签名：_____ 日期： 年 月 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：_____ 导师签名：_____

日期： 年 月 日

摘 要

电动汽车数量的快速增长给电动汽车充电市场带来困难与挑战。其中，电动汽车在并网时带来的额外需求会对电网造成很大的冲击。在现有供电策略下，电动汽车充电市场的电力供应量与电动汽车用户的电力需求不相匹配，这可能会导致充电市场各方利润损失和社会总福利损失。如何设计电动汽车充电汽车市场参与各方的策略使得各方利益最大化以及社会总福利最大化是智能电网的重要研究内容。传统的智能电网优化方案往往只包含供给侧与需求侧的单边模型而忽略了市场上的其他参与者，使得优化结果出现折损。基于此，本文基于历史数据搭建三方电动汽车充电市场模型，并求解规定目标下的各方最优策略以实现各方利益最大化的目标。

本文首先搭建三方充电市场模型，该模型主要包含充电汽车用户、集电商和电网三方参与主体。每种主体根据其对市场的观测结果，基于各自的收益制订策略并反馈给市场。其中，电动汽车充电用户在市场中观察集电商提供的售电价格，基于自身效用最大化原则选择自己的需求。电网在市场中观测用户的需求曲线，并且基于自身的收益及运营成本制订供给数量以及批发电价。集电商同时观测市场中所有信息，基于自身的运营效益制订零售电力价格。其次，本文提出两种假设目标以指导求解市场参与各方最优策略，并对求解结果的单方收益、社会福利、市场运行状况等因素做比较。其中，假设一的目标在单方总利润最大化的目标下各方参与者的决策，假设二的目标是在社会福利最大化的目标下参与各方的决策。

求解多目标优化的传统算法虽然求解速度快，但是很难应对即时的因素变动。针对充电市场多方决策问题，强化学习以及多智能体博弈是合适的求解工具。本文基于博弈论及强化学习的方案求解上述优化问题并提供模拟环境下的量化分析结果。

关键词：博弈论，多智能体强化学习，智能电网

ABSTRACT

The rapid growth in the number of electric vehicles(EVs) poses difficulties and challenges for the electric vehicle charging market. Besides, the additional demand brought by EVs when they are connected to the grid can cause a significant impact. Under the existing power supply strategy, the power supply in the EV charging market does not match the power demand of EV users, which may lead to loss of profit and total social welfare of all parties in the charging market. How to design a strategy to maximize the profit of all parties involved in the EV charging market and to maximize the total welfare of the society is an important part of the smart grid research. Traditional smart grid optimization schemes often consider only the unilateral model of supply and demand side but ignore the other participants in the market, resulting in compromised optimization results. Based on this, this thesis constructs a three-party EV charging market model based on historical data and solves the optimal strategy for each party under the specified objective to maximize the benefits of each party.

In this thesis, we first build a three-party charging market model, which consists of the following main participants: charging vehicle users, aggregators and the grid. Each participant makes strategy based on its own revenue and feeds back to the market based on its observation of the market. In the market, EV charging customers observe the price of electricity offered by the collector and choose their own demand based on their own utility maximization principle. The grid observes the demand curve of customers in the market and sets the supply quantity and wholesale tariff based on its own revenue and operating cost. The collector observes all the information in the market and sets the retail price based on its own operational efficiency. Second, this thesis proposes two hypothetical objectives to solve the optimal strategy for each party involved in the market, and compares the solution results in terms of single-party benefits, social welfare, and market operation conditions. Among them, hypothesis one aims at the decision of each participant under the objective of maximizing the total profit of a single party, and hypothesis two aims at the decision of each participant under the objective of maximizing social welfare.

Traditional algorithms for solving multi-objective optimization are fast but difficult to cope with immediate factor changes. For the multi-party decision problem in charging market, deep reinforcement learning and multi-intelligent games can solve this problem.

At the same time, multi-intelligent reinforcement learning can effectively deal with the multi-party game situation. welfare. This work will be validated under a virtual environment, the benefit and cost of each party under different assumptions will be quantified.

Keywords: Game Theory, Multi-Agent Reinforcement Learning, Smart Grid

目 录

第一章 绪 论

1.1 研究工作的背景与意义

电动汽车由于其经济环保的运行方式被认为是未来交通的一个主要方向[1]。截至 2019 年底，全世界有超过 100 个国家开始推广充电式电动汽车，全球电动汽车的市场占有量已经超过 700 万台，市场占有率约 3%。截至 2020 年底，全球电动乘用车的累计销量已经达到 1000 万辆，市场占有率约 4.8%。根据国际清洁交通委员会的预测，截止 2030 年，预期电动汽车销量将达到 2000 万辆。截止 2040 年，电动汽车市场的占有率将达到约 40%。随着大量的电动汽车涌入市场，电力市场的各方参与者将会面临巨大的挑战。

电力市场的用户总体上可以分为工业用电用户和家用用户两类。其中，工业用户的用电需求量较为稳定，且占有电力消费的较大份额。家用用户的电量需求相对比较灵活，占电力消费的较小份额。电动汽车用户作为一种特殊的购电用户，其购电需求大致上比较稳定，但是随着时段和价格的调整，其用电需求量在一定时间内也会有一定的改变。电动汽车在电力消费中占有的份额随着电动汽车市场的发展正在逐渐增加。电动汽车引入的能源需求会对电网造成很大的冲击[2]。在没有充电控制保护系统的情况下，超量电动汽车的自动充电可能会造成电网拥堵，引发一系列安全问题。举例来说，一台符合交流 2 级充电标准[3] 的 19.2kW 功率电动汽车直接并入电网时，其带来的充电负载相当于一个典型家庭用电功率[4] 的 20 倍。由于电动汽车充电需求与电网供电不平衡，当单位时间内的充电需求聚集而不协调时，电力负载带来的影响会更加严重，会对电网造成功率损耗增大、电压偏差增大等问题，影响电网运行的平稳性及运行损耗。与此同时，对于电力聚合商，现阶段电动汽车与充电站点的发展差异很大，其运营状况不佳，盈利困难。中国的新能源汽车发展市场规模扩大较快，但同时没有足够的充电桩数量满足用户的充电需求，导致众多充电企业仍没有实现盈利。对于电动汽车用户，现阶段其充电需求与电力聚合商的供应不匹配，往往会造成不友好的消费体验，例如高峰期无法匹配到合适的充电站，充电价格过高造成用户自身消费效用降低等。因此，在一个供需不平衡的市场下，如何科学设计电动汽车充电市场各方参与者的应对策略，使得自身收益最大化或是社会福利损失最小化是一个亟待解决的问题。

针对市场供需不平衡的调节问题，英国经济学家亚当·斯密在《国富论》[5] 中提出了“看不见的手”(the invisible hand) 的概念。在一个完全竞争的市场中，如果参与各方完全遵循市场机制做出决策，消费者遵循效用最大化原则选择自己的

需求曲线，生产者遵循利润最大化原则选择自己的供给曲线，则市场会自然收敛到出清状态。然而，在电动汽车充电市场的背景下，仅仅依靠市场机制并不能有效达到市场的纳什均衡点。第一，市场信息并非完全公开透明且具有时效性，参与各方并不能接收到完备的信息。第二，市场参与方对市场前景的判断并非完全理性，供给方会出于对基础设施投入的考量影响其供给价格和数量，导致市场供给曲线并不能真正反映出供给方的决策。第三，由于供需并未到达市场出清状态，市场机制调节的长期过程中会引起许多社会福利损耗，导致各方利益受损。基于此，一定程度下的市场干预，即合理设计电动汽车充电市场参与各方策略，对于市场的平稳运行是必要的[6]。

随着智能电网的建设，商业交易，信息流动和电力流动的不确定性和复杂性都在增加[7]。电力交易系统中包含了双向的能量流动，并伴随着生产者、消费者、输配电系统运营商和需求响应聚集者之间的信息流动[8]。这些因素从不同方面给电力系统带来了诸多问题与挑战。同时，随着智能设备的并网速度增加，信息量的指数级增长和数据的波动使得决策问题更难以用传统方法解决。因此，未来的智能电网需要一个能够实时监测、预测、安排、学习并做出生产决策的系统，这需要更高效和智能的解决方案，如深度强化学习。

强化学习是机器学习的一个分支领域[9]。其基本思想是使智能体与环境交互产生奖励，以最大化累计奖励值为目标对智能体执行的动作做序列化策略。因为强化学习机制的特点，其与其他机器学习分支领域的区别主要体现在其决策能力上。近年来，强化学习在许多领域上都有广泛的应用，例如推荐系统[10]、游戏[11]、智慧医疗[12]、智能电网[13]、智慧交通[14]等。对于电动汽车充电市场的三方决策模型，强化学习因为其互动奖励机制可以作为良好的求解方法。深度强化学习是一种数据驱动的方法，它是深度学习和强化学习的结合。这一研究领域已被应用于解决广泛的复杂的顺序决策问题，包括电力系统中的问题[15]。

1.2 国内外研究历史与现状

1.2.1 电力市场建模的研究

如图1-1所示，传统视角下，一个分层的电力市场可以分为批发电力市场（wholesale electricity market）和零售电力市场（retail electricity market）两部分。电力集成商在批发市场向电网运营商购电，在零售市场以零售价格向普通消费者售卖电力。层级模型通过信息流动和能量流动将服务提供商与电力公司和消费者结合起来。处于供给侧的将其电力提供给零售商，而零售商再将电力卖给服务提供商时，就存在着电力批发市场。同时，处于需求侧的消费者从存在相互竞争关系的

电力零售商中选择他们的供应商时，就存在零售电力市场。这些要素之间的平衡是一个复杂的博弈问题，而深度强化学习可以用来获得不完全信息下的最优策略。

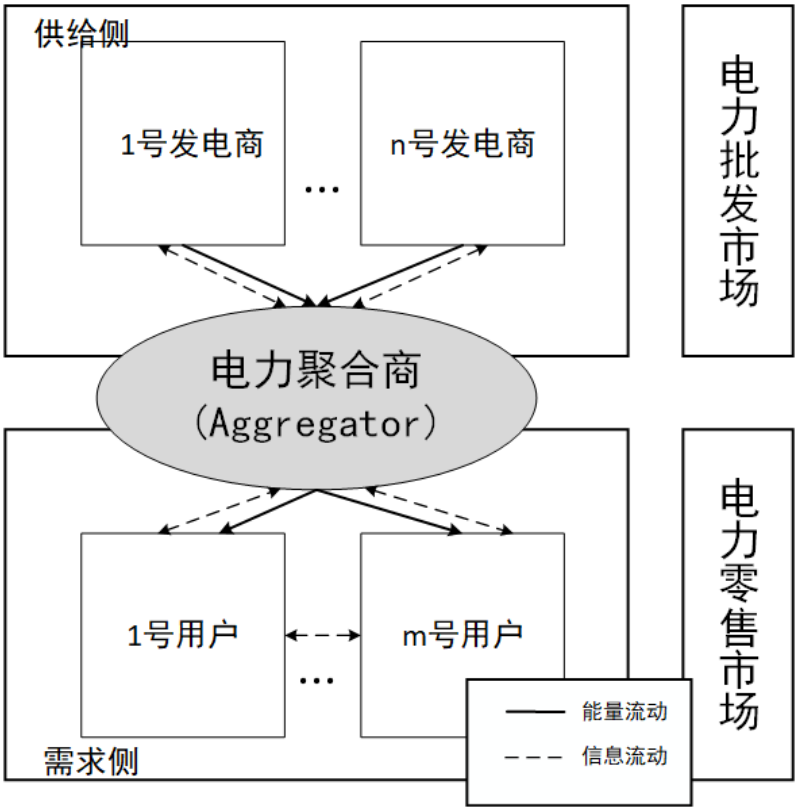

图 1-1 层级电力市场模型

目前学术界对于电动汽车充电定价方案的文献研究主要分为智能充电和动态定价两个主要方向。智能充电 [16]（Smart Charging, SC）是一种协调控制充电的手段，也是电动汽车并入智能电网和保证电力聚合商盈利运营的重要步骤。智能充电技术可以用来提高电网稳定性[17]、增加充电站利润[18]。同时，个体用户也可以利用智能充电技术减少其充电的损耗[19,20]。动态定价（Dynamic Pricing, DP）是指电力销售商动态调整终端用户即电动汽车充电服务的消费者面对的价格，以对电网运行条件的变化做出反应。例如，在高电价或高能源生产成本时期，分别提高充电价格。除此之外，动态定价可以增加用户对于市场反应的灵活性，因此有可能实现增加电网运行稳定性，提高用户满意度或降低电力聚合商的运营成本。国际可再生能源机构（International Renewable Energy Agency, IREA）认为智能充电和动态定价两个用户激励措施是释放电动汽车灵活性潜力的两个关键因素，也是未来大量电动汽车成功并网的必要条件[21]。

对于整个电力市场的参与者来说，一个合理的竞价策略可以给他们带来更多的利益和更低的成本。文献 [22] 设计了一个事件驱动的电力市场，用于本地配电

网中的能源交易。作者使用深度强化学习决定消费者的交易策略，以实现其利益最大化。与文献 [22] 类似，文献 [23] 研究了配电网中消费者对消费者的间接电力市场，设定用户之间可以进行二次能源交易，并使用强化学习来确定能源交易策略使各方利益最大化。文献 [24] 通过不完全信息的自适应强化学习算法提出了终端消费者之间的约束性能源交易博弈，最终，竞价策略收敛到纳什均衡状态。文献 [25] 研究了微电网中动态定价和能源消耗的电力市场模型，并且在该模型中应用强化学习来降低服务提供商的系统成本。

对于整个电力交易系统来说，纳什均衡和社会福利是博弈论的目标。文献 [24] 建立了能源交易的多领导者和多追随者 Stackelberg 博弈模型，并使用强化学习算法获得满足隐私政策约束下的均衡状态。文献 [26] 提出了一个考虑可再生能源发电和需求、用户电子设备电池电容量和交易历史的微电网能源交易博弈模型，并通过深度强化学习求解纳什均衡。此外，文献 [27] 提出了一个连续可再生能源渗透的电力市场模型，并通过模糊 Q-learning 实现了 IEEE-30 总线模型的测试系统。文献 [28] 研究了一个带有负荷服务实体投标和定价的分层电力市场，通过深度神经网络学习动态投标和价格响应函数，通过深度确定性策略梯度（Deep Deterministic Policy Gradient, DDPG）算法生成状态转换样本。

1.2.2 强化学习在智能电网中供需关系的研究

研究 [29, 30] 表明，智能电网的的优势在于它能够提高电力市场运行可靠性和需求侧的反应速度，并促使需求侧和公用事业提供者做出更高效的决策。因此，需求侧管理（Demand Side Management, DSM）是智能电网的一个重要组成部分 [31,32]。DSM 的完全整合需要通信系统、传感器、自动仪表、智能设备和专业处理器等手段和设备。利用智能计量和先进信息通信技术 (Information and Communications Technology,ICT) 的解决方案在能源管理中有利于节约能源和开发利用可再生能源 (Renewable Energy Sources, RES)。近年来，随着 ICT 技术的发展迭代，ICT 基础设施已经可以支持更有效的网络运行和更频繁的通信频率。这也为 DSM 带来了新的冲击，其需要更加动态的、即时的定价机制，同时需要考虑到波动 RES 的实时可用性 [33]，并实时跟踪供需平衡的变动。除此以外，DSM 也有利于消费者积极参与参与能源市场的市场调节机制。DSM 项目通常是由公用事业公司实施、在用户端管理能源消耗的项目 [34]，这些项目可以帮助电力市场以更有效的方式运行 [35]。公用事业公司和客户都可以从 DSM 项目中受益，从而在整个电力交易市场中削减峰值电力需求，平稳电力价格的波动 [36]。目前，专家和学者已经研究了大量的需求响应项目，业界也已经开始将需求响应纳入到智能电网的

建设之中，以更有效地利用现有能源，鼓励需求侧响应和激励电力零售商。

需求响应（Demand Response, DR）是智能电网的一个典型问题。它通过价格或激励措施保持客户的电力需求和公用事业公司的供应之间的平衡。为了提高电网的稳定性和转移峰值电力需求，DR 需要将消费者的需求曲线纳入控制范围。DR 项目是一类通过促进需求端的互动和响应，为电力系统运行和扩展以及市场效率提供广泛潜在利益的项目。这些项目包括保护和能源效率项目，燃料替代项目，以及住宅或商业负荷管理项目[37]。通过提高电力系统的鲁棒性，降低峰值需求，在长期视角下 DR 可以减少整个电网初期的资本成本投资和运营成本。

对于需求响应落地中带来的问题与挑战，深度强化学习是一种有效的解决方法，它利用数据驱动的模型来解决此类问题[38,39]。对于电力市场的消费者来说，最小化支出是其首要目标，而对于电力公司来说，其目标是利润最大化和保证电网平稳运行。为了解决基于实时激励的需求响应问题，文献 [40] 提出了一种深度强化学习方法，通过协助电力集成商从消费者方购买电力来平衡电力波动并保持电网的鲁棒性。文献 [41] 提出了一种深度强化学习方法，以解决需求响应计划下对供暖、通风和空调（Heating,Ventilation,Air-conditioning and Cooling，HVAC）系统的顺序优化决策问题。文献 [41] 基于博弈论方法，利用多智能体强化学习规划了一个自主和最优的 HVAC 用电调度策略，以最小化社会成本，达成社会福利最大化的目标。文献 [42] 提出了一个基于强化学习的需求响应计划最优定价方案，在学习过程中通过调节强化学习探索和利用的平衡，达到了负载服务实体的最优表现。文献 [43] 提出了一个住宅负荷调度的最优模型，考虑了消费者满意度、可再生能源和调度成本因素，并由强化学习算法求解纳什均衡。文献 [44] 提出了一种需求响应方法，通过强化学习算法和贝叶斯神经网络降低插电式电动汽车的长期充放电成本。

综上所述，深度强化学习在解决智能电网供需关系上具有以下优势：

（1）深度强化学习算法可以在不完全信息的基础上进行决策，而且决策可以在线实时进行；

（2）基于博弈论原理，深度强化学习可以求解系统利益最大化策略，降低交易成本；

（3）深度强化学习具有很强的迁移能力，可以应用于许多不同的场景。

同时，需求响应的挑战主要体现在以下几个方面：

（1）激励措施形式多样，包括经济激励、技术激励、环境激励、政策激励等，不同用户对激励措施的反应不同；

（2）需求响应通常伴随着负荷、电价等多种因素的变化，不同因素对结果的

影响程度不同；

（3）参与需求响应计划电气设备的控制方法和约束条件不同，从而使模型更加复杂；

（4）需求响应的过程往往伴随着消费者、电力集成商和电力公司之间的博弈过程，因此优化目标不同。

为了克服上述问题，深度强化学习方法可以在以下几点上进行探索：

（1）利用深度神经网络等方法提取消费者的行为特征并预测行为，作为优化控制的基础；

（2）选择合适的状态空间，包括电力价格、电网负荷等；

（3）充分利用历史数据和需求侧的反馈来弥补模型的不足。

1.2.3 解决电力交易系统问题中传统方法与深度强化学习方法的比较

由于电力交易系统中参与者的复杂性、要素不确定性和市场数据维度的增长性，传统方法在试图解决电力市场中决策和控制问题时经常遇到瓶颈。因此，学术界正在把目光投向研究解决此类问题的数据驱动方法。

电力交易系统中的许多问题可以被转化为顺序的决策任务。传统方法主要包括凸优化算法、规划算法和启发式算法。通过与深度强化学习算法的定性比较，这些方法的优点和缺点[45]被阐述如下:

（1）经典的数学方法，如 Lyapunov 优化算法 [46]。这些方法的优点是其拥有严格的数学模型，可以实现实时管理决策。然而，也因为这类方法依赖于明确的目标函数表达式，应用中很难从复杂现实世界的优化决策场景中抽象出来。此外，Lyapunov 优化算法所需的假设条件在复杂的高维数据场景中无法保证。

（2）规划方法，如混合整型规划 [47,48]，动态规划 [49,50]，和随机规划 [51,52]。这些方法可以解决包括序列优化在内的各种优化问题。然而，这种方法的每一次迭代都需要从初始状态开始重新计算，由于其计算成本太高，在某些情况下无法实现实时决策。此外，一些基于规划算法的场景依赖于对可再生能源发电和负荷的准确预测，这在实际场景中很难实现。

（3）启发式方法，如遗传算法（Genetic Algorithm, GA）[53]，蚁群优化算法（Ant Clony Optimization，ACO）[54]，粒子群优化算法（Particle Swarm Optimization, PSO）[55, 56] 。对于优化问题，尤其是非凸优化问题，启发式方法能够以一定的概率获得局部最优解，有利于解决大数据规模和复杂场景下的问题。但是，这种方法的鲁棒性较差，无法在数学上进行严格的证明。

与凸优化方法相比，深度强化学习并不需要对优化目标的准确表达式，因为

其并不依靠函数化的目标做迭代，而是依靠环境给出的奖励更新策略网络。此外，在高维度数据的处理上，深度强化学习由于利用深度神经网络对大规模数据的高效处理能力，可以处理更高维度的数据。与规划方法相比，深度强化学习根据当前的状态进行决策，因此可以进行实时和在线决策。与启发式方法相比，深度强化学习更加稳健，收敛结果稳定，更适合于决策问题。

总而言之，与传统方法相比，深度强化学习具有以下优势：

（1）神经网络作为函数近似器可以提取更多数值模型未考虑的数据特征；

（2）大多数深度强化学习算法不依赖于具体的模型，适用于无法制定模型的情况；

（3）电力市场中供需双方可以通过深度强化学习实现纳什均衡。

另一方面，电力市场有以下主要困难：

（1）分层的电力市场中存在多个实体，他们的目标不同，使得整体奖励函数难以定义和收敛；

（2）除了能源流动之外，实体之间还存在不完全的信息流动，因此需要数据驱动的方法解决；

（3）能源交易是一个序列决策问题，与典型的离散决策问题不同，电力市场需要实时决策。

针对这些问题，以下几个方面应该是深度强化学习对于解决电力交易市场供需问题的主要研究方向：

（1）构造博弈论模型，将不同的市场主体建模为不同的博弈主体；

（2）采用多智能体强化学习算法，对不同博弈主体对应建模为不同的智能体；

（3）由于博弈过程的复杂性，研究应从小规模的场景开始，逐步扩大场景的规模；

（4）提高模型对价格、电力等信息的整合和提取能力。

1.3 本文的主要贡献与创新

本文搭建电动汽车充电市场三方模型，基于深度强化学习和多智能体博弈算法探索不同优化目标下各方决策，实现市场均衡并做对比，探究不同优化目标下的各方收益及社会福利，并作对比实验探究市场最优的优化目标。本文主要贡献如下：

（1）本文搭建了基于多智能体环境 PettingZoo [57] 的一个电动汽车充电市场三方博弈模型。该模型根据市场参与者的特点对三方参与者，即电网、电力集成商和充电汽车用户建模，分别规定其对市场的观测空间，决策空间及决策收益，将市

场状态视为公共的资源池，受到各方决策的影响改变。每一个参与者只能从自己的视角中观察有限的信息，并做出对自己收益最大化的决策。其决策会影响市场状态，进一步影响其他参与者的决策。本文模型并不假定参与者的决策顺序，任何参与者都可以根据当下市场状态做出行动。

（2）本文设定了两个优化目标：社会总福利最大化和各方收益最大化。针对这两个优化目标，本文选择并设计多智能体强化学习算法求解三方最优策略并并达到市场均衡状态。

（3）本文对基于相同算法对于两个优化目标得出的市场均衡状态及达到均衡状态过程中特征的变化进行比较。具体来说，本文对均衡状态下社会总福利、电网收益、用户收益及电力聚合商收益进行比对。同时，本文通过调整模型参数，得到模型参数设定对于均衡状态的影响。其次，本文应用微观经济学供需均衡原理分析论证结论，并且探讨了本文研究在实际生产生活中的应用方案。

本文主要创新点如下：

（1）现存研究电力市场优化策略方案的工作大多从需求响应计划、需求侧改革等方面入手，主要着眼于“削峰填谷”（peak shift）的问题，通过讨论动态定价（dynamic pricing）等方法优化供给侧和需求侧的策略。然而，这些研究规定了市场的封闭性，是在讨论现有存量资源的分配问题，其本质上是零和博弈，一方的收益增加必须以另一方的收益降低为代价直到收敛至市场均衡。本文采用逆向思维，将电网的输配电成本考虑在市场模型之中，优化算法可以从源头出发降低发电成本，进一步影响市场电价，带来资源总量的扩张，跳出了封闭市场的假设，使共赢成为可能。

（2）现有研究电力交易市场供需问题的工作大多假设市场中供给侧与需求侧的相互博弈，缺乏系统性的考虑参加需求响应系统三方的博弈。本工作设计了一个三方博弈框架，将各方利益考虑在内，也考虑了如何使社会福利最大化的问题。同时，本文模型具有良好的扩展性，可以将问题延伸到多个同类主体体或多个不同主体参与者的市场模型中。

（3）现有研究电力市场均衡问题的工作大多从序贯博弈的角度出发，假定智能体之间的博弈存在次序，然而，在这种博弈中，先做出行动的一方具有先动优势，其余的参与者只能在首先行动的智能体之后做出决策，这限制了他们的决策空间，最终的策略可能并非全局最优而是局部最优解。本工作通过设计公用的市场状态资源池避免了智能体博弈的次序问题，决策主体并不受制于其他主体策略带来的决策空间缩减，而是可以以任意顺序进行博弈，规避了次优解的问题。

1.4 本论文的结构安排

本文的章节结构安排如下：

第二章搭建了基于多智能体环境的电动汽车充电市场的三方博弈模型，电网、电动汽车用户和电力集成商作为参与博弈的三方被建模成三个智能体，拥有不同的目标函数。同时，为了符合现实情况，本文规定每个智能体的观测都是有限的，并不能完全了解全部市场信息。各个主体的定义符合强化学习基本定义，即基于当前观察，依据自身策略做出行动，获得环境给予的奖励，并依据奖励更新自身策略。

第三章介绍了求解的基础算法和改进，求解社会总福利最大化目标下的各方策略及市场均衡点，并对各方收益、市场状态做比较分析。

第四章求解各方收益最大化目标下的各方策略及市场均衡点，与第三章结论做对比，并通过微观经济学原理对结果进行分析。

第五章总结了基于多智能体博弈求解电力交易市场各方策略的过程和结果，并展望电力市场运行机制建模和求解的研究方向。

第二章 电动汽车充电市场三方模型

2.1 电动汽车充电市场建模

传统的分层电力交易市场采用了主从博弈模型（Stackelberg game），将电力生产商即电网作为市场的领导者（Leader），将电力聚合商或终端用户设计为追随者（Follower）。其假设领导者已知追随者的所有决策空间，在考量所有追随者的策略之后做出权衡，选择自己的最优策略。追随者获悉领导者的决策之后，也选择自己的最优策略，达到市场均衡状态。

主从博弈模型在分配好参与方的角色以后分配决策顺序，由于模型设定，领导者决策在先，具有“先动优势”，能够准确获悉追随者后续的动作，保证自己获得最大利益。追随者只能在领导者做出决策以后，在有限的决策空间内选择自己满足自己当下最大利益而非全局最大利益的决策。

然而，主从博弈模型存在以下几点不足：

（1）主从博弈模型规定了博弈的次序，然而，在电力交易市场中，交易时间内每一个时间点都存在大笔交易量，并不符合博弈模型假设的先后决策顺序，而是以随机的顺序做出决策，并不完全满足模型假设情况；

（2）主从博弈模型对于参与博弈主体较少的情况下可以高效地求解市场均衡状态，但其迁移性与扩展性不足。智能电网正向着多主体、多设备、多线路的方向发展，电力交易市场参与者也不断增加，给电力交易市场带来了诸多不确定因素，主从博弈模型无法解决大规模且要素繁多的市场情况；

（3）主从博弈模型对于市场信息的整合能力稍显欠缺。电力市场中存在大规模的历史交易数据，而博弈模型无法高效整合利用这些历史数据，只能从模型求解出最优状态。与此同时；

（4）由于主从博弈是顺序博弈，现实情况下领导者无法准确获悉大规模追随者群体的所有决策空间，因此可能做出次优策略，影响收益，且顺序博弈规定追随者只能在领导者收益最大化的情况下做决策，实际导致追随者的决策空间缩小，错过其最优策略空间。

基于以上几点因素，本文设计了一个多智能体电力市场交易模型。与分层电力交易市场不同，本文模型将市场状态考虑成独立的公告资源池，参与各方通过自己的决策影响市场状态，从而影响其他各方的决策。参与者并不直接预设其他参与者的策略，而是根据自身行动影响市场状态后得到回报，基于回报优化自身策略，最终达到市场均衡状态。本文提出的模型考虑了市场中各方的互动，每方

基于各自对于市场状态的局部观测独立做出决策，如图2-1所示，PEV User 表示充电汽车用户方，Grid 表示智能电网方，Aggregator 表示电力聚合商方。中间的 Market State 表示当前市场上的信息状态。黄色箭头表示主体向环境做出的动作，如电动汽车用户制订需求曲线，聚合商制订零售价格。红色箭头表示主体可观测到的市场信息，蓝色箭头表示主体与市场状态交互带来的奖励。

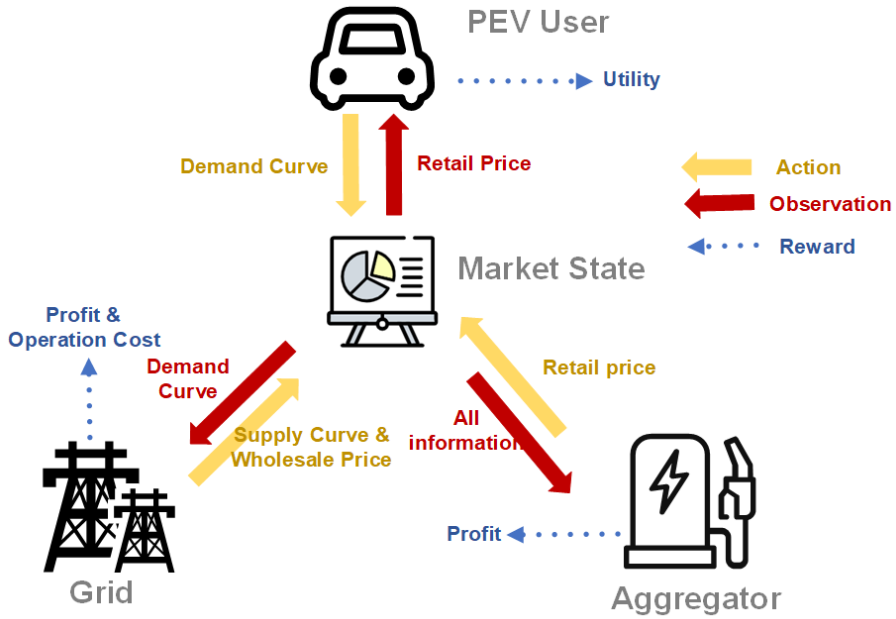

图 2-1 电动汽车充电市场三方模型

本文模型考虑如下三方参与主体：电网、电力集成商和充电汽车用户。市场状态（Market State）包括了市场上所能观察到的所有信息，包括供给曲线 ***SC***（Supply Curve）、需求曲线 ***DC***（Demand Curve）、电力批发价价格 ***WP***（Wholesale Price）和电力零售价格 ***RP***（Retail Price）。根据实际，并非每一方都能观测到市场上的所有信息。充电汽车用户观测感知到电力零售价格以后决定是否做出购电行动，产生新的需求曲线，同时获得自己的效用。电网观测到市场需求曲线之后，制订新的供给曲线及电力批发价格，同时获得自身的运营损失及收入。电力集成商可以观测到市场中所有信息，以批发价购入电力并以零售价格卖出。电力集成商可以制订零售价格，其回报为自身收益。

本文模型将每一天的交易划分为 24 个时间段 (time slot)，表示为数组 $(x_0, x_2, \ldots, x_{23})$，每一个时间段 x_n 代表一小时内的信息。市场状态每步更新一次，直到收敛状态。

本文模型具有以下几点特性：

（1）本文模型并不提前预设博弈次序，所有参与者依照自身对于市场的判，

而非直接基于其他主体可能做出的策略做出选择并影响市场。参与博弈主体之间不存在直接的交流，而是反馈至市场状态；

（2）本文模型将市场状态视为公共资源池，允许新的参与者加入市场博弈，具有良好的拓展性；

（3）本文模型对于参与博弈三方的定义并非具体到实际的个体，而是按照每一方决策者自身的特点和性质描述了其参与博弈的过程，可以视作一类相同性质实体的集合。例如，本文定义的电动汽车充电用户作为博弈的参与对象，并不指代某一位具体的电动汽车充电客户，而是代表效用函数相近的一类用户的集合。

2.2 充电汽车用户建模

电动汽车充电用户在时间步 t 可以获取市场在前一天的零售价格信息 $\boldsymbol{RP}_{t-1} = (RP_{t-1,x_0}, RP_{t-1,x_0}, ..., RP_{t-1,x_{23}})$, 从而做出当天的购买行为，形成新的需求曲线 $\boldsymbol{DC}_t = (DC_{t,x_0}, DC_{t,x_1}, ..., DC_{t,x_{23}})$

电动汽车充电用户在时间步 t 从参与市场获得的回报 R_t^{user} 可以定义为:

$$R_t^{user}(\boldsymbol{DC}_t, \boldsymbol{RP}_t) = U_t^{user}(\boldsymbol{DC}_t, \boldsymbol{RP}_t) - C_t^{user}(\boldsymbol{DC}_t, \boldsymbol{RP}_t) - S_t^{user}(\boldsymbol{DC}_t, \boldsymbol{RP}_t) \tag{2-1}$$

其中，U_t^{user} 代表用户由充电交易带来的满意度，由需求曲线决定；C_t^{user} 代表电力费用的支出，S_t^{user} 代表用户因需求变动引起的充电习惯变化。

本文模型假设满意度 U_t^{user} 遵循边际效益递减规律，如图2-2，Q 代表用户的消费量，MU 代表商品的边际效益 (Margin Utility)，每一单位带给用户的效用递减。在达到用户需求量后边际效益为负，TU 代表用户总体效用 (Total Utility)，在达到需求量之前，用户总体效用递增，在达到需求量之后总体效用减少。

本文以充电量（State of Charge,SoC）表示用户的消费水平。由于电动汽车充电用户电池容量有限，不会出现 SoC 达到需求量后继续消费的情况，在时间步 t, 时刻 x_n 用户的充电量定义为:

$$SoC_{t,x_n} = \frac{DC_{t,x_n}}{DC_{0,x_n}} \tag{2-2}$$

用户满意度由 SoC 与当前零售价格决定，即 $U_{t,x_n}^{user} = U(SoC_{t,x_n}, RP_{t,x_n})$。其中，$U(\cdot)$ 是用户对于充电服务的评价函数，为非凸函数。本文模型定义时间段内效用函数为:

$$U(SoC_{t,x_n}, RP_{t,x_n}) = RP_{t,x_n} \cdot DC_{0,x_n} \cdot (0.75 + lg(SoC_{t,x_n} + 1)) \tag{2-3}$$

一个时间步内，用户效益可以表示为该时间步内所有时间段效益的总和，即:$U_t^{user} =$

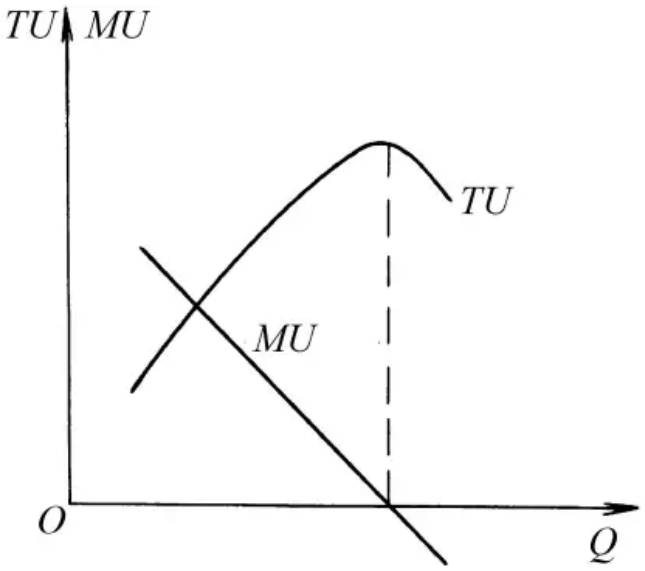

图 2-2 边际效益曲线与总效用曲线

$\sum_{n=0}^{23} U_{t,x_n}^{user}$。

时间步 t 内，电力费用的支出 C_t^{user} 根据实际充电消费电量与该时间的价格决定，时间段内的电费支出表示为：

$$\begin{aligned} C_t^{user} &= \sum_{n=0}^{23} C_{t,x_n}^{user} \\ &= \sum_{n=0}^{23} RP_{t,x_n} \cdot DC_{t,x_n} \end{aligned} \tag{2-4}$$

电动汽车充电用户做出新的需求曲线之后，需求变动会扰乱原有出行习惯的安排，带来负效用 S_t^{user}。在一天中不同的时间段，消费习惯引起的负效用在各个时间段内的程度并不完全相同。举例来说，大部分电动汽车用户会选择在下班后进行充电，在其他时间段内选择充电分布较少。因此，假设一个消费者由于改变充电策略，将原本在下班时要要充电量的一部分改变迁移至其他时段或是选择其他出行策略，这会带来比较大的负效用。相反，用户如果选择在其他时间段内改变充电策略，相比上述策略损失的效用较少。该部分负效用由初始需求曲线分布 $\boldsymbol{DC}_0$ 与当前时间步的需求曲线分布 $\boldsymbol{DC}_t$ 的差值决定。在某时间段内，无论是选择增加需求还是减少需求都会带来需求变动。因此，时间步 t 内 x_n 时段，该部分负效用定义为：

$$S_{t,x_n}^{user} = \alpha_t^2 \cdot RP_{t,x_n} \cdot |DC_{t,x_n} - DC_{0,x_n}| \tag{2-5}$$

其中，α_t 表示用户的需求弹性参数，随时间变动。需求弹性是指每单位商品

价格的变化引起消费者需求变动程度大小的指标，由商品的性质决定。在电动汽车充电市场的三方模型下，充电汽车用户在不同时段对于充电服务的质量需求并不完全相同[58], 根据优先级和需求特征，其需求分布可以被分为硬性需求和软性需求。在峰荷时间，消费对充电服务的需求最为迫切，所以定义 α_t 在该时刻为1，时间距离峰荷时间越远，消费者对于的充电服务的需求弹性越小。本工作根据文献 [25] 的研究，将 17，18，19 时定义为电动汽车充电服务的峰荷时间（peak hour）。α_t 的数值表示如图2-3：

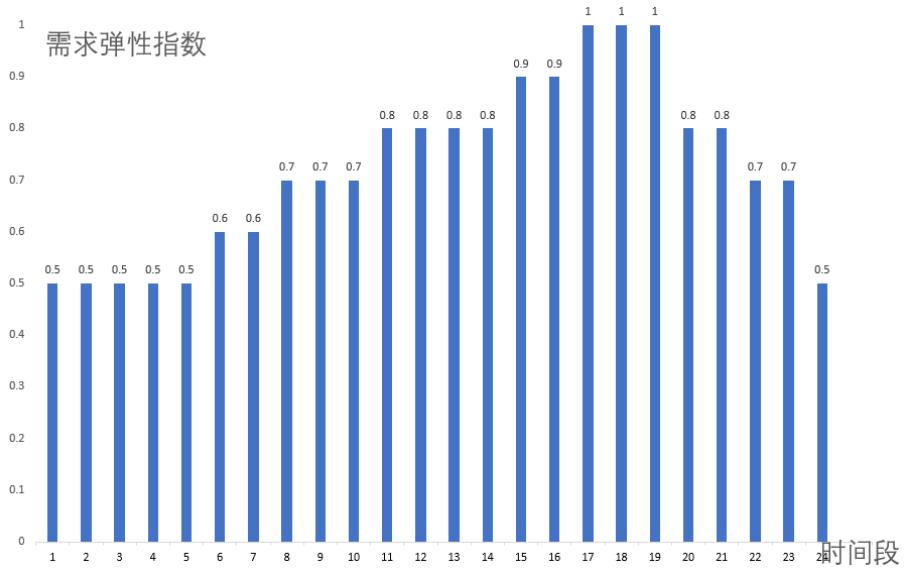

图 2-3 不同时间段下需求弹性参数设置

由式 (2-1)，用户在时间步 t, 时段 x_n 的目标函数定义为：

$$R_t^{user}(\boldsymbol{RP_t}, \boldsymbol{DC_t}) = \sum_{n=0}^{23} RP_{t,x_n}(0.75DC_{0,x_n} + DC_{0,x_n} \cdot lg(\frac{DC_{t,x_n}}{DC_{0,x_n}} + 1) - DC_{t,x_n} - \alpha^2 |DC_{t,x_n} - DC_{0,x_n}|) \tag{2-6}$$

此外，为了满足电动汽车充电需要，消费者需求函数需要满足约束：

$$\sum_{x=0}^{23} DC_{t,x_n} \geq 0.8 \sum_{x=0}^{23} DC_{0,x_n} \tag{2-7}$$

2.3 电力聚合商建模

电力聚合商，又称负荷聚合商 (load aggregator), 是一种聚合充电用户需求与市场信息的市场参与者。它可以作为电力大宗市场的买家，再以零售价格将充电服务销售给消费者。电力聚合商在电力市场中的作用主要体现电力系统的平稳运行与市场需求的调节上。从电力系统的运行方面来看，电力聚合商可以向电力系统

提供时段内的电力市场信息，使得电网在电压控制、供电线路控制等运行策略中有更多更合理的调节空间，从而降低电网的运行成本。根据《上海市电力发展“十二五”规划》，随着充电汽车的大量涌入带来电力负载高峰，电网如果如果通过增建发电厂，新建输电线路等措施来满足高峰供电需求，不仅成本较高，对环境也会造成较大的损害。电力聚合商作为需求响应的信息整合方，可以在前期优化发电站配电器与输电线路的设计，在中期降低电网发电、配电的成本[59]，最终提高总体电力市场的运行效率。

从市场需求调节方面来看，电力聚合商建立了需求响应计划中用户与电网之间的联系。单个电动汽车用户的需求量与电网的配电量之间存在数量级的差异，其负载弹性水平无法参与需求响应计划计划[60]。电力聚合商作为用户需求的整合平台，可以通过收集单位时间内用户的充电服务需求并加总整合，从而将其引入市场交易及需求响应机制，为整个市场提供更多信息从而优化市场参与各方的收益及市场效率。

在本工作搭建的电动汽车充电市场的三方模型中，因为电力聚合商在电力市场中担任中介和平台的角色，同时又是零售电力市场的卖家和价格制定者，本文模型规定聚合商的可以观测到所有市场信息，包括市场需求曲线 $\boldsymbol{DC}_t$、市场供给 $\boldsymbol{SC}_t$, 批发市场电力 $\boldsymbol{WP}_t$ 和零售市场电力价格 $\boldsymbol{RP}_t$。聚合商可以基于当前观测制订新的零售价格曲线 $\boldsymbol{RP}_t$。具体来说，聚合商根据观察到的市场需求曲线 $\boldsymbol{DC}_t$ 更新其购电行为，根据其他市场信息更新电力市场零售价格曲线。

电力聚合商在本文模型中设定的目标为自身收益最大化，其在时间步 t, 时段 x_n 内的收益 R^{agg}_{t,x_n} 可以定义为:

$$R^{agg}_{t,x_n} = \begin{cases} DC_{t,x_n} \cdot RP_{t,x_n} - SC_{t,x_n} \cdot WP_{t,x_n}, & SC_{t,x_n} \geq DC_{t,x_n} \\ SC_{t,x_n} \cdot RP_{t,x_n} - SC_{t,x_n} \cdot WP_{t,x_n}, & \text{otherwise} \end{cases} \tag{2-8}$$

同时，电力集成商还通过调整价格保证在每一个时间步内有利可图，即满足约束条件 $R_{t,x_n} \geq 0$。否则，为了保证自身不陷入亏损，电力集成商将选择停止营业造成市场失灵。电力集成商可以通过调整零售价格保证盈利，约束条件求解为：

$$\begin{cases} RP_{t,x_n} \geq \frac{SC_{t,x_n} \cdot WP_{t,x_n}}{DC_{t,x_n}}, & SC_{t,x_n} \geq DC_{t,x_n} \\ RP_{t,x_n} \geq WP_{t,x_n}, & \text{otherwise} \end{cases} \tag{2-9}$$

至此，电力集成商在每一个时间步内的目标函数为:

$$R_t^{agg}(\boldsymbol{DC}_t, \boldsymbol{SC}_t, \boldsymbol{RP}_t, \boldsymbol{WP}_t) = \begin{cases} \sum_{n=0}^{23} DC_{t,x_n} \cdot RP_{t,x_n} - SC_{t,x_n} \cdot WP_{t,x_n}, & SC_{t,x_n} \geq DC_{t,x_n} \\ \sum_{n=0}^{23} SC_{t,x_n} \left(RP_{t,x_n} - WP_{t,x_n}\right), & \text{otherwise} \end{cases} \tag{2-10}$$

2.4 电网建模

电网作为配电的核心，可以跨州、跨国家甚至将电力输送到广阔的外部地理区域，以电力的形式集中分配能源，通过为工业、服务业和消费者提供可靠的电力，发挥着核心的经济和社会作用，是现代社会生活中极其重要的的基础设施。随着世界在可持续发展的方向过渡，电网重要性在今天显得更加重要。能源需求和生产状况的变化、不断增加的可再生能源整合以及高压电网技术，对人类运营商在优化电力运输的同时避免停电构成了真正的挑战。

智能电网是现代化的电力配送、传输网络，其核心是将信息通信技术集成到电力系统中。智能电网使用数字信号收集供需端的电力使用信息，并利用这些信息对电力网络中的资源进行调整，比如电力生产、消耗以及电力网络结构。作为电力交易市场的供给侧，智能电网在提高电力市场运行效率方面占有重要的作用。

智能电网是一个复杂的拓扑结构，各地的能源设施，如火电、水电、风电、光电等设施分配在不同的区域，并以输配电线路连接至变压器，再连接到电力购买方。图2-4展示了 IEEE-14 总线系统，一个拥有 14 个主要节点的电网拓扑结构，它被广泛用于进行各种研究，如短路分析、负载流研究、互联电网问题等。

如果将电网拓扑图整体考虑到市场模型中，将会需要很大的动作空间和状态空间来描述电网中每一个节点的状态，不利于分析市场问题和求解过程。本文模型屏蔽电网运行细节，提取电网在电力交易市场中的特征，将电网建模成经济学模型的生产商。电网可以观察到市场上的电力需求曲线 DC_t, 制订自身的发电策略形成新的供给曲线 SC_{t+1}，并制订新时间步的电力批发价格 WP_{t+1}。

智能电网的优化目标 R^{grid} 可以被定义为售电所得 P 和运行损耗 OC 之差：

$$R_t^{grid} = P_t - OC_t \tag{2-11}$$

其中，售电 P 所得即为其以市场批发价格出售其生产曲线上的电力所得:

$$P_{t,x_n} = SC_{t,x_n} \cdot WP_{t,x_n} \tag{2-12}$$

在电力系统中，除了满足基尔霍夫电流定律（Kirchhoff’s current law）与基

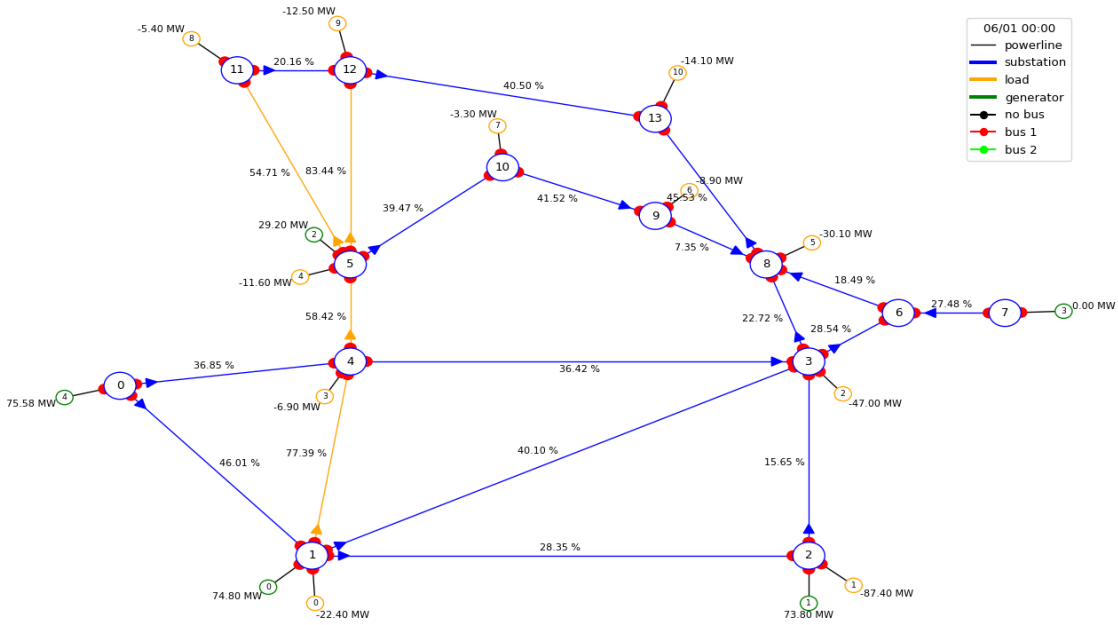

图 2-4 IEEE-14 电网拓扑图某时刻的运行状态

尔霍夫电压（Kirchhoff’s voltage law）定律外，为了保证电网负载平衡，电网中生产的电力总量必须符合消费的电力总量，否则电网的安全性将会受到影响。因此，本文设定在电动汽车充电交易市场上，电网作为生产商，在一个时间步内生产的电力总量与消费总量相近。因此，电网生产曲线 SC_t 需要满足以下两条约束条件：

$$\begin{aligned} 0.9\sum_{n=0}^{23} DC_{t,x_n} \le \sum_{n=0}^{23} SC_{t,x_n} \le 1.1\sum_{n=0}^{23} DC_{t,x_n} \\ 0.9DC_{t,x_n} \le SC_{t,x_n} \le 1.1DC_{t,x_n} \end{aligned} \tag{2-13}$$

在电网运行过程中，电力系统运行会不可避免地受到一些损耗，包括在电流在经过输电线路时产生地热损耗，需求变动时通过改变拓扑结构或是增减线路负荷带来地电路损耗等。因此，本文定义电网运行损耗 OC_t 为：

$$OC_t = m_{t,x_n}\sum_{n=0}^{23} SC_{t,x_n} + F(\boldsymbol{SC_t}) \tag{2-14}$$

其中，m_{t,x_n} 衡量了因为在电力输送过程中热损耗带来的损失，本文设定 $m_{t,x_n} = 0.05WP_{t,x_n}$。$F(\cdot)$ 衡量曲线波动程度的函数，代表因为电力波动所导致的线路操作、拓扑结构改变等损耗。$F(\boldsymbol{SC}_t)$ 用分段导数斜率的绝对值表示，每相邻两个时间段内调整造成的损耗以该时间段的电价均值的一部分衡量，表示如下：

$$F(\boldsymbol{SC}_t) = 0.2\sum_{n=1}^{23}(WP_{t,x_{n-1}} + WP_{t,x_n}) \cdot \left| SC_{t,x_n} - SC_{t,x_{n-1}} \right| \tag{2-15}$$

综上, 电网目标函数被定义为：

$$R_t^{grid}(\boldsymbol{SC}_t, \boldsymbol{WP}_t) = \\ 0.95\sum_{n=0}^{23} SC_{t,x_n} \cdot WP_{t,x_n} - 0.2\sum_{n=1}^{23}(WP_{t,x_{n-1}} + WP_{t,x_n}) \cdot \left|SC_{t,x_n} - SC_{t,x_{n-1}}\right| \tag{2-16}$$

2.5 本章小结

本章搭建了基于多智能体博弈框架的电动汽车充电市场三方博弈模型。本章从现实情况入手，抽象出各方参与者的特征，对模型总体框架的运行机制、模型参与者各方参与主体的观测、动作、回报函数做出说明并定义。

第三章 社会总福利最大化下的三方策略均衡点求解

本章将考虑一个理想化的假设：电动汽车充电市场中，三方参与者制订联合策略使得社会总福利最大化。在这个假设中，三方参与者目标一致，共享同一个相同的奖励函数，通过调整自身行动最大化社会总福利，是完全合作的智能体。本章工作采用多智能体强化学习算法求解市场最终状态并做分析。

3.1 多智能体强化学习算法选择与设计

3.1.1 博弈论与多智能体强化学习

博弈论 (Game Theory) 是现代经济学中一个重要的分支，基于参与博弈各方的收益预测其行为，研究各方的优化策略，使得某方参与者达成其个体收益最大化或是集体收益最大化目标。博弈行为可以定义为通过竞争、对抗或是合作取得收益的行为。具体来说，在博弈中参与各方被设定了不同的目标，他们并不仅仅会选择当前对自己收益最大的行为作为其行动策略，而是会根据其竞争对手的情况，预测对手的行动策略，在此基础上选择使自己收益最大化的策略。一般而言，参与博弈的各方需为理性决策者，满足以下四个条件：

（1）明白其当下所有可能的行动；

（2）明白其每一个行动带来的即时回报与远期回报；

（3）明白其行动会如何影响结果；

（4）其理性决策结果基于收益最大化假设，即参与博弈的理性决策者会在在任何条件下保持自私，永远为自身收益最大化为目标而行动。

与此同时，参与方会考虑长远利益，若其有能力预测到接下来的一系列各方决策过程，在当下可能并不会选择使自己情境收益最大化的行动。例如，在经典的棋牌游戏“斗地主”中，扮演农民的角色在下家农民队友仅剩一张牌时，可能会选择拆掉自己的牌以获得牌权，使队友赢得胜利从而获胜。博弈论主要研究参与博弈各方是否存在最合理的行动策略，以及求解该策略的数学理论及方法。目前，博弈论在统计学、经济学、生物学等学科中有广泛应用[61]。

然而，随着经济社会的发展，在很多现实问题中，参与博弈的主体越来越复杂，动作空间、状态空间越来越复杂，传统博弈论方法已经无法处理如此大规模的问题。例如，在智能电网的建设中，随着指数级增长的电子设备及可再生能源的并网，电网的控制节点及输配电线路也在不断增加，电网控制系统的决策主体也会不断增多。此时，即使博弈论在刻画决策主体的行动策略方面具有天然的优

势，但仅依靠传统博弈论方法已经无法适应当前电网发展建设，越来越多的专家学者开始将目光投向博弈论与人工智能技术结合的方向[62]。

强化学习属于人工智能的分支，其原理主要是利用智能体与环境交互得到奖励，根据累计奖励值优化其动作策略。图3-1是一个经典的单智能体强化学习框架。其中，Agent 表示与环境交互的智能体，Environment 为智能体所处的环境。智能体通过做出动作 Action 来影响环境，环境对于该动作做出反馈，向智能体提供新的状态 State 及奖励 Reward。智能体在时间 t 观测环境的状态 S_t，接收到环境的奖励 R_t，基于其策略做出动作 A_t 并影响环境，环境接收到智能体动作以后状态被改变并给予智能体奖励并在下一个时间步反馈给智能体。

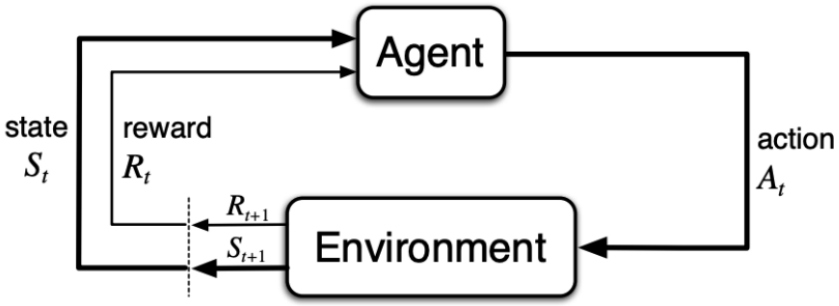

图 3-1 单智能体强化学习框架[9]

在单智能体强化学习体系中，马尔可夫决策过程（Markov decision process，MDP）详细描述了一个问题中智能体与环境交互的过程。在多智能体强化学习体系中，MDP 扩展到马尔可夫博弈 (Markov game) 或随机博弈（stochastic game）。马尔科夫博弈通常用一个元组表示：

$$(n, S, A_1, \cdots, A_n, T, \gamma, R_1, \cdots, R_n)$$

其中，n 表示参与博弈的主体数量；S 表示多智能体的环境状态；$A_1, \cdots, A_n$ 为每个智能体采取的动作；T 为状态转移函数，在当前状态下，当智能体的联合动作确定时转移到下一个状态的概率分布；γ 表示远期回报的折扣率；$R_1, \cdots, R_n$ 表示每个智能体通过环境所获得的回报。

多智能体强化学习属于机器学习的研究方向之一，主要研究多智能体策略的协同，竞争，演化等问题。多智能体强化学习与单智能体不同，由数个智能体同步或异步与环境进行交互，如图3-2，System 表示多智能体所处的环境，Agent 编号 1 至 N 为环境中参与交互的 N 个个体智能体。每个智能体都会基于自己对环境的观测，以自身的策略做出动作并获得奖励。深度强化学习利用了神经网络可以

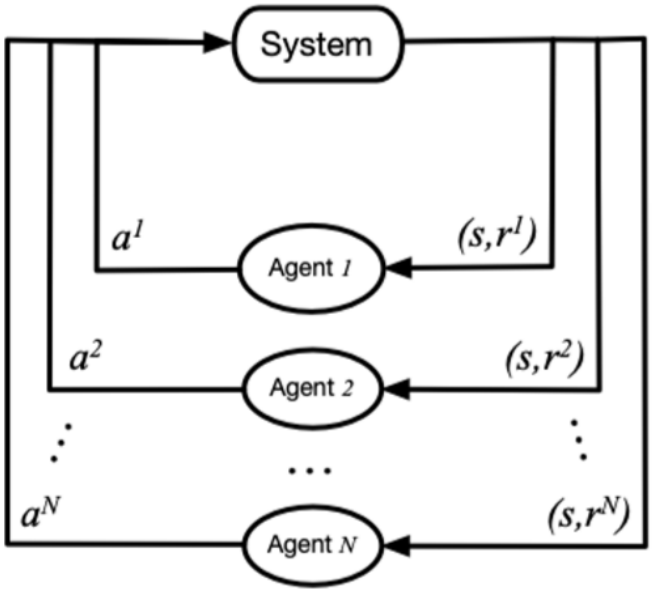

图 3-2 多智能体强化学习框架[9]

快速有效对大规模数据分析处理的特性，在求解多智能体博弈、决策，达成均衡策略方面具有天然的优势。基于此，本文采用多智能体深度强化学习算法对该博弈问题求解。

3.1.2 算法选择与设计

许多机器学习领域的学者提出了性能良好的多智能体强化学习算法，例如反事实多智能体策略梯度算法 (Counterfactual Multi-Agent Policy Gradients,COMA)[63]，多智能体深度置信梯度策略算法 (Multi-Agent Deep Deterministic Policy Gradient,MADDPG)[64] 和独立 Q 学习算法 (Independent Q-Learning,IQL)[65]。本文将对三种代表性多智能体强化学习算法做简介，并对其优缺点进行对比。最终，本文选择具有代表性的 MADDPG 算法进行解析，将其改进设计为本问题的求解方法。同时，COMA 算法与 IQL 算法将作为对比试验。

COMA 是一种基于策略梯度的多智能体深度强化学习算法，考虑到算法在高维动作状态空间中，智能体的联合策略比较难学到信息的特点，COMA 采用了分布式策略代替联合策略，允许各个智能体基于自身的观测和决策网络输出各自的动作。与此同时，为了在每一步中获得更多的信息，COMA 采用了一个中心化的 Critic 网络以对所有智能体的结果进行评价，在训练的过程中就可以获取全部智能体的信息。其次，本文设定了智能体的合作通信机制，所有智能体共享同一个奖励。为了解决奖励相同情况下，对于智能体贡献权重问题，COMA 设置了反事实基线（Counterfactual Baseline）来分配每一个智能体的权重。反事实基线技术其实

就是差分技术，先计算智能体在默认动作下获得的奖励，再计算执行其策略网络得到的动作下的奖励，通过两个奖励的差值判断该智能体对于整体任务的贡献。最后，为了保证 Critic 对每一个智能体的反事实基线进行快速的运算，COMA 为每一个智能体设置了优势函数（Advantage Function）, 在其他智能体保持固定动作的时候计算选中智能体的贡献。

COMA 算法在本文应用中有以下两点特性：

（1）COMA 算法中智能体动作网络相同，共享同一个奖励，但是在本文三方竞争模型中，设定目标为智能体各自收益最大化，COMA 算法无法设置高效的异质化的智能体。且该目标下无法设定有应用意义的目标函数。

（2）COMA 算法处理针对离散动作，学习的是随机策略。而本文模型为连续动作，COMA 算法对于数据的采样利用效率会相对较低，且无法探索完整的动作空间。

基于此，本文将对 COMA 算法的智能体架构进行修改，本文使用蒙特卡洛方法为 COMA 算法采样，并更改智能体的架构使其输入为符合文中模型的特征。本文拟将 COMA 算法作为 MADDPG 算法在社会福利最大化目标下的对比算法。

IQL 是一种简单地将 DQN 算法应用到多智能体领域的算法。对于每一个智能体，IQL 为其分配一个策略动作网络。每一个智能体的观测均为除其自身外其他智能体的行动结果，即 IQL 中的每一个智能体都会将其他智能体的动作结果视为其观测。IQL 的缺点显而易见：由于环境中存在其他智能体，该算法面对的是一个非稳态的环境，可能会遇到不收敛的问题。

IQL 算法在本文应用中有以下特性：

（1）由于 IQL 为每一个智能体分配网络，最大程度上避免了智能体之间的合作奖励设置难题。在合作任务上，可以将奖励设置为社会总福利。在竞争任务下，可以将奖励设置为智能体各自收益最大化。

（2）IQL 不设定智能体之间的通信。在本文模型中，三方参与者之间也是设定没有直接通信的，各方均从公共资源池中获取各自的观测。IQL 的设定十分契合本文

MADDPG 是一种基于演员-评论家（Actor-Critic, AC）算法的多智能体深度强化学习算法，将 AC 算法拓展到多智能体场景中。MADDPG 算法采用了中心化训练，去中心化执行的架构，其应用场景为智能体仅能观测到部分信息情况下的混合博弈模型中。

MADDPG 的架构如图3-3。在动作执行过程中，每一个智能体拥有一个独立的执行机（actor），输入其自身的观测信息 o，通过使用其自身的策略 π 得到自身

动作 a。此过程中，智能体无论是观测还是执行都只用到了自身的观测数据。在训练过程中，每一个智能体对应一个评论机（critic），但该 critic 状态评估价值 Q 所用到的信息并非只有其对应的 actor，而是用到了所有智能体的信息。MADDPG 架构内存在 N 个这样的中心化 critic

MADDPG 算法拥有以下三点特性：

（1）算法通过中心化训练，去中心化执行得到其最优策略学习完成后，对于每一个智能体，只需要观测局部信息，通过学习的策略就可以得到最优动作；

（2）不严格要求环境的动力学微分模型，且不要求智能体之间相互通信写作，而是通过观测全局的 critic 实现信息沟通；

（3）MADDPG 算法由于并不设定智能体之间的通信，所以其既能适应合作博弈，又能适用于竞争博弈。此外，在混合类型博弈中也发挥出了良好的效果。

因此，本文主要基于 MADDPG 的集中式决策，分散化执行的特点对其加以改进以适应本文研究的问题。在本文中，由于 COMA 与 IQL 算法各自的特点，本文将 COMA 设定为第三章合作模式的对比算法，将 IQL 设定为第四章竞争模式的对比算法。

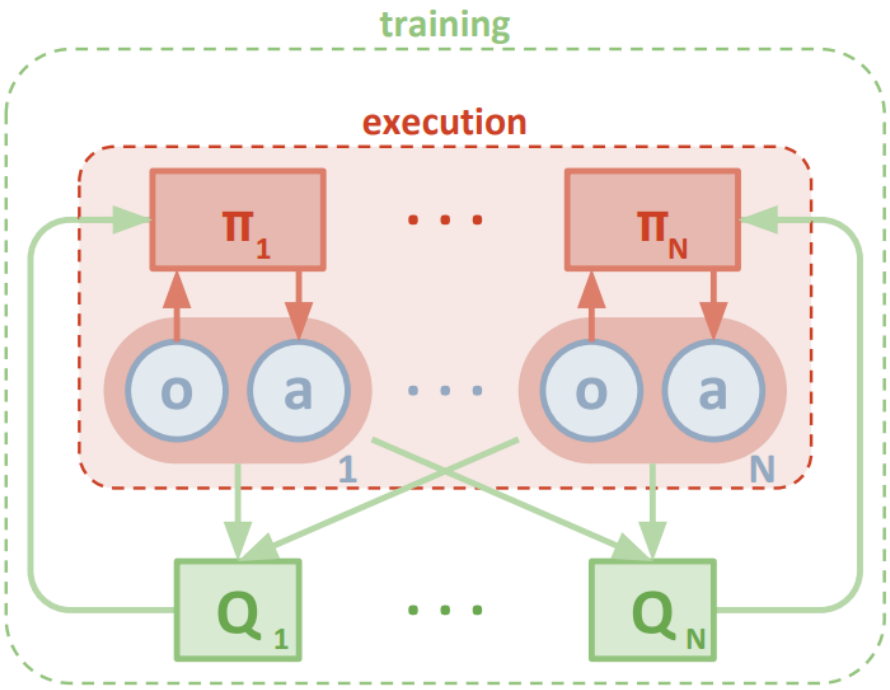

图 3-3 MADDPG 架构 [64]

3.2 社会福利最大化假设下的电动汽车充电市场强化学习模型

3.2.1 马尔可夫博弈描述

本文采用马尔可夫博弈将描述三方博弈电动汽车充电市场：

$$(\boldsymbol{S}, \boldsymbol{A}_{user}, \boldsymbol{A}_{agg}, \boldsymbol{A}_{grid}, \gamma, SW)$$

其中，参与博弈智能体数量为 3，分别为电网、电力聚合商和用户；

$\boldsymbol{S}$ 代表环境状态，即前文中定义的市场状态，包括电力市场批发零售价格与供给需求曲线，是一个 48 维向量：$\boldsymbol{S} = [\boldsymbol{DC}, \boldsymbol{SC}, \boldsymbol{RP}, \boldsymbol{WP}]$；

$\boldsymbol{A}_{user}$ 代表用户的动作。前文模型中定义了用户可以形成新的消费选择，向提供其供给曲线，是一个 12 维向量：$\boldsymbol{A}_{user} = \boldsymbol{DC}$；

$\boldsymbol{A}_{agg}$ 代表电力聚合商的动作，即为电力交易市场制订新的零售价格曲线，是 12 维向量：$\boldsymbol{A}_{agg} = \boldsymbol{RP}$；

$\boldsymbol{A}_{grid}$ 代表电网的动作，电网在市场中输配电力提供供给曲线，并且以批发价格售卖电力，是一个 24 维向量：$\boldsymbol{A}_{grid} = [\boldsymbol{SC}, \boldsymbol{WP}]$；

SW 为本章模型的目标函数，即体系的社会总福利，由市场所有参与者的收益构成：

$$SW = f(R_{user}, R_{agg}, R_{grid})$$

。$R_{user}, R_{agg}, R_{grid}$ 分别为三方的目标函数。本章将社会总福利函数定义为：

$$SW = \omega_1 R_{user} + \omega_2 R_{agg} + \omega_3 R_{grid} \tag{3-1}$$

其中，ω 为衡量参与主体对社会总福利贡献的权重。本文假设三方效用对社会福利贡献相同，通过微调参数 ω 数值控制三方福利水平。

3.2.2 算法流程

该问题定义下，智能体只能观测到部分信息，且其动作空间为连续动作空间，因此，本章基于 MADDPG 算法设计了智能体随机顺序参与博弈算法 (MADDPG with Random Order, MADDPG-RO)。算法描述如3-1：

算法 3-1 社会总福利最大化假设下的 MADDPG-RO 算法流程

for *episode = 1 to M* **do**
  为动作探索生成随机过程 $\mathcal{N}$;
  随机打乱智能体行动列表的顺序 [user, aggregator, grid];
  从环境中得到初始状态 s;
  for *t=1 to max-episode-length* **do**
    对于每个智能体 i, 根据当前策略与探索情况选择动作，得到联合动作 ;
    执行联合动作 $a = (a_1, a_2, a_3)$;
    从环境中得到奖励 $r = f(R_1, R_2, R_3)$ 和新时间步的状态 s';
    将四元组 (s, a, r, s') 存储到经验回放池中 $\mathcal{D}$;
    $s \leftarrow s'$;
    for *agent i=1,2,3* **do**
      从经验池 $\mathcal{D}$ 中随机采样 S 个样本 (s^j, a^j, r^j, s'^j);
      通过最小化如下损失函数更新 Critic $\mathcal{L}(\theta_i)$;
      利用采样的梯度策略更新 Actor;
    end
    对每个智能体 i 更新策略网络参数 θ_i ;
  end
end

3.3 算例结果分析

3.3.1 MADDPG-RO 算法结果分析

本章模型设定初始状态 $s_0 = (\boldsymbol{DC}_0, \boldsymbol{SC}_0, \boldsymbol{RP}_0, \boldsymbol{WP}_0)$。算法由此状态出发直到市场达到设定最大步数或均衡，并分析达到均衡时的收益。市场初始状态如图3-4。

初始状态的零售价格 $\boldsymbol{RP}_0$ 上下界由美国电力集成商圣地亚哥天然气与电力公司（San Diego Gas & Electricp Company）提供的夏季分时电价计划确定，本文取上下界平均值确定初始状态的零售价格曲线，如图3-5所示。其中，红线与黄线分别代表零售电价的上限与下限。为了保证市场有序运行，零售电价的调整不能超过绿色区域范围。为了保证电力聚合商有利可图，零售价格应同时满足约束条件（2-9）。

初始状态的电力市场批发价格 $\boldsymbol{RP}_0$ 由美国伊利诺伊州联邦爱迪生电力公司

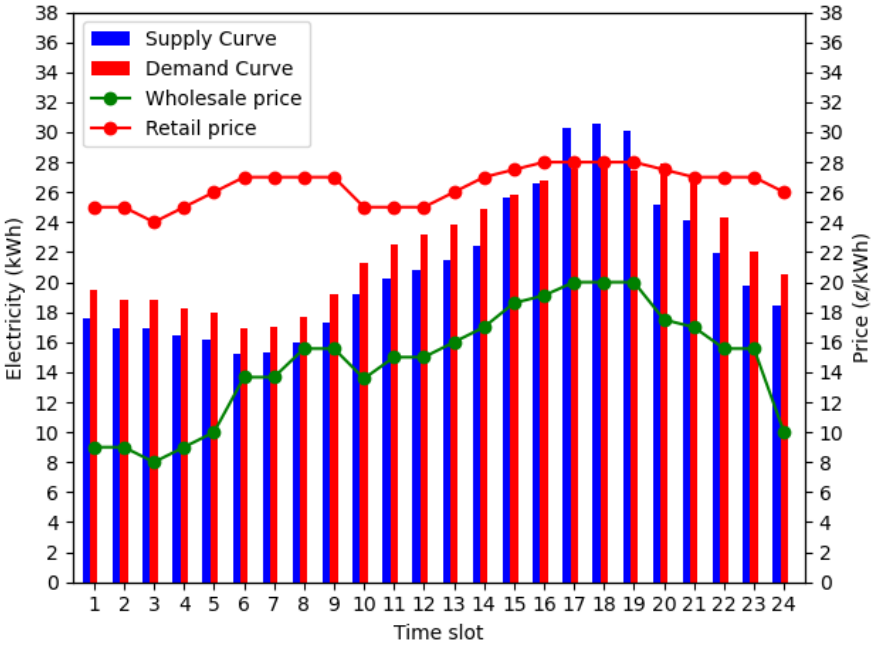

图 3-4 市场初始状态

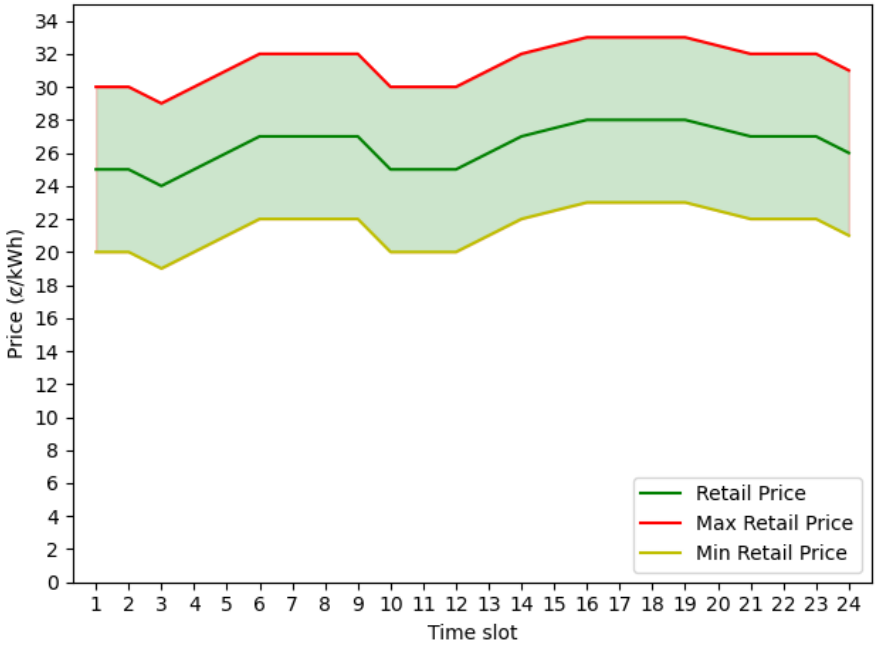

图 3-5 夏季分时电价零售市场价格

（Commonwealth Edison，ComEd）提供的交易数据确定，如图3-6，红线与黄线分别代表批发电价的上限与下限。为了保证市场有序，批发电价的调整不能超过绿色区域范围。

初始状态的需求曲线 $\boldsymbol{DC}_0$ 和供给曲线 $\boldsymbol{SC}_0$ 同样由 ComEd 数据提供。为了保

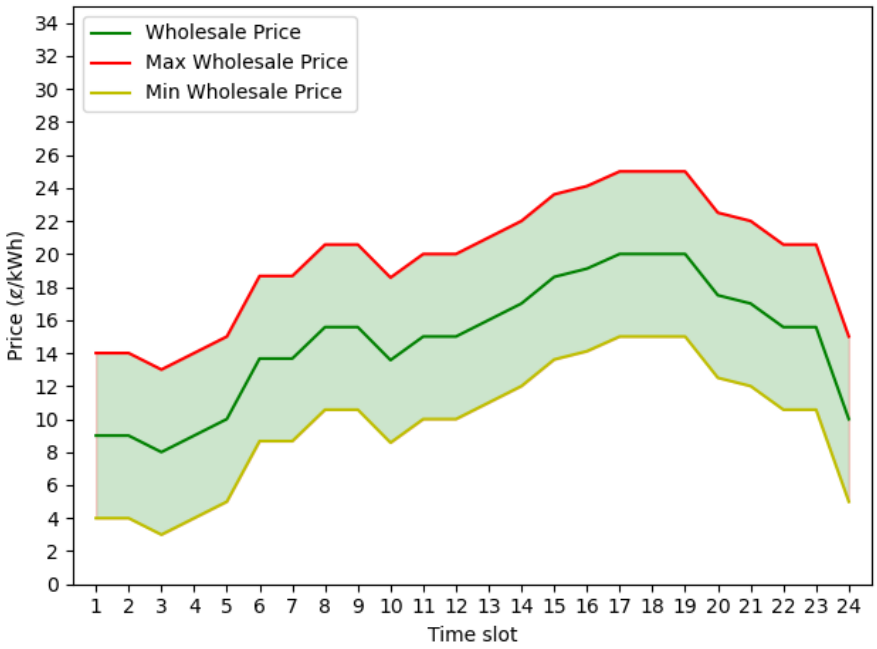

图 3-6 夏季分时电价批发市场价格

证市场能容纳足够的充电需求，模型设定每天需求总量不可以低于初始需求量的80%，即满足约束条件（2-7）；同时，供给曲线应满足约束条件（2-13）。

本章模型采用 MADDPG-RO 算法运行该算例。算法参数设定如表3-1:

表 3-1 MADDPG-RO 参数设定

参数	赋值
最大迭代步数 max episode length	30000
每步最大步长 max length per episode	50
随机探索概率 ε	0.3
经验回放池大小 $\mathcal{D}$	6400
经验池取样大小 batch size	64
远期回报折扣系数 γ	0.95
学习率 learning rate	0.01
三方权重分配 $(\omega_1, \omega_2, \omega_3)$	(1,1,1)

三方收益曲线与社会总福利曲线波动情况展示如下：（1）充电汽车用户效用分析如图3-7，蓝色折线为训练过程中用户收益曲线。为了便于观察，本文使用中值滤波器[66]对折现滤波得到橙色平滑曲线。在 12000 个时间步之前，训练过程中用户收益波动式下降。在 20000 步之后，收益曲线逐渐平缓趋于收敛，最终用户

收益收敛在 3100 左右。在社会总福利最大化的前提下，用户选择让渡一部分自己的收益以带给其他参与主体更多的效益以减少社会福利损失。

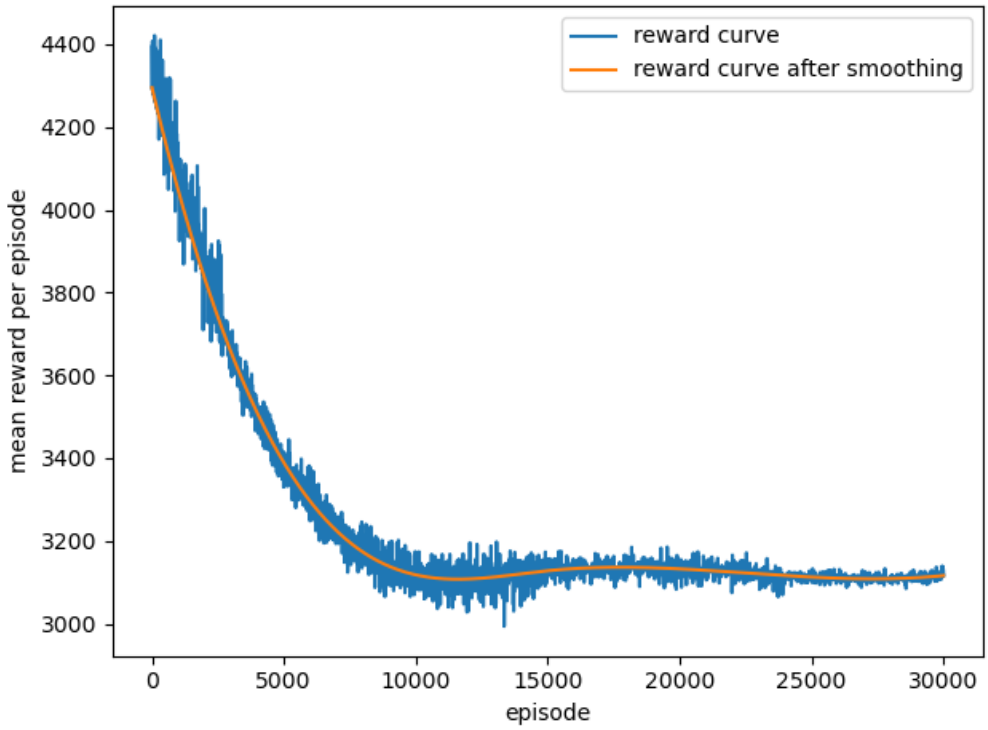

图 3-7 MADDPG-RO 训练过程中充电汽车用户效益

（2）电力聚合商效用分析如图3-8，训练过程在 12000 个时间步之前，电力聚合商的收益波动式上升。在 20000 步之后，收益曲线逐渐平缓趋于收敛，最终收益收敛在 9300 左右。

（3）电网效用分析如图3-9，训练过程在 12000 个时间步之前，电网的收益波动式上升。在 20000 步之后，收益曲线逐渐平缓趋于收敛，最终电网收益收敛在 8300 左右。

（4）社会总福利社会总福利曲线变动情况如图3-10所示，社会总福利在本节中被定义为三方收益总和。随着训练过程的进行，在 12000 步到 15000 步之间波动幅度较大，在 20000 步之后波动幅度减少，趋于收敛。

3.3.2 COMA 算法结果分析

使用 COMA 算法得到的聚合商收益，电网收益，用户收益曲线如图3-11，3-12，3-13。与 MADDPG-RO 算法的结果类似，用户会折损一部分自身收益以达到社会总福利最大化到的目标，其余两方收益均大幅上升，趋于稳定。使用 COMA 算法得到的社会总福利曲线变动如图4-8

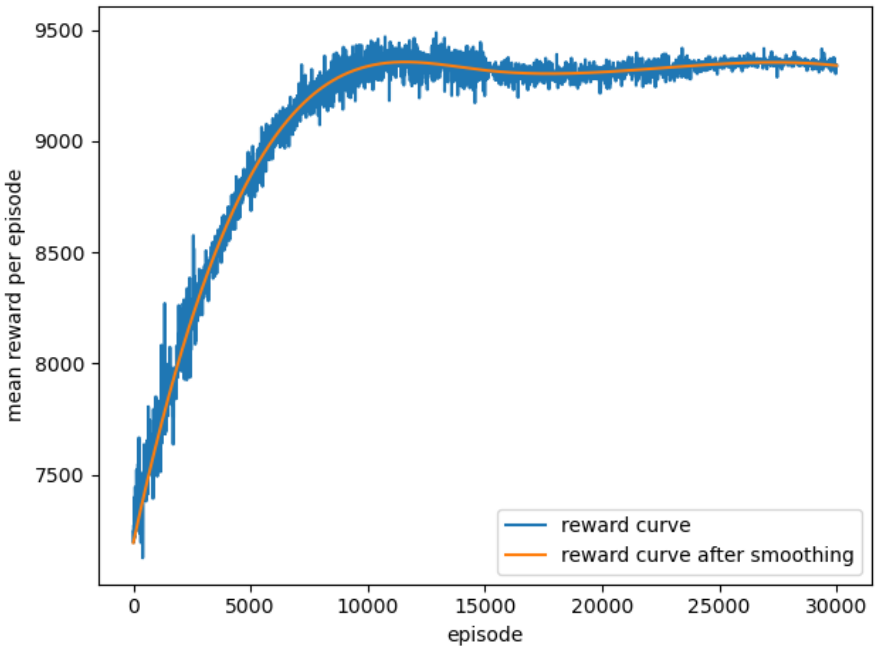

图 3-8 MADDPG-RO 训练过程中电力聚合商效益

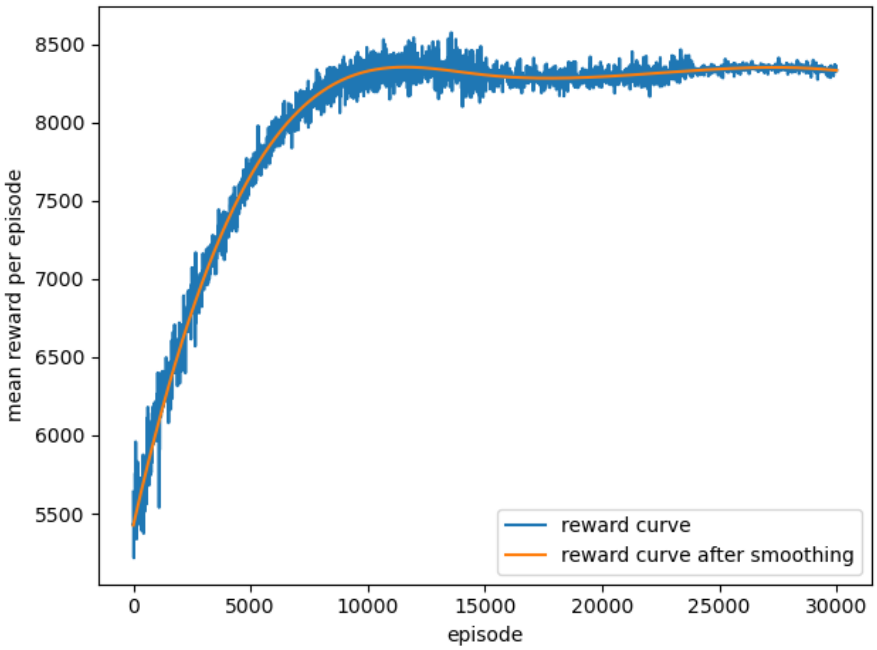

图 3-9 MADDPG-RO 训练过程中电网效益

3.4 结论

在本文三方博弈框架中，采用 MADDPG-MO 算法求解以社会福利最大化为优化目标的条件下三方主体博弈策略，结果如表3-2所示。在社会福利最大化目标下，

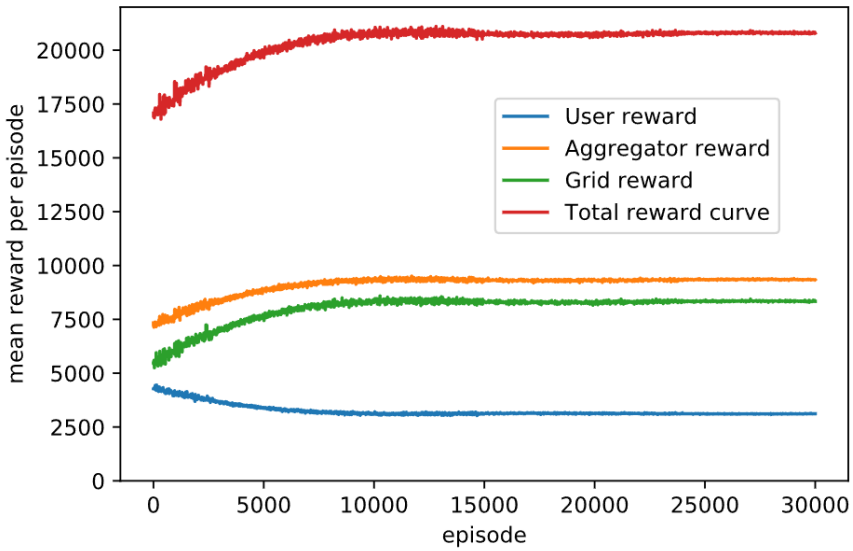

图 3-10 MADDPG-RO 训练过程中社会总福利

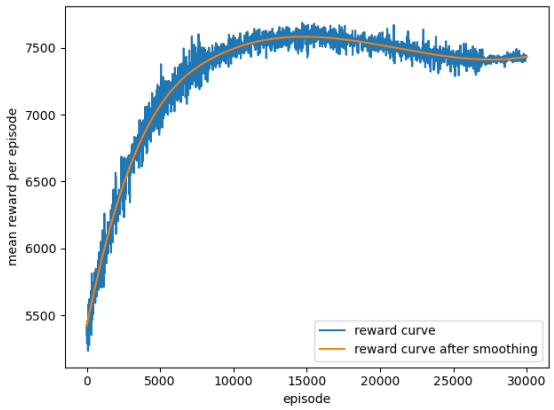

图 3-11 COMA 算法的聚合商收益变动

MADDPG-MO 算法得到的策略使得社会总福利提高了 22.91%, 其中电力聚合商从该目标下的优化策略中获益最多，收益增加了 53.59%; 电网收益上升了 29.83%; 然而电动汽车充电用户的效益受损，下降了 27.43%。最终市场状态如图3-15。

采用 COMA 算法求解以社会福利最大化为优化目标的条件下三方主体博弈策略，结果如表3-3所示。在社会福利最大化目标下，COMA 算法得到的策略使得社会总福利提高了 21.71%, 其中电力聚合商收益增加了 63.97%; 电网收益上升了 27.22%; 然而电动汽车充电用户的效益受损，下降了 6.77%

需求曲线 ***DC*** 变动如图3-16，充电市场用户选择将自己在需求高峰期的部分需求转移到其他非需求高峰期。在距离高峰期越近的时间点，用户选择转移的需求量越多。

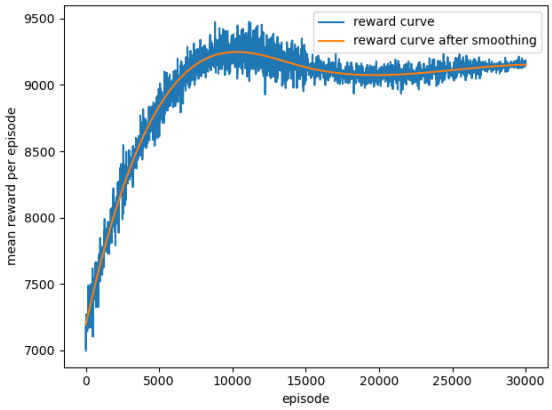

图 3-12 COMA 算法的电网收益变动

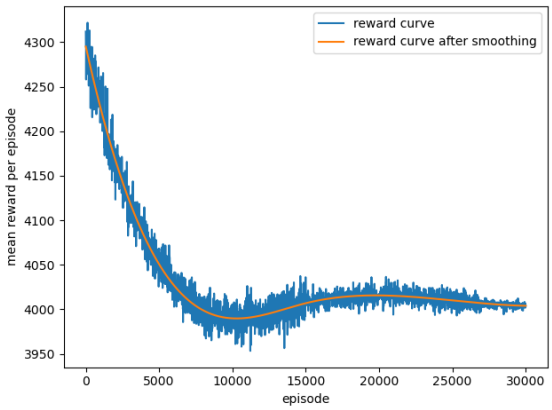

图 3-13 COMA 算法的用户收益变动

供给曲线 ***SC*** 变动如图3-17，由于客户需求变动，电网可以选择调整自己的发电策略，在高峰期减少部分发电量的供给，在非高峰期增加供给。

零售价格曲线 ***RP*** 变动如图3-18，电力集成商通过提高高峰时段电动汽车充电价格的手段抑制用户的需求，在其他所有非高峰时段均降低零售价格使用户在该时段可以选择更多的充电服务。

批发价格曲线 ***W*** 变动如图3-19，电网通过平衡供给曲线降低了成本，整体电价水平降低。在需求高峰时刻通过提高批发价格促使零售价格上升，使消费者降低该时间段的购买需求。

根据上文数据结果，MADDPG-MO 策略使得电力需求高峰时刻的零售价格与批发价格上升，该时段的用户需求下降，用户选择将在该部分的充电需求转移到其他时刻。对于用户来说，充电需求带来改变带来的出行习惯、充电策略的改变增加了其不满意度，同时，需求高峰时刻的零售电力价格上涨也使得用户需要在该时段支付更高的费用满足自己的充电服务需求，这导致了用户总体收益下降。

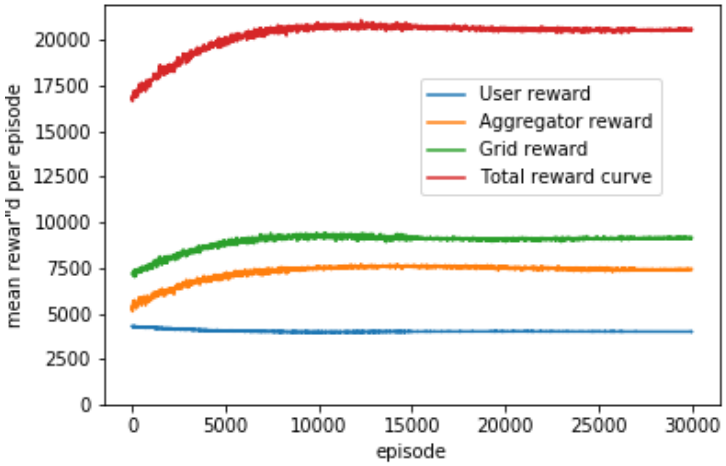

图 3-14 COMA 算法的社会总福利变动

表 3-2 MADDPG-RO 社会福利最大化目标下的前后收益对比

状态	用户效益 R_{user}	电力聚合商效益 R_{agg}	电网效益 R_{grid}	社会总福利 SW
初始状态	4295.24	5425.86	7194.27	16915.38
最终状态	3116.70	8333.43	9340.5	20790.63
变化率	-27.43%	53.59%	29.83%	22.91%

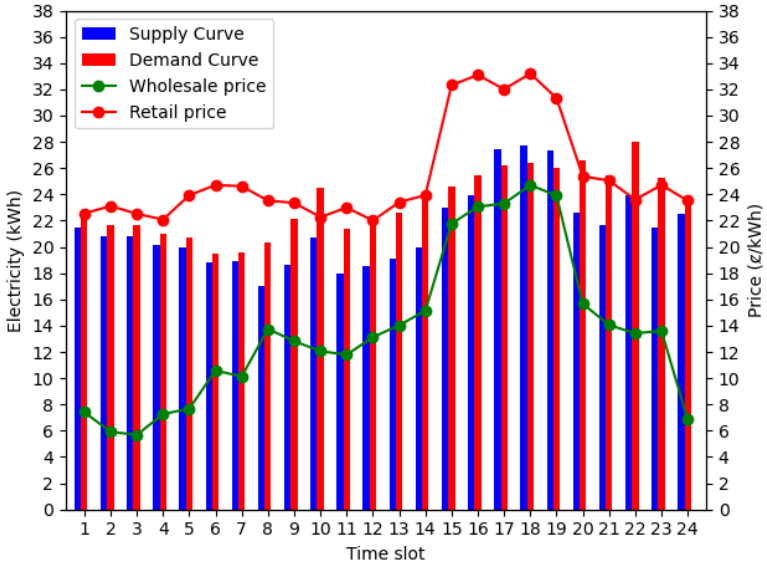

图 3-15 MADDPG-RO 求解所得市场终止状态

对于电网来说，需求高峰的转移使得其电力供给曲线变得更加平滑，调整电网拓扑结构的频率和电量减少，大大降低了其运营成本，同时，在电力需求高峰

表 3-3 COMA 社会福利最大化目标下的前后收益对比

状态	用户效益 R_{user}	电力聚合商效益 R_{agg}	电网效益 R_{grid}	社会总福利 SW
初始状态	4295.24	5425.86	7194.27	16915.38
最终状态	4004.3	7432.26	9152.5	20588.06
变化率	-6.77%	36.97%	27.22%	21.71%

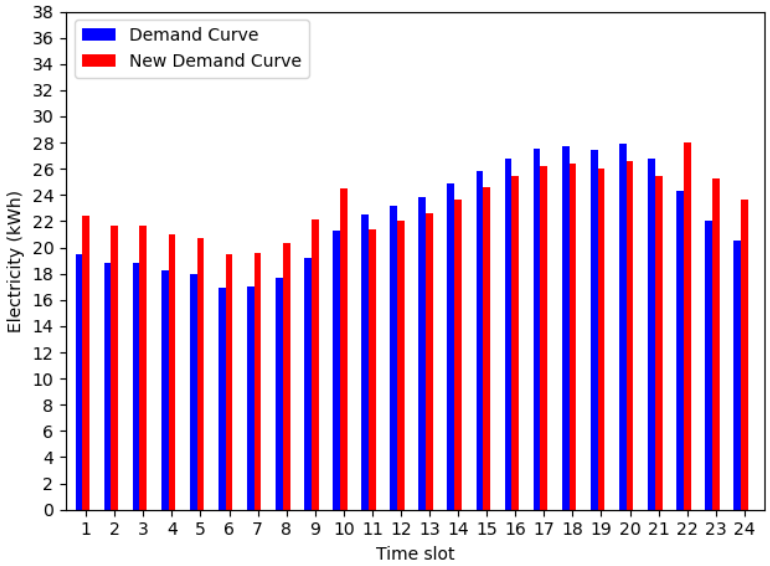

图 3-16 市场初始状态与最终状态需求曲线对比

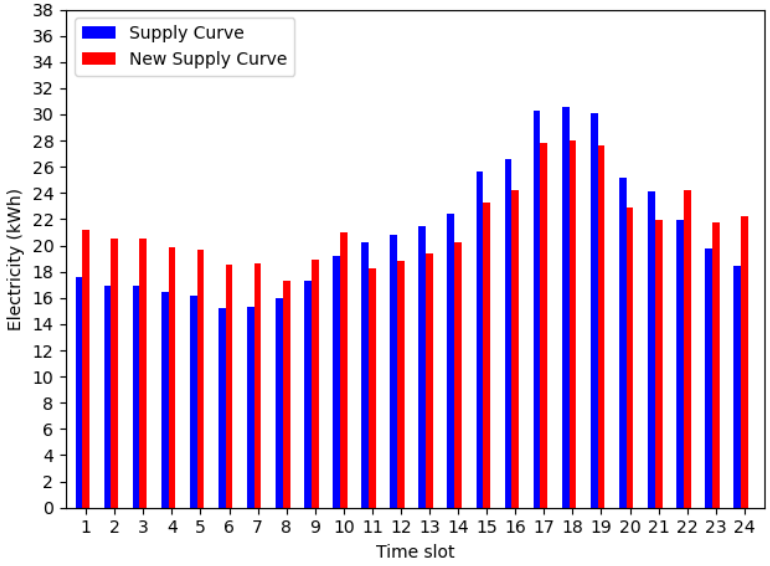

图 3-17 市场初始状态与最终状态供给曲线对比

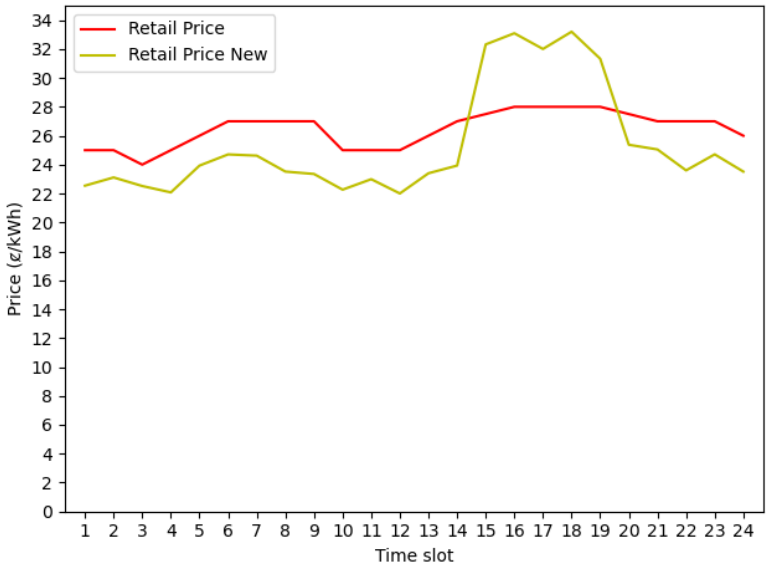

图 3-18 市场初始状态与最终状态零售价格对比

图 3-19 市场初始状态与最终状态批发价格对比

时段提高批发价格提升了其售电收入，总体提升了电网效益。

对于电力聚合商来说，降低电力需求低谷时段的零售价格，提升电力需求高峰时段的价格使其售电收入大大增加。

3.5 本章小结

本章以社会福利最大化为目标，设计多智能体强化学习算法 MADDPG-RO 求解三方博弈策略，最终使得社会总福利上升 22.91%

第四章 各方收益最大化下的三方策略均衡点求解

与上一章相反，本章将考虑一个理想化的假设：电动汽车充电市场中，三方参与者各自制订策略使得自身收益最大化并达到市场的纳什均衡状态，即任意一方改变策略都不会使得任意一方收益增加的状态。三方参与者拥有各自的目标函数，通过调整自身行动最大化自身收益，是完全竞争的智能体。本章工作采用多智能体强化学习算法求解市场最终状态。

4.1 各方收益最大化假设下的电动汽车充电市场强化学习模型

4.1.1 马尔可夫博弈描述

与前文类似，本文采用马尔可夫博弈将描述三方博弈电动汽车充电市场：

$$(\boldsymbol{S}, \boldsymbol{A}_{user}, \boldsymbol{A}_{agg}, \boldsymbol{A}_{grid}, \gamma, R_{user}, R_{agg}, R_{grid})$$

由于本章研究内容为完全竞争下的各方策略，所以参与市场三方主体不再共享目标函数，而是制定策略以最大化自身的收益，达到市场均衡。

即该假设下，各智能体的关系为竞争关系，每个智能体的目标为追求自身利益最大化，最终达到市场均衡。

4.1.2 算法流程

该问题定义下，智能体只能观测到部分信息，状态、动作空间为连续动作空间，本章仍利用 MADDPG 算法中心化训练，去中心化执行的特性将其作为求解算法，并考虑到智能体参与博弈顺序的随机性，仍采用 MADDPG-RO 算法求解，描述如算法4-1。为了比较不同目标设定下的结果，本章仍采用前述状态作为市场初始状态，本章不做赘述。MADDPG-RO 算法参数设定如表4-1。

此外，根据本章智能体奖励的设定，本章将设置对比算法 IQL，为每个智能体设置各自的奖励函数为其参与市场的收益。最后，本章将对两种算法结果进行对比。

算法 4-1 各方收益最大化假设下的 MADDPG-RO 算法流程

for *episode = 1 to M* **do**
  为动作探索生成随机过程 $\mathcal{N}$;
  随机打乱智能体行动列表的顺序 [user, aggregator, grid];
  从环境中得到初始状态 s;
  for *t=1 to max-episode-length* **do**
    对于每个智能体 i, 根据当前策略与探索情况选择动作，得到联合动作 ;
    执行联合动作 $a = (a_1, a_2, a_3)$;
    从环境中得到奖励 $r = (r_1, r_2, r_3)$ 和新时间步的状态 s';
    将四元组 (s, a, r, s') 存储到经验回放池中 $\mathcal{D}$;
    $s \leftarrow s'$;
    for *agent i=1,2,3* **do**
      从经验池 $\mathcal{D}$ 中随机采样 S 个样本 (s^j, a^j, r^j, s'^j);
      通过最小化如下损失函数更新 Critic $\mathcal{L}(\theta_i)$;
      利用采样的梯度策略更新 Actor;
    end
    对每个智能体 i 更新策略网络参数 θ_i;
  end
end

表 4-1 MADDPG-RO 参数设定

参数	赋值
最大迭代步数 max episode length	30000
每步最大步长 max length per episode	50
随机探索概率 ε	0.3
经验回放池大小 $\mathcal{D}$	6400
经验池取样大小 batch size	64
远期回报折扣系数 γ	0.95
学习率 learning rate	0.01

4.2 算例分析

4.2.1 充电汽车用户效益

电动汽车充电用户在训练过程中奖励的变化如图4-1所示。在训练过程的前5000 步，用户主体的收益在原地大幅波动，在 5000 步到 25000 步消费者效益大幅上升，最终收益曲线震荡幅度减小，收敛到 4750 左右。

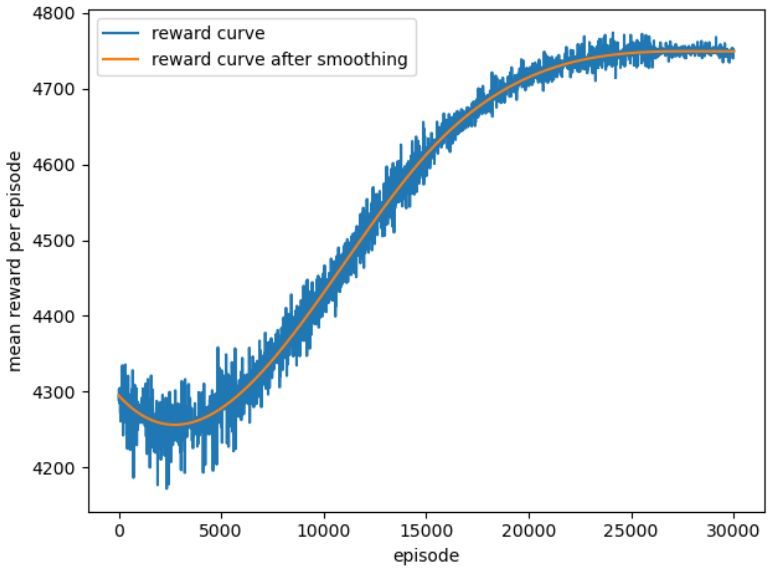

图 4-1 MADDPG-RO 训练过程中充电汽车用户效益

4.2.2 电力聚合商效益

电力聚合商在训练过程中奖励的变化如图4-2所示。在训练过程的前 15000 步，其收益曲线大幅震荡上升. 在 15000 到 25000 步，光滑后的曲线趋于平缓，但奖励震荡幅度仍然较大。早 25000 步之后，奖励震荡幅度减少，曲线趋于收敛，最终电力聚合商收益收敛到 8300 左右。

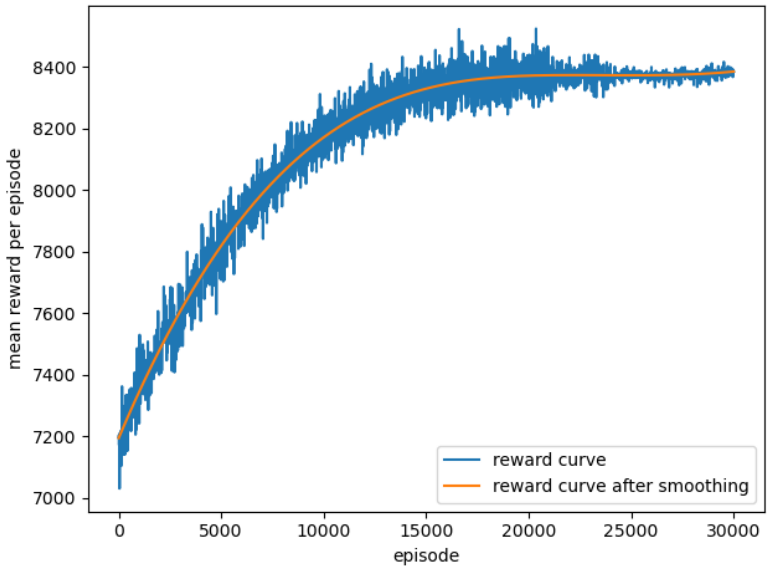

图 4-2 MADDPG-RO 训练过程中电力聚合效益

4.2.3 电网效益

电网在训练过程中奖励的变化如图4-3所示。在前 7000 步，电网的收益大幅震荡，逐渐增加。在 7000 步至 20000 步，收益曲线小幅增加.20000 步以后趋于收敛，最终收敛到 6700 左右。

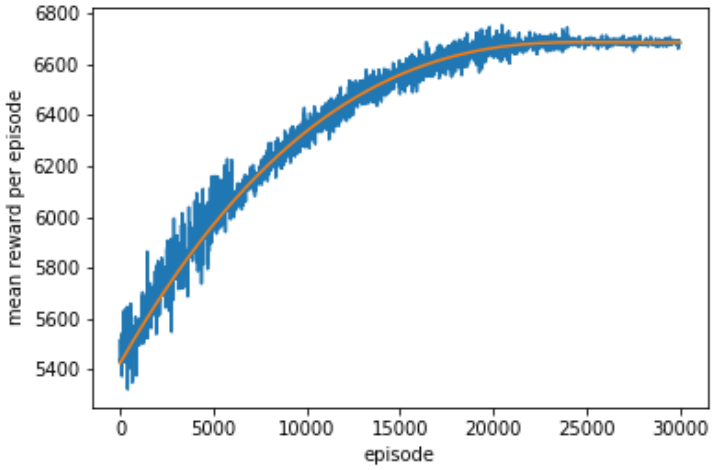

图 4-3 MADDPG-RO 训练过程中电网效益

4.2.4 社会总福利

社会总福利在训练过程中的变动如图4-4所示，是三方参与者收益的加和。在训练过程的前 10000 步左右，社会总福利逐渐波动上升，最终收敛于 18000 左右。可以看出消费者效用增量较少，三方参与者的收益均在上升。

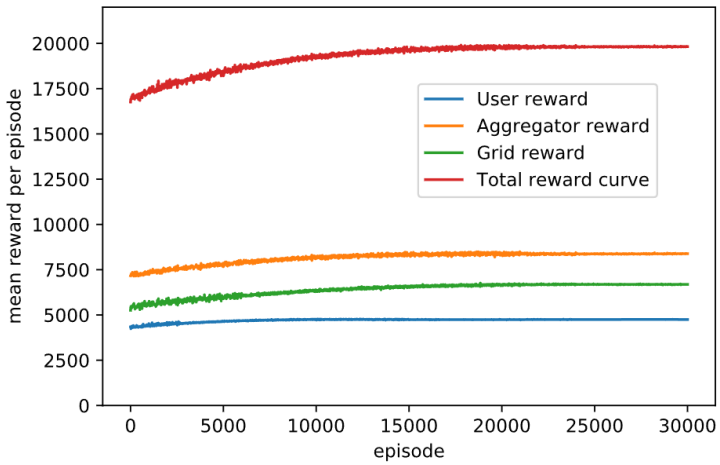

图 4-4 MADDPG-RO 算法训练过程中社会总福利

4.2.5 IQL 算法对比试验结果

使用 COMA 算法得到的聚合商收益，电网收益，用户收益曲线如图4-5，4-6，4-7。与 MADDPG-RO 算法的结果类似，三方收益均波动上升趋于稳定。使用 iql 算法得到的社会总福利曲线变动如图4-8

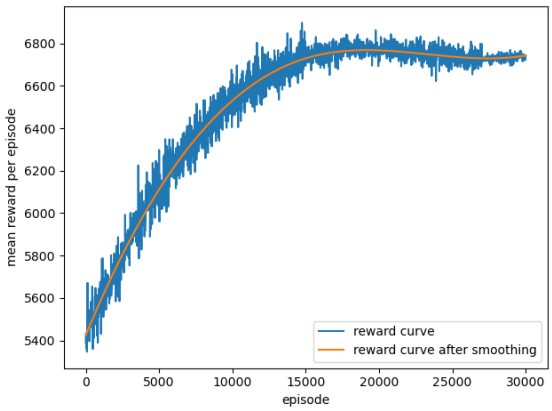

图 4-5 IQL 算法的聚合商收益变动

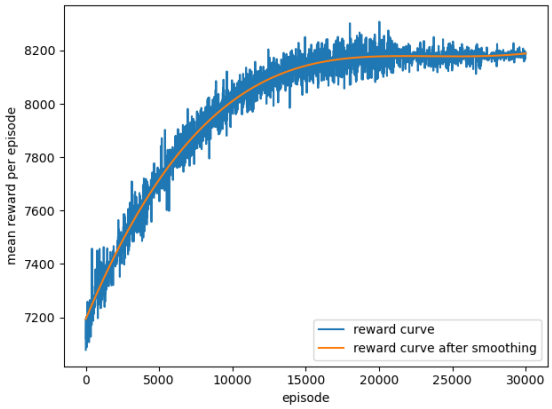

图 4-6 IQL 算法的电网收益变动

4.3 结论

在本文三方博弈框架中，采用 MADDPG-MO 算法求解以各方收益最大化为优化目标的条件下三方主体博弈策略，结果如表4-2所示。在社会福利最大化目标下，MADDPG-MO 算法得到的策略使得社会总福利提高了 17.19%，其中，电网收益增幅最多，达到 23.27%，电力集成商收益增加 16.56%，用户效益增加 10.57%。最终市场状态如图4-9。

采用 IQL 算法算法求解以各方收益最大化为优化目标的条件下三方主体博弈

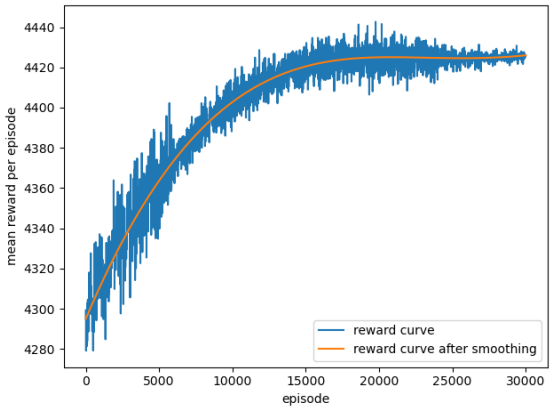

图 4-7 IQL 算法的用户收益变动

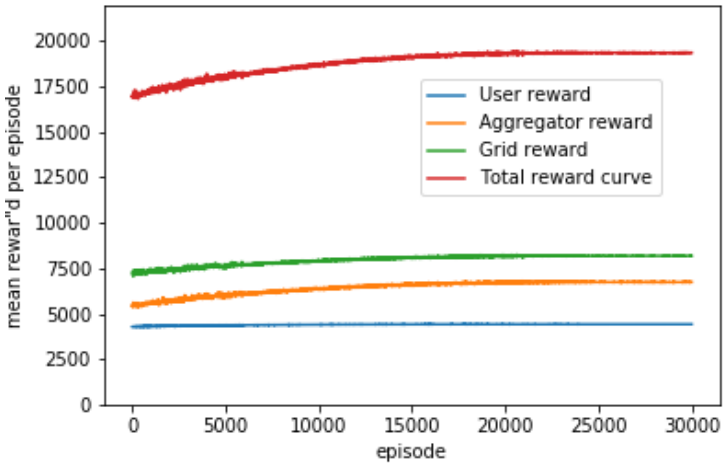

图 4-8 IQL 算法的社会总福利变动

策略，结果如表4-3所示。在社会福利最大化目标下，IQL 算法得到的策略使得社会总福利提高了 14.44%，其中，电力聚合商收益增幅最多，达到 24.3%，电网收益增加 13.83%，用户效益增加 3.05%。

需求曲线 ***DC*** 变动如图4-10，在各方收益最大化假设下，充电市场用户仍然选择将自己在需求高峰期的小部分需求转移到其他非需求高峰期。相比社会总福利最大化的求解结果，虽然此时零售电力价格上涨，但涨幅并不大，用户为了避免充电习惯改变带来的负效用，在考虑零售电价的情况下对出行曲线稍作更改。

供给曲线 ***SC*** 变动如图4-11，由于客户需求变动，电网可以选择调整自己的发电策略，在高峰期减少小部分发电量的供给，在非高峰期增加一部分电力供给。

零售价格曲线 ***RP*** 变动如图4-12，电力集成商在高峰时段提高电力零售价格，在其他所有非高峰时段均小幅降低零售价格使用户在该时段可以选择更多的充电

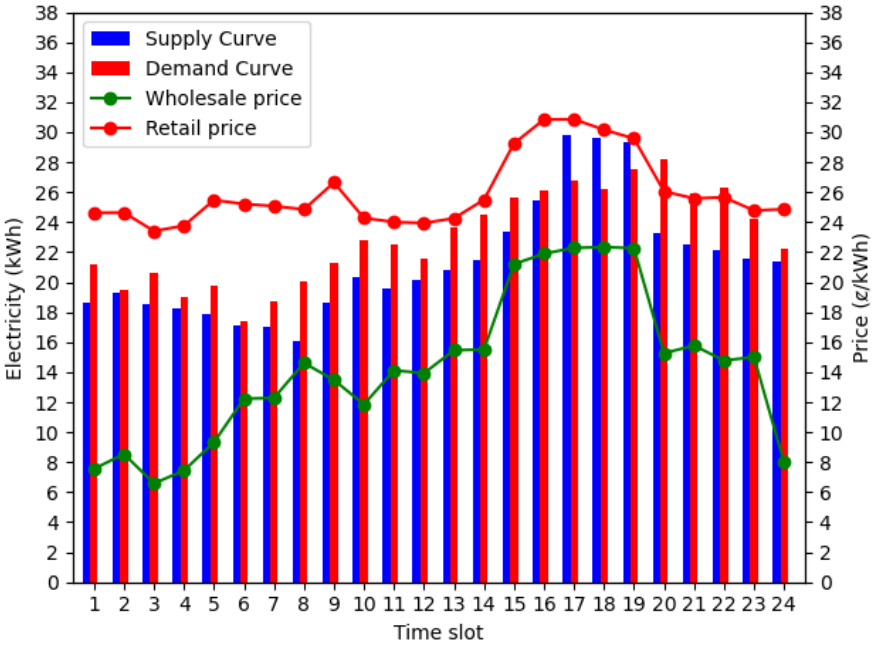

图 4-9 MADDPG-RO 求解所得市场终止状态

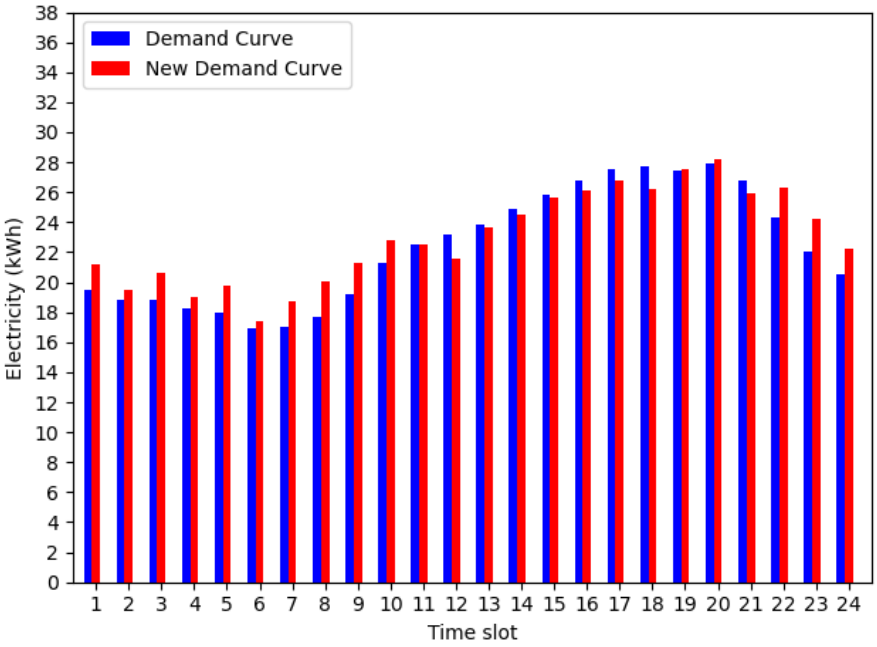

图 4-10 市场初始状态与最终状态需求曲线对比

服务，同时保证自身盈利水平。

批发价格曲线 ***WP*** 变动如图4-13，电网通过平衡供给曲线降低了成本，整体电价水平降低。在需求高峰时刻通过提高批发价格促使零售价格上升，使消费者降低该时间段的购买需求。

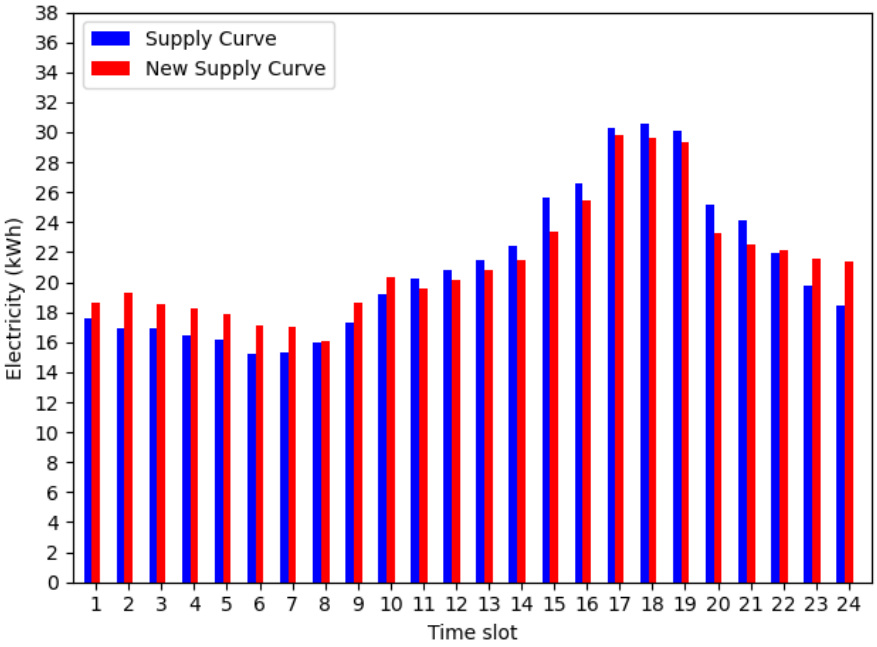

图 4-11 市场初始状态与最终状态供给曲线对比

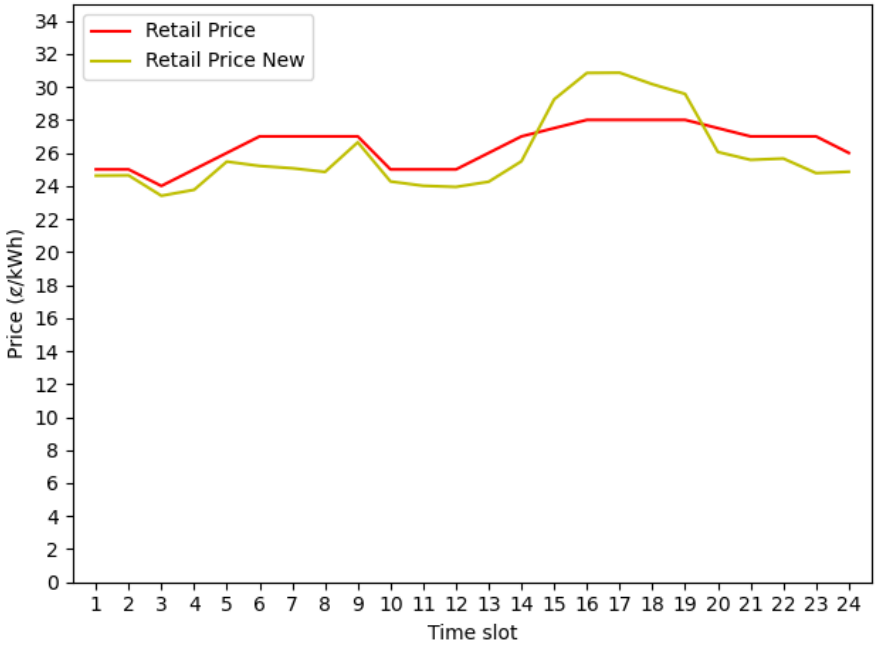

图 4-12 市场初始状态与最终状态零售价格对比

与第三章结论类似，MADDPG-MO 策略使得电力需求高峰时刻的零售价格与批发价格上升，该时段的用户需求下降，用户选择将在该部分的充电需求转移到其他时刻。然而，与社会福利最大化目标不同，在三方收益最大化的假设下，用户并不会为了最小化社会福利损失降低自身收益，这表现在面对零售价格变动时，

图 4-13 市场初始状态与最终状态批发价格对比

用户考虑自身出行习惯改变带来的不满意度以及在需求高峰时刻充电的效用，对价格提升的反应并不明显，只是略微降低需求，在保证自身收益的情况下合理分配自己的需求曲线。

对于电网来说，需求曲线的峰荷与峰谷的变动幅度不如社会福利最大化假设下的变动幅度，因此，为了保证满足需求，其供给曲线也相应减少了变化幅度。电网虽然仍能由于更加平滑的供给曲线降低成本，但收益幅度大大下降了。

对于电力聚合商来说，相比社会福利最大化假设，用户并不会根据零售曲线大幅减少高峰时刻的充电需求，因此，聚合商制订零售价格时仅需要对非高峰时间的价格略微调整，保证自身收益。

表 4-2 MADDPG-RO 算法求解下三方收益最大化目标下的前后收益对比

状态	用户效益	电力聚合商效益	电网效益	社会总福利
初始状态	4295.24	5425.86	7194.27	16915.38
最终状态	4749.36	6688.74	8385.5	19823.56
变化率	10.57%	23.27%	16.56%	17.19%

表 4-3 IQL 算法求解下三方收益最大化目标下的前后收益对比

状态	用户效益	电力聚合商效益	电网效益	社会总福利
初始状态	4295.24	5425.86	7194.27	16915.38
最终状态	4426.54	6743.28	8189.44	19358.23
变化率	3.05%	24.3%	13.83%	14.44%

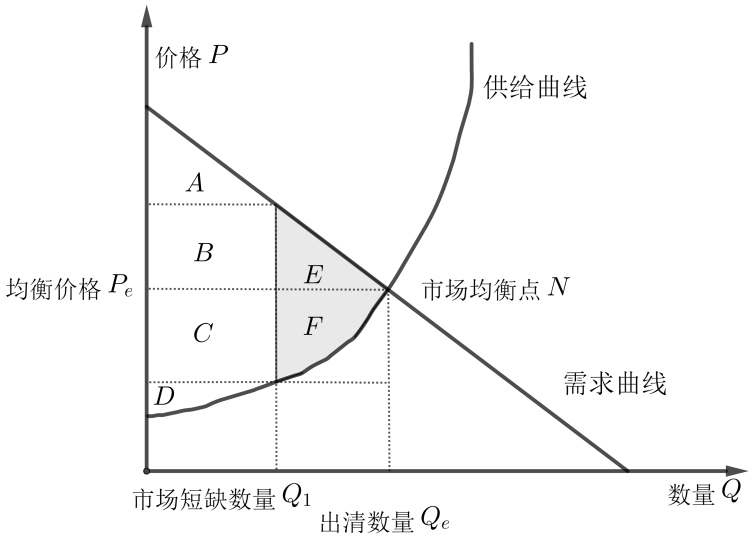

图 4-14 供给短缺下的市场状态

4.4 不同目标下的市场均衡状态对比及微观经济学分析

在社会福利最大化为目标的市场下，MADDPG-RO 算法求解消费者自身收益折损 27.43%，但电网收益得到 53.59% 的大幅度提升，社会总福利提升 22.91%。在三方主体各自最大化自身收益为目标时，消费者收益提升 10.57%，社会总福利提升 17.19%。以下从微观经济学角度分析该结果可能出现的原因。

在本章以各自利益最大化为目标的完全竞争市场中，生产者利润最大化，消费者追求效用最大化。生产者通过改变定价或是决定产量影响市场状态，消费者通过支付货币完成对商品的投票。在双方博弈的过程中，市场会逐渐达到均衡状态，任意一方单独改变策略都不会使任意一方的境况变得更好。如图4-14，假设生产者选择生产图中图中 Q_1 数量的商品，此时市场处于供不应求的状态，消费者将会购买全部数量的商品并获得消费者剩余 A, B, C，生产者获得生产者剩余 D，此时社会福利损失为 $E+F$。然而，此时并没有到达市场出清的状态，生产者若果选择生产更多数量的商品，其生产者剩余三角形 D 会继续扩大，直到到达市场均衡点 N，此时消费者剩余为 $A+B+E$，生产者剩余为 $C+D+F$ 市场出清，消费者

选择消费商品的数量与生产者生产的数量相等。

在电力交易市场中，生产者只有电网一方的情况下，其余各方都均为消费者。在第四章各方收益最大化模型假设下，市场处于完全竞争状态，由于消费者改变需求曲线的分布会带来利益损耗，在市场价格不够低的情况下仍选择将需求集中在原有需求分布上，电网只能在保持供给曲线的分布下对电力批发价格曲线调整以追求自身利润。在该假设下，市场由原状态向均衡状态移动，各方利益均上升，达到均衡点时社会福利最大化。

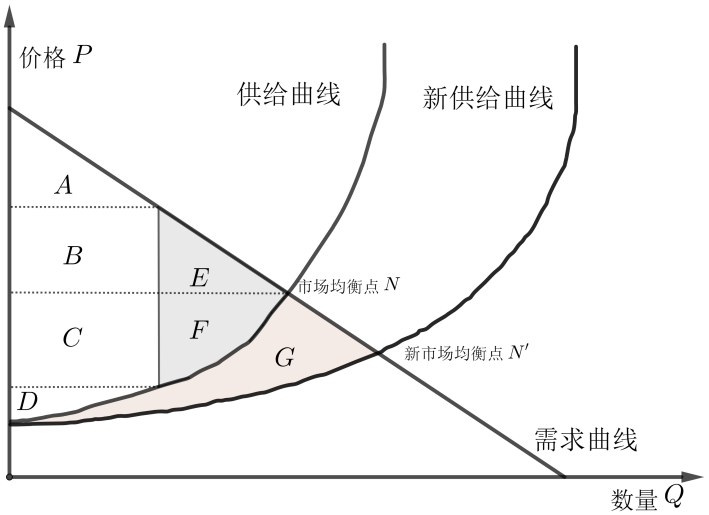

图 4-15 供给曲线移动后的市场状态

然而在第三章社会福利最大化的市场机制假设下，用户愿意为了社会福利最大化让渡收益，电网可以通过改变自身发电曲线直接降低运行成本，改变市场的均衡状态. 如图4-15，生产者的单位商品生产成本降低将导致整体供给曲线下移，市场达到新的均衡点 N'，此时市场比原市场的社会福利多了 G 的面积，总体社会福利增加。

4.5 本章小结

本章模型以三方参与者收益最大化为目标，设计竞争性的多智能体强化算法求解三方博弈策略，最终参与三方收益均有提升，电动汽车用户收益增加了 10.57%，电力聚合商收益增加了 23.27%，电网收益增加 16.56%，社会总福利增加 17.19%。同时，本章通过微观经济学理论分析结果。

第五章 全文总结与展望

5.1 全文总结

本文以电动汽车充电市场为研究背景，搭建了一个电动汽车充电市场的三方交易模型，对三方参与者的特点及其对市场的交互行为建模，规定每一方对市场的观测，其行为对环境造成的影响以及环境中获得的奖励，并依据现实情况设定约束条件。本文设定社会福利最大化和各方收益最大化的目标函数，基于改进的MADDPG 算法求解这两种目标函数下的各方最优策略及市场均衡，最终分别得到了 22.91% 和 17.19% 的社会福利增加，对结果进行比对和分析。

本文用微观经济学视角对本文结论进行分析得到结论：如果三方共同目标使追求社会福利最大化，博弈主体以合作方式参与市场互动，电动汽车用户会选择折损自身利益对需求做出较大的变动，将高峰的充电需求迁移至非高峰时段，使电网发电曲线趋于平缓，降低发电成本，电网的供给曲线下移，形成新的市场均衡点，带来更多的社会总福利。如果目标是三方各自追求自己的利益最大化，用户会根据价格小幅调整需求变动，不会选择将许多高峰时刻的需求改变，三方只能追求当前市场的均衡点，三方收益均有上升，但社会总福利变动小于前假设情况。

5.2 后续工作展望

多智能体博弈、深度强化学习在智能电网、电力市场机制设计领域都有很大的研究潜力。许多问题值得进一步探索，本文提出以下展望：

（1）本文搭建的三方模型为高度抽象的博弈论参与者模型，是一个群体的平均表示。然而，群体策略并不完全适用于个人策略，举例来说，电动汽车充电用户对于充电服务的评价不同，不同特征的用户如果采用相同的策略参与市场会带来很大的损失。因此，如何将本文模型细化到不同特征的参与者，将参与博弈的主体增加，设计各自参与市场的策略，是值得研究的方向。

（2）本文设定各方在每一个时间步都会参与博弈。如果存在某一个智能体在某一步采取了非常差的策略，会导致整个市场的无效率。因此，在博弈模型之外单独设计市场管控模型可能会更好引导博弈主体参与市场，提升市场的效率。

（3）本文设定三方参与者之间出于隐私保护考虑，不存在沟通，每一方只能从当前市场状态获取信息。联邦学习的特性可以结合强化学习的优点被应用到市场建模中。如果将博弈模型设计成联邦学习的模式，每方主体仅上传其策略网络参数，由决策中心更新主体网络并下发至各个参与主体，这样能提高智能电网交

易中的安全性和隐私性。

致 谢

在攻读硕士学位期间，首先衷心感谢我的导师张彦如教授，她在科研上给了我很大的帮助，从论文选题到最终完成，她始终耐心地指导启发我，我所取得的成就与张彦如老师的无私关怀分不开。张彦如教授学识渊博且为人和蔼可亲，每次组会上都会分享自己在学习和工作中的见解，在生活中跟组里的小伙伴关系都很好，经常组织各种课余活动，比如办 homeparty、滑雪、吃火锅、吃柴火鸡、吃烤鱼等。在张彦如教授的指导下，我不仅取得了现在的成绩，还度过了愉快的三年研究生生涯。在此向张彦如教授致以衷心的感谢。

感谢 IntelliGame 实验室的各位成员。王岩博士后从我进组对未来感到迷茫的时候以他丰富的人生经验指导我走出迷雾。在各类竞赛中，我从陈维龙师兄和李金豪师弟身上学到了许多东西，他们是对我代码能力提升贡献最大的两个人，也是我的榜样。感谢我的舍友庄岩，他养的布偶猫虽然只陪伴了我 4 个月，之后就送到了他的女朋友白薇家里，但在这四个月里我意识到了除了科研，生活也一样重要。黄承浩师弟和陈典师弟和我互相鼓励，每次在我碰到困难的时候都能互相鼓励对方。IntelliGame 实验室的各位成员无论是在科研还是生活上都在我三年研究生生涯中留下了很多的影响。

感谢父亲张振宇和母亲刘霞。虽然他们不在我身边，但我随时都能感受到他们在大多数方面对我的支持和理解。正是有了坚强的后盾，我才能在至今 19 年的求学生涯全身心投入而没有后顾之忧。如果没有他们，我至今为止的所有成就将毫无意义。

由衷感谢百忙之中审阅、评议这份论文的各位老师！

以上，感谢硕士三年遇到的大家！在你们的关心和帮助下，我感受到了生活的乐趣。今后我会更加努力学习、认真投入到下一个人生阶段。祝你们身体健康，家庭和睦！

参考文献

[1] Cao Y, Tang S, Li C, et al. An optimized ev charging model considering tou price and soc curve[J]. IEEE Transactions on Smart Grid, 2011, 3(1): 388-393.

[2] Lopes J A P, Soares F J, Almeida P M R. Integration of electric vehicles in the electric power system[J]. Proceedings of the IEEE, 2010, 99(1): 168-183.

[3] Yilmaz M, Krein P T. Review of battery charger topologies, charging power levels, and infrastructure for plug-in electric and hybrid vehicles[J]. IEEE transactions on Power Electronics, 2012, 28(5): 2151-2169.

[4] Ardakanian O, Rosenberg C, Keshav S. Distributed control of electric vehicle charging[C]. Proceedings of the fourth international conference on Future energy systems, 2013: 101-112.

[5] Smith A. The wealth of nations: An inquiry into the nature and causes of the wealth of nations[M]. Harriman House Limited, 2010.

[6] Keynes J M. The general theory of employment[J]. The quarterly journal of economics, 1937, 51(2): 209-223.

[7] He X, Chu L, Qiu R C, et al. A novel data-driven situation awareness approach for future grids—using large random matrices for big data modeling[J]. IEEE Access, 2018, 6: 13855-13865.

[8] He X, Ai Q, Qiu R C, et al. A big data architecture design for smart grids based on random matrix theory[J]. IEEE transactions on smart Grid, 2015, 8(2): 674-686.

[9] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. MIT press, 2018.

[10] Zhao X, Xia L, Zhang L, et al. Deep reinforcement learning for page-wise recommendations[C]. Proceedings of the 12th ACM Conference on Recommender Systems, 2018: 95-103.

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[12] Yu C, Liu J, Nemati S, et al. Reinforcement learning in healthcare: A survey[J]. ACM Computing Surveys (CSUR), 2021, 55(1): 1-36.

[13] Zhang D, Han X, Deng C. Review on the research and practice of deep learning and reinforcement learning in smart grids[J]. CSEE Journal of Power and Energy Systems, 2018, 4(3): 362-370.

[14] Tang X, Qin Z, Zhang F, et al. A deep value-network based approach for multi-driver order dispatching[C]. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019: 1780-1790.

[15] Zhang Z, Zhang D, Qiu R C. Deep reinforcement learning for power system applications: An overview[J]. CSEE Journal of Power and Energy Systems, 2019, 6(1): 213-225.

[16] Wang Q, Liu X, Du J, et al. Smart charging for electric vehicles: A survey from the algorithmic perspective[J]. IEEE Communications Surveys & Tutorials, 2016, 18(2): 1500-1517.

[17] Waraich R A, Galus M D, Dobler C, et al. Plug-in hybrid electric vehicles and smart grids: Investigations based on a microsimulation[J]. Transportation Research Part C: Emerging Technologies, 2013, 28: 74-86.

[18] Mehta R, Srinivasan D, Trivedi A. Optimal charging scheduling of plug-in electric vehicles for maximizing penetration within a workplace car park[C]. 2016 IEEE Congress on Evolutionary Computation (CEC), 2016: 3646-3653.

[19] Rotering N, Ilic M. Optimal charge control of plug-in hybrid electric vehicles in deregulated electricity markets[J]. IEEE Transactions on Power Systems, 2010, 26(3): 1021-1029.

[20] Iversen E B, Morales J M, Madsen H. Optimal charging of an electric vehicle using a markov decision process[J]. Applied Energy, 2014, 123: 1-12.

[21] Limmer S. Dynamic pricing for electric vehicle charging—a literature review[J]. Energies, 2019, 12(18): 3574.

[22] Chen T, Su W. Local energy trading behavior modeling with deep reinforcement learning[J]. Ieee Access, 2018, 6: 62806-62814.

[23] Chen T, Su W. Indirect customer-to-customer energy trading with reinforcement learning[J]. IEEE Transactions on Smart Grid, 2018, 10(4): 4338-4348.

[24] Wang H, Huang T, Liao X, et al. Reinforcement learning in energy trading game among smart microgrids[J]. IEEE Transactions on Industrial Electronics, 2016, 63(8): 5109-5119.

[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] Xiao X, Dai C, Li Y, et al. Energy trading game for microgrids using reinforcement learning[C]. International Conference on Game Theory for Networks, 2017: 131-140.

[27] Salehizadeh M R, Soltaniyan S. Application of fuzzy q-learning for electricity market modeling by considering renewable power penetration[J]. Renewable and Sustainable Energy Reviews, 2016, 56: 1172-1181.

[28] Xu H, Sun H, Nikovski D, et al. Deep reinforcement learning for joint bidding and pricing of load serving entity[J]. IEEE Transactions on Smart Grid, 2019, 10(6): 6366-6375.

[29] Forte V J. Smart grid at national grid[C]. 2010 Innovative Smart Grid Technologies (ISGT), 2010: 1-4.

[30] Potter C W, Archambault A, Westrick K. Building a smarter smart grid through better renewable energy information[C]. 2009 IEEE/PES Power Systems Conference and Exposition, 2009: 1-5.

[31] Vos A. Effective business models for demand response under the smart grid paradigm[C]. 2009 IEEE/PES Power Systems Conference and Exposition, 2009: 1-1.

[32] Zhong J, Kang C, Liu K. Demand side management in china[C]. IEEE PES General Meeting, 2010: 1-4.

[33] Saffre F, Gedge R. Demand-side management for the smart grid[C]. 2010 IEEE/IFIP Network Operations and Management Symposium Workshops, 2010: 300-303.

[34] Mohsenian-Rad A H, Wong V W, Jatskevich J, et al. Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid[J]. IEEE transactions on Smart Grid, 2010, 1(3): 320-331.

[35] Oh H, Thomas R J. Demand-side bidding agents: Modeling and simulation[J]. IEEE Transactions on Power Systems, 2008, 23(3): 1050-1056.

[36] Nguyen D. Demand response for domestic and small business consumers: A new challenge[C]. IEEE PES T&D 2010, 2010: 1-7.

[37] York D, Kushler M. Exploring the relationship between demand response and energy efficiency: A review of experience and discussion of key issues[C]. ACEEE, 2005: 35-44.

[38] Vázquez-Canteli J R, Nagy Z. Reinforcement learning for demand response: A review of algorithms and modeling techniques[J]. Applied energy, 2019, 235: 1072-1089.

[39] Siano P. Demand response and smart grids—a survey[J]. Renewable and sustainable energy reviews, 2014, 30: 461-478.

[40] Lu R, Hong S H. Incentive-based demand response for smart grid with reinforcement learning and deep neural network[J]. Applied energy, 2019, 236: 937-949.

[41] Hao J. Multi-agent reinforcement learning embedded game for the optimization of building energy control and power system planning[J]. arXiv preprint arXiv:1901.07333, 2019.

[42] Ghasemkhani A, Yang L. Reinforcement learning based pricing for demand response[C]. 2018 IEEE International Conference on Communications Workshops (ICC Workshops), 2018: 1-6.

[43] Remani T, Jasmin E, Ahamed T I. Residential load scheduling with renewable generation in the smart grid: A reinforcement learning approach[J]. IEEE Systems Journal, 2018, 13(3): 3283-3294.

[44] Chiş A, Lundén J, Koivunen V. Reinforcement learning-based plug-in electric vehicle charging with forecasted price[J]. IEEE Transactions on Vehicular Technology, 2016, 66(5): 3674-3684.

[45] Wang H, Li C, Li J, et al. A survey on distributed optimisation approaches and applications in smart grids[J]. Journal of Control and Decision, 2019, 6(1): 41-60.

[46] Shi W, Li N, Chu C C, et al. Real-time energy management in microgrids[J]. IEEE Transactions on Smart Grid, 2015, 8(1): 228-238.

[47] Zachar M, Daoutidis P. Microgrid/macrogrid energy exchange: A novel market structure and stochastic scheduling[J]. IEEE Transactions on Smart Grid, 2016, 8(1): 178-189.

[48] Ordoudis C, Pinson P, Morales J M. An integrated market for electricity and natural gas systems with stochastic power producers[J]. European Journal of Operational Research, 2019, 272(2): 642-654.

[49] Zéphyr L, Anderson C L. Stochastic dynamic programming approach to managing power system uncertainty with distributed storage[J]. Computational Management Science, 2018, 15(1): 87-110.

[50] Duchaud J L, Notton G, Darras C, et al. Power ramp-rate control algorithm with optimal state of charge reference via dynamic programming[J]. Energy, 2018, 149: 709-717.

[51] Nguyen H T, Le L B, Wang Z. A bidding strategy for virtual power plants with the intraday demand response exchange market using the stochastic programming[J]. IEEE Transactions on Industry Applications, 2018, 54(4): 3044-3055.

[52] Mahmutoğulları A İ, Ahmed S, Çavuş Ö, et al. The value of multi-stage stochastic programming in risk-averse unit commitment under uncertainty[J]. IEEE Transactions on Power Systems, 2019, 34(5): 3667-3676.

[53] Megantoro P, Wijaya F D, Firmansyah E. Analyze and optimization of genetic algorithm implemented on maximum power point tracking technique for pv system[C]. 2017 international seminar on application for technology of information and communication (iSemantic), 2017: 79-84.

[54] Srikakulapu R, Vinatha U. Optimized design of collector topology for offshore wind farm based on ant colony optimization with multiple travelling salesman problem[J]. Journal of Modern Power Systems and Clean Energy, 2018, 6(6): 1181-1192.

[55] Li H, Yang D, Su W, et al. An overall distribution particle swarm optimization mppt algorithm for photovoltaic system under partial shading[J]. IEEE Transactions on Industrial Electronics, 2018, 66(1): 265-275.

[56] Gu H, Yan R, Saha T K. Minimum synchronous inertia requirement of renewable power systems[J]. IEEE Transactions on Power Systems, 2017, 33(2): 1533-1543.

[57] Terry J K, Black B, Grammel N, et al. Pettingzoo: Gym for multi-agent reinforcement learning[J]. arXiv preprint arXiv:2009.14471, 2020.

[58] Jin M, Feng W, Marnay C, et al. Microgrid to enable optimal distributed energy retail and end-user demand response[J]. Applied Energy, 2018, 210: 1321-1335.

[59] Faria P, Vale Z, Soares J, et al. Demand response management in power systems using particle swarm optimization[J]. IEEE Intelligent Systems, 2011, 28(4): 43-51.

[60] Valero S, Ortiz M, Senabre C, et al. Methods for customer and demand response policies selection in new electricity markets[J]. IET generation, transmission & distribution, 2007, 1(1): 104-110.

[61] Myerson R B. Game theory: analysis of conflict[M]. Harvard university press, 1997.

[62] Yang Y, Wang J. An overview of multi-agent reinforcement learning from game theoretical perspective[J]. arXiv preprint arXiv:2011.00583, 2020.

[63] Foerster J, Farquhar G, Afouras T, et al. Counterfactual multi-agent policy gradients[C]. Proceedings of the AAAI conference on artificial intelligence, 2018.

[64] Lowe R, Wu Y I, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in neural information processing systems, 2017, 30.

[65] Tampuu A, Matiisen T, Kodelja D, et al. Multiagent cooperation and competition with deep reinforcement learning[J]. PloS one, 2017, 12(4): e0172395.

[66] Brownrigg D R. The weighted median filter[J]. Communications of the ACM, 1984, 27(8): 807-818.

攻读专业硕士学位期间取得的成果

[1] (5/6). Plug-in electric vehicle charging with multiple charging options: A systematic analysis of service providers’ pricing strategies[J]. IEEE Transactions on Smart Grid, 2020, 12(1):524-537. **JCR-1**

[2] (2/6).Deep Reinforcement Learning for Optimal Power Flow with Renewables Using Spatial-Temporal Graph Information[J]. IEEE J. Selected Areas on Communications(submitted)