Empowering Peer-to-peer Energy Trading in Smart Grid via Deep Reinforcement Learning

Tianyi Chen

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

August 2022

Abstract

Electricity is traditionally generated in large, centralised power plants, resulting in high transmission costs and high emissions. Recently, small-scale renewable generation has become more and more popular due to the low carbon energy policy. Microgrids (MG) have been employed to address the challenges arising from the presence of a high share of distributed energy resources in local regions of modern energy systems. Furthermore, the peer-to-peer (P2P) energy trading paradigm with its improved system efficiency and reduced greenhouse gas emissions benefits the MGs more than traditional energy trading strategies. Considering the intermittent nature of renewable generation and hard-predicted local consumption patterns, a P2P energy trading system must cope with uncertainty, scale, and reliability issues in real-time. In order to ensure a fast and optimised energy trading operation and settlement, an automated decision-making system is necessary. However, traditional optimisation methodologies may not be able to produce results in a real-time manner when dealing with large data sets and an increased level of uncertainty in P2P energy trading schemes. Recent energy coupling technologies can be integrated by different power carriers to form a multi-energy microgrid (MEMG), resulting in economic and environmental benefits. A MEMG consists of DERs, energy coupling technologies, local active loads and energy storage systems (ESSs). By connecting multiple MEMGs, the distribution network can be made more efficient and reliable. In addition to the challenges posed by the intermittent nature of DERs, there are also additional obstacles related to the stability and operational safety of the network of multiple MEMGs that vary according to the deployment, including the size and type of DERs.

To address the above challenges, this thesis utilises Deep reinforcement learning (DRL), as a decision-making learning algorithm, to automatically derive optimal P2P energy trading policies for MGs participating in a local energy trading market. Furthermore, this thesis investigates the external P2P energy trading problem and internal energy conversion problem within interconnected residential, commercial and industrial MEMGs. The problem is solved by a novel multi-agent deep reinforcement learning (MADRL) method. Finally, this thesis studies a P2P energy trading and energy conversion framework based on the highly efficient double auction (DA) market. A novel DA-MADRL method is proposed, which not only inherits the ability of MADRL to perform well in a multi-agent environment with various uncertainties and also addresses privacy concerns of the MEMGs.

Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank my supervisor, Professor Shengrong Bu for her invaluable advice, continuous support, and patience during my PhD study. Thanks for numerous discussion and idea exchange in her cosy office or through video meeting during the pandemic. Her sharp advice and remarkable help concluding all the publications contained in this thesis. I really appreciate it!

I would like to express my gratitude to my examiners, Dr Chenghong Gu and Dr Yao Sun, for an interesting discussion we had during the viva examination and for their valuable suggestions on my thesis.

I would like to thank Professor Zhu Han, Professor Xue Liu, Professor F. Richard and Professor Goran Strbac for their advice and knowledge in revising my journal papers. I also would like to thank Dr Dawei Qiu for his insight and inspiration in the field. I also appreciate my colleagues' technical and mental support from Dr James Oyedokun, Dr Fauzun Abdullah Asuhaimi, Yushen Miao and Gaddafi Almannouny from James Watt School of the Engineering of the University of Glasgow.

Finally, I would like to express my gratitude to my parents, Ning Chen and Chunmei Fu, and my love Yuying Nong. Without their tremendous understanding and encouragement in the past four years, it would be impossible for me to complete my PhD study.

Declaration

This thesis is submitted in accordance with the rules for the degree of Doctor of Philosophy at the University of Glasgow. This thesis does not include work forming part of a thesis presented successfully for another degree.

Some parts of Chapter 5 of this thesis result from collaboration with Dawei Qiu, a postdoctoral research associate at Imperial College London. Specifically, the double auction market in Section 5.2.3 is based on ideas developed by Dawei Qiu. Also, the abstracted Q-value function of the proposed MARL method described in Section 5.3.2 originated from Dawei Qiu. Furthermore, the Python implementation of the ZI MILP method that led to the related experimental results in Section 5.4 was done by Dawei Qiu. All other results therein are claimed as original.

Publications

The following publications contain some of the results presented in this thesis.

1. Tianyi Chen and Shengrong Bu, “Realistic Peer-to-Peer Energy Trading Model for Microgrids using Deep Reinforcement Learning,” *IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe’2019)*, Bucharest, Romania. (This paper is based on some of the results in Chapter 3.)

2. Yushen Miao, Tianyi Chen, Shengrong Bu, Hao Liang and Zhu Han, "Co-optimizing Battery Storage for Energy Arbitrage and Frequency Regulation in Real-time Markets Using Deep Reinforcement Learning," *Energies*, vol. 14, no. 24: 8365, Dec. 2021.

3. Tianyi Chen, Shengrong Bu, Xue Liu, Jikun Kang, F. Richard Yu and Zhu Han, “Peer-to-Peer Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning,” *IEEE Transactions on Smart Grid*, vol. 13, no. 1, pp. 715-727, Jan. 2022. (This paper is based on some of the results in Chapter 4.)

4. Dawei Qiu, Tianyi Chen (Corresponding author), Goran Strbac and Shengrong Bu, “Coordination for Multi-Energy Microgrids Using Multi-Agent Reinforcement Learning,” *IEEE Transactions on Industrial Informatics*, to be published. (This paper is based on some of the results in Chapter 5.)

5. Tianyi Chen and Shengrong Bu, “Detection and Mitigation False Data Injection Attacks on Local Energy Trading in a Multi-energy Microgrid using Deep Reinforcement Learning,” *IEEE Transactions on Smart Grid*, to be submitted in August 2022.

List of Abbreviations

To facilitate quick reference, the abbreviations used throughout the thesis is listed below. Other symbols are also defined in the text where applicable.

ADMM	Alternating Direction Method of Multipliers
AEMS	Automatic Energy Management System
AI	Artificial Intelligence
CHP	Combined Heat and Power
CNN	Convolutional Neural Network
COM	Commercial
DA	Double Auction
DDPG	Deep Deterministic Policy Gradient
DER	Distributed Energy Resource
DL	Deep Learning
DNN	Deep Neural Network
DNO	Distribution Network Operator
DNO	Distribution Network Operators
DQN	Deep Q-Network
DR	demand response
DRL	Deep Reinforcement Learning
EES	Electric Energy Storage
EHP	Electric Heat Pump

EL	Electric Load
EMS	Energy Management System
ESS	Energy Storage Systems
FC	Fuel Cell
FiT	Fit-in Tariff
FMDP	Finite Markov Decision Process
GB	Gas Boiler
GHG	Greenhouse Gas
HL	Heat Load
HP	Heat Pump
ICT	Information and Communication Technology
IDDPG	Independent Deep Deterministic Policy Gradient
IDQN	Independent Deep Q-Network
ILP	Integer Linear Programming
IND	Industrial
IRL	Independent Reinforcement Learning
LMP	Locational Marginal Pricing
LP	Linear Programming
MA	Multi-Agent
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
MADRL	Multi-Agent Deep Reinforcement Learning
MATD3	Multi-Agent Twin Delayed Deep Deterministic Policy Gradient
MDP	Markov Decision Process
MEMG	Multi-Energy Microgrid
MES	Multi-Energy System

MG	Microgrid
MGCC	Microgrid Central Controller
MILP	Mixed-Integer Linear Programming
NE	Nash Equilibrium
NLP	Nonlinear Programming
P2P	Peer-to-Peer
POMDP	Partially Observable Markov Decision Process
PV	Photovoltaics
ReLU	Rectified Linear Unit
RES	Residential
RL	Reinforcement Learning
SA	Single-Agent
TD3	Twin Delayed Deep Deterministic Policy Gradient
TES	Thermal Energy Storage
ToU	Time of Use
WE	Water Electrolyser
WG	Wind Generator
ZI	Zero Intelligence

Chapter 1

Introduction

1.1 Background

Unlike many other products, electricity cannot be stored in large quantities since it is unique. For this reason, the supply and demand of electricity must always be balanced. Before 1990, the government had monopolised the supply of energy in the UK. A fully privatised energy market was established in 1999, and since then, the matching of supply and demand has primarily been carried out by energy trading. Regardless of the market type, there are wholesale markets as well as retail markets in the UK. The term "energy trading" usually refers to buying and selling energy on the wholesale market, where generators and producers make deals with retailers to sell energy. Consumers can then compare energy providers on retail energy markets and determine which supplier best meets their needs.

The bulk of electricity has traditionally been produced in centralised large power stations located near fuel sources or far away from the condensed population centres. This has resulted in high costs for transmission and large amounts of greenhouse gas (GHG) emissions. In recent years, the effects of climate change have become increasingly noticeable. The frequency and intensity of hurricanes, droughts, fires and flooding have increased drastically over the last few decades [1]. Global climate change is widely recognised to be driven by carbon dioxide emissions. Therefore, reduced emissions are urgently needed to prevent climate change's most damaging impacts [2].

In recent years, governments worldwide have been promoting low-carbon transitions to make the energy system more climate-friendly and more energy-efficient. In Scotland, for example, the Scottish government has taken steps [3] to reduce carbon emissions by exploring the potential of Scotland's renewable energy resources to meet the heat, transport, and electricity needs of its people.

As a result of this growing necessity, the use of small renewable power plants connected to local distribution networks rather than to transmission networks is increasing. In the UK, for example, distributed generation increased from 14.5 GW in 2011 to 47.4 GW in 2019 [4].

However, the previous forecasts of distributed generation capacity for 2019 were 25 GW in 2011, 41 GW in 2015, and 44 GW in 2018 [5]. It shows that distributed generation can continue to surpass long-term growth forecasts under the right economic and political circumstances.

While renewable generators may participate in traditional wholesale markets, there is a potential problem incorporating renewable energy into the grid due to the intermittency and the level of distribution of renewable generation. As a result, the traditional energy trading scheme would not be the appropriate solution in this circumstance. With the help of smart grid infrastructure and energy storage systems (ESSs), peer-to-peer (P2P) energy trading has become a viable option at the distribution level of the power grid.

P2P energy networks are distributed energy networks that let users share portions of their energy resources with each other. Additionally, P2P energy networks include interconnected platforms that serve as an online marketplace for consumers and suppliers to directly exchange energy without the involvement of a middleman. Increasing numbers of energy end-users, historically passive consumers of energy, are now becoming prosumers due to the rapid adoption of distributed energy resources such as electric vehicles and rooftop solar systems. However, current market arrangements in the energy sector, such as feed-in tariffs, limit the potential value of the distributed energy resources and create transmission loss and congestion. In P2P energy trading systems, for example, solar panel owners (residential and commercial) can sell excess energy to locals who prefer renewable energy over fossil fuels. This eliminates the need for central authorities [6], complex business models and tech equipment and allows for small-scale production and consumption in the local community. An example of such a P2P energy system being utilised in a real-world setting can be seen in the Brooklyn Microgrid [7].

There are three main benefits for P2P energy trading, which are shown in the following.

- Improve System Efficiency and Agility
  Energy is used up at or near the point of production in P2P energy trading. Therefore the product will meet the user's need locally, thus reducing distance-related transmission losses and avoiding congestion. There has been a rise in distributed energy resources (DERs) and ESSs, which enable different forms of distributed generation and decrease dependency on the main power grid. It also allows the utility companies to provide additional ancillary services, which improves the efficiency of the power grid and enhances the reliability of equipment and reduces interruption costs to consumers.

- Reduction in GHG/CO_2 Emissions
  An essential part of reducing GHG emissions is to maximise the usage of DERs, which means saving power and reducing the emissions of waste gases produced on-site. DERs can be effectively utilised through P2P energy trading methods to meet local energy needs, thus reducing the chance of wasting renewable energy.

- Cost Minimisation

There are many benefits to be gained by end-users through P2P energy trading, including the reduction of energy transportation costs and energy supply costs, or even the possibility of maximising profits by selling their surplus energy to their local peers.

A P2P energy trading system offers many benefits, but it also presents many challenges. Due to the fact that P2P energy trading operates without a central controller, it is quite difficult to convince consumers to trust such a system [8]. In addition, given the large number of users in a P2P energy system, it is challenging to model the decision-making process associated with various energy trading parameters and conflicts of interest among the users in the system [6]. Moreover, electricity exchange differs from other exchanges of goods because fully decentralised P2P energy trading could have detrimental effects on the power system [9]. The reason is that P2P energy trading takes place on an electricity network, which has inherent network limitations, e.g., energy balance constraints and power flow constraints. Therefore, it is critical to figure out how to trade energy in the P2P network without compromising the network's security. Last but not least, different stakeholders may require different grid services in order to meet their objectives. Therefore, innovative pricing schemes and energy trading policies are necessary to keep up with individual requests and maximise the welfare of the network at the same time [10].

1.2 Motivation

P2P energy trading relies on balancing local supply and demand to make an excellent network system for trading energy. Hence, the accuracy of supply and demand prediction will be crucial towards creating a robust system for P2P energy trading. As mentioned earlier, in a P2P energy trading network, the end user's demand fluctuates with increasing randomness, which cannot be efficiently predicted compared to a macro grid. In addition to this, the intermittent nature of renewable energy generation is also a significant obstacle to finding the optimal energy trading policy. Thus, it is critical that the P2P energy system is designed, analysed, and planned to operate and settle rapidly and accurately with a huge amount of uncertain energy data.

Previous research on P2P energy trading has mostly been solved by traditional constrained optimisation methods such as mixed-integer linear programming (MILP) [11] and alternating direction method of multipliers (ADMM) [12]. Those methods have proven to be very useful for a lot of complex tasks considering multiple factors and constrictions. While there are some advantages, MILP ignores the assumption of non-linearity of the system [11], and the ADMM can only handle regular and convex problems [12], which is unrealistic in many cases. Furthermore, they require complete information on DERs and the local power system. These methods are more appropriate for day-ahead planning than real-time operation when electricity production and consumption are continually changing.

DRL is an advanced machine learning algorithm that is designed to produce autonomous agents to achieve artificial intelligence (AI) [13]. DRL combines deep learning (DL) with re-

inforcement learning (RL), where RL is a mathematical framework for experience-driven behaviour learning, and DL uses deep neural networks as approximators for RL functions. There have been several successful applications of DRL algorithms to video games and board games, including the most famous one, AlphaGo. In addition to being better players than humans, DRL algorithms can also solve real-world decision-making problems. It is challenging to make decisions by human and traditional optimisation methods given the many continuous data sets involved in P2P energy trading, such as stochastic and uncertain renewable generation and load demand. With DRL, making an optimal decision in P2P energy trading will be possible. Therefore, it is critical to understand the feasibility of using DRL algorithms to solve P2P energy trading problems.

Furthermore, MES which can integrate electricity networks with other energy sectors, has gained increased visibility in recent years. As energy coupling technologies such as hydrogen fuel cells and heat pumps have progressed rapidly, MES can provide the power grid with greater flexibility and reliability. It is not surprising that P2P energy trading will surely benefit from maximal flexibility through MES. However, correctly modelling and quantifying the P2P energy trade and understanding the flexibility of MES can prove to be a challenging undertaking. Modelling P2P energy trading with MES involves temporal, multi-vector interactions on different networks (e.g., electricity, heat and gas) in response to uncertain energy generation and demand [14]. It also includes potential conflicting energy trading and energy conversion policies of end-users. Therefore, it is important to understand how P2P energy trading and energy conversion can facilitate the MG in a holistic way. Furthermore, the difficulty in modelling multi-agent interactions among the P2P energy traders by DEL methods need to be addressed properly.

Finally, end-users in a fully decentralised P2P energy trading market can benefit from complete autonomy. However, it is unlikely to guarantee the maximum level of social welfare, and economic prosperity and safety concerns remain. On the other hand, local energy trading with a centralised operator can lead to theoretical welfare maximisation, but it can also bring in privacy concerns for participants. This is because the operator will be able to acquire detailed models and information about the customers' system and DERs. Thus, preserving the privacy of traders and finding an efficient and safe market design is crucial for P2P energy trading.

1.3 Contributions

Research contributions are arranged according to the specific chapters in which they appear in this thesis. This thesis aims to present methodologies for implementing DRL in P2P energy trading markets to achieve automatic operation of MGs. Further, these contributions are described as follows:

Chapter 3 Contributions

1. In this work, we formulate a realistic energy trading model for MGs with a set of critical physical constraints. An MG needs to make a trading strategy and negotiate with other MGs only based on its generation, demand and energy storage level. The physical constraints like transmission losses and power limits at some nodes of the system may affect the strategy of an MG.

2. We also set a flexible utility function for each MG to evaluate its strategy, which consists of not only trading profits but also the battery wear cost, demand penalty, and optional social factors.

3. We proposed a novel DRL method for the MG to derive better energy trading strategies based on the system states and the utility function. Using deep Q-network (DQN) and an experience replay mechanism [15], the algorithm can speed the Q-learning rate and update the loss function with continuously collected new states and rewards instead of updating the model at the end of each episode.

Chapter 4 Contributions

1. To the best of our knowledge, this is the first work to consider P2P energy trading, energy conversion and multi-vector energies together in a holistic way. A new P2P energy trading and energy conversion scheme is established for interconnected residential, commercial and industrial MEMGs. A two-stage problem consisting of P2P energy trading and energy conversion process is formulated as a partially observable Markov decision process (POMDP).

2. A MADRL approach MATD3 is proposed to optimise P2P energy trading and energy conversion policies of MEMGs in real-time. The proposed method combines MADRL framework in [16] with twin delayed deep deterministic policy gradient algorithm (TD3) [17] further to improve the performance of the MA actor-critic algorithm. The original MADRL framework has been modified particularly for our P2P energy trading and energy conversion problem and also for stabilizing the learning process. To our best knowledge, this is the first paper using multi-agent DRL models for P2P energy trading. Our proposed MATD3 approach can be used to choose the optimal actions within continuous action space and enables all the MEMGs to learn their policies simultaneously to achieve the best goal individually.

Chapter 5 Contributions

1. The flexibility due to the P2P energy trading among different MEMGs and the coupled energy conversions in each MEMG is explored. The examined problem is complex because of various system dynamics and uncertainties. A DA market-based P2P energy trading framework has been proposed to obtain good performance with privacy preservation. To

the best of our knowledge, this is the first work to adopt the DA market mechanism to a local energy community with multiple MEMGs.

2. A novel DA-MATD3 method is proposed, which inherits the ability of the MATD3 to perform well in a multi-agent environment with various system dynamics and uncertainties and addresses privacy concerns using a DA market framework. Specifically, the DA-MATD3 method integrates the key information of the DA market into the state-of-the-art MATD3 algorithm by connecting the critic networks of the agents with the DA market order books. To the best of our knowledge, this is the first work to integrate the DA market information into the MATD3 algorithm.

1.4 Thesis Outline

The rest of this thesis is organised as follows:

Chapter 2 firstly introduces the general structure of P2P energy trading. This chapter then discusses the market mechanisms of electricity trading markets, including wholesale, retail, and P2P energy trading. In addition, this chapter provides an overview of enabling technologies for P2P energy trading. Furthermore, this chapter describes the participants in the P2P energy trading market. In addition, this chapter covers five common technical solution approaches used in the P2P energy trading. Last but not least, this chapter collects and discusses the related work in the area of P2P energy trading.

Chapter 3 integrates DRL with a realistic P2P energy trading model to address a decision-making problem for MGs in the local energy market. First, this chapter develops an hour-ahead P2P energy trading model with a set of critical physical constraints. Next, the decision-making process of energy trading is modelled as a Markov decision process, which is used to find optimal strategies for MGs using DRL. In particular, this chapter modifies a DQN to aid the MGs in utilising resources and constructing better energy trading policies. Finally, The case study discusses the performance of the proposed DQN-based P2P energy trading strategy and analysed the impact of generation and demand uncertainties, P2P price ratio and virtual penalty on the average daily utility and power plant schedule.

Chapter 4 examines the external P2P energy trading problem and the internal energy conversion problem within interconnected residential, commercial and industrial MEMGs. These problems include data with high dimensionality and high uncertainty, so making decisions about them is a complex process. Therefore, this chapter proposes a MADRL approach combining the multi-agent actor-critic algorithm and the twin delayed deep deterministic policy gradient algorithm. Next, this chapter describes the changes to the original MADRL framework and the neural network architecture of the proposed methodology. Finally, this chapter discusses the impact of energy conversion and energy trading, the cost and emissions implications of carbon taxes, and the scalability of the proposed approach.

Chapter 5 presents a DA market-based P2P energy trading framework. MEMGs can therefore schedule their own components and trade energy with other MEMGs in the DA market. The chapter then formulates the P2P energy trading and energy conversion problem as a POMDP and proposes a MADRL solution by leveraging DA market information to improve stability while protecting privacy. Finally, several case studies illustrate the performance of different DRL methods and the benefits of the proposed method and P2P energy trading in the DA market.

Chapter 6 outlines the major contributions of this thesis and summarises the relevant findings drawn from the case studies undertaken throughout this thesis. Furthermore, some topics for future research are also proposed in this chapter.

Chapter 2

Literature Review

2.1 P2P Energy Trading Structure

P2P energy trading is defined as directly trading energy between consumers and prosumers within the local electricity distribution system, based on the concept of the "P2P economy" (also called sharing economy) [18]. In the P2P energy trading model, the peers are end users with DERs, flexible loads, and ESSs. By eliminating conventional energy suppliers from the process, peers buy and sell energy directly. In order to understand the P2P energy trading system, the system can be separated into two main components: the physical energy system and the virtual energy trading system.

The physical energy system primarily comprises a distribution network and data sensing system.

- *Distribution Network*: A distribution network transfers physical energy between peers and controls the system's power flow. This network is usually provided and maintained by distribution network operators (DNO), companies that own and operate cables and towers for distributing electricity. As an alternative, it could also be implemented by leveraging a separate microgrid distribution network in conjunction with the main power grid. This microgrid distribution network can be isolated from the main grid during times of emergency, increasing flexibility. Defining the grid connection point is essential to balance demand and supply. Phasor measurement units installed in the grid connection points can be used to monitor and evaluate power system information [6].

- *Data Sensing System*: Processable and digital data are essential for the development of P2P energy trading systems. Therefore, the data sensing system that generates and monitors physical data bridges the gap between physical and virtual energy trading systems. Data sensing technologies, such as smart meters, are crucial to help achieve P2P energy trading. We lacked high-quality data to determine user consumption and generation patterns for a long time. Fortunately, smart meters have been widely deployed in recent

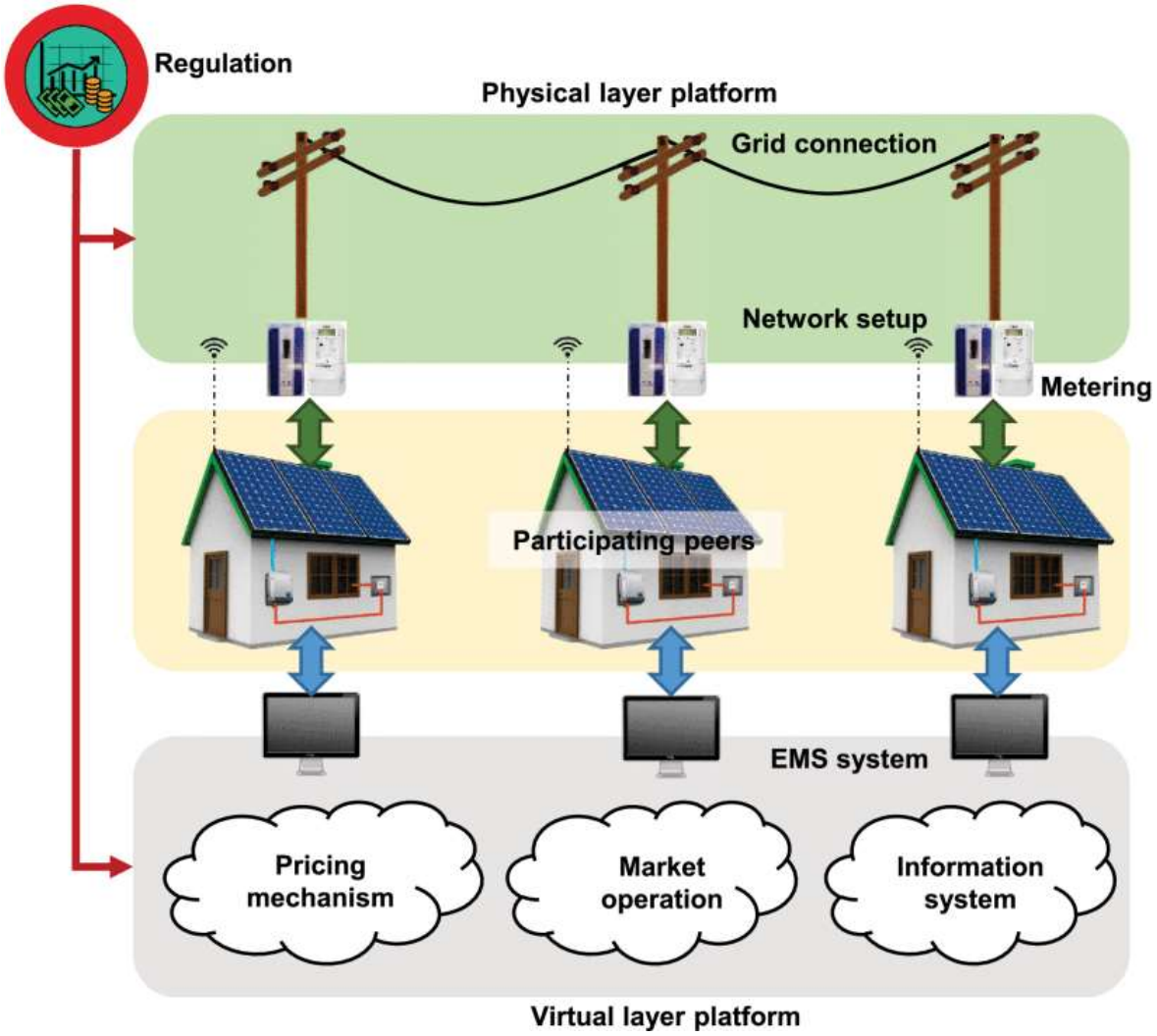


Figure 2.1: An illustration of the physical layer and virtual layer platforms of a P2P energy network.

decades, allowing for the collection of enormous amounts of fine-grained electricity data. In addition to smart meters, other monitors or sensors can retrieve P2P market data (trading volume, trading time, supply reliability) and other auxiliary information, such as social media and weather data.

The virtual layer of the P2P energy trading system provides the technical infrastructure for the local P2P energy market, which is composed of an information and communication technology (ICT) system, a financial trading system, and an energy management system (EMS).

- *ICT System*: The ICT system allows end users to find the most appropriate energy market and enables them to communicate more effectively. It is thought that a secure ICT environment will lead to trading peers having equal access to the information, the privacy of traders is protected, and the market will run smoothly.

- *Financial Trading System*: A financial trading system determines how energy is traded be-

tween peers and suppliers in the P2P energy market. One key component of the financial trading system is the pricing scheme. The energy price of P2P energy trading is determined depending on the selected market mechanism. However, an appropriate pricing scheme should reflect the energy demand and supply information of the local community to maximise the social welfare of all parties in the market.

- *EMS*: As a decision-making system in the P2P energy market, the EMS helps end-users determine the optimal energy trading policies and control their flexible load. Users can set the policies of the EMS based on their energy states and market information such as generation, demand, price, and energy source.

During the P2P energy trading period, the EMS will choose its energy trading policies based on the current system states and market information on behalf of the end-users. The end-users will then negotiate with each other and make energy trading deals. After completing the trading arrangement, a payment is made, and then the energy is subsequently exchanged in the physical network. A demonstration of the physical and virtual layer of the P2P energy trading system and its components is illustrated in Figure 2.1 [19].

2.2 Market Mechanism

According to the different market mechanisms, the P2P energy trading market can be classified into three categories: centralised market, decentralised market, and community market.

2.2.1 Centralised Market

In a centralised framework, a central operator is responsible for coordinating the trading decisions and the communication between users. In the end, after the trading process is complete, the operator distributes the revenue to the end-users based on the predetermined rules [20]. A centralised market is a market in which each participant does not communicate or negotiate with other participants regarding energy trading. However, peers can affect the choice of energy trading parameters by deciding how much energy and what price they would like to trade with the coordinator. It aims to maximise the social welfare of its participants, which is one of the critical advantages of the centralised market. However, it is necessary to have a complete, mathematical model of the energy system to support P2P energy trading [21]. In addition, the central coordination of a market can potentially compromise the privacy of end-users due to direct control over their DERs and flexible loads. The framework of the centralised market is shown in Figure 2.2 [22].

There have been a number of studies that suggest or discuss centralised P2P energy trading markets. The authors in [23] proposed two centralised markets, the "Flexi User" and the "Pool Hub," that examine the role and value of centralised and decentralised batteries. The

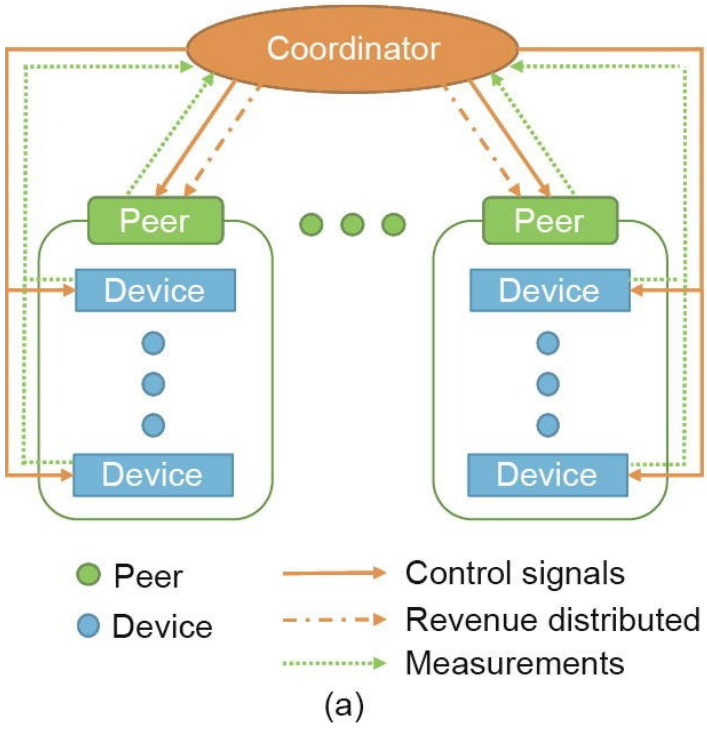


Figure 2.2: Demonstration of P2P energy trading in a centralised market.

author in [24] developed a P2P energy exchange platform named "Smart elecTricity Exchange Platform" (STEP) which assumed a set price for the centralised P2P energy markets. The authors in [25] designed a near-optimal algorithm named “Energy Cost Optimization via Trade” (ECO-Trade) for centralised P2P energy trading markets.

2.2.2 Decentralised Market

The second type focuses on designing a decentralised framework that will enable peers to manage their own resources independently and directly exchange energy with each other. Therefore, a decentralised market system involves both a decentralised decision-making process and a decentralised information exchange. Decentralised markets have an advantage over centralised markets in which end users have the option of deciding when, whether and how much energy to participate in P2P energy trading [19]. In addition to being incredibly flexible, decentralised markets are also highly scalable [22]. However, It has been observed that decentralised markets are inefficient and social welfare does not reach its full potential because of the lack of centralised oversight [26, 27]. Moreover, decentralised markets present a more significant challenge to service providers. For example, third parties, such as network operators and policy regulators, are not always certain how to manage the volume of energy that can be exchanged within a given community. On top of that, it is difficult to maintain network constraints of the power system alongside a decentralised P2P energy market. As a result of these factors, some network operators may have to curtail their loads and block their end-users from the network to maintain

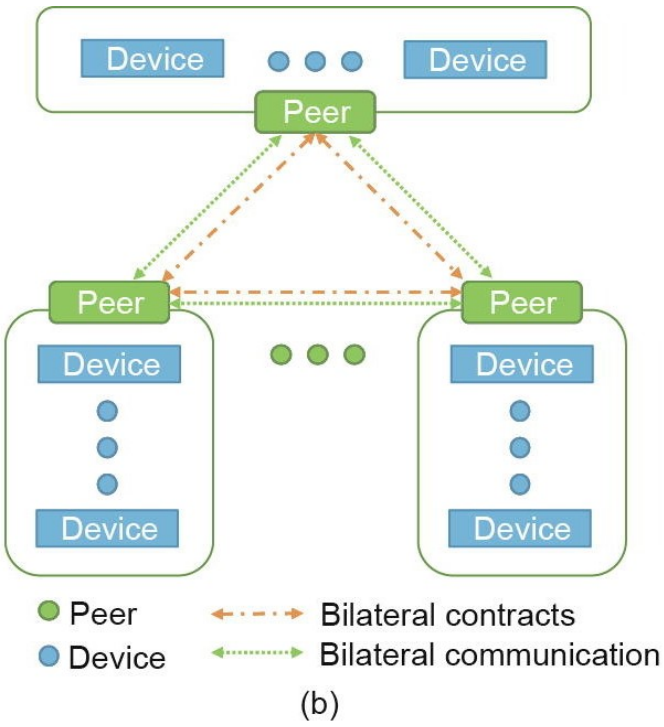


Figure 2.3: Demonstration of P2P energy trading in a decentralised market.

the grid's reliability [26]. The decentralised market framework is shown in Figure 2.3 [22].

A smaller number of studies have been published on decentralised P2P energy trading markets compared with centralised markets. In [27], a full decentralised bilateral contract network that can be used in forward and real-time P2P energy markets was proposed. The authors in [28] constructed a fully decentralised P2P market where the preferences of customers can be respected while social welfare is maximised. In [29], an innovative demurrage mechanism was designed for a blockchain-based energy trading marketplace where power consumption can be allocated to times with surplus local generation. Furthermore, in [30], the authors presented a distributed economic load dispatch algorithm that synchronises limited renewable energy sources, storage devices, and flexible loads in microgrids.

2.2.3 Community Market

Lastly, the community market is a hybrid of the centralised market and the decentralised market, combining the strengths of both. A community market has decentralised energy trading systems, while the communications between traders and users are more tightly controlled than in a fully decentralised market. Efficient transactions are carried out through a community manager among end-users in this market. In contrast to centralised markets, community managers cannot directly control how energy is exported or imported by the different participants in the market. Alternatively, a community manager may influence end-users indirectly by selecting appropriate pricing signals for P2P trading activities [20]. In this way, end-users in a community market can

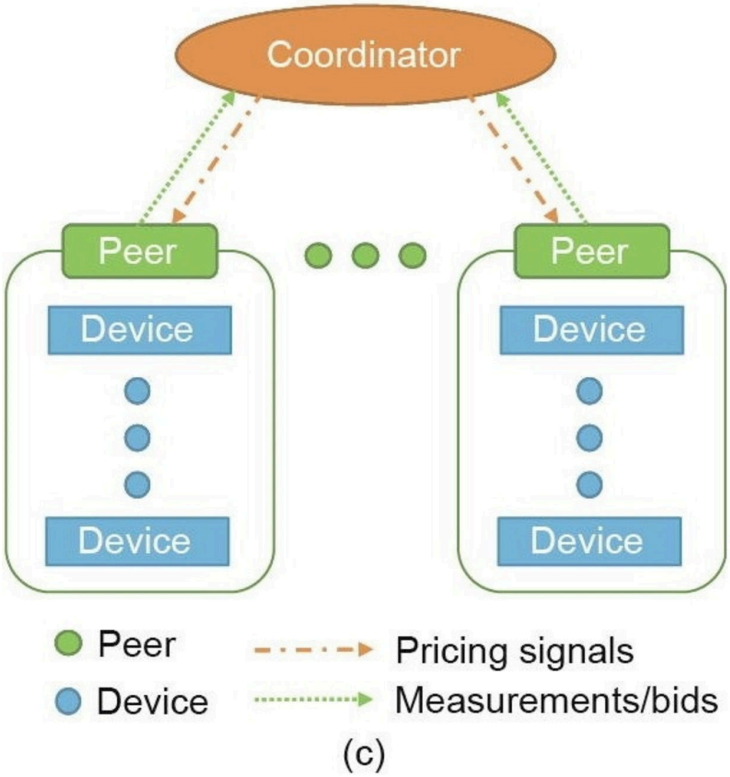


Figure 2.4: Demonstration of P2P energy trading in a community market.

share limited information, therefore maintaining a higher level of privacy [31]. Additionally, the community market mechanism will protect end users' autonomy to make their own energy trading decisions. In the literature on community market-based energy [32, 33], one of the primary focuses is designing pricing schemes for P2P energy trading which provide energy services to a range of end-users within the local community. A reasonable pricing scheme will also encourage many end-users to participate in energy trading by giving them related incentives. A community market framework is shown in Figure 2.4 [22].

There have been several studies that suggest P2P energy trading can be achieved through community markets. In [34], P2P energy trading among plug-in hybrid electric vehicles (PHEVs) was carried out using an iterative double-auction mechanism to maximise social welfare. The authors in [35] proposed three pricing mechanisms for P2P energy trading in a community microgrid, including bill sharing and mid-market rates. Using non-cooperative games, evolutionary games, and Stackelberg game theories, the authors in [32] developed an iterative pricing mechanism for P2P energy trading in a community market. An event-driven community market was proposed [36] in which brokers determine occasional market open rates based on a double-auction model in response to an event or request.

2.3 Enabling Technologies

2.3.1 Distributed Energy Resources

Solar Energy

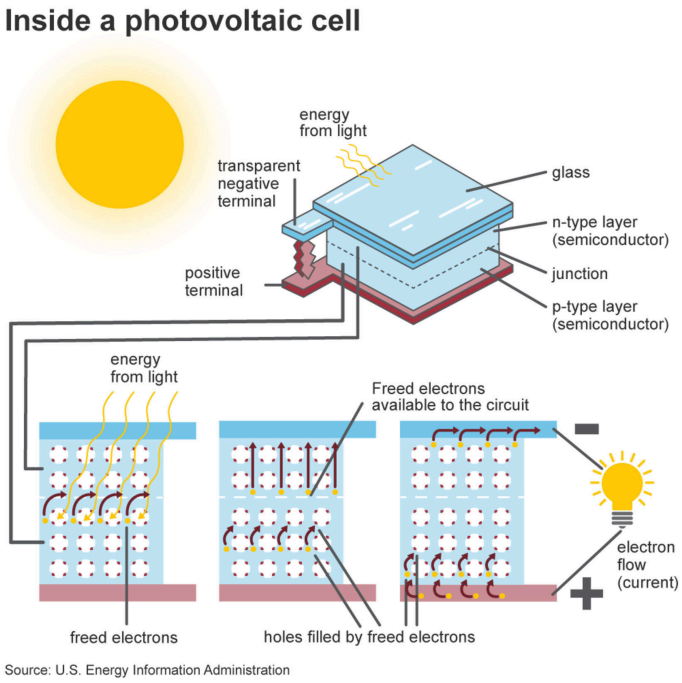


Figure 2.5: A basic diagram of photovoltaic solar cell.

The use of solar energy includes the use of radiant heat and light from the Sun to generate electricity, heat water, and for architectural purposes. Photovoltaic (PV) systems, also known as solar power systems, use photovoltaic technology to convert solar energy into usable power. A basic diagram of a PV cell is illustrated in Figure 2.5 [37]. The photocurrent of a PV cell I_{PH} is described as [38]:

$$I_{\mathrm{PH}} = \left[I_{\mathrm{SC}} + K_I \left(T_{\mathrm{C}} - T_{\mathrm{Ref}}\right)\right] \lambda, \tag{2.1}$$

where I_{SC} denotes the short-circuit current of the PV cell at a $25°\mathrm{C}$ and $1\ \mathrm{kW/m^2}$, K_I represents the short-circuit current temperature coefficient, T_{Ref} is the cell's reference temperature, and λ is the solar insolation in $\mathrm{kW/m^2}$.

PV systems are distributed electricity generation systems. Typically, a grid-connected PV system is connected to the main power grid and supplies energy directly to the grid. Residential or commercial buildings can directly use solar energy from a PV system before or after the revenue measurement point. In terms of size, these systems range anywhere between residential (less than 10 KW) and large solar power plants (up to tens of MW).

Research on P2P energy trading focuses on residential households, which usually have rooftop solar panels. Therefore, most P2P trading mechanisms are designed to take solar power as the primary renewable energy source into consideration. P2P energy trading's ability to reduce energy costs may be its most significant advantage when encouraging consumers to install PV systems and participate in P2P energy markets. Cost savings for using solar panels for P2P energy trading will depend on the energy trading price and energy trading policy adopted by each participating end-user [39].

Wind Power

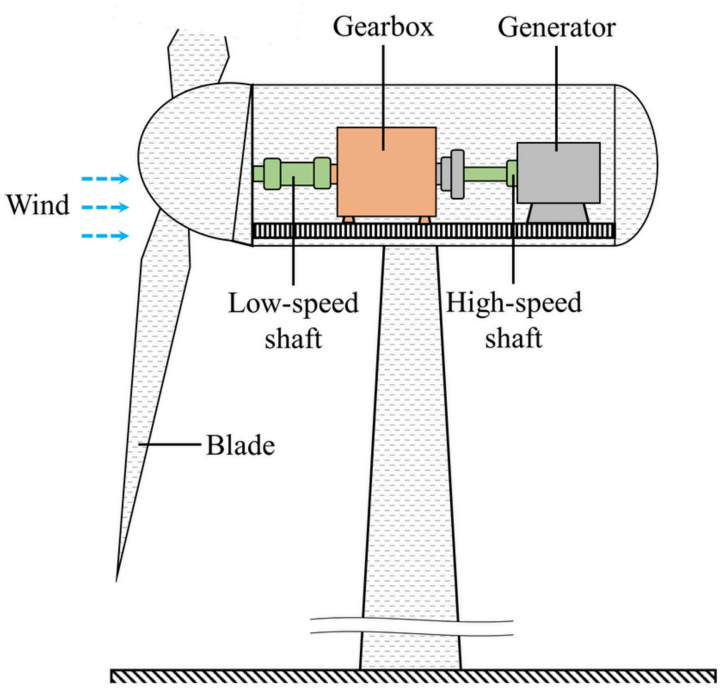


Figure 2.6: A basic diagram of wind turbine.

The use of wind turbines for generating electricity is known as wind power or wind energy. Utilising wind power as an energy source is sustainable, renewable, and ecologically friendly compared to burning fossil fuels. A basic diagram of a wind turbine is illustrated in Figure 2.6 [40]. The power generated by a wind turbine P_m is described as [41]:

$$P_m = \frac{1}{2}\rho\pi R^2 v_w^3 C_p, \tag{2.2}$$

where ρ denotes the air density, R denotes the radius of the rotor, v_w denotes the wind speed and C_p represents the power coefficient based on the turbine design. Many wind turbines are grouped together in wind farms connected to electrical transmission lines. Unlike other renewable energies, wind power is highly viable. Hence, power management techniques such as excess capacity management, building geographically distributed turbines, exporting and im-

porting, adding grid storage, and demand response are utilised to match supply and demand. Another way of addressing intermittent wind power is predicting the weather so that the power system can anticipate production fluctuations in advance. In some cases, using wind power may require upgrading the grid as it becomes more prevalent in a region.

Wind energy P2P trading studies are relatively few compared to solar energy. The reason is that wind turbines generally are not installed in residential homes but are used as part of a larger wind farm within a MG [18]. However, there is some evidence that electricity generated from wind turbines can be shared among participants in a local community through P2P energy trading. For example, a P2P energy trading system enhances the resilience and stability of multiple MGs with wind turbines [42]. In [43], a distributed optimisation scheme is proposed to improve the efficiency of wind power generation under dynamic conditions using a P2P collaboration model. In [44], a P2P energy trading framework is designed to provide reservation of wind power for demand response using a stochastic decision-making approach. Finally, [45] stresses the importance of P2P energy trading for renewable energy, including wind energy, offering a feasible socio-cultural prospect of renewable energy.

Hydrogen Fuel

Hydrogen fuel is a zero-carbon fuel when burned with oxygen if made in a carbon-neutral manner. The green hydrogen (CO-free) can be generated from water electrolysers and used in fuel cells to produce electricity or in internal combustion engines to produce energy. The working mechanism of water electrolyser and fuel cell are explained in the following sections. Currently, residential hydrogen is in an early stage of development. However, the use of hydrogen fuel to generate electricity and heat in the residential sector shows great promise due to its versatility as an energy source for low-carbon emission vehicles in the future.

With the penetration of hydrogen into the energy system, researchers are now examining the opportunity for P2P energy trading with the hydrogen sector. For example, a P2P energy market is designed to trade electricity and hydrogen [46]. The authors assumed that the market participants have hydrogen vehicles and hydrogen storage systems. In [47], a P2P energy trading system for networked MGs containing fuel cell vehicles is proposed using a multi-agent approach. Furthermore, the hydrogen storage and solar panels are examined for optimal size and timing in a P2P home energy trading scheme are studied in [48].

2.3.2 Energy Storage Systems

Electrical Storage System

The electrical storage system (ESS) or battery storage systems is used to stored and released the renewable energy, such as solar energy and wind power, when needed. In large-scale plants that provide electricity grids with a reliable supply of renewable energy, lithium-ion batteries, also

commonly used in mobile phones and electric vehicles, are currently the predominant storage technology. Comparing to other energy storage systems, ESS has the advantages of flexible configuration of power and energy according to different application requirements. The electrical storage system has a fast response time, is not restricted by external conditions such as geographical resources, and is suitable for large-scale applications, making it irreplaceable in terms of network operation and grid operation assistance.

The energy transition function of an ESS is shown as [49],

$$E_e^{t+1} = \eta_e^{ES} \cdot E_e^t + P_e^{ES,t} \cdot \left(I_{(P_e>0)} \cdot \eta_{e,ch}^{ES} - \frac{I_{(P_e \leq 0)}}{\eta_{e,dis}^{ES}} \right) \cdot \Delta t, \tag{2.3}$$

where $P_e^{ES,t}$ denotes the charging or discharging rate of an ESS, η_e^{ES} defines the self discharge rate of an ESS, and $\eta_{e,ch}^{ES}$ and $\eta_{e,dis}^{ES}$ denote the charging and discharging efficiency of an ESS. The power and energy limitations are shown as,

$$P_e^{ES^{min}} \leq P_e^{ES,t} \leq P_e^{ES^{max}}, \tag{2.4}$$

$$0 \leq E_e^{t+1} \leq B_e, \tag{2.5}$$

where $P_e^{ES^{min}}$ and $P_e^{ES^{max}}$ denote the power capacity and B_e is the energy capacity of an ESS.

Thermal Storage System

The thermal storage system (TSS) uses thermal storage materials as a medium to store thermal energy such as solar heat, geothermal heat, industrial waste heat and release the heat when needed.TSS is used to solve the problems caused by the mismatch between the supply and demand of thermal energy in time, space or intensity, maximising the energy utilisation of the whole system. There are several types of thermal storage materials, including water, air, oil, bedrock, brick, and concrete. Material selection is usually based on a material's heat capacity and the amount of storage space available. There are three types of TSS: sensible heat storage, latent heat storage, and thermochemical storage [50]. In sensible heat storage, heat is stored by raising the temperature of storage material, whether it is solid, liquid, gaseous, or supercritical. In Latent Heat Storage, the phase transition of the material is used to store heat. In most cases, the solid-liquid phase change is achieved by melting and solidifying a material. Thermochemical energy storage involves a chemical reaction with high energy being generated. A reverse reaction should be able to release the products of the reaction, and the heat can be retrieved when it occurs.

The energy transition function of an TSS is defined as [51],

$$E_{th}^{t+1} = \eta_{th}^{TS} \cdot E_{th}^{t} + Q_{h}^{TS,t} \cdot \left(I_{(Q_h>0)} \cdot \eta_{th,in}^{TS} - \frac{I_{(Q_h \leq 0)}}{\eta_{th,out}^{TS}} \right) \cdot \Delta t, \tag{2.6}$$

where $Q_h^{TS,t}$ denotes the heat flow rate of an TSS, η_{th}^{TS} defines the heat dissipation rate of an ESS, and $\eta_{th,in}^{TS}$ and $\eta_{th,out}^{TS}$ denote the inflow and outflow efficiency of an TSS. The heat power and energy limitations of an TSS are shown as,

$$Q_h^{TS^{min}} \leq Q_h^{TS,t} \leq Q_h^{TS^{max}}, \tag{2.7}$$

$$0 \leq E_{th}^{t+1} \leq B_{th}, \tag{2.8}$$

where $Q_h^{TS^{min}}$ and $Q_h^{TS^{max}}$ denote the heat power capacity and B_{th} is the heat energy capacity of an TSS.

Hydrogen Storage System

Hydrogen energy is the cleanest of all known energy sources. The product of the hydrogen use process is water, which can genuinely achieve zero emissions and no pollution. It is seen as one of the most promising energy sources for application. For hydrogen energy to be utilised, the hydrogen storage system (HSS) is essential. There are two established technologies of storing the hydrogen gas: compressed hydrogen and Liquefied hydrogen. A compressed hydrogen tank increases the density of hydrogen gas by applying pressure to it. For liquefied hydrogen storage, similar to liquefied natural gas stored at -162 degrees Celsius, hydrogen is liquefied by reducing its temperature to -253 degrees Celsius [52].

The energy transition function of an HSS is defined as [53],

$$E_{h_2}^{t+1} = \eta_{h_2}^{HS} \cdot E_{h_2}^{t} + Q_{h_2}^{HS,t} \cdot \left(I_{\left(Q_{h_2}>0\right)} \cdot \eta_{h_2,in}^{HS} - \frac{I_{\left(Q_{h_2} \leq 0\right)}}{\eta_{h_2,out}^{HS}} \right) \cdot \Delta t, \tag{2.9}$$

where $Q_{h_2}^{HS,t}$ denotes the hydrogen flow rate of an HSS, $\eta_{h_2}^{HS}$ defines the hydrogen dissipation rate of an HSS, and $\eta_{h_2,in}^{HS}$ and $\eta_{h_2,out}^{HS}$ denote the inflow and outflow efficiency of an HSS. The hydrogen power and energy limitations of an HSS are shown as,

$$Q_{h_2}^{HS^{min}} \leq Q_{h_2}^{HS,t} \leq Q_{h_2}^{HS^{max}}, \tag{2.10}$$

$$0 \leq E_{h_2}^{t+1} \leq B_{h_2}. \tag{2.11}$$

where $Q_{h_2}^{HS^{min}}$ and $Q_{h_2}^{HS^{max}}$ denote the hydrogen power capacity and B_{h_2} is the hydrogen energy capacity of an HSS.

2.3.3 Energy Converters

Fuel Cell

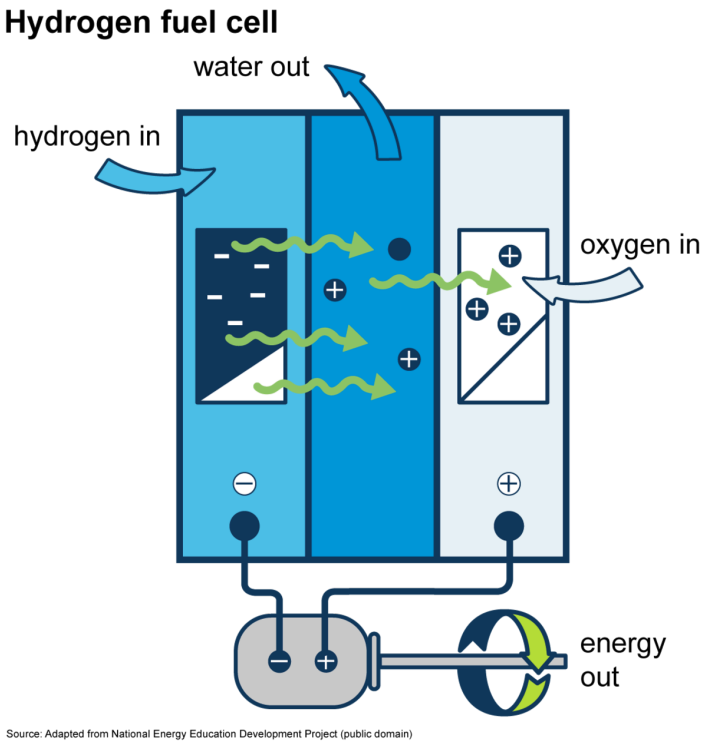


Figure 2.7: A basic diagram of hydrogen fuel cell.

Fuel cells are the main converters to convert hydrogen to electricity and heat. The benefits of hydrogen fuel cells include low emissions, no noise, and high energy conversion efficiency. It has been demonstrated that hydrogen fuel cell technology can be used on a large scale in vehicles, portable power generation, and stationary power stations [54]. A basic diagram of a hydrogen fuel cell is illustrated in Figure 2.7 [55]. If the fuel is CO-free, the overall reaction in the anode catalyst is described as [56]:

$$\mathrm{H_2} \longrightarrow 2\mathrm{H^+} + 2\mathrm{e^-}. \tag{2.12}$$

The proton in reaction (2.12) moves through the polymer electrolyte and participates in oxygen reduction in the cathode catalyst layer:

$$2\mathrm{H^+} + 2\mathrm{e^-} + \frac{1}{2}\mathrm{O_2} \longrightarrow \mathrm{H_2O_{(1)}}, \tag{2.13}$$

which result in the overall reaction in the fuel cell:

$$\mathrm{H_2} + \frac{1}{2}\mathrm{O_2} \longrightarrow \mathrm{H_2O_{(l)}} + \text{ heat } + \text{ electrical energy}. \tag{2.14}$$

The energy convert functions of fuel cell is shown as,

$$P_e^{FC,t} = \eta_e^{FC} \cdot Q_{h_2}^{FC,t}, \tag{2.15}$$

$$Q_h^{FC,t} = \eta_h^{FC} \cdot Q_{h_2}^{FC,t}, \tag{2.16}$$

where $Q_{h_2}^{FC,t}$ represents the input hydrogen flow of the fuel cell, $P_e^{FC,t}$ and η_e^{FC} denote output electricity power and electricity conversion efficiency, and $Q_h^{FC,t}$ and η_h^{FC} represent the output heat flow and heat conversion efficiency of the fuel cell.

Water Electrolyser

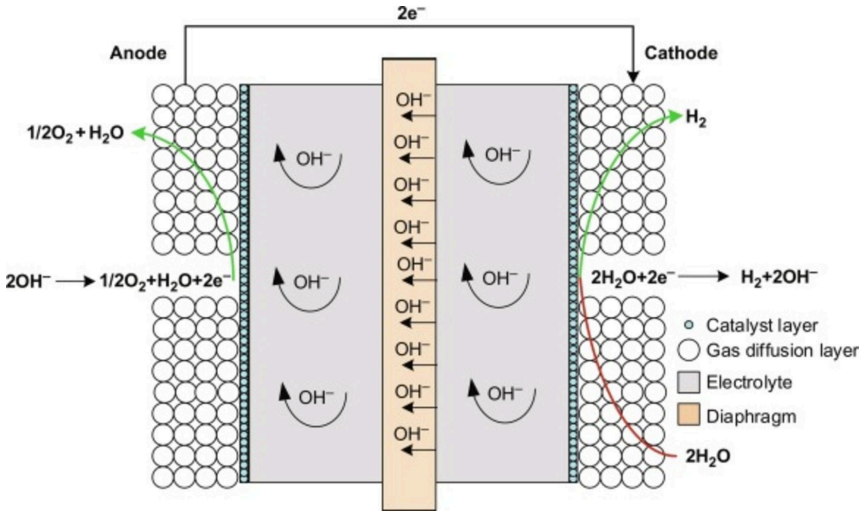


Figure 2.8: A diagram of an alkaline water electrolyser.

Electrolysis of water is a mechanism that uses electricity to electrolyse water into oxygen and hydrogen. As the name implies, water electrolyser requires electricity and the production of hydrogen from renewable energy sources (wind, solar, hydro, tidal, geothermal, biomass, etc.) that do not emit carbon dioxide is known as "green hydrogen" and is highly regarded as the ultimate clean energy technology. Alkaline water electrolysis was the first commercialised water electrolysis system, and it remains the most widely used. A potassium hydroxide (KOH) or sodium hydroxide (NaOH) solution is used as an electrolyte solution between two electrodes. A diagram of an alkaline water electrolyser is given in Figure 2.8 [57].

The energy convert functions of water electrolyser is shown as,

$$Q_{h_2}^{WE,t} = \eta^{WE} \cdot P_e^{WE,t}, \tag{2.17}$$

where $P_e^{WE,t}$ represents the input electricity power of the water electrolyser, $Q_{h_2}^{WE,t}$ and η^{WE} denote output hydrogen flow and hydrogen conversion efficiency of the water electrolyser.

Gas Boiler

Gas boilers are boilers that are fuelled by gas, traditionally natural gas but some boilers can be fuelled by hydrogen gas as well. Gas boilers are the most economical boilers compared to oil boilers and electric boilers, so most end-users choose gas boilers for space heating and water heating. A boiler is composed of a feed water system, a steam system, and a fuel system. To meet the steam demand, the feed water system automatically regulates the water supply to the boiler. Furthermore, steam is collected and controlled by the steam system in the boiler. In order to direct steam to the point of use, a piping system is used. Finally, fuel systems include all equipment used to generate heat from the fuel.

The energy convert functions of gas boiler is shown as,

$$Q_h^{GB,t} = \eta_{ng}^{GB} \cdot Q_{ng}^{GB,t}, \tag{2.18}$$

where $Q_{ng}^{GB,t}$ represents the input natural gas flow of the gas boiler, $Q_h^{GB,t}$ and η_{ng}^{GB} denote output heat flow and heat conversion efficiency of the gas boiler.

Heat Pump

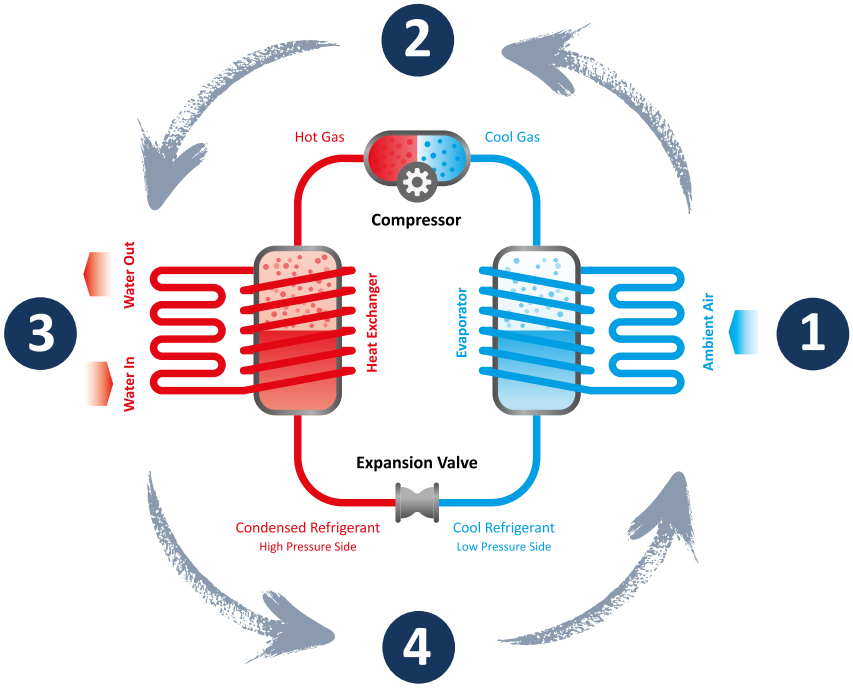


Figure 2.9: A diagram of a heat pump.

The heat pump opens up a new era of energy savings. Unlike steam engines, heat pumps are based on the inverse Carnot cycle, using mechanical energy to convert low temperature heat into high temperature heat. Heat pumps are a great way to save energy, with efficiencies of over 300%, making them more energy efficient than electric heating and gas boilers. Heat pumps

consist of four main components: an evaporator, a compressor, a condenser, and an expansion device. Each of these components is filled with a refrigerant, which circulates through them. An evaporator extracts heat from a waste heat source or ambient air. During the condenser's operation, the heat is delivered at a higher temperature to the consumer. For the compressor to work, electricity is needed, and the added heat will be available in the condenser. A diagram of an heat pump is given in Figure 2.9 [58].

The energy convert functions of heat pump is shown as,

$$Q_h^{HP,t} = \eta^{HP} \cdot P_e^{HP,t}, \tag{2.19}$$

where $P_e^{HP,t}$ represents the input electricity power of the heat pump, $Q_h^{HP,t}$ and η^{HP} denote output heat flow and heat conversion efficiency of the heat pump.

Combined Heat and Power

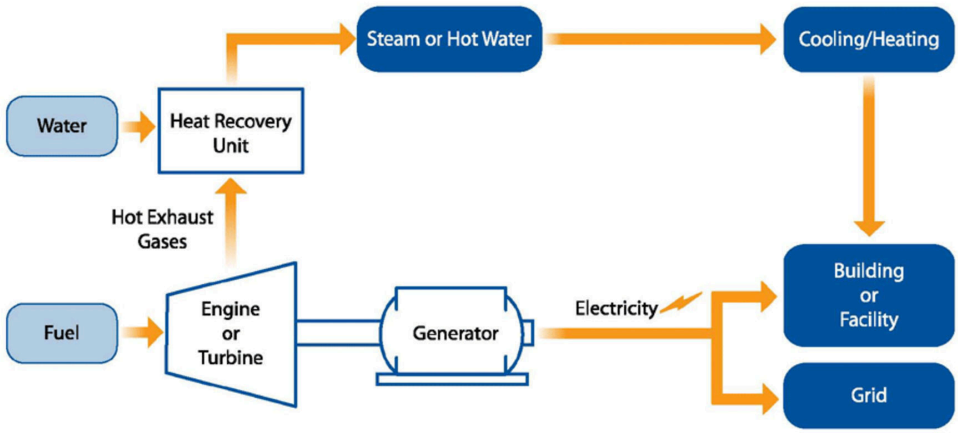


Figure 2.10: A diagram of a CHP system.

Combined Heat and Power (CHP) turbines or power plants can generate both electricity and heat, improving the thermodynamically efficiency of the fuel. In separate electricity production, most energy has to be discarded as waste heat. Hoever, in CHP some of this thermal energy is put to use. For conventional generators the efficiency of the fuel are only about 30%, up to 70% of the fuel energy is converted into useless heat, and CHP generation can reuse 30% of the heat energy along with the electricity production, bringing the total fuel efficiency to 60% [59]. A CHP system usually consists of a combustion turbine or reciprocating engine with a heat recovery system. It produces electricity by burning fuel (natural gas, oil, or biogas) and captures heat from the turbine or engine through heat recovery devices. In order to provide useful thermal energy, steam or hot water is generally generated from this heat. A diagram of a CHP system is given in Figure 2.10 [60].

The energy convert functions of CHP is shown as,

$$P_e^{CHP,t} = \eta_e^{CHP} \cdot Q_{ng}^{CHP,t}, \tag{2.20}$$

$$Q_h^{CHP,t} = \eta_h^{CHP} \cdot Q_{ng}^{CHP,t}. \tag{2.21}$$

where $Q_{ng}^{CHP,t}$ represents the input natural gas flow of the CHP, $P_e^{CHP,t}$ and η_e^{CHP} denote output electricity power and electricity conversion efficiency, and $Q_h^{CHP,t}$ and η_h^{CHP} represent the output heat flow and heat conversion efficiency of the CHP.

2.3.4 Data Sensing Systems

When it comes to P2P energy trading, a fine resolution of the data is critical to determine optimal trading strategies. Based on demand, generation, and market conditions provided by the smart meter, each prosumer determines whether or not to share his or her energy with others within the community [6]. A smart meter is a type of electricity or gas meter that reads and store the data automatically and provides the user with accurate readings. Communication between prosumers can also be done in a smart meter using any communication protocol that is appropriate.

During the past decade, smart meters have been deployed around the world. By the end of September 2021, there were 26.4 million smart meters installed in the UK [61]. It is possible to collect vast amounts of fine-grained electricity data thanks to the widespread popularity of smart meters. In addition to detecting bad data attacks, smart meter data can also be used for load forecasting, load management, and P2P energy trading. The authors in [62] compared seven existing techniques to forecast load demand using smart meter data, including linear regression, neural network, support vector machine and their variants. Many studies using cluster algorithms to mine the sociodemographic information from the massive smart meter data [63, 64].

2.3.5 ICT Systems

P2P energy trading requires a great deal of communication and information during the negotiation, delivery, and settlement stages. The coordinator and peers engage in a lot of bilateral communication in centralised P2P energy trading markets. In contrast, the peers still need a lot of bilateral communication among themselves in markets that are decentralised. A study of P2P energy trading and its impact on ICT infrastructure is therefore important [18].

Communication architectures for P2P energy trading and sharing between prosumers in microgrids were compared by the authors in [65]. Their evaluation based on IEEE Standard 1547.3-2007 found that both structured and unstructured P2P communication protocols offered robust performance and promise for supporting prosumers in communicating in a P2P market. In [66], the authors investigated the requirements for ICT infrastructures for P2P energy trading bidding and control systems. Models and simulations of existing and private communication

networks with different features, such as bandwidth and medium, were proposed and conducted using OPNET. The study showed that the P2P energy trading system can be enabled with existing ICT infrastructure, such as broadband network connections and GPRS smart metering networks, without requiring additional large investments.

2.4 Participants

2.4.1 Prosumers

A growing number of distributed generators and energy storage systems with smart energy management systems enable residential consumers to generate electricity and send it back to the distribution system. This expands the role of residential consumers from mere consumers to prosumers. Therefore, prosumers are participants in the P2P energy market capable of producing, consuming, and providing demand response [67]. The objective of the prosumer is to minimise its energy cost C_{PS}^t,

$$C_{PS}^t = \rho_{ToU}^t \cdot q_{P2P}^{+,t} - \rho_{FiT}^t \cdot q_{P2P}^{-,t}, \tag{2.22}$$

$$g^t + q_{P2P}^{+,t} + q_{P2P}^{-,t} - d^t - e^t = 0, \forall t. \tag{2.23}$$

where $q_{P2P}^{+,t}$ and $q_{P2P}^{-,t}$ and $q_{grid}^{+,t}$ are the trading quantity of the prosumer with respect to buying from the retailer and selling to the retailer; ρ_{ToU}^t and ρ_{FiT}^t denotes time of use (ToU) tariffs when buying energy from the retailer and fit-in tariff (FiT) when selling energy to the retailer at time t. The constraint that the prosumer must balance its generation and demand with the energy trading quantity and the charging/discharging amount of its energy storage is shown in (2.23), where g^t is the generation power of the prosumer at time t, e^t describes the charging ($e^t > 0$) or discharging ($e^t > 0$) amount of the energy storage system.

2.4.2 Generation Companies

Generation companies often participate in the wholesale market traditionally. Thanks to the P2P energy trading scheme, it's more profitable for some generation companies with distributed generations to trade with the local consumers directly. To simplify the representative model, we assume the generation company owns a single generation. Thus the quadratic cost function of the generation company model is defined as

$$C_G^t = c_G + l_G \cdot g^t + q_G \cdot {g^t}^2, \tag{2.24}$$

where g^t is the generation power of the generation company at time t, c_G, l_G and q_G are denoted as constant, linear and quadratic cost coefficient of the generation company. It is better to model

the cost as a linear function if the generation company only has renewable generation assets where the quadratic cost coefficient q_G is equal to 0. Renewable generation usually involves an initial installation cost of the generator and a linear operation and maintenance cost [68] which can be expressed as c_G and l_G in (2.24). The linear marginal cost function of the generation company model can then be expressed as the derivatives of the quadratic cost function, which is shown as

$$M_G^t = l_G + 2 \cdot q_G \cdot g^t. \tag{2.25}$$

The power limit of the generation company is shown as

$$0 \leq g^t \leq g^{\max}, \forall t. \tag{2.26}$$

To avoid non-linearity of the model, the quadratic cost function can be approximated by a piece-wise cost function, consisting of a number of generation blocks [48]. As the quadratic term has been removed, the cost function of each time block is linear and the marginal cost is constant, leading to a step-wise linear marginal curve. The approximated cost, marginal cost and the power limit of the generation company at each time block b are defined as

$$C_G^b = l_G^b \cdot g^b \tag{2.27}$$

$$M_G^b = l_G^b, \tag{2.28}$$

$$0 \leq g^b \leq g^{\max}, \forall b. \tag{2.29}$$

2.4.3 Energy Retailers

In P2P energy market, the prosumers may not be self-sufficient trading with each other. In addition, they often lack the abilities or energy scales to trade in the wholesale market. Therefore, the energy retailers are still essential in the P2P energy market. The energy retailers can be seen as a representative of a large majority of the energy consumers in the wholesale market, buying energy at the wholesale prices and selling it to their contracted users at certain retail prices [8]. Equation (2.30) describes the goal of a energy retail that is to maximise its profit by setting the retail buying price (also referred to time-of-use price) $\rho_{retail}^{+,t}$ and the retail selling price (also called feed-in tariff) $\rho_{retail}^{-,t}$ offered to its customers and its trading quantity d_{grid}^t in the wholesale market.

$$r_R^t = \rho_{retail}^{+,t} \cdot d^t - \rho_{retail}^{-,t} \cdot g^t - \rho_{grid}^{+,t} \cdot d_{grid}^t, \tag{2.30}$$

$$d^t - g^t - d_{grid}^t = 0, \forall t. \tag{2.31}$$

where d^t and g^t are the demand and generation of the costumers of the retailer, $\rho_{grid}^{+,t}$ denotes the wholesale buying price and the retailer must balance its customers' generation and demand with the trading quantity from the wholesale market shown in (2.31).

To prevent the customers of the retailers been coerced into unfair contracts, the local regulators set the price cap of the retail prices based on the wholesale prices to promote the fairness of the market and rights of the customers [49]. Such retail price constraints is shown as

$$\rho_{retail}^{+,t}, \rho_{retail}^{-,t} \in \mathcal{R}, \forall t. \tag{2.32}$$

2.4.4 Microgrids

The concept of MG has been proposed to reconcile the contradictions between large grids and DERs and make full use of the value and benefits that DERs can bring to the grid and consumers. A MG is a low-voltage network using DERs, energy storage devices, and controllable loads together, with capacities ranging from a few hundred KW to several MW, capable of operating independently in an emergency re-connected to the grid after recovery. MGs comprise several types of DERs such as solar panels, wind turbines, microturbines, and thermal power plants, each of which is a form of distributed generation, with energy reserves from the energy storage system. There are several types of loads that supply power to MGs, including residential, commercial, and industrial buildings. These loads are connected to the low voltage network at the point of common coupling [69].

In a P2P energy market, multiple connected MGs within the local community can cooperate to improve the efficiency and resilience of the distribution network further. The objective of the MG is to negotiate with each other to minimise its energy operational cost C_{MG}^t, which is defined as

$$C_{MG}^t = \rho_{P2P}^{+,t} \cdot q_{P2P}^{+,t} - \rho_{P2P}^{-,t} \cdot q_{P2P}^{-,t} - \rho_{grid}^{+,t} \cdot q_{grid}^{+,t}, \tag{2.33}$$

$$g^t + q_{P2P}^{+,t} + q_{P2P}^{-,t} + q_{grid}^t - d^t - e^t = 0, \forall t. \tag{2.34}$$

where $q_{P2P}^{+,t}$, $q_{P2P}^{-,t}$ and $q_{grid}^{+,t}$ are the trading quantity of the MG with respect to buying from other MGs, selling to other MGs and buying from the main grid; $\rho_{P2P}^{+,t}$, $\rho_{P2P}^{-,t}$ and $\rho_{grid}^{+,t}$ denotes the P2P energy buying price, selling price and the wholesale buying price at time t. The constraint that MG must balance its generation and demand with the energy trading quantity and the charging/discharging amount of its energy storage is shown in (2.34).

2.5 Technical Approaches

Recent studies have identified five general approaches to designing P2P energy trading schemes as the main contributors to their design. They are game theory, auction theory, constrained optimisation, blockchain and deep reinforcement learning.

2.5.1 Game-theoretic Approaches

Game theory is both a new branch of modern mathematics and an important discipline of operations research. It is a mathematical theory and method for studying phenomena of a combative or competitive nature. Game theory considers the predicted and actual behaviour of individuals in a game and investigates their optimisation strategies. Furthermore, game theory has a wide range of applications in finance, securities, biology, economics, international relations, computer science, political science, military strategy and many other disciplines. In general, there are two main categories in game theory: cooperative and non-cooperative games.

Cooperative Game

Cooperative games study how the benefits of cooperation are distributed when people agree to cooperate, i.e. the distribution of benefits. Games that are cooperative take a collaborative approach or a compromise. Compromise enhances the interests of both parties in the compromise and the interests of society as a whole because it generates a cooperative surplus. These relationships and approaches lead to this surplus, in part. The distribution of the cooperative surplus between the parties to the game depends on the balance of power and skill of the parties [70]. There needs to be a compromise negotiated between the parties to the game to reach a consensus to cooperate. Here, the distribution of the cooperative surplus is both a result of the compromise and a condition for it. In general, there are two main categories in cooperative game: canonical coalition game, coalition formation game and coalitional graph game.

a) Canonical coalition game: Canonical coalition games are never detrimental to any player when a grand coalition is formed with all players. In such a game, one of the main objectives is to determine whether a grand coalition can be formed. This is to determine whether the grand coalition is stable and to determine how to distribute the gains of the coalition among the players. Canonical coalition games are commonly solved using the core concept [70]. In addition, shapley value, Kernel, nucleolus, and strong epsilon-core distribution methods are popular for revenue distributions.

b) Coalition formation game: This game explores the network coalitional structure by playing a static coalition formation game. On the other hand, a dynamic coalitional game will be affected by changes in the environment. This includes the number of players or variations in the network topology. Thus, the primary objective of this type of dynamic game is to study how coalitional structures form through interactions between the players and inquire into the properties of those structures as well as their adaptability to changes in the environment [6].

c) Coalitional graph game: Games that use coalition graphs focus on connecting communication between players. It is critical to develop low-complexity distributed algorithms that help players build network graphs and understand their properties [70].

Non-cooperative Game

In a non-cooperative game, everyone's actions are treated as individual actions within a strategic environment. In that strategic environment, each player makes autonomous decisions, independent of others [6]. We usually refer to this as a game. It is common for games to contain both elements of conflict and cooperation, rather than only one. As a result, conflict and cooperation are intertwined.

Energy trading schemes have typically been designed with static and dynamic non-cooperative games. Static games occur when each player takes one action, either simultaneously or at different times. As opposed to static games, dynamic games require players to take time into consideration when making decisions. Dynamic games involve multiple players acting at multiple times and influencing each other's decisions. Two popular solution concept of a non-cooperative game will be discussed are Nash equilibrium and Stackelberg game.

a) Nash equilibrium: The Nash equilibrium occurs when all players are following their Nash equilibrium strategies and do not unilaterally deviate from their actions in a non-cooperative game [71]. For instance, let S_i be the set of all possible strategies for player i, where $i = 1, \ldots, N$. Consider a strategy profile $s^* = (s_i^*, s_{-i}^*)$, which is a set of strategies for each player, where s_{-i}^* denotes the $N-1$ strategies of all the players except player i. Let $u_i(s_i, s_{-i}^*)$ be player i's payoff as a function of the strategies. The strategy profile s^* is defined as a Nash equilibrium if

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*) \text{ for all } s_i \in S_i. \tag{2.35}$$

It is possible for a game to have more than one Nash equilibrium. The equilibrium can still be weak if the players choose different strategies depending on their preferences, despite the unique equilibrium. A strict Nash equilibrium is defined as that one strategy is the unique best response if the inequality is strict, which is shown as [72],

$$u_i(s_i^*, s_{-i}^*) > u_i(s_i, s_{-i}^*) \text{ for all } s_i \in S_i, s_i \neq s_i^* \tag{2.36}$$

b) Stackelberg game: There is another non-cooperative game called the Stackelberg game that has been widely used to design P2P trading in the literature [6]. Stackelberg games involve a hierarchy of a leader and his followers, with the leader moving first. In game theory, the players in this game are leaders and followers, competing on quantity. Stackelberg leaders are sometimes called market leaders. Stackelberg's model can be solved to determine the subgame Nash equilibrium. Based on the strategies of other players, the Stackelberg equilibrium represents the optimal strategy profile for each player, so each player reaches its own Nash equilibrium in the subgame. Stackelberg equilibrium occurs when neither the leader nor his followers have any incentive to deviate from their strategies [73].

2.5.2 Auction-theoretic Approaches

In auction theory, researchers study how auction markets induce predictable outcomes by incentivising bidders to behave in them. It is used to design auction markets in the real world. In auctions, sellers are able to raise higher revenues while buyers can procure at lower costs. An economic equilibrium is achieved when the price between the buyer and seller is agreed upon and cleared by the market [74]. Market failure can be addressed through auction rules, designed by auction theorists. They are designed to encourage optimal bidding strategies across a variety of informational settings. A common method applied auction theory is double auction.

Double Auction

A double auction involves a market of a number of buyers and sellers submitting the price at which they would like to buy or sell an item [75], which is then quickly processed by computer and matched against each other's bids. A typical example of a double auction is the stock market, where many buyers and sellers gather together, resulting the changes of supply and demand situation at any time.

Market efficiency depends on sellers and buyers reporting their reservations and bids honestly during the double auction process. Individual rationality and incentive compatibility must therefore be satisfied by auction mechanisms [20]. A double auction scheme is considered to possess individual rationality when all other prosumers in the auction are choosing the strategies that they will employ, providing that the utility received by a prosumer cannot be improved otherwise.

2.5.3 Constrained Optimisation Approaches

Constrained optimisation methods are numerical algorithms that seek solutions to linear or nonlinear programming problems with constraints. In general, there are two main categories in constrained optimisation: linear programming (LP) and nonlinear programming (NLP).

Linear Programming

An LP method involves solving a mathematical model with linear relationships to achieve the best result. In more formal terms, LP involves the optimization of a linear objective function under linear equality constraints and linear inequality constraints. An LP problem can be expressed in a canonical manner as follows [76]:

$$\begin{aligned} &\text{Maximise} && \boldsymbol{c}^T\boldsymbol{x}, \\ &\text{subject to} && \boldsymbol{A}\boldsymbol{x} <= \boldsymbol{b}, \\ &\text{and} && \boldsymbol{x} >= 0, \end{aligned} \tag{2.37}$$

where vector $\boldsymbol{x}$ are the variables to be determined, $\boldsymbol{A}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are the coefficient matrix and vectors. The function $\boldsymbol{c}^T\boldsymbol{x}$ to be maximised is called the objective function and the inequalities $\boldsymbol{Ax} <= \boldsymbol{b}$ and $\boldsymbol{x} >= 0$ are known as the constraints that need to be satisfied.

Two popular method can be used to solve LP problem are mixed integer linear programming (MILP) and alternating direction method of multipliers (ADMM).

a) MILP: MILP is a special case of linear programming. If we change the range of values of the constrained variables from a continuous real space to a discontinuous integer space, then the problem changes from an LP problem to an integer linear programming (ILP) problem. ILP problems are very common in combinatorial optimisation problems. Furthermore, if a number of the variables contain both continuous and discrete variables, then these problems are called mixed integer linear programming problems [11]. MILP problems are much more common in real-world problems, especially in the scenario of P2P energy trading that we are interested in. An MILP problem can be expressed similarly to a LP problem mathematically:

$$\begin{aligned} &\text{Maximise} && \boldsymbol{c}^T\boldsymbol{x}, \\ &\text{subject to} && \boldsymbol{Ax} + \boldsymbol{s} = \boldsymbol{b}, \\ & && \boldsymbol{s} >= 0, \\ & && \boldsymbol{x} >= 0, \\ &\text{and} && \boldsymbol{x} \in \mathbb{Z}^n, \end{aligned} \tag{2.38}$$

where some decision variables are not discrete.

b) ADMM: ADMM is a simple method for solving decomposable convex optimisation problems and is particularly effective in solving large-scale problems. Using the ADMM algorithm, the objective function of the original problem can be equivalently decomposed into a number of solvable sub-problems, each of which can then be solved in parallel, and finally the solutions of the sub-problems can be co-ordinated to obtain a global solution to the original problem [12].

Nonlinear Programming

NLP is the problem of solving an optimisation defined by a set of equations and inequalities/constraints consisting of a set of unknown real functions, accompanied by an objective function to be maximised or minimised, except that some of the constraints or objective functions are non-linear. It is a subfield of optimisation problems dealing with non-linear problems. It is possible to represent the NLP problem in (2.37) with a nonlinear objective function and/or nonlinear constraints.

2.5.4 Blockchain Approaches

Blockchain technology, which was first introduced in [77], is an advanced database mechanism that allows information to be shared transparently across an enterprise network. A blockchain

database stores data in blocks, while the databases are linked together in a chain. The data is consistent over time because the users cannot delete or modify the chain without network consensus. The users can therefore use blockchain technology to create immutable ledgers to keep track of orders, payments, accounts and other transactions. The system has built-in mechanisms to block unauthorised transaction entries and create consistency in the shared view of these transactions.

Traditional database technology poses many challenges for recording financial transactions. To avoid potential legal problems, a credible third party is needed to oversee and verify transactions. The existence of such a central authority not only complicates transactions, but also creates a single point of vulnerability. If that central data port is compromised, both parties could suffer losses [78]. Blockchain can mitigate such problems by creating a decentralised, tamper-proof system to record transactions. Blockchain could create a ledger for the buyer and seller respectively. All transactions would have to be approved by both parties and would be updated in real time in both ledgers. Any corruption in the historical transactions would result in the entire ledger being corrupted. These properties of blockchain technology allow it to be used in various industry sectors, including the creation of digital currencies such as Bitcoin. Hence, blockchain is a fundamental component of a P2P energy network because of its decentralised properties. Several blockchain-based platforms have been established recently for P2P energy trading.

Smart Contracts

In simple terms, a smart contract is a deterministic program that will perform a specific task when certain conditions are met [79]. Thus, smart contract systems usually follow the conditional statement "if ..... then .....". In general, smart contracts can be defined as applications or programs that run in the blockchain. Typically, they function as a digital protocol that follows specific rules for enforcement. These rules are predefined by computer code and are replicated and enforced by all network nodes [80]. Blockchain smart contracts support the creation of trustless protocols. This means that both parties to the contract make commitments via the blockchain without the need to know or trust each other. Both parties determine that the contract will not be executed if the conditions are not met. In addition, the use of smart contracts eliminates the need for intermediaries, thus significantly reducing operational costs [78].

Consortium blockchain

The consortium blockchain network is overseen by a group of organisations. A number of pre-selected organisations share responsibility for maintaining the blockchain and determining access to data [34]. Consortium blockchain networks are often preferred for industries where many of these organisations share a common goal and can benefit from shared responsibility. When P2P energy trading is implemented, energy transaction records are encrypted and then

uploaded to authorised nodes. Transactions are audited by the authorised nodes and recorded in the shared ledger by an algorithm. The consortium blockchain is accessible by participants and authorised nodes.

2.5.5 Reinforcement Learning Based Approaches

RL is a mathematical framework aiming to find optimal actions to maximise a reward signal in an environment. RL is the third fundamental machine learning technique after supervised and unsupervised learning. In contrast to supervised learning, RL does not require labels for inputs and outputs, nor does it require the exact correction of non-optimal solutions. Unlike other computational approaches, it focuses on learning through trial and error rather than modelling the environment or providing adequate supervision [81].

Markov Decision Process

Markov Decision Process (MDP) is core framework in reinforcement learning, which can be seen as a mathematical description of RL. MDP consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a state transition operator $\mathcal{T}$ and an immediate reward function $\mathcal{R}$ [82],

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}\}\,. \tag{2.39}$$

In MDP:

- The state $s \in \mathcal{S}$ represents the information available to the agent. The state is assumed to be determined by some preprocessing system that is nominally part of the environment.
- The action $a \in \mathcal{A}$ can be discrete or continuous depending on the use case of the agent.
- The state transition operator $T(s' \mid s, a) = \Pr\{s^{t+1} = s' \mid s^t = s, a^t = a\}$ is the transition probability from state to the next state given the selected action.
- The reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a mapping from Cartesian product of the state and action space to numerical values, which is the rewards of the agent.

Even when an MDP determines the system's dynamics, it is common for the agent not to know the system's state completely. Thus, a generalisation of MDP, partially observable Markov Decision Process (POMDP), is introduced for more general modelling of real-world sequential decisions. POMDP consists of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a set of observations $\mathcal{O}$, a state transition operator $\mathcal{T}$, an emission probability $\mathcal{E}$ and an immediate reward function $\mathcal{R}$ [16],

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{E}, \mathcal{R}\}\,. \tag{2.40}$$

POMDP further augments the definition of MDP with two additional objects:

- The observation $o \in \mathcal{O}$ represents the agent's knowledge of the state. In general, it is given by a sensor model maintained by the agent.
- The emission probability $E(o \mid s) = \Pr\{o^t = o \mid s^t = s\}$ is the probability of the observation given the system state.

To abstract an environment to an MDP or a POMDP, the environment must possess the Markov property. The Markovian property describes the conditional probability distribution of the following state, which is determined only by the current state and action. The environment's transition function can then be defined only specifying the current state and action, which is shown as [82],

$$T(s' \mid s, a) = \Pr\left\{s^{t+1} = s' \mid s^t = s, a^t = a\right\}, \forall s, a, s'. \tag{2.41}$$

Using this equation, we can predict all future states and expected rewards based on the current state rather than the entire previous state. The Markov property is essential in reinforcement learning because the expected future value is related only to the current state [81]. Therefore, the system state must be informative to represent the environment's dynamics effectively.

During reinforcement learning, the agent's objective is formulated as a unique reward signal that passes from the environment to the agent. In RL, we care about most is the expectation. Therefore, the value functions, functions of states, represent how good the agent will perform under a policy by utilising the expectations of the whole return. For MDPs, the state-value function $V_\pi(s)$ is defined in (2.42), and the objective of RL can be formally defined in (2.43) [81].

$$V_\pi(s) = \mathbb{E}_\pi\left[G(t) \mid s^t = s\right] = \mathbb{E}_\pi\left[\sum_{\tau=0}^{\infty} \gamma^\tau R(t+\tau+1) \mid s^t = s\right], \tag{2.42}$$

$$V_\pi^*(s) = \max_\pi \mathbb{E}_{a'\sim\pi, s'\sim\mathcal{T}}\left[\sum_{\tau=0}^{\infty} \gamma^\tau R(t+\tau+1) \mid s^t = s\right], \tag{2.43}$$

where $\gamma \in [0, 1]$ (usually close to 1, i.e. 0.9) denotes the discounted factor representing the uncertainty of the future reward and τ is the interval of each time step.

Q-Learning

Inspiring by the [83] behaviour learning, Christopher Watkins proposed the core prototype of Q-learning in 1989 [84] which massively enhance the feasibility of RL. Q Learning is a value-based reinforcement learning algorithm. "Q" refers to quality, representing the expected future reward given a specific state-action pair. Therefore, the main idea of the algorithm is to build a Q-table to store all the Q-values. The agent can then choose the action at every state based on the Q-value, which provides the highest benefit. Q-learning is also a model-free reinforcement learning algorithm, which means it does not need a complete understanding of the environment model. The main advantage of Q-learning is the ability to solve optimal strategies for Markov

processes using the Bellman equation and the time-difference (TD) method, which is a blend of Monte Carlo and dynamic programming.

For MDPs, the action-value function $Q_\pi(s,a)$ is defined as

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[G(t) \mid s^t = s, a^t = a\right] = \mathbb{E}_\pi\left[\sum_{\tau=0}^{\infty} \gamma^\tau R(t+\tau+1) \mid s^t = s, a^t = a\right]. \quad (2.44)$$

If the agent acts optimally recursively, it will receive the optimal total discounted future reward according to the Bellman Optimality Equation shown in (2.45) [84].

$$\begin{aligned} Q^*(s,a) &= \mathbb{E}\left[R(t+1) + \gamma \max_{a'} Q^*\left(S^{t+1}, a'\right) \mid s^t = s, a^t = a\right] \\ &= \sum_{s'} T\left(s' \mid s, a\right)\left[r + \gamma \max_{a'} Q^*\left(s', a'\right)\right]. \end{aligned} \quad (2.45)$$

Having Q^* makes choosing optimal actions easier. With Q^*, the agent does not need to search one-step-ahead every time compared to using v: for any state s, it can simply find any action cached in the result that maximises $Q^*(s,a)$.

The Q-learning algorithm is an adaptation of the Q-value iteration algorithm to the situation with no prior knowledge of the environment. Q-learning utilise the temporal difference learning to gradually improve the estimates of the Q-value through multiple iterations. The one-step Q-learning is defined as [84]

$$Q^{new}(s,a) \leftarrow Q(s,a) + \alpha \cdot \Big(R(s,a) + \gamma \cdot \max_{a'} Q\left(s', a'\right) - Q(s,a)\Big) \quad (2.46)$$

where α is learning rate,(s', a') is the next state-action pair, $\max_a Q\left(s', a'\right)$ is the estimate of optimal future value. The maximise of $Q(s', a')$ makes Q-learning an off-policy learner, which learns the optimal policy independently of which policy the agent currently chooses. The Q-learning algorithm utilise a Q-table to keep track of all Q-values. Once the estimates of the Q-values become accurate, the optimal policy can be yield by choosing the action has the highest Q-value. This optimal policy is also called the greedy policy.

Deep Q-Learning

Developed by DeepMind [85], a DQN is essentially a DNN used to estimate the Q-value function. And Deep Q-learning is referred to approximate Q-learning using a DQN. In general, DRL is formed by combining RL as the framework and DNNs as the approximators. Deep Q-learning was the first DRL algorithm which is able to solve a wide range of computer games (even higher than human level in some games) by combining RL and DNNs at scale.

In deep Q-learning, the input of the DQNs is the system state and the output are the estimated Q-values of every possible actions. To train the DQNs, we need the transitions (s, a, r, s') similar

as in Q-learning. The predicted Q-value for a state-action pair (s, a) from DQN needs to close to the reward given at the next state plus the discounted value of act optimally from then on, as discussed in the Bellman equation (2.45). Thus, the target Q-value y is defined as [85]

$$y = r + \gamma \cdot \max_{a'} Q_\theta\left(s', a'\right). \tag{2.47}$$

The DQNs can then be trained by minimising the squared error between target Q-value and predicted Q-value using one-step Gradient Descent algorithm. However, naively training the deep Q-learning algorithm can bring significant stability issues causing the performance suddenly drops (also called catastrophic forgetting). To address the above issue, DeepMind used two main solutions: fixed Q-value targets and experience replay.

Fixed Q-Value Target

In the basic deep Q-learning algorithm, the prediction and setting target are made by the same DQN model, which leads to huge instabilities. To address this issue, the researchers from DeepMind used a separate DQN model to estimate the target. The target model has the same architecture as the online model but with frozen parameters to keep stable. For every C iterations (C is a hyperparameter tuned by the user, but in general it is a large number), the parameters of the target model are copied from the online model. Since the target model updates much less frequently than the online model, the target estimations are more stable than the basic Q-learning. After using the target network, the loss function of the deep Q-learning is defined as [85]

$$\mathcal{L}\left(\theta\right) = \mathbb{E}_{s\sim T_\pi, a\sim\pi, r\sim R}\left[\left(r + \gamma \cdot \max_{a'} Q_{\theta'}\left(s', a'\right) - Q_\theta\left(s, a\right)\right)^2\right], \tag{2.48}$$

where $Q_{\theta'}$ is denoted as the target Q-network.

Experience Replay

To address the correlated experience, DeepMind proposed a experience replay buffer which is used to store the past experience (s, a, r, s'). When updating the networks, a mini-batch of past experience will be randomly selected to train the DQN, results in less variance than updating a single transition. By replaying the experience, the agent can explore an extensive range of state-actions previously encountered; otherwise, the agent rewrites its experience with new ones. In general, the larger size of the replay buffer, the less possibility an correlated experience occurs.

2.5.6 Deep Deterministic Policy Gradient

Deep Deterministic Policy Gradient (DDPG) algorithm, proposed in [15], adapt the ideas underlying the success of Deep Q-Learning to the continuous action domain. Since the action space is continuous, the optimal Q-value $Q^*(s, a)$ is differentiable with respect to the action. Thus, the optimal actions can be approximated by an efficient, gradient-based policy function $\pi(s)$, and

the next optimal Q-value can be approximated as [15]

$$\max_{a'} Q_{\theta'}\left(s', a'\right) \approx Q_{\theta'}\left(s', \pi_{\phi'}(s')\right). \tag{2.49}$$

In DDPG, it has four networks, a Q-network Q_θ, a deterministic policy network π_ϕ, a target Q-network $Q_{\theta'}$, and a target policy network $\pi_{\phi'}$. Different from the deep Q-learning algorithm, the target networks in DDPG are time-delayed copies of their original networks that slowly update toward the learned networks. Using these target networks greatly improve stability in learning. The target networks are updated by polyak averaging [86]:

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', \tag{2.50}$$

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi', \tag{2.51}$$

where ϕ denotes the parameters of the policy network and $\tau \ll 1$ is a hyperparameter.

The deterministic policy network is used to choose optimal actions in certain states. A random noise sampled from a Gaussian distribution is added to the actor policy to increase exploration. The Q network is pretty similar to deep Q-learning, expect that the target value of Q-network has been changed to $y = r + \gamma \cdot Q_{\theta'}\left(s', \pi_{\phi'}(s')\right)$. The Q-network of DDPG is trained by minimising the loss function with stochastic gradient descent is shown in (2.52). Since the optimal Q-value is presumably differentiable with respect to the action, the deterministic policy network is updated using sampled policy gradient which is shown in (2.53) [15].

$$\mathcal{L}\left(\theta\right) = \mathbb{E}_{s\sim T_\pi, a\sim\pi, r\sim R}\left[\left(r + \gamma \cdot Q_{\theta'}\left(s', \pi_{\phi'}(s')\right) - Q_\theta\left(s, a\right)\right)^2\right] \tag{2.52}$$

$$\nabla_\phi J \approx \mathbb{E}_{s\sim T_\pi, a\sim\pi}\left[\nabla_a Q_\theta\left(s, a\right)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi\left(s\right)\right] \tag{2.53}$$

2.6 Related Work

This thesis has the following three technical chapters, one which deals with P2P energy trading in the electricity sector for microgrids (MGs) using deep reinforcement learning (DRL), one which incorporates multi-energy system (MES) into P2P energy trading for multi-energy MGs (MEMGs) using multi-agent DRL (MADRL), and one which considers the privacy issue in a P2P energy trading for MEMGs participated in a double-auction (DA) market. A literature review of how other researchers model and solve P2P energy trading problems is divided by these three categories is presented in the following.

2.6.1 P2P Energy Trading for Microgrids using Deep Reinforcement Learning

Some P2P energy trading models have been proposed to solve the renewable energy dilemma, e.g., game-theoretic approaches and contract networks for P2P energy trading. The authors in [71] explored how geographically distributed storage units interact and trade energy using noncooperative game theory as a framework. Using a non-cooperative Stackelberg game model, the authors in [87] formulated an event-driven energy trading problem between MGs, which can be either consumers or providers based on their energy generation status. In [88], the problem of energy trading between smart grid prosumers and a grid power company is formulated as a single-leader, multiple-follower Stackelberg game. A two-layer non-cooperative game-theoretic approach is presented in [89] to provide decision support during load restoration on energy price for utilities and on energy dispatch for MGs. In [73], a cooperative Stackelberg game-based peer-to-peer (P2P) energy trading scheme is proposed to help a centralized power system reduce its peak-hour electricity demand. In [27], bilateral contract networks are proposed as a scalable design for peer-to-peer energy trading. It will reduce the need for investment in upstream generation and transmission infrastructure, increase network efficiency, and ensure energy security. In [90], An energy hub's profit-driven behaviour is modelled using energy contracts in order to account for market clearing problems in the strategic bidding of the energy hub.

However, making decisions based on the massive amount of data and unpredictable renewable generation in P2P energy trading by using conventional optimised techniques is problematic. DRL techniques, combined with deep neural networks and reinforcement learning (RL) techniques, could be powerful tools for addressing such P2P energy trading issues since they can solve the decision-making problems by learning from the high-dimensional historical data. DRL/RL have been used in the area of smart grids to optimise the operation of MGs [91], energy management [92] and storage planning [93]. There is also some recent research using DRL for P2P energy trading, where a large amount of uncertainty data can be directly learned by DRL to make the decisions in the real world. For example, a local energy trading problem for prosumers was formulated as an MDP and was solved by using deep Q-learning to maximise prosumer's daily economic benefit [94]. A DQN-based MG trading game was formulated to improve the utility of the MG without knowing information about other MGs [95]. However, the physical constraints in a distributed renewable energy system were not considered in these papers, and their study was limited to a typical day of the P2P energy trading, where in reality the trading behaviours change throughout the year.

2.6.2 P2P Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning

The literature on the P2P energy trading can be classified into five techniques based on the approaches adopted: game theory, auction theory, constrained optimisation, blockchain and deep reinforcement learning (DRL). In [18,71,73,88,96–100], game theory is used to address the P2P energy trading problems in the electricity sector. Some of this work [18, 97, 99, 100] considers trading among multiple MGs, while [71, 73, 88, 96, 98] considers trading between prosumers. In [94,101–104], game-theoretic approaches are used to solve the P2P energy trading in a multi-energy setting. However, only in [104] did the authors model the P2P energy trading among multiple MEMGs.

Auction-theoretic approaches are used in recent literature to address the P2P trading problems between prosumers in the electricity sector. A continuous double auction methodology is proposed in [26] to assess the impact of peer-to-peer transactions on the network and guarantee energy exchange without violating network constraints. Based on auction theory, the authors [105] in proposes a new local energy market design that integrates P2P energy trading and probabilistic locational marginal pricing. In [106], a discriminatory continuous double auction market mechanism is proposed to gain a better understanding of the community P2P electricity trading with household distributed photovoltaic. Only in [107], the authors proposed an auction mechanism for energy trading in a multi-energy district.

Constrained optimisation methods are numerical algorithms and have been applied to P2P energy trading under different market and system constraints in many literatures. The authors in [10] proposed a decentralised constrained method to attribute P2P allocation costs through exogenous network charges in several alternative ways. In [23], a linear programming optimisation model was developed for a small community in London, UK, to represent P2P interactions with storage. A distributed price-directed optimisation mechanism is used in [108] to coordinate the trading of energy between prosumers with heterogeneous preferences on a P2P energy market platform. Based on a primal-dual gradient method, the authors in [109] proposed a decentralised P2P energy trading scheme for electricity markets with a high penetration of distributed energy resources. In [110], a fully decentralised P2P market based on the consensus alternating direction method of multipliers theory is proposed to determine the quantities and prices of traded energy and reserve for the agents.

Blockchain has been used to enable secured and decentralised energy trading in the electricity sector by many researchers. The authors in [29] proposed a new demurrage mechanism to reduce the redemptive value of energy-backed tokens in blockchain electricity markets. The proposed model in [34] deploys consortium blockchains to facilitate the localised P2P trading of electricity among Plug-in Hybrid Electric Vehicles. A general, blockchain-based smart contract

was presented in [80] that captures the key elements needed for P2P energy trading that can help formalise decentralisation.

The game-theoretic and auction-theoretic models are mainly solved by traditional constrained optimisation methods such as mixed-integer linear programming (MILP) [11] and alternating direction method of multipliers (ADMM) [12]. Those methods are useful for a lot of complex tasks considering multiple factors and constraints. However, the MILP method assumes linear relationships among factors [11], while ADMM assumes that the problems are regularised and convex [12], which are unrealistic in many cases.

DRL, combined with deep neural networks (DNNs) and reinforcement learning (RL) techniques, can be powerful tools for addressing the P2P energy trading issues in the network of multiple MEMGs using the trial-and-error mechanism without any extensive feature engineering. In [36, 111], deep Q-learning is used in their corresponding electricity trading problems. In [112], a convolutional neural network (CNN) is used to predict the MG utility while helping the Q-learning algorithm choose the optimal policy for the MG to trade electricity. Deep Q-learning has two major pitfalls: it cannot do well when the environment has a colossal number of actions in continuous action space [15]; it tends to overestimate the Q-value [113].

Previous work on P2P energy trading mainly focuses on the electricity sector. Some work has been done in a multi-energy setting but does not consider both external P2P energy trading and internal energy conversion process. The literature also lacks modelling of different types of MEMGs participating in the P2P energy trading. Considering the varieties of MGs such as residential, commercial and industrial is essential for P2P energy trading and energy conversion within a local community since the energy generation patterns and energy coupling technologies of different kinds of MGs complement others' demand. Also, the existing work on P2P energy trading only uses single-agent DRL algorithms.

2.6.3 Coordination for Multi-Energy Microgrids in Double Auction Market Using Multi-Agent Deep Reinforcement Learning

So far, the existing literature on the P2P energy trading framework of multiple MEMGs can be classified into two categories. The first one focuses on the design of a centralised framework that employs a central operator to manage all local resources [114]. Although such a framework provides a theoretical solution for social welfare maximisation, it exhibits various drawbacks in practice. Specifically, the central operator needs to acquire mathematical models and collect all technical parameters of local resources, thereby raising privacy concerns. The second one focuses on the design of a decentralised framework that allows the MEMGs to manage their own resources independently with limited information exchange, preserving their privacy. Currently, alternating direction method of multipliers (ADMM) [104, 115], Lagrangian relaxation [116, 117], consensus algorithm [118] and bilateral contract [119] are popular methods in

decentralised framework for solving coordination management of multiple MEMGs. However, the optimality of solutions is not guaranteed under such a decentralised framework without a central coordinator [114].

To this end, double auction (DA) market [120] is a kind of framework that takes advantage of both centralised and decentralised frameworks, which is potential to be considered to form local coordination of MEMGs. More specifically, an auctioneer, as a third-party coordinator, is responsible for clearing the market to ensure the market efficiency, which is close to optimal in a centralised framework [106]. On the other hand, MEMGs can manage their resources independently and submit only the bidding information (i.e., price-quantity bids) to the auctioneer. As such, the privacy can be preserved that is similar to the decentralised framework. However, MEMGs in DA market are faced with a complex quotation decision process. Thus, an appropriate trading strategy is challenging to select in such a complicated market environment. Zero Intelligence (ZI) is a fundamental trading strategy adopted by traders in DA market [121]. Specifically, ZI selects the price bid uniformly at random values between Feed-in Tariff (FiT) and Time-of-Use (ToU) and runs a day-ahead self-optimisation problem for quantity bid submitted to the DA market. However, the randomised price bid does not capture the market dynamics. Furthermore, pre-optimised trade quantity decision requires the complete MEMG mathematical models, technical parameters, and accurate forecasting information of uncertainties, which are generally impractical in real-world applications [26].

In view of the above drawbacks in ZI strategy, reinforcement learning (RL) [81] is a model-free and data-driven control method to study the sequential decision-making problem, where the agents within MEMGs gradually learn the optimal trading strategies by utilising experiences acquired from their repeated interactions with the environment (MEMGs and DA market), without a *prior* knowledge of MEMGs. In addition, RL as an online learning method can make use of increasing data acquired from the environment to learn the optimal control strategies and to cope with the uncertainties that are encapsulated in the data [49].

Previous works have successfully applied various RL methods to energy management problems in power systems, as reviewed in [122]. The majority of them, however, only consider the energy management problem of a single entity, e.g., a smart energy-hub [123] and a residential multi-energy home [49], and employ single-agent reinforcement learning (SARL) methods. On the other hand, the research efforts on the application of multi-agent deep reinforcement learning (MADRL) on power systems are still sparse, particularly for our studied MEMGs coordination management problem. The most straightforward approach to solving a multi-agent problem is independent RL (IRL) that each agent trains its independent control policy depending on the local information.

Independent deep Q-network (IDQN) [111] and independent deep deterministic policy gradient (IDDPG) [124] have been applied to the energy management problems of the multiple MGs, where each agent treats others as part of the environment and learns its own policy without

considering others' policies. However, directly applying IRL methods to a multi-agent setting is problematic, since the environment appears non-stationary from the view of every agent [16]. To overcome this issue, multi-agent deep deterministic policy gradient (MADDPG), an extension of IDDPG to a multi-agent setting, has been proposed to address the energy trading problem among the MGs [125]. Each agent in MADDPG trains a centralised Q-value function (critic) with access to all agents' observations and actions to stabilise the training performance. During the execution, the decentralised actor of each agent makes decisions based on its local observation value. However, MADDPG mainly suffers from 1) privacy concern: knowing the local observations and actions of all other agents; and 2) stability concern: the learned Q-values may be overestimated, which can lead to the sub-optimal polices [126].

2.7 Conclusion

In this literature review chapter, an overview of existing research in P2P energy trading has been provided. First, the P2P energy trading structure has been discussed with specific emphasis on the components of the physical energy system and the virtual energy trading system. Second, a classification of P2P energy market has been organised into centralised market, decentralised market, and community market that have been developed by the research papers. Third, enabling technologies including distributed energy resources, energy storage systems, energy converters, data sensing systems and ICT systems have been explained in detail in terms of working mechanism and mathematical formulation. Fourth, participants of the P2P energy trading have been identified into prosumers, generation companies, retailers and MGs based on the use cases of P2P energy trading literature. Furthermore, core technical approaches including game theory, auction theory, constrained optimisation, blockchain and deep reinforcement learning that have been extensively used in the literature have been identified and summarised. Finally, this chapter have provided a literature review of how other researchers model and solve P2P energy trading problems with respect to the following three technical chapters.

Chapter 3

Peer-to-Peer Energy Trading for Microgrids using Deep Reinforcement Learning

3.1 Introduction

Renewable energy resources have been exploited to solve the foreseeable fossil fuel shortage problem in the past decade. Although renewable energy is sustainable, it brings significant challenges to the stability and operational safety of a large power network due to its intermittent and location-variant nature. As a result, microgrids (MGs) have been proposed to address these challenges by coordinating the control of distributed energy resources (DER), local active loads and energy storage systems (ESSs) within certain regions. Within a MG, the distributed renewable energy sources, such as wind power and solar energy, can switch traditional energy consumers to prosumers. Multiple MGs located in a large area can be networked to improve the efficiency and reliability of the distribution network further. However, since the installed DERs in MGs belong to different owners, it is not realistic to directly control or operate them by a central authority. Recently, P2P energy trading has emerged as a novel paradigm for decentralised energy market designs. P2P energy trading allows the end-users to join the trading without a central authority unit [6].

The question we want to answer in this chapter is: How can we decide optimal P2P energy trading policy for MGs despite the need to process high-dimensional generation and demand data? The answer is deep reinforcement learning (DRL), which directly learn the optimal policies using error and trial based on the historical data without requiring the full dynamics of the system model. However, given the physical constraints of P2P energy trading, how can we implement the mathematical and systematic procedures of DRL? The answer is to use a mathematical framework called Markov decision process (MDP) which create a virtual environment for the DRL algorithm. But then another question arises, which DRL algorithm is appropriate

for P2P energy trading and what practical modifications do we need to make in the DRL algorithm when we apply it to our particular P2P energy trading problem? The answer is detailed throughout the next pages of this work.

In this chapter, we formulate a realistic energy trading model for MGs with a set of critical physical constraints. An MG needs to make a trading strategy and negotiate with other MGs only based on its generation, demand and energy storage level. The physical constraints like transmission losses and power limits at some nodes of the system may affect the strategy of an MG. We also set a flexible utility function for each MG to evaluate its strategy, which consists of not only trading profits but also the battery wear cost, demand penalty, and optional social factors. Deep reinforcement learning is used to train the agent as an MG to derive better strategies based on the states and the utility function. Using DQN and an experience replay mechanism [15], the algorithm can speed the Q-learning rate and update the loss function with continuously collected new states and rewards instead of updating the model at the end of each episode. Last but not least, we choose one-year real-world data sets to test the algorithm during four seasons.

3.2 P2P Energy Trading Model

3.2.1 Problem Setting

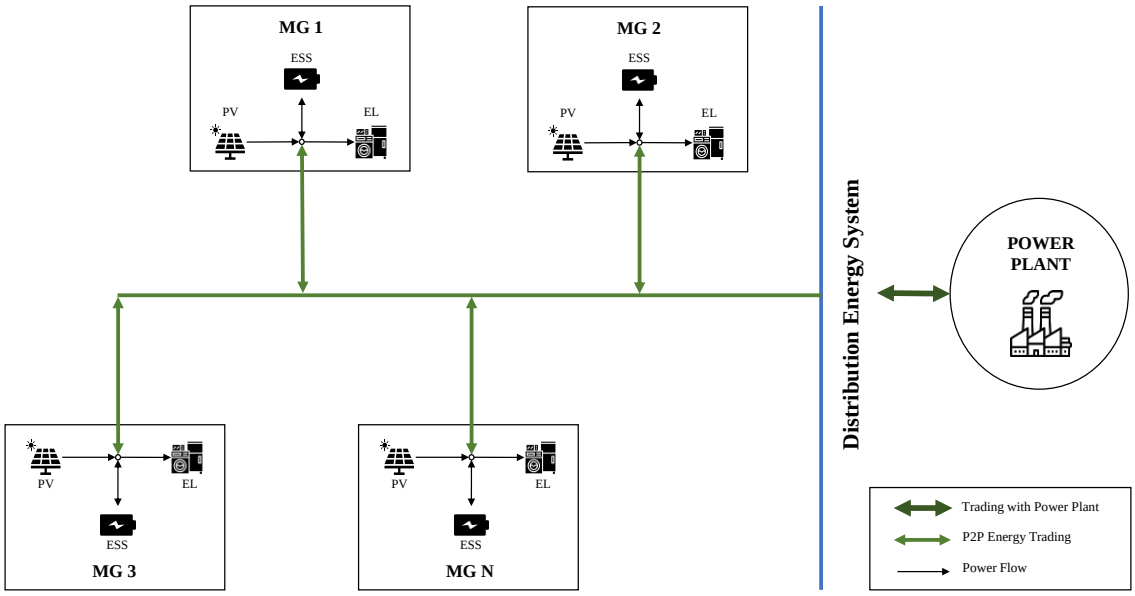


Figure 3.1: The framework of P2P energy trading model among MGs.

In this chapter, we focus on a local energy community, which is composed of a group of MGs connected with each other in a distribution network. In detail, the set of components of the

proposed MGs includes 1) electric load (EL); 2) renewable generators, e.g., solar photovoltaic (PV); and 3) electric storage systems (ESS). As illustrated in Figure 3.1, the options of each MG to supply its consumption loads are diverse. First, MGs can manage their own installed renewable energy resources to supply the energy demand of EL. Second, MGs can trade their electricity with each other in the P2P energy market. We assume that the MGs decide an agreed P2P price for all MGs, and then negotiate the amount of electricity traded among themselves. Third, MGs are allowed to buy/sell their unbalanced electricity with the power plant at the grid buy/sell prices. Moreover, MG can make full use of its generators and storage system so that it can decide how to charge or discharge the battery and whether to turn down some of its generators if needed. Finally, the MGs can observe the generation and demand meter and its battery level at each trading block. The decision-making problem is processed for each trading block/hour across a daily horizon, with the objective of maximising the utility of the MG. At each trading block, each MG manage its trading strategies based on 1) the agreed P2P trading price signal and 2) local information of its energy demand, renewable generation and the energy level of the ESS.

3.2.2 Markov Decision Process & System Objective

The P2P energy trading problem is formulated as a POMDP to maximise the utility of the MG. The POMDP consists of a set of states, a set of observations, a set of actions, and a reward function.

System States and MG Observations

In this work, We assume the energy trading takes place in the local hour-ahead P2P energy market, in which each trading block has one hour. At the beginning of each trading block, the MG will forecast its renewable generation and load demand based on historical data in the trading block. The amount of renewable energy of MG i in trading block t is denoted as $R_i\left(t\right)$, and the estimated generation is denoted as $\hat{R}_i\left(t\right)$. The actual and estimated amount of energy demand of MG i in trading block t are denoted as $D_i\left(t\right)$ and $\hat{D}_i\left(t\right)$, respectively. The remaining battery level of MG i at the beginning of trading block t is denoted as $S_i\left(t\right)$. Finally, the average price of P2P energy trading at the beginning of trading block t is defined as $\rho_{p2p}(t)$.

The The system state of MG i at time t is thus defined as

$$\boldsymbol{s}_i(t) = \left[R_i(t), D_i(t), S_i(t), \rho_{p2p}(t)\right], \tag{3.1}$$

Since the part of the actual system states (i.e. renewable generation and load demand) are not observed by the MG, the observation of MG i at time t is defined as,

$$\boldsymbol{o}_i(t) = \left[\hat{R}_i(t), \hat{D}_i(t), S_i(t), \rho_{p2p}(t)\right]. \tag{3.2}$$

MG Action - Trading Strategy

The strategy list of MG i is denoted as

$$\boldsymbol{x}_i(t) = [x_{ij}(t)]_{1\leq j\leq N, i\neq j} = [x_{i1}(t), x_{i2}(t), \ldots x_{iN}(t)], \tag{3.3}$$

where $x_{ij}(t)$ is the intended amount of energy trading from MG i to MG j in trading block t. If $x_{ij}(t) > 0$, which means MG i want to buy energy from MG j; if $x_{ij}(t) < 0$, which means MG i want to sell energy to MG j. Since MGs often have conflicting trading intentions, e.g., $x_{ij}(t) \times x_{ji}(t) > 0$, trading negotiations have been made, which resulting in actual trading action,

$$\boldsymbol{a}_i(t) = [a_{ij}(t)]_{1\leq j\neq i\leq N} = [a_{i1}(t), a_{i2}(t), \ldots a_{iN}(t)], \tag{3.4}$$

where $a_{ij}(t) > 0$ means MG i buy energy from MG j; $a_{ij}(t) < 0$ means MG i sell energy to MG j. MGs only have a deal when one of them wants to sell energy and another wants to buy energy. It is clear that the actual energy trading might not be the same as the intention, therefore MGs need to buy or sell energy to the power plant to realize their strategy in trading block t. The amount of energy trading with the power plant in trading block t is denoted by $a_{ii}(t)$, which is the difference between the sum of $x_{ij}(t)$ and $a_{ij}(t), i \neq j$. Note that, the reason we denote it by $a_{ii}(t)$ is for algorithm convenience and we can use the vacant position $a_{ii}(t)$ to represent trading with the power plant and making just single list of $\boldsymbol{a}_i(t)$. The actual amount of energy trading of MG i is shown in (3.5).

$$a_{ij}(t) = \begin{cases} \frac{x_{ij}(t)}{|x_{ij}(t)|} \cdot \min\left(|x_{ij}(t)|, |x_{ji}(t)|\right), & \text{if } x_{ij}(t)\cdot x_{ji}(t) < 0,\ \forall i \neq j. \\ 0, & \text{if } x_{ij}(t)\cdot x_{ji}(t) \geq 0, \forall i \neq j. \\ \sum_{j=1}^{N} x_{ij}(t) - \sum_{j=1, j\neq i}^{N} a_{ij}(t), & \forall i = j. \end{cases} \tag{3.5}$$

MG Reward Function

The reward function can help an MG evaluate the strategies that have been created in order to produce better strategies later. The reward or utility of MG i performing energy trading in trading block t, denoted as $u_i(t)$, depends on the trading profits, wear cost of the ESS, penalty if local demand is not met and virtual penalty if the MG wants to fulfill a certain goal. The local P2P market price can be dynamically changing, however, for encouraging MGs to trade energy with each other, the P2P energy trading price is set as

$$\rho^{-}_{grid}(t) \ll \rho^{-}_{p2p}(t) \approx \rho^{+}_{p2p}(t) \ll \rho^{+}_{grid}(t) < \rho_{retail}(t) \tag{3.6}$$

where $\rho^{-}_{grid}(t)$ is the price MG selling energy to the power plant at time t, $\rho^{-}_{p2p}(t)$ represents the P2P energy selling price to other MGs at time t, $\rho^{+}_{p2p}(t)$ represents the P2P energy buying price from other MGs at time t, $\rho^{+}_{grid}(t)$ denotes the price MG buying energy from the power

plant at time t and $\rho_{retail}(t)$ is the price MG selling energy to the local consumers at time t. In our system model, we use same P2P energy selling price and buying price, thus we denote the P2P energy price as $\rho_{p2p}(t) = \rho^{-}_{p2p}(t) = \rho^{+}_{p2p}(t)$ and we assume the P2P energy price can be determined by the grid prices and a P2P price ratio α_{p2p}, which is shown as,

$$\rho^{-}_{grid}(t) = \rho_{p2p}(t) \cdot (1 - \alpha_{p2p}), \tag{3.7}$$

$$\rho^{+}_{grid}(t) = \rho_{p2p}(t) \cdot (1 + \alpha_{p2p}). \tag{3.8}$$

The reward/utility function is expressed as

$$u_i(t) = \sum_{j=1, j\neq i}^{N} a_{ij}(t) \cdot \rho_{p2p}(t) + a_{ii}(t) \cdot \left(I_{(a\leq 0)} \cdot \rho^{-}_{grid}(t) - I_{(a_{ii}>0)} \cdot \rho^{+}_{grid}(t)\right) + \rho_{retail}(t) \cdot p_{L,i}(t) \cdot T - c_w \cdot |S_i(t+1) - S_i(t)| - C_{pen} - C_{vir}, \tag{3.9}$$

where

$$C_{pen} = C_p p_{dif}(t), \tag{3.10}$$

$$C_{vir} = C_v a_{ii}(t). \tag{3.11}$$

The first term in the right-hand side of (3.9) is the trading profit of MG i trading with other MGs, the second term is the trading profit of MG i trading with the power plant, the third term is retail profit, the rest are energy storage wear cost and other penalties. The demand penalty C_{pen} happens when $\sum p_{ij} + p_{RE,i} < p^{min}_{ES,i} + p_{L,i}$, where $p_{dif} = p^{min}_{ES,i} + p_{L,i} - \sum p_{ij} + p_{RE,i}$ and C_p is the penalty coefficient. To be noticed that if $\sum p_{ij} + p_{RE,i} > p^{max}_{ES,i} + p_{L,i}$, MG i is assumed to reduce their generation output or sell the surplus electricity to the grid to balance the demand without penalty. The virtual penalty C_{vir} is optional, and its existence is to make the algorithm believe achieving some goal is beneficial even though it might not be economically optimal. In this work, the objective of MG i is to maximise the trading profits while also minimising the dependence on the power plant. Thus, the virtual penalty can be set as (3.11), where C_v is a virtual penalty coefficient. The virtual penalty can be also set to achieve other social welfare goals for the MG.

System Problem

As each MG does not know energy generation and demand information of other MGs, MG i will choose its trading strategies $\boldsymbol{x}_i(t)$ based on the estimated generation $\hat{R}_i(t)$, energy demand $\hat{D}_i(t)$, current storage level $S_i(t)$ and P2P energy price $\rho_{p2p}(t)$. Therefore, the reward function can also be written as

$$u_i(t) = u\left(\hat{R}_i(t), \hat{D}_i(t), S_i(t), \rho_{p2p}(t) | \boldsymbol{x}_i(t)\right). \tag{3.12}$$

The goal is to maximise the expected total utility which is the sum of all future utilities based on the optimal policy $\pi\left(\boldsymbol{x}_i(t)|\hat{R}_i(t), \hat{D}_i(t), S_i(t), \rho_{p2p}(t)\right)$, which can be shown as

$$\mathbf{P1}: \max_{\pi} U_{i_\pi}(t) = \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau u_i(t+\tau+1)\right]. \tag{3.13}$$

The trading policies made by MGs could be based on naive intention (Trading surplus or needed energy of trading block t without thinking about the future), board resolution, or an automatic energy management system (AEMS). In this paper, MG i will use deep Q-learning algorithm as part of an AEMS to derive better strategies over time.

3.2.3 Physical Constraints

For MG i, it will send or receive $a_{ij}\,(t)$ (kWh) energy in trading block t, which means that it will send or receive $p_{ij}(t) = \frac{a_{ij}(t)}{T}$ (kW) active power in trading block t, where T is equal to 1 hour. In this model, we assume that there is no transmission loss between MGs since MGs are located near each other in a local community. The physical constraints can be written as

$$p_{ij}(t)^{\min} \leq p_{ij}(t) \leq p_{ij}(t)^{\max}, \tag{3.14}$$

$$p_{ES,i}(t)^{\min} \leq p_{ES,i}(t) \leq p_{ES,i}(t)^{\max}, \tag{3.15}$$

$$0 \leq S_i(t+1) \leq B, \tag{3.16}$$

$$\sum_{j=1}^{N} p_{ij}(t) + p_{RE,i}(t) = p_{ES,i}(t) + p_{L,i}(t), \tag{3.17}$$

where $p_{RE,i}\,(t)\,,\ p_{ES,i}\,(t)\,,\ p_{L,i}\,(t)\,,$ and B are power from renewable generators, ESS, load device (kW) and capacity of the ESS (kWh) respectively.

The first three components are hard constraints, where (3.14) limits the power that MG i can receive from other MGs or power plant, (3.15) limits the power when charging or discharging the ESS battery, and (3.16) means that at the end of trading block t, the remaining ESS level cannot surpass its capacity. Constraint (3.17) means the MG must balance the energy generation and consumption in trading block t. When charging the ESS, $p_{ES,i}\,(t) > 0$; when discharging the ESS, $p_{ES,i}\,(t) < 0$. In order to derive $S_i\,(t+1)$ in (3.16), the ESS is modeled as,

$$S_i(t+1) = S_i(t) + E_{ch}\eta_{ch} - \frac{E_{dis}}{\eta_{dis}}, \tag{3.18}$$

where $E_{ch}(E_{dis})$, $\eta_{ch}(\eta_{dis})$ are the energy charging (discharging from) the battery and the charge (discharge) efficiency. Since charge and discharge action will degrade the condition of the batteries in the ESS, we consider the ESS wear cost, which will affect the energy trading strategies of the MGs. The empirical wear cost efficiency c_w (\$/kWh) [127] is shown as

$$c_w = \frac{C_{rep}}{S_b Q_b \sqrt{\eta_{rt}}}, \tag{3.19}$$

where C_{rep} is the replacement cost of the ESS, S_b is the battery size of the ESS, Q_b (kWh) is the total transferable energy during the life cycle of a battery unit in the storage and η_{rt} is the battery round-trip efficiency which is equal to the square of the storage discharge efficiency.

3.3 Deep Reinforcement Learning and Solution Algorithm

P2P energy trading involves a large number of continuous data sets in which are made up of stochastic and uncertain data like renewable generation and load demand, so making a decision by human or conventional optimisation methods would be challenging. Reinforcement learning (RL) is a mathematical framework aiming to find optimal strategy to maximise the objective function through trial and error without modelling the system model, making it a suitable solution to solve P2P energy trading problem. However, directly applying RL algorithms such as Q-learning is problematic and inefficient.

The problem for Q-learning is that the state-action pairs are often discrete, and it needs to sample all actions repeatedly (find all Q-values in the Q-table), which means that it cannot tackle the problems with high-dimensional state-action space and continuous data sets. The Q-table simply cannot store the estimation of every Q-value if the number of single Q-value is almost infinity. The solution is to find a parameterised value function [128] $Q_\theta(s, a)$, where θ is a manageable number of parameters that approximate the Q-values. Before, the function is usually approximated by the combinations of linear functions of extracted features of the state. However, it requires a lot of feature engineering. Until 2015, deep neural networks (DNN) are considered to approximate the Q-values to empower the RL to next stage, which is classified as deep reinforcement learning (DRL) algorithms. With DRL, making optimal decisions in P2P energy trading could be possible.

3.3.1 Deep Q-learning

Deep Q-learning, consists of DNN and Q-learning, which is also called Deep Q-Network (DQN) algorithm. The idea of deep Q-learning is to approximate the Q-values using DNN since the basic Q-learning cannot tackle the problems with high-dimensional state-action space and continuous data sets.

In the DQN updating function (3.21), the Q value function,

$$Q\left(s,a\right)=\mathbb{E}\left[U_i(t)|(s,a)\right] \tag{3.20}$$

is replaced by a parameterised value function [128] $Q\left(s,a;\theta\right)$, where θ is the parameters that define the Q-values, $\max_a Q\left(s',a';\theta\right)$ is the estimate of optimal future value of the next state.

$$Q^{new}(s,a;\theta)\leftarrow Q(s,a;\theta)+\alpha\left(R(s,a)+\gamma\max_{a'}Q\left(s',a';\theta\right)-Q(s,a;\theta)\right). \tag{3.21}$$

The parameters of the DQN θ is updated by the minimising the difference of target Q-value $R(s,a)+\gamma\max_{a'}Q\left(s',a';\theta\right)$ and Q-value $Q\left(s',a';\theta\right)$ using gradient descent. However, the target Q-value is actually estimated by the current parameters θ. During each iteration, the updated Q-values will move closer to the target Q-values, while the target Q-values will be moving in the same direction. As a result, the optimisation can easily chasing the bad target Q-values, making the training of DQN very unstable.

To overcome the above issue, one method is to obtain the target Q-values from a totally separate network, which is the target network. The target network is a clone of the DQN. The target network's parameters are frozen with those of the original DQN. The target network's parameters are updated to copy the parameters of the original DQN after a certain amount of training timestep. By updating the target Q-network much less often than its original version, a substantial portion of the instability is reduced.

The another method to improve the stability of the training is to use an experience replay buffer. Instead of using the most recent set of transitions (s,a,r,s'), the updates of parameters are made on a randomly selected mini-batch of the transitions in an experience replay buffer, results in less variance than just updating a single tuple. This experience replay technique allows the algorithm to explore a large range of previous state-action space; otherwise, DNN tends to rewrite them with new experiences. The updates equation of parameters and details of experience replay will be shown in the energy trading algorithm section.

3.3.2 Deep Q-Network based P2P Energy Trading algorithm

The information flowchart of our proposed DQN-based method is illustrated in Figure 3.2. To implement the DQN-based P2P energy trading, the agent need to input the MG observations into DQN at the beginning of the trading block t. As stated in Section 3.2.2, the observed state before trading block t is $\left[\hat{R}_i\left(t\right),\hat{D}_i\left(t\right),S_i\left(t\right),\rho_{P2P}(t)\right]$. As the state in trading block t is not fully observable, we formulate an experience sequence $\phi(t)$ consisting of the current observations and last actual generation, demand and trading deals, which is shown as,

$$\phi_i(t)=\left(R_i(t-1),D_i(t-1),\boldsymbol{a}_i(t-1),\hat{R}_i(t),\hat{D}_i(t),S_i(t),\rho_{P2P}(t)\right). \tag{3.22}$$

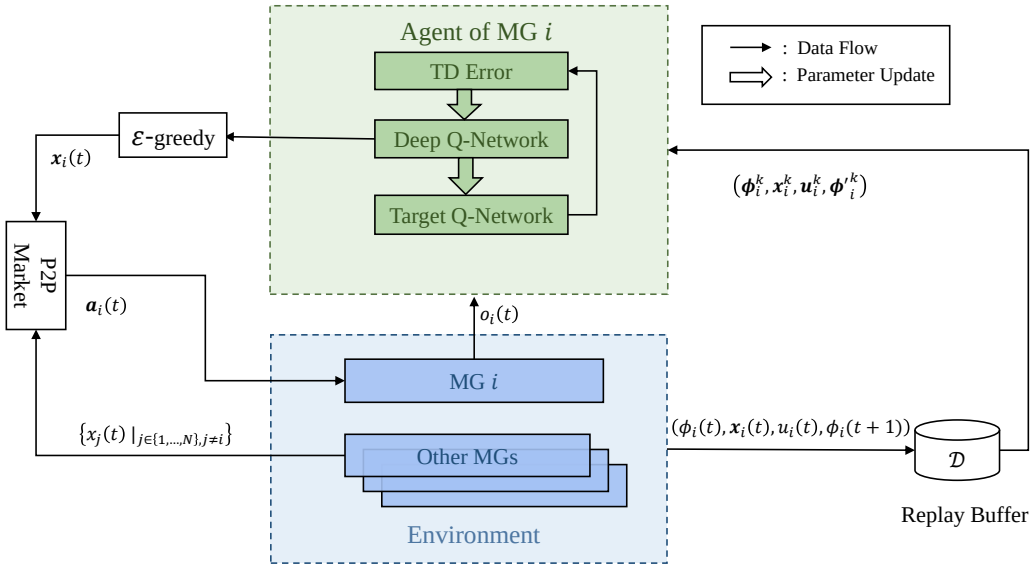


Figure 3.2: Schematic of the DQN-based P2P energy trading model for MG i.

This experience transition helps the MG understand the hidden dynamics of the system model and P2P energy trading with other MGs, which is more representative of the true system states than only the current observations.

Sine the DQN is the approximate the Q-values with the inputs (i.e., the experience sequence) and its parameters $\boldsymbol{\theta}_t$ in trading block t, the output of the DQN is $Q\left(\phi_i(t), \boldsymbol{x}_i\left(t\right); \boldsymbol{\theta}_t\right)$. Furthermore, the P2P energy trading strategy for MG i is chosen based on ε-greedy policy and the Q-values, which is shown in (3.23). With probability ε, the strategy is selected randomly, otherwise selecting the strategy that maximises the Q-value.

$$\Pr\left(\boldsymbol{x}_i(t)\right) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|X|}, & \text{if } \boldsymbol{x}_i(t) = \arg\max_{x\in X} Q\left(\phi_i(t), \boldsymbol{x}_i(t); \boldsymbol{\theta}_t\right), \\ \frac{\varepsilon}{|X|}, & \text{otherwise,} \end{cases} \tag{3.23}$$

where $|X|$ denotes the number of actions the MG can make in terms of P2P energy trading strategy. This ε-greedy policy ensures the MG to explore the choices of trading strategy rather exploit the current optimal strategy when in the training process.

After selecting the P2P energy trading strategy, MG i will receive the reward in trading block t as evaluated in (3.9) and observe the new observation $\boldsymbol{o}_i(t)$. Then, the preprocessed new experience sequence $\phi_i(t+1)$ along with the past experience sequence, actions and reward form the transition $\left(\phi_i(t), \boldsymbol{x}_i\left(t\right), u_i\left(t\right), \phi_i(t+1)\right)$ which is stored in the experience replay buffer $\mathcal{D}$. The next step is to sample a random mini-batch of k transitions $\left(\boldsymbol{\phi}_i^k, \boldsymbol{x}_i^k, \boldsymbol{u}_i^k, \boldsymbol{\phi'}_i^k\right)$ from $\mathcal{D}$. Finally, the parameters $\boldsymbol{\theta}_t$ are updated by minimising the loss function shown in (3.24) using stochastic gradient descent. Note that, the parameters θ_t^- are the parameters of the target Q-network and are only update every C iterations to reduce the risk of divergence. The training

Figure 3.3: Flowchart of implementing our proposed DQN-based P2P energy trading method for MG i.

process of DQN-based P2P energy trading for MGs is shown in Algorithm 1. The flowchart of implementing our proposed DQN-based P2P energy trading method for MG i is shown in Figure 3.3.

$$\mathcal{L}\left(\theta_t\right) = \mathbb{E}_{\left(\phi_i, \boldsymbol{x}_i, u_i, \phi_i'\right) \sim U(\mathbb{D})}\left[\left(u_i + \gamma \max_{\boldsymbol{x}_{i'}} Q\left(\phi_i', \boldsymbol{x}_i'; \boldsymbol{\theta}_t^-\right) - Q\left(\phi_i, \boldsymbol{x}_i; \boldsymbol{\theta}_t\right)\right)^2\right]. \tag{3.24}$$

3.4 Numerical Simulation

3.4.1 Case Study Setup

In this chapter, the deep Q-learning algorithm for P2P energy trading was simulated by using real-world data from Pecan Street Inc. [129], which consists of 1-year electricity generation

Algorithm 1: Deep Q-Learning for P2P Energy Trading

Initialize $\gamma,\ \theta_1$ and replay memory $\mathcal{D}$ to capacity N_{max}
for $t \in \mathbf{T}$ **do**
  Forecast $\hat{R}_i(t), \hat{D}_i(t)$ and observe $S_i(t)$
  Form experience sequence $\phi_i(t)$
  Input $\phi_i(t)$ with θ_t and get $Q(\phi_i(t), \boldsymbol{x}_i(t); \theta_t)$
  Choose trading strategy $\boldsymbol{x}_i(t)$ using ε-greedy via (3.23)
  for $j \in \mathbb{N}$ **do**
    Receive the intended energy $x_{ji}(t)$ from MG j
  end
  Calculate $a_{ij}(t)$ via (3.5) and Check constrain $p_{ij}(t)$
  Observe actual generation $R_i(t)$ and demand $D_i(t)$
  Calculate constrain $p_{ES,i}(t)$ via (3.17)
  if $p_{ES,i}(t)$ *not in constraint (3.15)* **then**
    $p_{ES,i}(t) = p_{ES,i}^{limit}(t)$
  end
  Calculate $S_i(t+1)$ via (3.18)
  if $S_i(t+1)$ *not in constraint (3.16)* **then**
    $S_i(t+1) = S_i^{limit}(t)$
  end
  Calculate Penalty using (3.10), (3.11)
  Observe the electricity price $\rho_{grid}, \rho_{p2p}, \rho_{retail}$
  Calculate utility $u_i(t)$ via (3.9)
  Store transition $(\phi_i(t), \boldsymbol{x}_i(t), u_i(t), \phi_i(t+1))$ in $\mathcal{D}$
  Sample a mini-batch of transitions $\left(\boldsymbol{\phi}_i^k, \boldsymbol{x}_i^k, \boldsymbol{u}_i^k, \boldsymbol{\phi'}_i^k\right)$ from $\mathcal{D}$
  Calculate loss function $\mathcal{L}(\theta_t)$ via (3.24)
  Update DQN parameters $\boldsymbol{\theta}_t$ by gradient descent
  Update target Q-network parameters $\boldsymbol{\theta}_t^-$ to $\boldsymbol{\theta}_t$ every C iterations
end

and demand data at 1-hour resolution from 100 households located in Mueller, Austin, Texas. The 100 households were divided into three groups as three MGs, the PV generation of the households was aggregated properly to work as a sufficient renewable generator for local MG. We choose MG1 as our agent using the proposed DQN-based method, which consists of 30 households. The typical daily PV generation and local demand for MG 1 in four seasons are shown in Figure 3.4. The training data are separated into four seasons, each sub training set contains 1-month collected electricity generation and demand data with added Gaussian noise to represent the uncertainties. We use the original collected data as each sub test set to test the performance of the examined methods.

In this chapter, we will focus on finding optimal P2P energy trading for one MG, which is MG 1. Modelling multiple MGs involves multi-agent interactions which is out of the scope of this chapter and will be addressed in following chapters. The other two MGs are dummy agents in our environment and will choose their trading strategies based on the rule-based method. The rule-based P2P energy trading strategy will be explained below. Also, the P2P electricity prices followed hourly Locational marginal pricing (LMP) records from ISO New England Inc. [130]. The system parameters are given in Table 3.1. with ESS parameters given in Table 3.2.

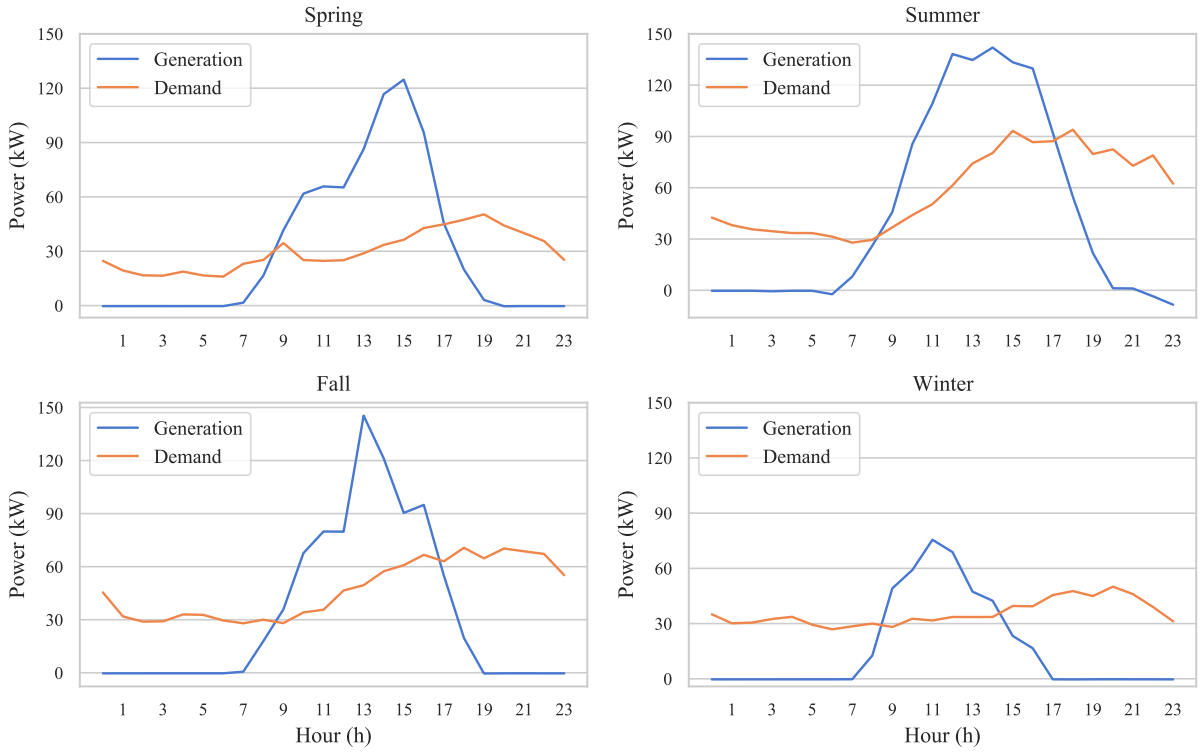


Figure 3.4: PV generation and local demand for MG 1 in four seasons.

The neural network architecture of (target) DQN has two hidden fully connected (FC) layers with Rectified Linear Unit (ReLU) as activation functions. The number of neuron units is 128 per layer. During training of the proposed method, the Adam optimiser is selected to optimise the neural networks with learning rates of 10^{-3} for the DQN. The discount factor γ, the target

Table 3.1: System Parameters

Parameters	**Values**	
Power limit (kW)	$-150 \leq p_{12} \leq 150$	$-200 \leq p_{13} \leq 200$
Penalty coefficient	$C_p = 0.3$	$C_v = 0.2$
Electricity price	$\alpha_{p2p} = 0.2$	$\rho_{retail} = 1.8\rho_{P2P}$

Table 3.2: Battery Parameters

Battery Model	**A**	**B**	**C**
Capacity	300 kWh	400 kWh	500 kWh
Rated Power	80 kW	100 kW	130 kW
Wear Cost	0.009$/kWh		
Efficiency	$\eta_{ch} = \eta_{dis} = 0.9$		

update parameter C, exploration probability ε, the size of the replay buffer, the mini-batch size m, and the full episodes are set to be 0.9, 120, 0.01, 10^5, 256, and 1000, respectively. The examined DQN-based method have been implemented using Tensorflow 1.80 in Python. The case studies have been carried out on a computer with a 4-core 3.80 GHz Intel Core i7-4790 processor, an NVIDIA GeForce GTX 1070 graphics card and 12 GB of RAM.

3.4.2 Performance Evaluation

To demonstrate the proposed of our proposed DQN-based P2P energy trading algorithm, the following trading strategies are compared:

The Rule-based Strategy

As the action space is impossible for a basic Q-learning, we design a rule-based trading strategy as a benchmark. The rule-based P2P energy trading strategy calculates the difference between the estimated energy demand and generation in the trading block t, and then sells the surplus electricity or buy the needed energy equally from/to other MGs. For example, if the MG 1 forecasts its generation 60 Kwh higher than its demand in the trading block t, the rule-based trading strategy is then selling 30 Kwh of energy to both MG 2 and MG3.

The DQN-based Strategy

The MG 1 will use our proposed DQN-based P2P energy trading strategy, and the action space of the trading quantity in each hour is from -150 kWh to 150 kWh. In order to correctly work in a DQN-based method, we discretised the action space with 30 Kwh of energy trading quantity

between adjacent trading actions. Thus, the number of total strategies of MG 1 with the other 2 MGs is $11 \times 11 = 121$. To be noted, the distance of energy trading quantity between the actions is a trade-off between performance and computational cost.

The Q-learning Strategy

The MG 1 will use basic Q-learning P2P energy trading strategy. It will not use DNN to map the Q-value to the observations of the MG, instead, it will store the Q-values with respect to each observation-action pair. However, it is not possible to store every Q-values for each observation-action pair since the observation space is continuous in P2P energy trading. Thus, we need to discretise the observation space into a number of bins. For this simulation, the bin is 84 and the number of all possible observation-action pairs is 10164.

The Approximate Q-learning Strategy

The approximate Q-learning strategy is a modification to the basic Q-learning strategy to reduce the size of the Q-table and allow generalisation to unvisited observations of the MG. In this strategy, the observations $\hat{R}_i(t)$ and $\hat{D}_i(t)$ is replaced by $\hat{R}_i(t) - \hat{D}_i(t)$ since the difference of the generation and demand is key to decide the P2P energy trading strategies rather than their absolute value. By doing so, the dimension of the observation space reduces by one.

Results

The learning curves of P2P energy trading problem for MG 1 in a typical winter day using each strategy are illustrated in Figure 3.5. The figure shows that the proposed DQN-based strategy can increase the utility of the MG and outperform other strategies. The DQN-based P2P energy trading strategy achieves average hourly utility of 4.019 dollars for the MG 1. In relative terms, The DQN-based P2P energy trading strategy increase the utility of the MG 122.9%, 125.8% and 116.2% compared to that of rule-based, Q-learning and approximate Q-learning strategies. The Q-learning and approximate Q-learning did not perform very well and achieve average hourly utility roughly equal to the rule based strategy. This is because the discretisation cannot fully represent the dynamics of the observation space and the algorithms perform poorly in unvisited observation-action pairs. The proposed DQN-based strategy, however, using DNN to learn a good representation of the mapping of Q-value to observation-action pairs thanks to the strong approximation ability of the neural network. The figure also demonstrates that the DQN-based P2P energy trading algorithm converges fast around 800 episodes while Q-learning and approximate Q-learning were very unstable during the training process.

The computational performance of the compared RL strategies is illustrated in Table 3.3 in terms of training and execution time. The average CPU time per episode is the highest in DQN-based strategy since a lot of the time spent on the training of the neural network which involves

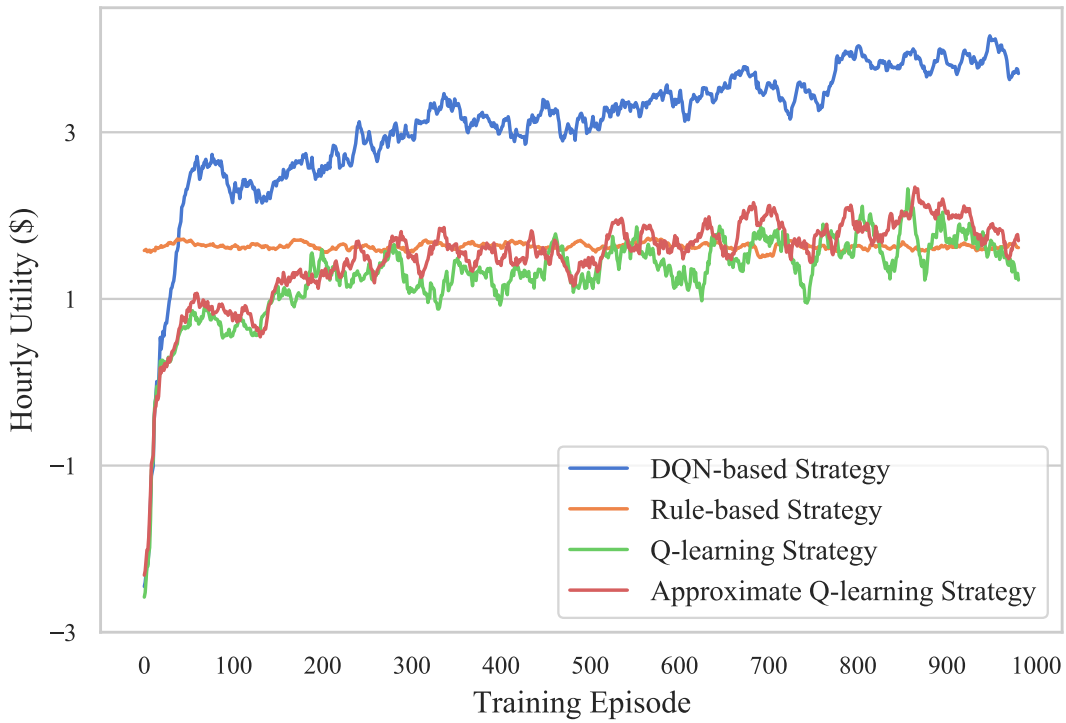


Figure 3.5: Learning curves of P2P energy trading problem for MG 1 under examined strategies.

the forward inference and backpropagation of a large number of parameters. The approximate Q-learning achieve fastest training and execution time since it used manual feature engineering to reduce the dimension of the observation space. Therefore, the number of all possible observation-action pairs is smaller than in the Q-learning strategy. For execution, the CPU time of Q-learning and approximate Q-learning strategies are smaller than the DQN-based strategy since the methods only involves directly retrieve the Q-values in the Q-tables. However, the execution time of the DQN-based strategy is still very small and in the order of milliseconds.

Table 3.3: Computational Performance of The RL Strategies

Method	**Q-learning**	**Approx. Q**	**DQN-based**
CPU time per episode (s)	1.76	1.33	3.89
Total CPU time (h)	0.48	0.37	1.08
CPU time at execution (ms)	1.77	1.42	2.83

3.4.3 Impact of Uncertainties in Renewable Generation and Demand

In this section, we investigate the impact of uncertainties in renewable generation and demand. In our P2P energy trading problem, the MG need to use the renewable generation and energy demand estimation of the next trading block to decide the trading decisions. However, the

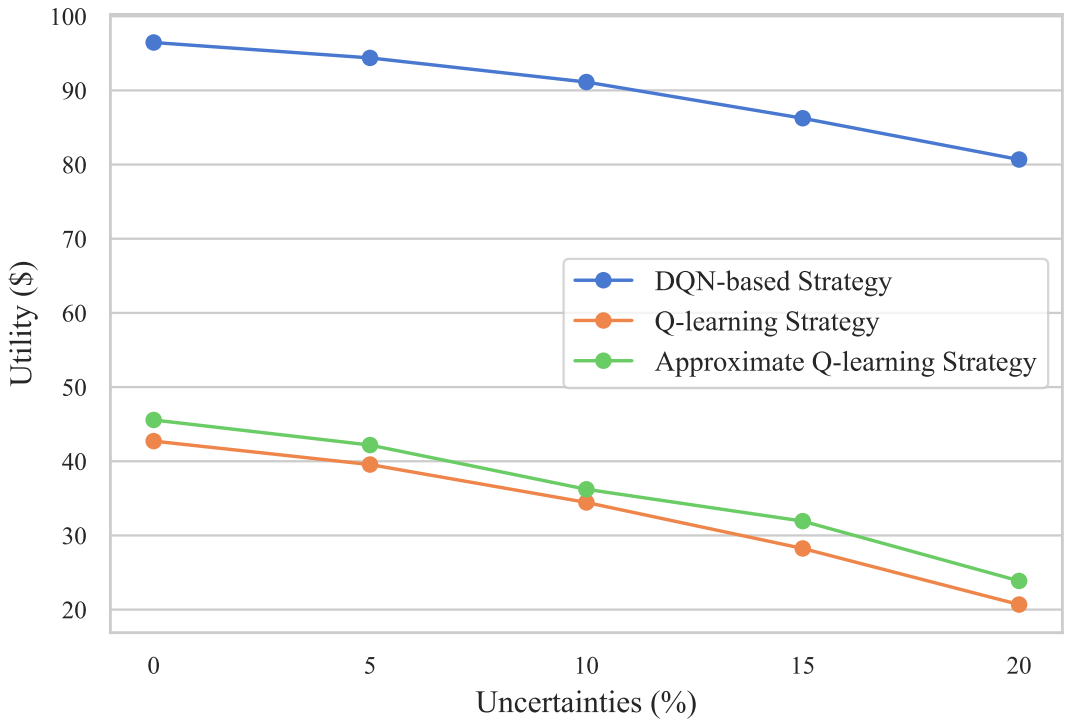


Figure 3.6: Average daily utilities for MG 1 under different uncertainty levels.

prediction can vary from the true states. In this simulation, we add a error term ϵ to represent the uncertainty of the renewable generation and demand, such that,

$$\hat{R}_i(t) = R_i(t) \cdot (1 \pm \epsilon), \tag{3.25}$$

$$\hat{D}_i(t) = D_i(t) \cdot (1 \pm \epsilon). \tag{3.26}$$

The uncertainty is considered in five levels from 0% to 20%, which will be tested by the examined RL strategies. As shown in Figure 3.6, when the uncertainty level increases from 0 to 20 %, the average daily utilities drops from 96.45 dollars to 80.68 dollars for DQN-based strategy, 42.71 dollars to 20.54 dollars and 45.57 dollars to 23.86 dollars for Q-learning and approximate Q-learning strategies, respectively. It can be showed that our proposed DQN-based are more robust in terms of relative utility decrease compare to the other Q-learning strategies. It is because DNN in our proposed method are more flexible in deciding suitable energy trading actions with unseen observations. The experience sequence $\phi(t)$ proposed in 3.22 also help the algorithm to take previous prediction errors into considerations to make robust P2P energy trading decisions. Furthermore, Figure 3.7 shows that our proposed DQN-based strategy copes the uncertainty by largely increasing the trading amount of energy with the power plant to mitigate the possibilities of imbalance penalties.

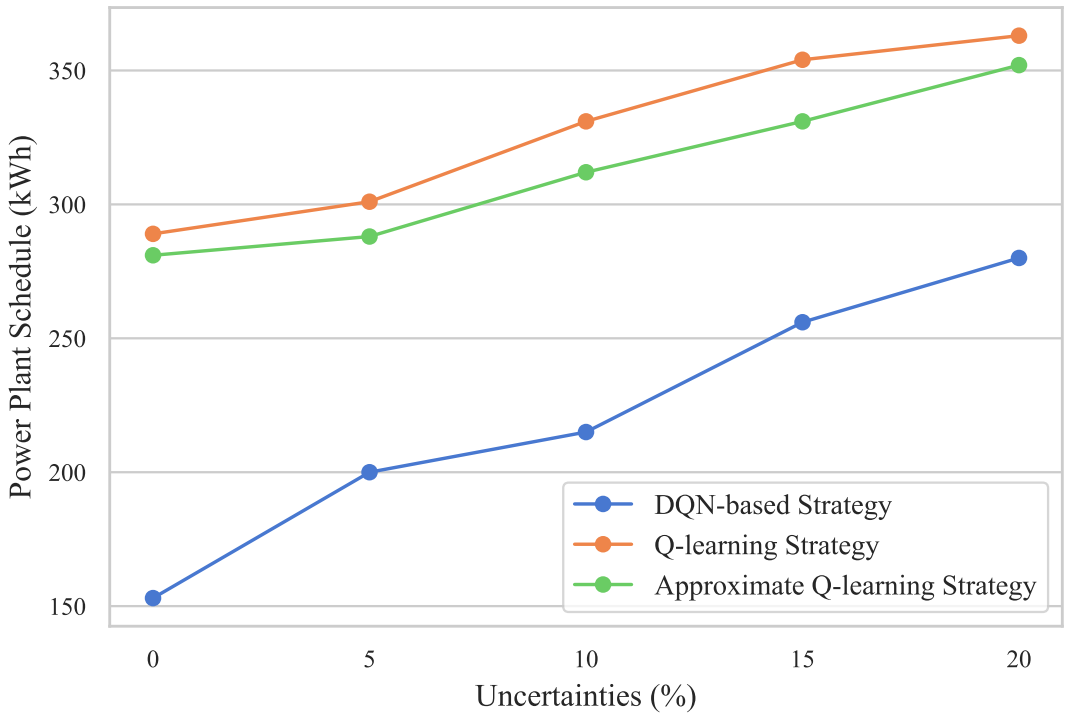


Figure 3.7: Average daily power plant schedules for MG 1 under different uncertainty levels.

3.4.4 Impact of P2P Price Ratio

P2P energy price is a big factor to the P2P energy trading problem. The difference between grid selling price and grid buying price is a critical motivation of using P2P energy trading. The P2P price ratio represents how significant the difference is. As shown in Figure 3.8, the average daily utilities increase when the P2P price ratio increases for all RL strategies. For the proposed DQN-based strategy, the average daily utilities increases from 93.14 dollars to 110.37 dollars; for Q-learning and approximate Q-learning strategies, the average daily utilities increase from 40.36 dollars to 50.68 dollars and 43.61 dollars to 52.16 dollars, respectively. The reason is that trading in P2P energy market is more profitable than trading with the power plant when the P2P price ratio is high. It can also be proved in Figure 3.9 where the power plant schedules are reduced by the MG as more energy are traded with other MGs. It also can be observed that the power plant schedule can not be further reduced by increasing P2P energy ratio. It is due to the conflict interests of energy trading for the participated MGs in the P2P energy market.

3.4.5 Impact of Virtual Penalty

The virtual penalty is a parameter used to train the P2P energy trading strategy, which makes the algorithm believe achieve some goal is beneficial even though it migh not be economically optimal. For example, in this simulation, we want to reduce energy dependency on power plant. Therefore, the virtual penalty can be set as (3.11), which is proportional to the amount of the

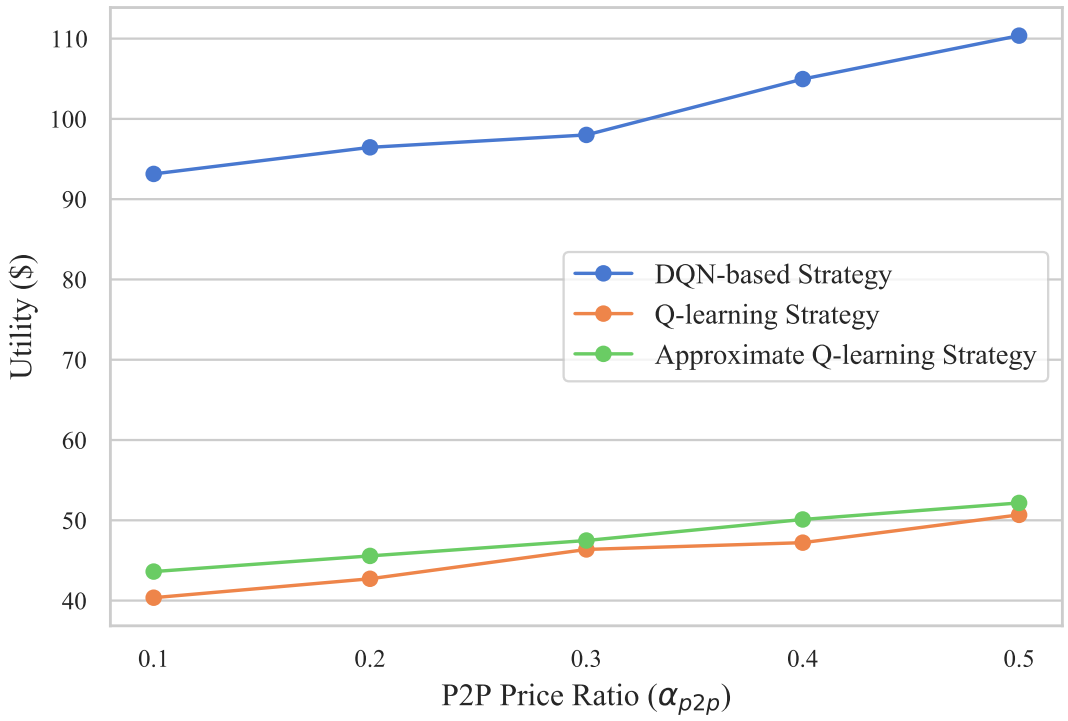


Figure 3.8: Average daily utilities for MG 1 under different P2P price ratios.

trading energy with the power plant. As shown in Figure 3.11, the average power plant schedules decreases when increasing the virtual penalty coefficient. For the proposed DQN-based strategy, the average power plant schedules decreases from 153 kWh to 50 kWh. For Q-learning and approximate Q-learning strategies, the average power plant schedules decrease dramatically from 289 kWh to 101 kWh and 281 kWh to 98 kWh, respectively. The reason is that there are more room for P2P energy trading for the sub-optimal Q-learning and approximate Q-learning strategies.

The virtual penalty pushes the MG to choose P2P energy trading more often than trading with the power plant, however, increasing the risk of causing energy imbalance. As shown in Figure 3.10, when the virtual penalty coefficient increases from 0 to 4, the average daily utilities drops from 96.45 dollars to 70.36 dollars for DQN-based strategy, 42.71 dollars to 2.39 dollars and 45.57 dollars to 10.97 dollars for Q-learning and approximate Q-learning strategies, respectively. It can be observed that our proposed DQN-based strategy can mitigate the risk of energy imbalance while other Q-learning strategies fail to meet the energy balance constraint when the coefficient is very large. The results also show that choosing suitable virtual penalty coefficient such as $C_v = 2$ can reduce the power plant schedule substantially while minimising the utility losses for our proposed DQN-based strategy.

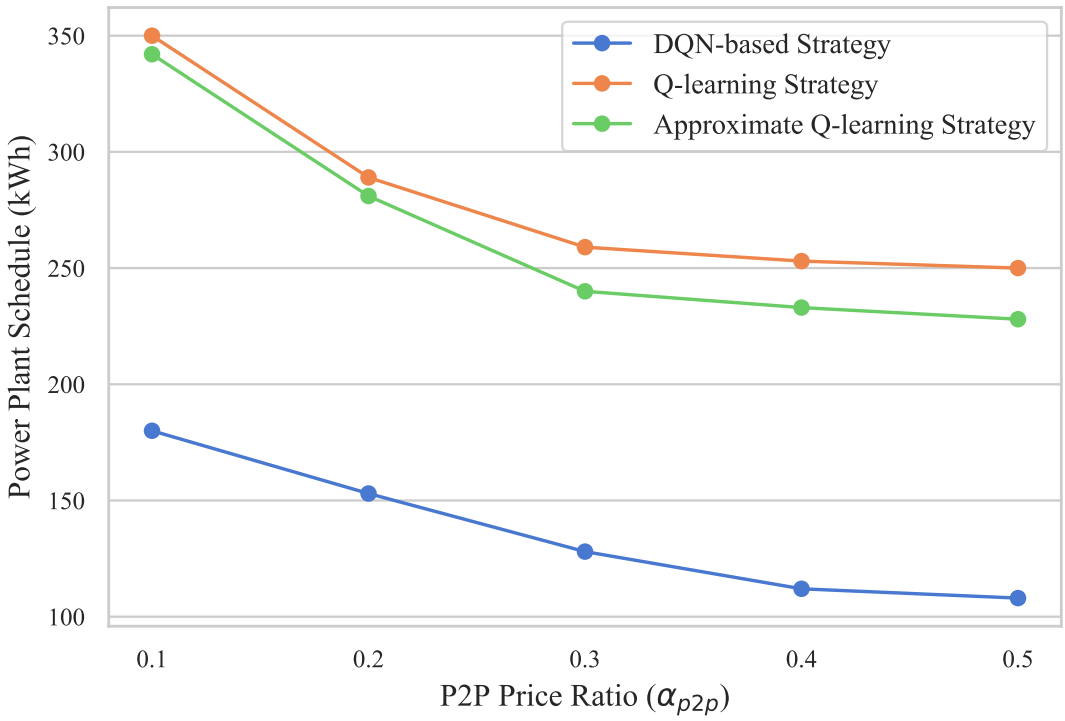


Figure 3.9: Average daily power plant schedules for MG 1 under different P2P price ratios.

3.4.6 Effect of Battery Size and Seasonal Changes on Utility

ESS is essential in P2P energy trading. Therefore the impact of different battery sizes (shown in Table 3.2.) during the four seasons is studied and shown in Figure 3.12. With no battery, the daily utility of MG 1 is always lower than having a battery. The flexibility of the battery provides the MG with more room to choose an optimal P2P energy trading strategy considering P2P energy price and other MG strategies. However, the result shows that larger battery size is not always better. The larger-size battery may result in a massive amount of charge and discharge and resource waste as we consider charge and discharge rate and battery wear cost. In this system model of MG 1, having a battery Model B is most beneficial to the MG. The figure also shows a interesting fact when comparing the utilities of MG 1 using the proposed DQN-based method in different seasons. The utilities of MG 1 are about the same in spring and winter; while in summer and autumn, the utilities drop. This is because MGs have higher energy demand in summer and fall. Therefore, there are less P2P energy trading deals been made as MGs are busy meeting their own demand during these seasons. In addition, we found adding the virtual penalty C_{vir} can reduce the power plant schedule by 82% although it is not economically beneficial.

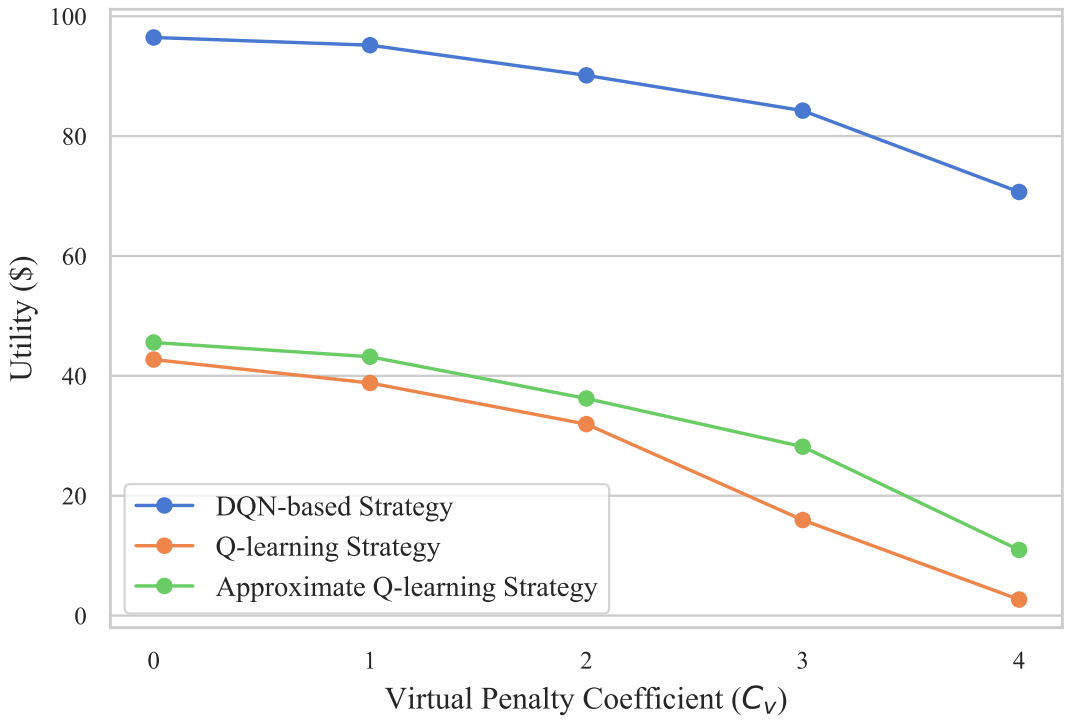


Figure 3.10: Average daily utilities for MG 1 under different virtual penalty coefficients.

3.5 Conclusion

In this chapter, we proposed a P2P energy trading model for MGs using MDP and DRL. With several essential physical constraints, the model can be better adapted for real situations. The simulation was performed using 1-year real generation and demand data, showing that the proposed DQN-based P2P energy trading strategy can choose better P2P energy trading strategies than other Q-learning and rule-based strategies. Furthermore, the impact of generation and demand uncertainties, P2P price ratio and virtual penalty on the average daily utility and power plant schedule were analysed and discussed. The proposed DQN-based P2P energy trading strategy showed great potential to be a robust and generalised method in the real-world situations. We also tested the effect of battery size on daily utility and analysed the scalability of the proposed method across four seasons. Furthermore, this model can help MGs to choose the most suitable battery setting and achieve their own social goals.

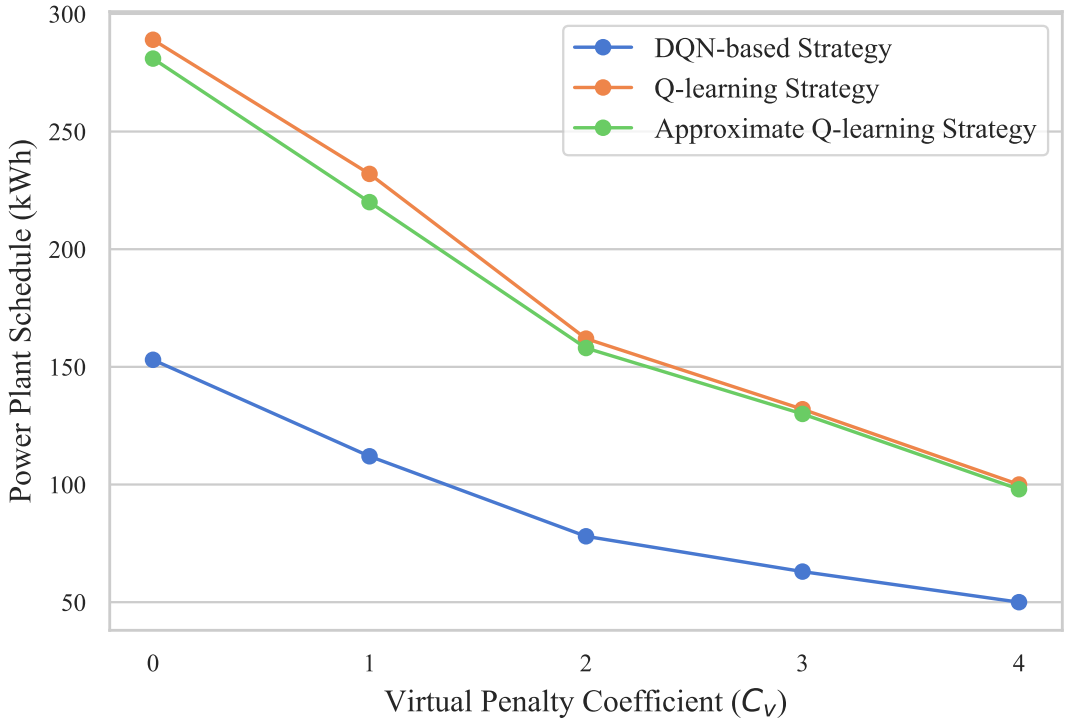


Figure 3.11: Average daily power plant schedules for MG 1 under different virtual penalty coefficients.

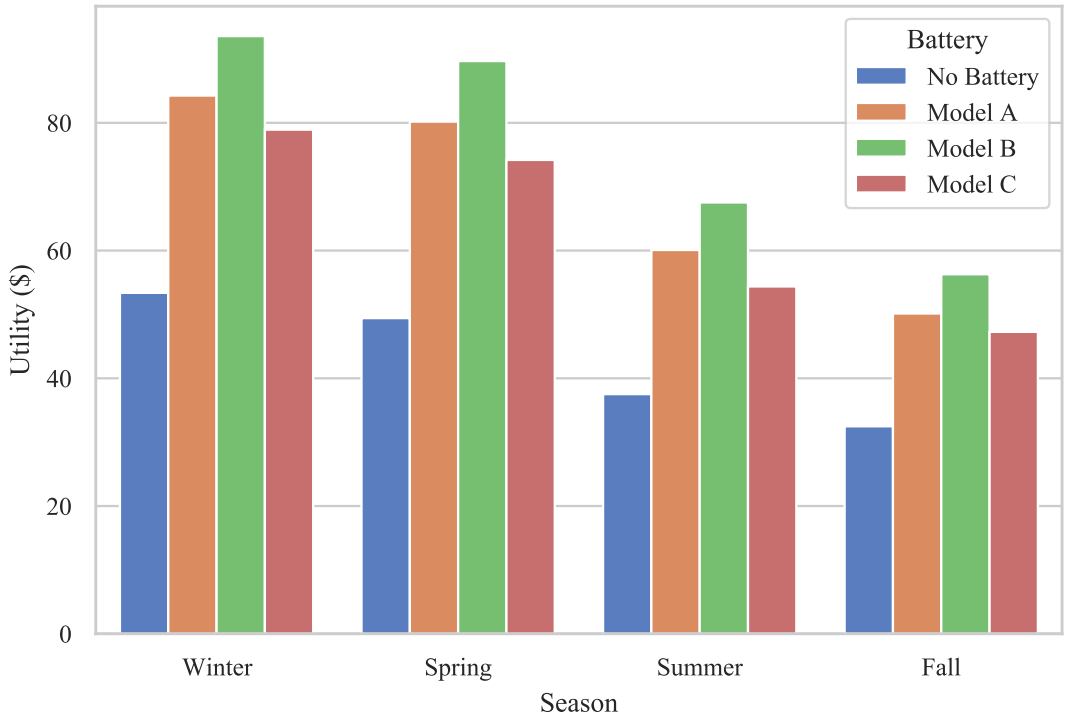


Figure 3.12: Average daily utilities with different battery models.

Chapter 4

Peer-to-Peer Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning

4.1 Introduction

Microgrids (MG) are used to address the challenges arising from having a high share of distributed energy resources (DERs) within a local region in modern energy systems. At the distribution network level, a multi-energy microgrid (MEMG) consists of DERs, energy coupling technologies, local active loads and energy storage systems (ESSs). The recent energy coupling technologies, such as hydrogen fuel cells (FCs), water electrolysers (WEs) and electric heat pumps (HPs), can be integrated by multiple energy carries together to benefit the energy systems economically and environmentally [131]. Multiple MEMGs can be networked further to improve the efficiency and reliability of the distribution network. However, besides the primary challenges posed by the intermittent nature of DERs, there is an additional difficulty in the stability and operational safety for the network of multiple MEMGs because the deployment including the size and type of DERs varies by location. It is also not realistic to directly control or operate those DERs by a central authority since they may belong to different owners. Peer-to-peer (P2P) energy trading has emerged as a novel paradigm for decentralised energy market designs. P2P energy trading allows the end-users or MGs to join the trading without a central authority unit [6] and offers an opportunity to produce and sell energy at the edge of the network. Correctly modelling and quantifying the P2P energy trading as well as understanding the flexibility of MEMGs are complicated tasks. It involves not only temporal, multi-vector interactions on different networks (e.g., electricity, heat and gas) in response to uncertain energy generation and demand, but also includes potential conflicting trading and operating policies of MGs.

This chapter aims to determine the optimal external energy trading and internal energy conversion policies for multiple MEMGs (MEMGs) using multi-agent deep reinforcement learning (MADRL). Specifically, the following questions need to be answered: 1) Can we integrate MES with P2P energy trading and how to model this problem? The answer is that we build a P2P energy trading model consisting of three types of MEMGs: residential MEMGs, commercial MEMGs and industrial MEMGs within a local community. Furthermore, we model this problem using a mathematical framework, POMDP.

2) Could deep Q-learning algorithm handle this more complicated P2P energy trading and energy conversion problem involving a continuous space of actions? If not, what is the most appropriate algorithm for this type of problem? The answer is no, but we introduce a state-of-the-art DRL algorithm TD3 to solve the P2P energy trading and energy conversion problem with highly uncertain data and continuous actions.

3) Is it possible to adjust the algorithm to handle the non-stationary environment, in which several MEMGs are exchanging energy with one another causing changing policies? The answer is to introduce a multi-agent framework that utilises the centralised training and decentralised execution to solve the non-stationary environment and maintain the independent objective of each MEMG.

4) What modifications can we make to the chosen DRL model to ensure that when the DRL agents choose their actions, the physical constraints of the P2P energy trading and energy conversion model are satisfied? The answer is that we make modifications to the original MADRL framework for our P2P energy trading and energy conversion problem and also for stabilising the learning process.

The remainder of the chapter is organised as follows. Section 4.2 formulates the problem of P2P energy trading and energy conversion for interconnected residential, commercial and industrial MEMGs. Section 4.3 proposes the MATD3 method. Section 4.4 presents a case study to evaluate the effectiveness of the proposed model. Section 4.5 draws the conclusion.

4.2 System Model

4.2.1 System Overview

Figure 4.1 shows the P2P energy trading paradigm among N MEMGs located in residential, commercial and industrial areas within a local community. These MEMGs can not only trade electricity with the main power grid and buy natural gas as fuel from the external networks but also trade electricity and heat among themselves. We assume there are heat networks built within local areas so that the MGs are more willing to trade heat instead of gas sources. Each MG includes renewable generators, a multi-energy system (MES) and electricity and heat load. The energy flow of the residential, commercial and Industrial MEMGs is illustrated in Figure 4.2,

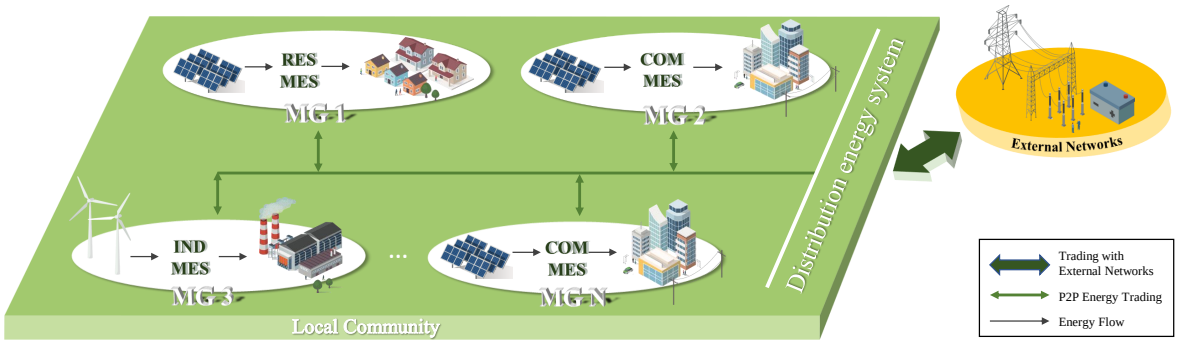


Figure 4.1: The framework of P2P energy trading among multiple Multi-energy MGs. RES, COM and IND stand for residential, commercial and industrial.

4.3 and 4.4.

Residential MEMG

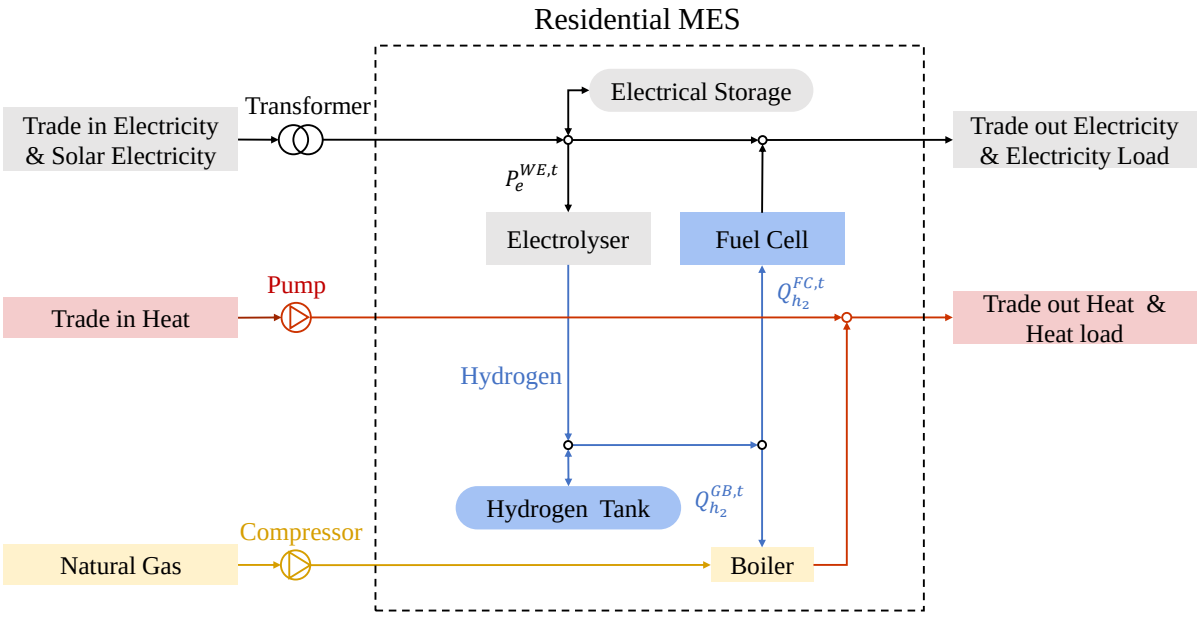


Figure 4.2: The energy-flow-diagram of the residential MES.

Solar panels are installed at the residential houses, and the electrical storage system can store any excess electricity. Solar power can be used to produce hydrogen with the help of a water electrolyser. Hydrogen can be converted to electricity and heat using a fuel cell [132] or generate heat using a boiler. Natural gas is a standby fuel to cover the necessary heat demand.

Hydrogen fuel has great potential to be used in the residential sector, because of the versatility to generate electricity, heat and serve as energy source for vehicles with low carbon emissions in the future, even though the residential usage of hydrogen is still on the initial stage. Corporate investment [133] and hydrogen infrastructure development [134] are likely to reduce

hydrogen cost. Governments have invested in clean and safe hydrogen heating systems in residential building [135]. E.g., A project has been approved in Scotland to heat homes with 100 per cent green hydrogen [136]. These investments can further accelerate residential usage of hydrogen.

The water electrolyzers are used as a backup system to provide extra electricity and heat when necessary, because hydrogen generated by a water electrolyzer can be stored in a hydrogen tank indefinitely until needed. Other energy storage systems, such as batteries, lose energy over time, and have to be recharged periodically [137]. The cost of electrolyzers is continually declining [138], and the efficiency of electrolyzers is improving [139].

Commercial MEMG

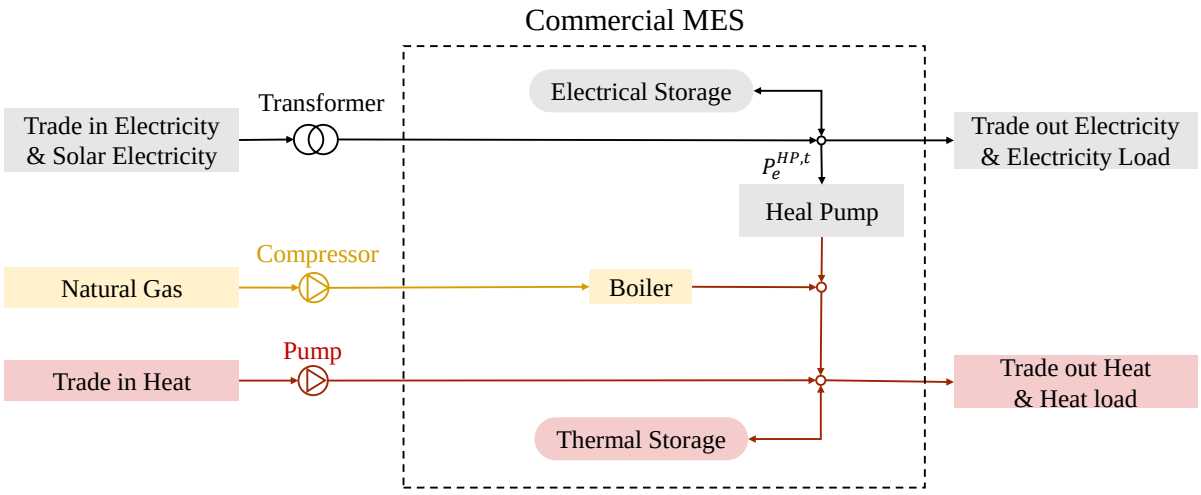


Figure 4.3: The energy-flow-diagram of the commercial MES.

For commercial MG, the primary heat load is for space heating. Therefore, the heat pump, converting electricity into heat, is a better choice for space heating than the gas boiler, which is only a standby resource. The MG has solar panels installed on the their buildings and a natural gas supply as well. In addition, the MG has a thermal storage system to store excess thermal energy for later use.

Industrial MEMG

A combined heat and power (CHP) generator provides electrical and thermal energy simultaneously to meet electricity and heat demand, which is used to improve energy efficiency. There are also wind turbines on-site to provide additional electricity generation. Moreover, this MG is equipped with electrical and thermal storage systems.

Within a local community, residential MEMGs, industrial MEMGs and commercial MEMGs may have different generation and demand patterns and consist of different energy converters.

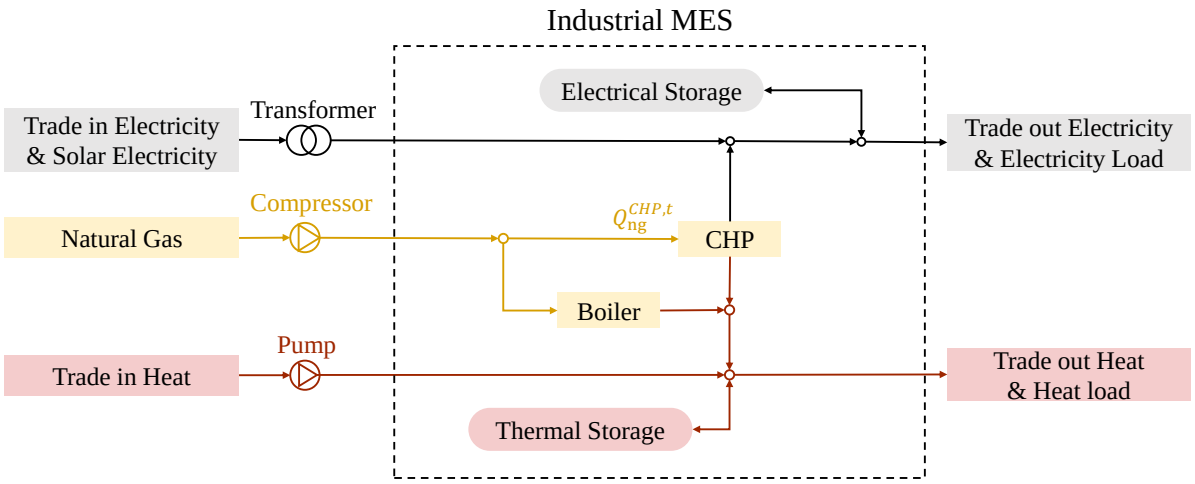


Figure 4.4: The energy-flow-diagram of the industrial MES.

We deliberately exclude hydrogen system in commercial and industrial MEMGs is to demonstrate the generality and how these MEMGs can work together through P2P energy trading and energy conversion to meet their demand with lowest energy cost and carbon emissions.

4.2.2 Markov Decision Process & System Objective

The P2P energy trading and energy conversion problem is formulated as a POMDP to minimise the operation cost of each MEMG. A POMDP consists of a set of states, a set of observations, a set of actions, a set of reward functions and a set of state transition functions.

System States and MG Observations

The system states,

$$\boldsymbol{s}^t = \left\{ s_1^t, \ldots, s_N^t \right\}, \tag{4.1}$$

describe the configurations of all MGs at time t. The Microgrids will choose energy trading and energy conversion actions based on the P2P electricity price and their hourly energy profile including renewable generation, energy demand and energy level of the energy storage system. Therefore, the system state of MG i at time t is defined as

$$s_i^t = \left[G_i^t, D_{e,i}^t, D_{h,i}^t, E_i^t, \rho_{P2P,e}^t \right], \tag{4.2}$$

where G_i^t is the renewable generation of MG i between time t and time $t+1$, $D_{e,i}^t$ and $D_{h,i}^t$ represents the electricity and heat demand of MG i between time t and time $t+1$, E_i^t includes electrical storage energy level $E_{e,i}^t$, level of hydrogen stored in the tank $E_{h_2,i}^t$ and thermal storage energy level $E_{th,i}^t$ at time t, and $\rho_{P2P,e}^t$ is the P2P electricity price at time t. The natural gas price is not considered in the system state, since it is fixed within a month. Since the MGs can not ob-

tain the true generation and demand at the beginning of each time slot, they need to forecast their generation and demand. Random Gaussian noise is added into the true states to represent the estimated generation and demand (i.e., the observation values), which can effectively represent the uncertainty of the differences between actual states and estimations [140]. The observation of MG i at time t is defined as,

$$o_i^t = \left[\hat{G}_i^t, \hat{D}_{e,i}^t, \hat{D}_{h,i}^t, E_i^t, \rho_{P2P,e}^t\right], \tag{4.3}$$

where the hat symbol indicates that the variable is an estimation of the true system state.

MG Actions

The system actions,

$$\boldsymbol{a}^t = \left\{a_1^t, \ldots, a_N^t\right\}, \tag{4.4}$$

describe the actions of all MGs at time t. The actions of MG i at time t is defined as

$$a_i^t = \left[x_i^t, y_i^t\right], \tag{4.5}$$

where x_i^t are P2P energy trading actions and y_i^t are energy conversion actions. These actions will be described in detail in Subsection 4.2.3. Each MG will choose actions based on their observations.

Reward Functions

The system reward functions,

$$\boldsymbol{r}^t = \left\{r_1^t, \ldots, r_N^t\right\}, \tag{4.6}$$

describe the reward functions of all MGs at time t. The reward functions can be used to calculate the MGs' revenue (cost is treated as negative revenue) after taking actions $\boldsymbol{a}^t$ and then evaluate the MGs to choose better policies. The reward function of MG i at time t is formulated as

$$r_i^t = r_{P2P,i}^t - C_{eco,i}^t - C_{pen,i}^t - C_{th,i}^t - C_{carbon,i}^t, \tag{4.7}$$

which includes P2P energy trading profit $r_{P2P,i}^t$, economic cost $C_{eco,i}^t$, electricity penalty $C_{pen,i}^t$, discomfort cost $C_{th,i}^t$ and environmental cost $C_{carbon,i}^t$ at time t.

The P2P energy trading profit is described as

$$r_{P2P,i}^t = \sum_{j=1, j\neq i}^{N} \sum_{u\in U} z_{ij,u}^t \cdot \left(I_{(z_{ij,u}\leq 0)} \cdot \rho_{P2P,u}^{-,t} - I_{(z_{ij,u}>0)} \cdot \rho_{P2P,u}^{+,t}\right), \tag{4.8}$$

where $U = \{e, h\}$ includes electricity (denoted e) and heat (denoted h). For each u, the amount

of P2P energy traded is $z^t_{ij,u}$, and $\rho^{-,t}_{P2P,u}$ and $\rho^{+,t}_{P2P,u}$ represent the selling and buying price at time t, respectively.

The economic cost consists of wholesale electricity cost and natural gas cost, which is expressed as

$$C^t_{eco,i} = z^t_{ii,e} \cdot \left(I_{(z_{ii,e}>0)} \cdot \rho^{+,t}_{grid,e} - I_{(z_{ii,e}\leq 0)} \cdot \rho^{-,t}_{grid,e}\right) + z^t_{ii,h} \cdot \rho_{gas}, \tag{4.9}$$

where $z^t_{ii,e} > 0$ represents buying electricity from the external network (i.e., the main grid) in the wholesale market by MG i at time t, $z^t_{ii,e} \leq 0$ means MG i sells electricity in the wholesale market at time t and $z^t_{ii,h}$ denotes the amount of natural gas bought by MG i from the external network at time t; $\rho^{+,t}_{grid,e}$, $\rho^{-,t}_{grid,e}$ and ρ_{gas} refer to as the buying and selling price offered by the main grid and natural gas price at time t, respectively. In this work, the relationship between P2P electricity prices and electricity prices of the main grid is limited as

$$\rho^{-,t}_{grid,e} \ll \rho^{-,t}_{P2P,e} \approx \rho^{+,t}_{P2P,e} \ll \rho^{+,t}_{grid,e}. \tag{4.10}$$

In the electricity wholesale market, the price that the MGs buy from the main grid is usually higher than the price that MGs sell to the main grid, since there are transaction costs due to transmission loss [141]. The higher the transaction costs, the larger the difference between the buying price and the selling price [141]. For the P2P trading market, the selling price is set to be the same as the buying price, since the transaction costs are negligible within a local distribution network [19]. The P2P electricity price is set between the main grid buying and selling prices to encourage P2P energy trading. We assume that the MGs decide an agreed P2P price for all of the MGs, and then negotiate the amount of electricity traded among themselves. This method has been used in [20, 25, 112, 142]. This assumption is made because the combined dynamics of energy trading, energy conversion and multi-vector energies are considered as a whole. The P2P electricity price is set as

$$\rho^t_{P2P,e} = \alpha_{P2P} \cdot (\rho^{+,t}_{grid,e} - \rho^{-,t}_{grid,e}) + \rho^{-,t}_{grid,e}, \tag{4.11}$$

where $\alpha_{P2P} \in (0, 1)$ is the price coefficient.

The electricity penalty happens when the electricity load supply is short between time t and time $t+1$ [143], which is shown in (4.12). The discomfort cost occurs when the thermal demand of local consumers are not met [100], which is shown in (4.13).

$$C^t_{pen,i} = \alpha_e \cdot \left(D^t_{e,i} - L^t_{e,i}\right), \tag{4.12}$$

$$C^t_{th,i} = \alpha_h \cdot \left(D^t_{h,i} - L^t_{h,i}\right)^2 + \alpha_h \cdot \left(D^t_{h,i} - L^t_{h,i}\right), \tag{4.13}$$

where $L^t_{e,i}$ is real electricity load supplied by MG i between time t and time $t+1$ and α_e represents penalty coefficient; $L^t_{h,i}$ is the heat load of MG i between time t and time $t+1$ which

the consumers are actually provided with and α_h is the heat sensitivity coefficient. In this model, the renewable generation curtailment is considered since electricity consumption and generation needs to be balanced. However, extra network charges of renewable generation curtailment are not considered. Therefore the penalty only occurs when $D^t_{e,i} > L^t_{e,i}$. The penalty terms and the penalty coefficients have been designed based on [143] to obtain good performance of the proposed method [144].

The environmental cost is the economic penalty caused by the CO_2 emissions from the natural gas combustion and electricity bought from the main grid [145], which is expressed as

$$C^t_{carbon,i} = \alpha^{CO_2} \cdot \left(\beta^{gas} \cdot z^t_{ii,h} + \beta^e \cdot z^t_{ii,e}\right), \tag{4.14}$$

where β^{gas} and β^e denote carbon intensity (CI) which are the emission rate of CO_2 related to the natural gas combustion and the bought net electricity. The carbon tax price denoted α^{CO_2} converts the carbon emissions into economic penalty.

State Transition Functions

After executing the system actions $\boldsymbol{a}^t$, the system states $\boldsymbol{s}^t$ will transfer to $\boldsymbol{s}^{t+1}$ based on the state transition functions. The transition functions of the energy level of storage are shown in Subsection 4.2.4. However, the transition functions of the renewable generation and energy load are not available. We will use our proposed DRL algorithm to learn from the real-world datasets without knowing the complete state transition functions of the system.

System Problem

The system problem for MG i is to find optimal policy $\pi(x^t_i, y^t_i|\hat{G}^t_i, \hat{D}^t_{e,i}, \hat{D}^t_{h,i}, E^t_i, \rho^t_{P2P,e})$ at time t to maximise its expected total rewards (same as minimising the expected total operation cost) which summarises discounted future rewards over the time horizon T, formulated as

$$\mathbf{P1}: \max_{\pi} R^t_{i_\pi} = \mathbb{E}\left[\sum_{\tau=0}^{T} \gamma^\tau \cdot r_i^{t+\tau+1}\right], \tag{4.15}$$

where γ is the discount factor.

4.2.3 Two-stage System Process

The MEMGs' operation process contains two stages: P2P energy trading and energy conversion stages. We assume the external P2P energy trading take place in an hour-ahead P2P energy market, in which each MG can buy or sell the desired energy for the next hour. After the real energy trading deals have been made, the process moves to the internal energy conversion stage.

P2P Energy Trading Stage

Before the trading begins, MG i uses its observations o_i^t and its policy to choose trading actions to seek for possible deals. The trading actions of MG i at time t is denoted as

$$x_i^t = \left[x_{ij}^t\right]_{1\leq j\neq i\leq N} = \left[x_{i1}^t, x_{i2}^t, \dots x_{iN}^t\right], \tag{4.16}$$

where $x_{ij}^t = [x_{ij,e}^t, x_{ij,h}^t]$ are the intended amounts of energy trading (including electricity and heat) between MG i and MG j at time t. If $x_{ij}^t > 0$, which means MG i wants to buy energy from MG j; if $x_{ij}^t < 0$, which means MG i wants to sell energy to MG j. MGs often have conflicting trading intentions, e.g., $x_{ij}^t \times x_{ji}^t > 0$. Therefore, trading negotiations have been made resulting in real deals of energy trading,

$$z_i^t = \left[z_{ij}^t\right]_{1\leq j\neq i\leq N} = \left[z_{i1}^t, z_{i2}^t, \dots z_{iN}^t\right], \tag{4.17}$$

where $z_{ij}^t > 0$ means MG i buys energy from MG j; $z_{ij}^t < 0$ means MG i sells energy to MG j. MGs only have a deal when one of them wants to sell energy and another wants to buy energy. Note that the actual energy trading might not be the same as the intention, and therefore, MGs need to trade energy with external networks to realise their intended trading actions. The amount of energy traded with external networks at time t are denoted as z_{ii}^t. The actual amount of P2P energy trading of MG i are shown as

$$z_{ij}^t = \begin{cases} \frac{x_{ij}^t}{|x_{ij}^t|} \cdot \min\left(\left|x_{ij}^t\right|, \left|x_{ji}^t\right|\right), & \text{if } x_{ij}^t \cdot x_{ji}^t < 0,\ \forall i \neq j. \\ 0, & \text{if } x_{ij}^t \cdot x_{ji}^t \geq 0, \forall i \neq j. \\ \sum_{j=1}^N x_{ij}^t - \sum_{j=1, j\neq i}^N z_{ij}^t, & \forall i = j. \end{cases} \tag{4.18}$$

Energy Conversion Stage

The complexity of an MES is due to the flexibility of exchanging different energy vectors, achieved by managing the energy converters such as fuel cell, heat pump and CHP. The MG needs to consider all available information, including the energy trading results. For residential MG i, the conversion actions y_i^t consist of inflow vector of the water electrolyser $P_e^{WE,t}$, inflow vector of the fuel cell $Q_{h_2}^{FC,t}$ and inflow hydrogen of the boiler $Q_{h_2}^{GB,t}$, as shown in Figure 4.2. For commercial MG i, the conversion action y_i^t is the inflow vector of the heat pump $P_e^{HP,t}$ as shown in Figure 4.3. For industrial MG i, the conversion action y_i^t is the inflow vector of the CHP $Q_{ng}^{CHP,t}$, as shown in Figure 4.4.

4.2.4 Physical Constraints

Energy Converters Constraints

Energy convert functions are used to show the energy conversion mapping from inflow energy to outflow energy through the energy converters [145]. The convert functions are defined as follows,

$$P_e^{FC,t} = \eta_e^{FC} \cdot Q_{h_2}^{FC,t}, \tag{4.19}$$

$$Q_h^{FC,t} = \eta_h^{FC} \cdot Q_{h_2}^{FC,t}, \tag{4.20}$$

$$Q_{h_2}^{WE,t} = \eta^{WE} \cdot P_e^{WE,t}, \tag{4.21}$$

$$Q_h^{GB,t} = \eta_{h_2}^{GB} \cdot Q_{h_2}^{GB,t}, \tag{4.22}$$

$$Q_h^{GB,t} = \eta_{ng}^{GB} \cdot Q_{ng}^{GB,t}, \tag{4.23}$$

$$Q_h^{HP,t} = \eta^{HP} \cdot P_e^{HP,t}, \tag{4.24}$$

$$P_e^{CHP,t} = \eta_e^{CHP} \cdot Q_{ng}^{CHP,t}, \tag{4.25}$$

$$Q_h^{CHP,t} = \eta_h^{CHP} \cdot Q_{ng}^{CHP,t}. \tag{4.26}$$

Equations (4.19)-(4.21) denote the convert functions of fuel cell and water electrolyser in the residential MGs, where $Q_{h_2}^{FC,t}$, $P_e^{FC,t}$ and $Q_h^{FC,t}$ denote hydrogen inflow, electricity outflow and heat outflow of the fuel cell at time t; η_e^{FC} and η_h^{FC} represent the electricity and heat conversion coefficient of the fuel cell; $P_e^{WE,t}$, $Q_{h_2}^{WE,t}$ and η^{WE} denote electricity inflow, hydrogen outflow and conversion coefficient of the water electrolyser at time t.

Equations (4.22)-(4.23) refer to the convert functions of the gas boiler with hydrogen or natural gas input, where $Q_{h_2}^{GB,t}$, $Q_{ng}^{GB,t}$ and $Q_h^{GB,t}$ denote hydrogen inflow, natural gas inflow and heat outflow of the gas boiler at time t; $\eta_{h_2}^{GB}$ and η_{ng}^{GB} represent the hydrogen and natural gas conversion coefficient of the gas boiler. The energy conversion process of heat pump is denoted in (4.24), where $P_e^{HP,t}$, $Q_h^{HP,t}$ and η^{HP} represent electricity inflow, heat outflow and conversion coefficient of the heat pump at time t.

The convert functions of CHP are denoted in (4.25)-(4.26), where $Q_{ng}^{CHP,t}$, $P_e^{CHP,t}$ and $Q_h^{CHP,t}$ denote natural gas inflow, electricity outflow and heat outflow of the CHP at time t; η_e^{CHP} and η_h^{CHP} represent electricity and heat conversion coefficient of the CHP.

Energy Storage Systems Constraints

The dynamic energy level of the storage systems depends on their inherent constraints, shown as follows,

$$E_e^{t+1} = \eta_e^{ES} \cdot E_e^t + P_e^{ES,t} \cdot \left(I_{(P_e>0)} \cdot \eta_{e,ch}^{ES} - \frac{I_{(P_e \leq 0)}}{\eta_{e,dis}^{ES}} \right) \cdot \Delta t, \tag{4.27}$$

$$P_e^{ES^{min}} \leq P_e^{ES,t} \leq P_e^{ES^{max}}, \tag{4.28}$$

$$0 \leq E_e^{t+1} \leq B_e, \tag{4.29}$$

$$E_{th}^{t+1} = \eta_{th}^{TS} \cdot E_{th}^t + Q_h^{TS,t} \cdot \left(I_{(Q_h>0)} \cdot \eta_{th,in}^{TS} - \frac{I_{(Q_h \leq 0)}}{\eta_{th,out}^{TS}} \right) \cdot \Delta t, \tag{4.30}$$

$$Q_h^{TS^{min}} \leq Q_h^{TS,t} \leq Q_h^{TS^{max}}, \tag{4.31}$$

$$0 \leq E_{th}^{t+1} \leq B_{th}, \tag{4.32}$$

$$E_{h_2}^{t+1} = \eta_{h_2}^{HT} \cdot E_{h_2}^t + Q_{h_2}^{HT,t} \cdot \left(I_{\left(Q_{h_2}>0\right)} \cdot \eta_{h_2,in}^{HT} - \frac{I_{\left(Q_{h_2} \leq 0\right)}}{\eta_{h_2,out}^{HT}} \right) \cdot \Delta t, \tag{4.33}$$

$$Q_{h_2}^{HT^{min}} \leq Q_{h_2}^{HT}(t) \leq Q_{h_2}^{HT^{max}}, \tag{4.34}$$

$$0 \leq E_{h_2}^{t+1} \leq B_{h_2}. \tag{4.35}$$

Equations (4.27)-(4.29) show the characteristics of the electrical storage system. Equation (4.27) explains the transition function of energy level of electrical storage, where $P_e^{ES,t}$ is the charging or discharging power of electrical storage; η_e^{ES}, $\eta_{e,ch}^{ES}$ and $\eta_{e,dis}^{ES}$ represents the self decay rate, charging coefficient and discharging coefficient of electrical storage. Equation (4.28) shows the limits the electrical power when charging or discharging the electrical storage and (4.29) is the capacity limitation, where B_e is the capacity of electrical storage.

Equations (4.30)-(4.32) indicate the limits of the thermal storage system. Equation (4.30) shows the transition function of energy level of thermal storage, where $Q_h^{TS,t}$ is the inflow or outflow heat power of thermal storage; η_{th}^{TS}, $\eta_{th,in}^{TS}$ and $\eta_{th,out}^{TS}$ represent the self decay rate, inflow coefficient and outflow coefficient of thermal storage. Equation (4.31) limits the inflow and outflow heat of the thermal storage system and the energy level of thermal storage is bounded by (4.32), where B_{th} is the capacity of thermal storage. Similarly, the transition function of energy level of hydrogen tank is formulated in (4.33), where $Q_{h_2}^{HT,t}$ is the inflow or outflow hydrogen

of hydrogen tank; $\eta_{h_2}^{HT}$, $\eta_{h_2,in}^{HT}$ and $\eta_{h_2,out}^{HT}$ represent the self decay rate, inflow coefficient and outflow coefficient of hydrogen tank.

The hydrogen gas flow limitation and hydrogen tank capacity limitation are described in (4.34)-(4.35), where B_{h_2} is the capacity of hydrogen tank. If the power of an energy storage system is greater than 0, it means charging the storage or the energy is flowing into the storage. If an energy storage system's power is less than 0, it means discharging the storage.

Energy Balance Constraints

For the energy networks of an MEMG working correctly, the MG must balance the energy generation and consumption between time t and time $t+1$. The energy balance constraints for residential, commercial and industrial MEMGs are formulated as follows,

$$z_{e,i}^t + G_i^t + P_{e,i}^{FC,t} \cdot \Delta t = P_{e,i}^{ES,t} \cdot \Delta t + P_{e,i}^{WE,t} \cdot \Delta t + D_{e,i}^t, \tag{4.36}$$

$$z_{h,i}^t + Q_{h,i}^{GB,t} \cdot \Delta t = D_{h,i}^t, \tag{4.37}$$

$$Q_{h_2,i}^{WE,t} \cdot \Delta t = Q_{h_2,i}^{HT,t} \cdot \Delta t + Q_{h_2,i}^{FC,t} \cdot \Delta t + Q_{h_2,i}^{GB,t} \cdot \Delta t, \tag{4.38}$$

$$z_{e,i}^t + G_i^t = P_{e,i}^{ES,t} \cdot \Delta t + P_{e,i}^{HP,t} \cdot \Delta t + D_{e,i}^t, \tag{4.39}$$

$$z_{h,i}^t + Q_{h,i}^{HP,t} \cdot \Delta t + Q_{h,i}^{GB,t} \cdot \Delta t = Q_{h,i}^{TS,t} \cdot \Delta t + D_{h,i}^t, \tag{4.40}$$

$$z_{e,i}^t + G_i^t + P_{e,i}^{CHP,t} \cdot \Delta t = P_{e,i}^{ES,t} \cdot \Delta t + D_{e,i}^t, \tag{4.41}$$

$$z_{h,i}^t + Q_{h,i}^{CHP,t} \cdot \Delta t + Q_{h,i}^{GB,t} \cdot \Delta t = Q_{h,i}^{TS,t} \cdot \Delta t + D_{h,i}^t. \tag{4.42}$$

Equations (4.36)-(4.38) indicate that residential MG i must balance the electricity, heat and hydrogen energy, respectively. The electricity and heat networks of commercial MG i are constrained in (4.39)-(4.40). Equations (4.41)-(4.42) describe the energy balance equations of industrial MG i with electricity and heat distribution networks.

4.3 Proposed Multi-Agent Deep Reinforcement Learning Based Approach

A MATD3 approach is proposed to solve the P2P energy trading and energy conversion problem formulated in (4.15). TD3 is a model-free, off-policy actor-critic algorithm which uses DNNs to learn policies in high-dimensional, continuous state-action spaces. The MATD3 approach adopts the form of centralised critics to ease training and decentralised actors to ensure all MEMGs are operating independently.

4.3.1 Twin Delayed Deep Deterministic Policy Gradient Algorithm

TD3 [17] was proposed to solve the overestimation and high variance problems lied in deep Q-learning [85] and deep deterministic policy gradient (DDPG) [15] algorithms. The novelty of the TD3 algorithm is that it considers the propagated function approximation error during both policy and value updates. As a corollary, TD3 mitigates the risk that the algorithm keeps track on bad Q-values. It also manages to directly address the high variance estimates when updating the policy network.

To solve the overestimation problem, TD3 adopts the idea of double Q-learning [146]. In TD3, the critic consists of two Q-networks (Q_{θ_1} and Q_{θ_2}) and their target networks ($Q_{\theta_1'}$ and $Q_{\theta_2'}$), and the actor is formed by a deterministic policy network π_ϕ and its target network $\pi_{\phi'}$. The target networks are time-delayed copies of their Q-networks, which greatly improve stability in learning [17]. To update the TD3 networks, the Q networks in critic minimise the loss via (4.43), where p_π is the state distribution, π and R are distribution of the policy and reward function, and y^t is the target value. The deterministic policy network in actor is updated using sampled policy gradient which is shown in (4.44), i.e.,

$$\mathcal{L}(\theta) = \mathbb{E}_{s^t\sim p_\pi, a^t\sim\pi, r^t\sim R}\left[\left(Q_\theta\left(s^t, a^t\right) - y^t\right)^2\right], \tag{4.43}$$

$$\nabla_\phi J \approx \mathbb{E}_{s^t\sim p_\pi}\left[\nabla_a Q_\theta(s,a)\,|_{s=s^t, a=\pi_\phi(s^t)} \nabla_\phi \pi_\phi(s)\,|_{s=s^t}\right], \tag{4.44}$$

where

$$y^t = r^t + \gamma \min_{j\in\{1,2\}} Q_{\theta_j'}\left(s^{t+1}, \tilde{a}^{t+1}\right), \tag{4.45}$$

$$\tilde{a}^{t+1} = \pi_{\phi'}(s^{t+1}) + \tilde{\epsilon},\ \tilde{\epsilon} \sim \mathrm{clip}(\mathcal{N}(0,\tilde{\sigma}), -c, c). \tag{4.46}$$

The critic will choose the minimum target value between the two target Q-networks as in (4.45), where $\tilde{a}^{t+1}$ is the clipped target action. The minimum operation results in low-variance value estimations and makes the algorithm more stable. To address the high variance problem, TD3 updates the policy networks once every several Q-value updates. By sufficiently delaying the

policy updates, TD3 allows the Q-network to produce lower Q-values, and hence less chance of a mistake being exploited. TD3 algorithm also adds target policy noise as shown in (4.46) when forming the target, where $\tilde{\epsilon}$ is the clipped Gaussian noise and c is the edge value. This target policy regularisation technique will smooth but keep close to the original target action, which helps the algorithm remain stable and converge fast in the stochastic domain.

4.3.2 Multi-Agent Twin Delayed Deep Deterministic Policy Gradient Approach

As our P2P energy trading and energy conversion model is in an MA environment (each MG is an agent), a naive approach is to directly apply TD3 algorithm to learn each agent's policy independently. However, the environment is no longer static from the view of each agent since the agents are learning their own policy independently. $P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t, \boldsymbol{a}^t, \pi_1, \ldots, \pi_N\right) \neq P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t, \boldsymbol{a}^t, \pi'_1, \ldots, \pi'_N\right)$ for any $\boldsymbol{\pi} \neq \boldsymbol{\pi}'$, which violates the Markov assumption. Therefore, this naive approach has difficulty in learning good policies. Instead, we adopted the concept of centralised training with decentralised execution in [16], where the training of critic take consideration of the actions and observations of all the agents in the environment but the actor of each agent choose actions only based on its own observations. The centralised Q-value function of MG i, $Q_{\theta_i}\left(o_1^t, \ldots, o_N^t, a_1^t, \ldots, a_N^t\right)$, takes observations and actions of all MGs as inputs instead of only its own. The critics are learned by their rewards, where the reward functions can be different from each other, allowing both competitive and collaborative multi-agent settings. The main reason for using a centralised critic is that the environment is stationary if all the actions of the agents are known, i.e.,

$$P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t, a_1^t, \ldots, a_N^t, \pi_1, \ldots, \pi_N\right) = P\left(\boldsymbol{s}^{t+1}|\boldsymbol{s}^t, a_1^t, \ldots, a_N^t, \pi'_1, \ldots, \pi'_N\right), \tag{4.47}$$

even with different sets of MGs' policies $\boldsymbol{\pi} \neq \boldsymbol{\pi}'$. The actor works in a decentralised way, to ensure that only local information is used when executing policies.

The centralised Q-value function is updated as

$$\mathcal{L}\left(\theta_i\right) = \mathbb{E}_{\boldsymbol{o}^t, \boldsymbol{a}^t, \boldsymbol{r}^t, \boldsymbol{o}^{t+1}}\left[\left(Q_{\theta_i}\left(o_1^t, \ldots, o_N^t, a_1^t, \ldots, a_N^t\right) - y_i^t\right)^2\right], \tag{4.48}$$

where

$$y_i^t = r_i^t + \gamma \min_{j \in \{1,2\}} Q_{\theta'_{ij}}\left(o_1^{t+1}, \ldots, o_N^{t+1}, \tilde{a}_1^{t+1}, \ldots, \tilde{a}_N^{t+1}\right), \tag{4.49}$$

$$\tilde{a}_i^{t+1} = \pi_{\phi'_i}(o_i^{t+1}) + \tilde{\epsilon}_i,\ \tilde{\epsilon}_i \sim \mathrm{clip}(\mathcal{N}(0, \tilde{\sigma}_i^2), -c_i, c_i). \tag{4.50}$$

The gradient of the policy network can then be written as

$$\nabla_{\phi_i} J \approx \mathbb{E}_{\boldsymbol{o}^t, \boldsymbol{a}^t}[\nabla_{a_i} Q_{\theta_i}\left(o_1^t, \ldots, o_N^t, a_1^t, \ldots, a_N^t\right)|_{o_i = o_i^t, a_i = \pi_{\phi_i}(o_i^t)} \nabla_{\phi_i} \pi_{\phi_i}\left(o_i\right)|_{o_i = o_i^t}]. \tag{4.51}$$

In the implementation, mini-batches are used to train the networks rather than a single transition of data. An experience replay buffer $\mathcal{D}$ is used to stabilise the learning. The experience replay is implemented as a fixed-size circular buffer that stores the agent's most recent transitions data. The agent can then use the uniformly sampled mini-batch for training which avoids the temporal correlations of the agent's experience. Thus, the experience replay buffer significantly improves the sample efficiency and stability of the learning algorithm. In the proposed MATD3 approach, a centralised replay buffer is used, including the transition of observations, actions, rewards, and next observations of all MGs.

4.3.3 Implementation of The Proposed Method

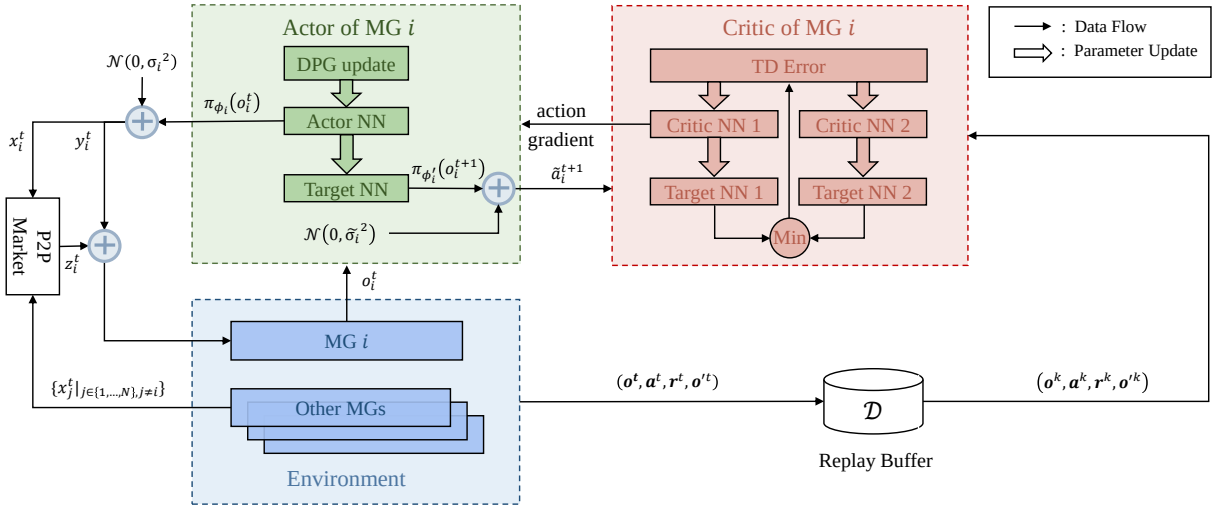


Figure 4.5: Schematic of the MADRL agent training and execution process. NN stands for neural network.

The schematic of our proposed approach is illustrated in Figure 4.5. For each MG i, it firstly receives its observations o_i^t at time t. The actor of MG i will then choose P2P energy trading actions x_i^t and energy conversion actions y_i^t based on o_i^t and its policy π_{ϕ_i}. A random noise sampled from a Gaussian distribution is added to the actor to increase exploration. During the P2P energy trading stage, MG i will negotiate with other MGs and get real energy trading deals z_i^t. After that, y_i^t and z_i^t are used to operate MG i in the energy conversion stage. MG i will then receive the reward r_i^t and observations of next states o_i^{t+1}. Finally, the transition of observations, actions, rewards and next observations of all MGs $(\boldsymbol{o}^t, \boldsymbol{a}^t, \boldsymbol{r}^t, \boldsymbol{o'}^t)$ will be stored in the replay buffer $\mathcal{D}$, where

$$\boldsymbol{o}^t = \{o_1^t, \dots, o_N^t\}, \tag{4.52}$$

$$\boldsymbol{a}^t = \{a_1^t, \dots, a_N^t\}, \tag{4.53}$$

$$\boldsymbol{r}^t = \{r_1^t, \ldots, r_N^t\}, \tag{4.54}$$

$$\boldsymbol{o'}^t = \{o_1^{t+1}, \ldots, o_N^{t+1}\}. \tag{4.55}$$

For the centralised training, each MG will sample a random mini-batch of size m $(\boldsymbol{o}^k, \boldsymbol{a}^k, \boldsymbol{r}^k, \boldsymbol{o'}^k)$ from $\mathcal{D}$. The parameters of the critic θ_i will be updated by minimising the sample loss via (4.48), and the actor will be updated using sampled policy gradient according to (4.51). The target networks of MG i will then be updated using the following equations

$$\theta'_{i1} \leftarrow \tau\theta_{i1} + (1-\tau)\theta'_{i1}, \tag{4.56}$$

$$\theta'_{i2} \leftarrow \tau\theta_{i2} + (1-\tau)\theta'_{i2}, \tag{4.57}$$

$$\phi'_i \leftarrow \tau\phi_i + (1-\tau)\phi'_i, \tag{4.58}$$

where $\tau \ll 1$ is the target update parameter. Thus, the target values change slowly which greatly improves the stability of learning. Each episode contains T time steps, and the training process repeats M times to ensure the algorithm converges. The proposed MATD3 algorithm is shown in Algorithm 2. The flowchart of implementing our proposed MATD3 solution method for each agent is shown in Figure 4.6.

4.3.4 Modifications to The Original Framework

The original MADRL framework has been modified particularly for the P2P energy trading and energy conversion problem and for stabilising the learning process. The modifications of the original MADRL framework include as follows:

TD3 Agent Customisation

In the original MADRL framework, the activation function of the output layer in the actor networks is a hyperbolic tangent or sigmoid function. In the proposed MATD3 method, for each MG, the activation function of the output layer in the actor networks is customised to provide the requisite output shape of the actor in terms of energy trading and energy conversion actions, since the range of values for the energy trading actions and energy conversion actions can be very different.

State/Observation Normalisation

For each MG, the components of the observation vector have different magnitudes. Normalising the observations can prevent bias and speed up the training process [147].

Algorithm 2: MATD3-based P2P Energy Trading and Energy Conversion in Interconnected MEMGs

Initialise $\gamma, \tau, \theta_{i1}, \theta_{i2}, \phi_i$ and replay buffer $\mathcal{D}$
for *episode = 1 to M* **do**
  Initialise random process $\mathcal{N}$ for action exploration
  for *t = 1 to T* **do**
    For each MG i, forecast $\hat{G}_i^t$, $\widehat{D}_{e,i}^t$, $\widehat{D}_{h,i}^t$, and observe E_i^t and $\rho_{P2P,e}^t$ to form o_i^t
    Choose P2P energy trading actions x_i^t and energy conversion actions y_i^t w.r.t. the current policy π_{ϕ_i}
    P2P energy trade with other MGs, and get the real energy trading deals z_i^t via (4.18)
    Convert energy based on z_i^t and y_i^t, and get reward r_i^t and new observations o_i^{t+1}
    Store $(\boldsymbol{o}^t, \boldsymbol{a}^t, \boldsymbol{r}^t, \boldsymbol{o'}^t)$ of all MGs in $\mathcal{D}$
    $\boldsymbol{o}^t \leftarrow \boldsymbol{o'}^t$
    for *MG i = 1 to N* **do**
      sample a random mini-batch of size m $(\boldsymbol{o}^k, \boldsymbol{a}^k, \boldsymbol{r}^k, \boldsymbol{o'}^k)$ from $\mathcal{D}$
      Update critic parameters θ_{i1} and θ_{i2} by minimising the loss via (4.48)
      Update actor parameter ϕ_i every two critic updates via (4.51)
    end
    Update target network parameters for each MG i via (4.56)-(4.58)
  end
end

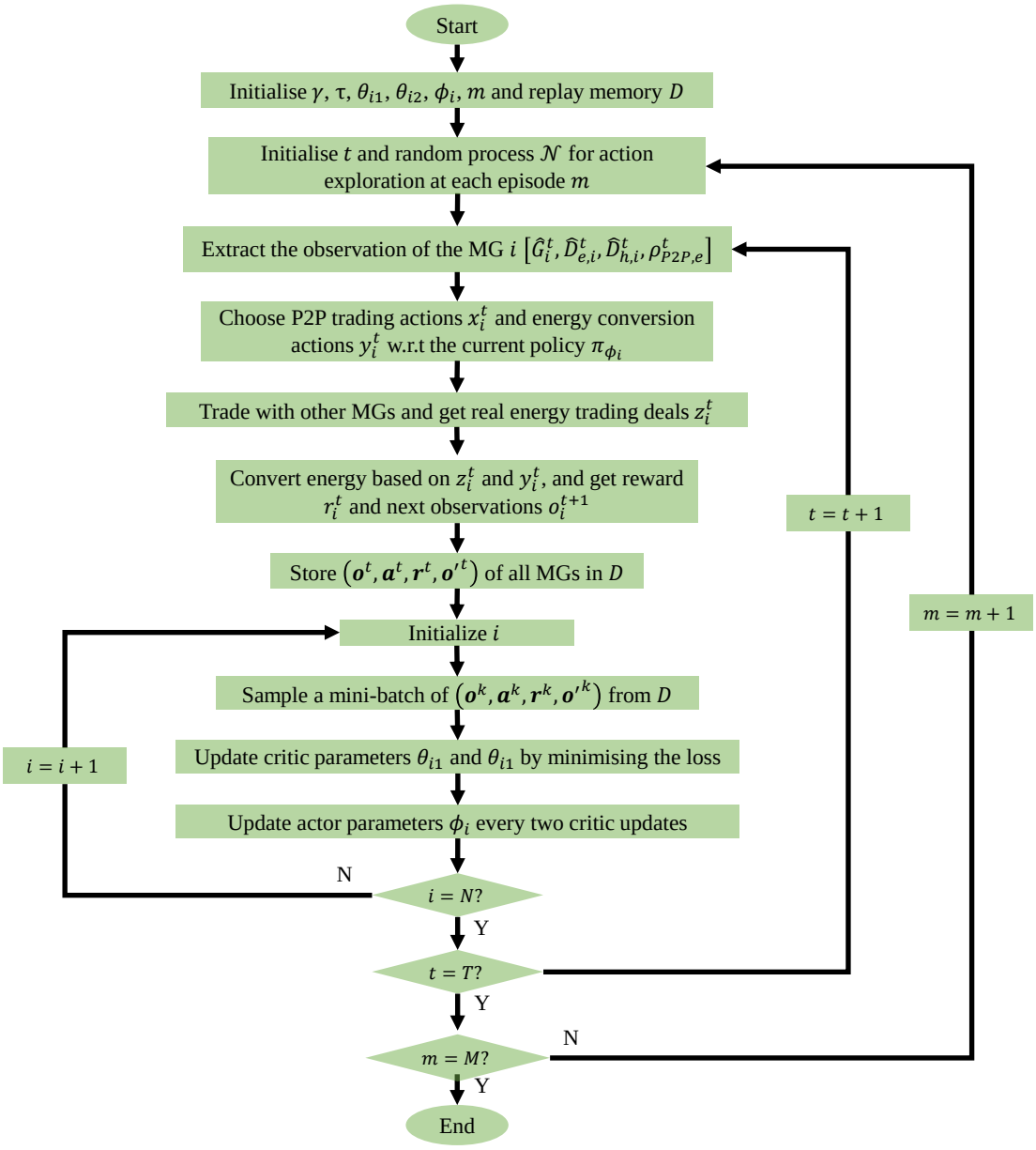


Figure 4.6: Flowchart of implementing our proposed MATD3 method for each agent.

Reward Scaling

Reward values obtained from the reward function cannot be used directly by the agent, since the learning process might not be stable due to the wide range of reward values [148]. Therefore, the reward is sampled from the reward functions to calculate the distribution of the reward, and then the z-score of the new reward (i.e., the standardised reward) can be calculated based on the distribution. This scaling of the reward and setting of a lower bound of the z-score make our learning process stable.

Network Architecture

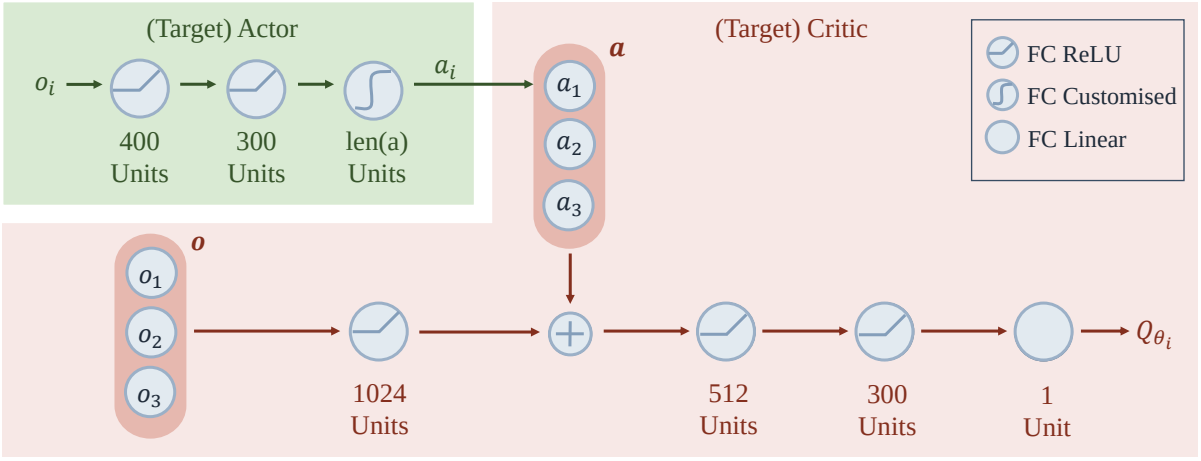


Figure 4.7: The neural network architecture of (target) actor and (target) critic for each agent.

The neural network architecture of (target) actor and (target) critic for each agent are presented in Figure 4.7. The fully connected (FC) layers use the Rectified Linear Unit (ReLU) or the customised function as activation functions. The number of neuron units is labelled below each layer. Compared to the original MADRL framework, a hidden layer has been added to the observations before concatenating with the actions.

4.4 NUMERICAL SIMULATION

4.4.1 Case Study Setup

The proposed MATD3 approach is simulated in a 3-MEMG model including a residential MEMG, a commercial MEMG and an industrial MEMG. Three real-world datasets containing renewable generation and energy demand data at 1-hour resolution are used to train our model, where MG 1 uses data [129] from residential households located in Mueller, Austin, Texas; MG 2 uses data [149] from a commercial data warehouse located in Mueller, Austin, Texas; and MG 3 uses data [150] from a power plant at trial site Aachen/Cologne, Germany[1]. The parameters of energy converters are given in Table 4.1. The electricity price offered by the main grid follows the hourly locational marginal pricing from ISO New England Inc. [130] and the natural gas prices follow the monthly Natural Gas Industrial Price from US Energy Information Administration [151]. Also, the carbon tax price α^{CO_2} is set to 0.0316 \$/kg, while the carbon intensities of natural gas and grid electricity are $\beta^{gas} = 0.245$ kg/kWh and $\beta^{e} = 0.683$ kg/kWh, respectively [152].

[1]We cannot find any industrial MG dataset in the same location as the previous two, however, in our setting we assume these three MG are in the same local area.

Table 4.1: Efficiencies and Capacities of Energy Converters

DER	Efficiency	Capacity (kW/kWh)	Location
WE	$\eta^{WE} = 80\%$	$\bar{P}_e^{WE,t} = 150$	MG 1
FC	$\eta_e^{FC} = 30\%$ $\eta_h^{FC} = 55\%$ [153]	$\bar{Q}_{h_2}^{FC,t} = 330$	MG 1
GB	$\eta^{GB} = 90\%$	$\bar{Q}_{ng}^{GB,t} = 1500$	MG 1,2 & 3
HP	$\eta^{HP} = 300\%$	$\bar{P}_e^{HP,t} = 150$	MG 2
CHP	$\eta_e^{CHP} = 45\%$ $\eta_h^{CHP} = 40\%$	$\bar{Q}_{ng}^{CHP,t} = 900$	MG 3

During training of the proposed method, the Adam optimiser is selected to optimise the neural networks with learning rates of 10^{-5} and 10^{-4} for each actor and critic. The discount factor γ, the soft update parameter τ, the size of the replay buffer, the mini-batch size m, and the full episodes are set to be 0.99, 0.01, 10^5, 100, and 5000, respectively. The exploration noise is generated from a standard Normal distribution with a scale set to be 0.05. The target noise is generated from a clipped Normal distribution with parameters μ, σ and clipped scale set to be 0, 0.1 and 0.25. The examined DRL methods have been implemented using Tensorflow-GPU 2.20 in Python. The case studies have been carried out on a computer with a 4-core 3.80 GHz Intel Core i7-4790 processor, an NVIDIA GeForce GTX 1070 graphics card and 12 GB of RAM.

4.4.2 Performance Evaluation

To demonstrate the effectiveness of our proposed scheme and MATD3 algorithm, the following methods are compared:

The Rule-based Method

The MGs do not use any energy converters, and they only trade energy with external networks. The rule-based operating policy calculates the difference between the estimated energy demand and generation for the trading time slot, and then sells the surplus electricity or buy the needed energy.

SATD3-SEP

The SATD3-SEP method has the same configuration as rule-based one, except that it uses three independent TD3 agents to find the trading actions of each MG with external networks. There-

fore, the agents only use their own observations, actions, rewards and next observations to train their critic networks.

SATD3

The SATD3 method will use our system model for P2P energy trading and energy conversion. However, the agents for the three MEMGs are independent TD3 agents.

MATD3

The MEMGs will use our proposed P2P energy trading and energy conversion scheme, and the optimal energy trading and conversion policies will be selected by our proposed MATD3 method.

As the TD3 algorithms use Gaussian noises to increase exploration, we need to test the algorithms without any added noise. To validate the performance of examined methods, We train each method for 5×10^3 episodes and test them every 20 episodes.

Results

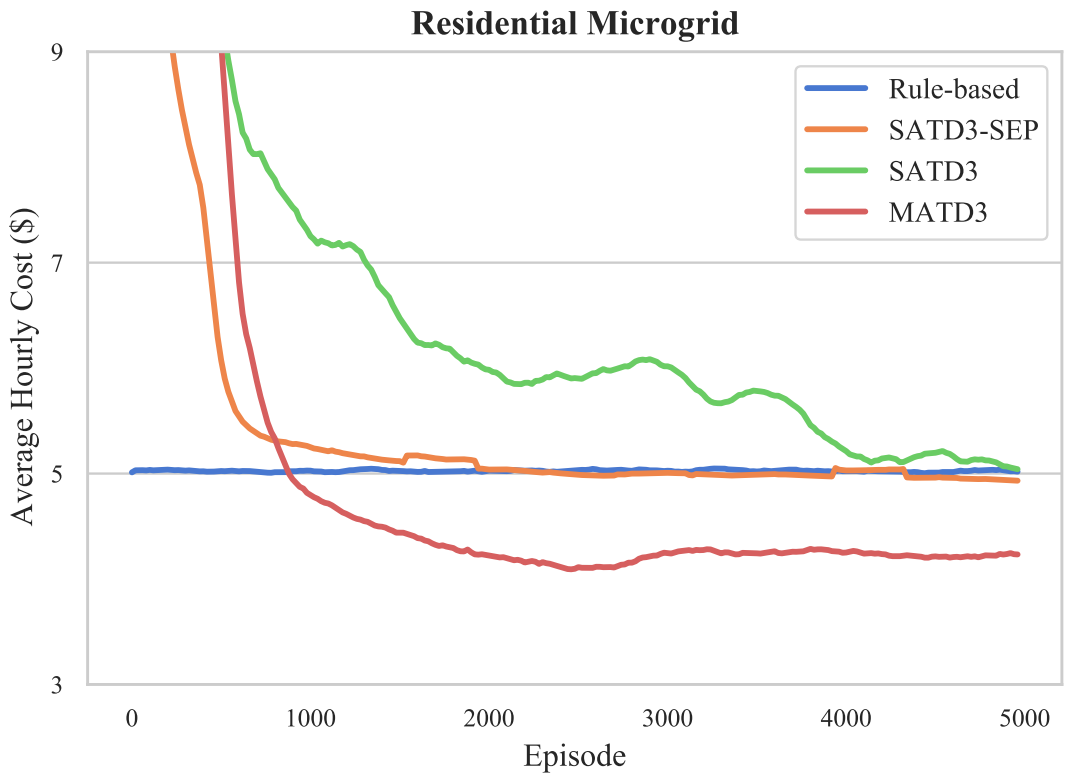


Figure 4.8: Learning curves of P2P energy trading problem for the residential MEMG under examined methods. Curves are smoothed for visual clarity.

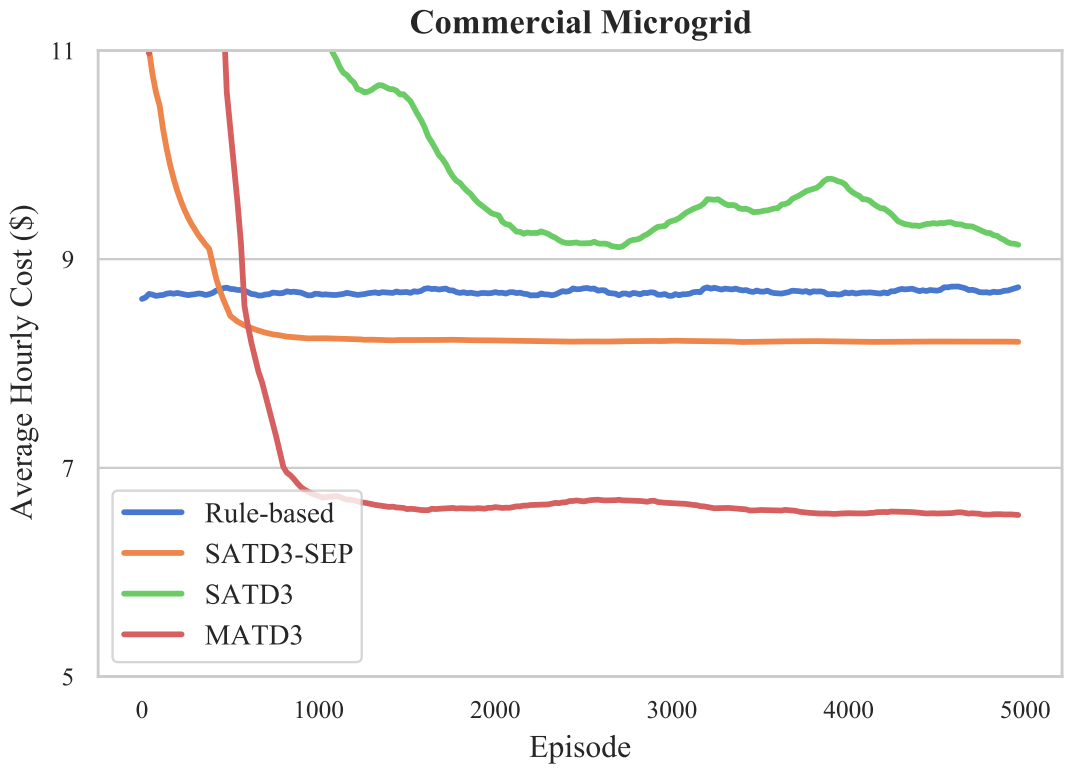


Figure 4.9: Learning curves of P2P energy trading problem for the commercial MEMG under examined methods. Curves are smoothed for visual clarity.

The average hourly operation costs of each MG in a typical winter day using each method are illustrated in Figure 4.8, 4.9 and 4.10. The industrial MG has the highest operation cost due to its highest demand. The figure shows that the proposed MATD3 approach can reduce the operation cost and outperform other methods. The MATD3 approach achieves average hourly costs of 4.119, 6.566 and 9.230 US dollars in the residential MG, commercial MG and industrial MG, respectively. In relative terms, MATD3 reduces the costs 18.2%, 16.5% and 18.1% compared to those of SATD3, SATD3-SEP and rule-based methods for the residential MG; 27.8%, 20% and 24.8% compared to those of SATD3, SATD3-SEP and rule-based methods for the commercial MG; and 23.1%, 13.1% and 20.3% compared to those of SATD3, SATD3-SEP and rule-based methods for the industrial MG. SATD3 did not perform well and failed to converge within 5000 episodes, because directly applying the algorithm into an environment with three interacting MGs violates the Markov assumption. The SATD3-SEP method reduced commercial and industrial MGs' costs compared to the rule-based method. However, SATD3-SEP performs comparably to the rule-based method for the residential MG. These results are due to the fact that the residential MG only has electrical storage, while the commercial and industrial MG have electrical storage and thermal storage.

Furthermore, the mean and the standard deviation of the average hourly cost of the community (all MGs combined) over 100 adjacent episodes for the examined methods are presented in

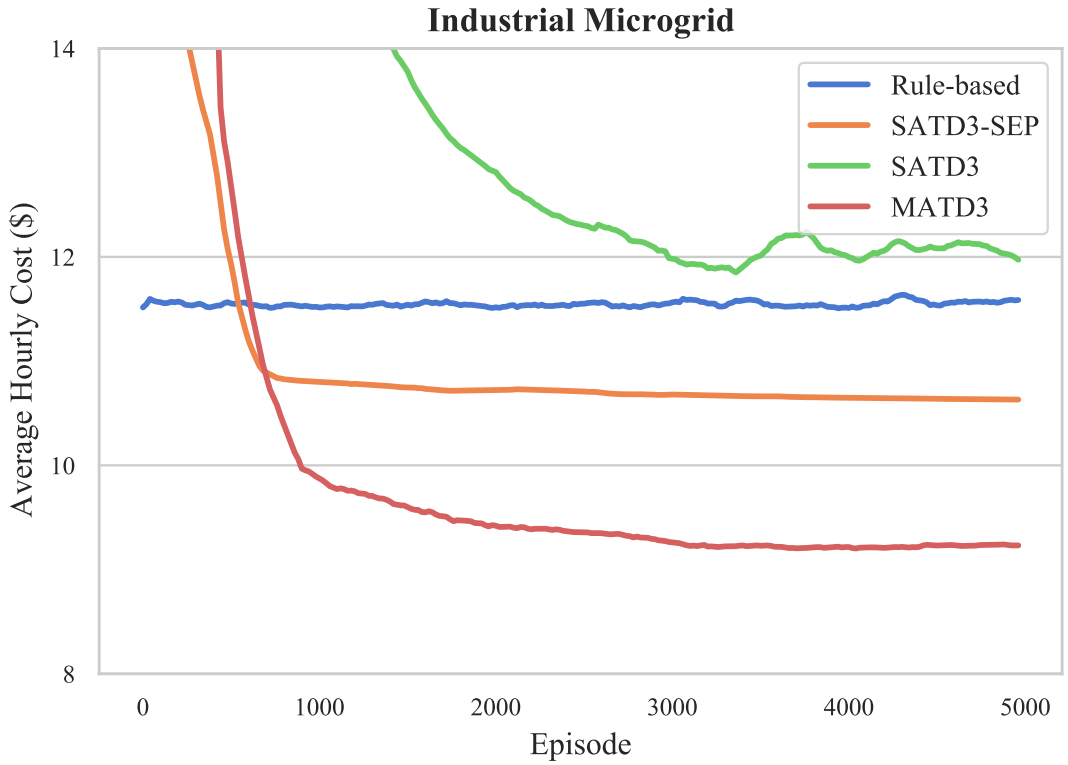


Figure 4.10: Learning curves of P2P energy trading problem for the industrial MEMG under examined methods. Curves are smoothed for visual clarity.

Table 4.2. From the table, similar results can be observed as in Figure 4.8. Specifically, SATD3-SEP and MATD3 converge around 1500 and 3000 episodes, respectively, since the standard deviation is smaller than that of the rule-based method and the mean hardly changes afterwards. However, the cost of the SATD3 method continues to swing during the training.

The computational performance of the compared DRL methods is illustrated in Table 4.3 in terms of training and execution. The average CPU time per episode is the highest in MATD3 since the method involves interactions among all three MGs, and each agent trains its critic using the information from all of the MGs. The total CPU time required to reach convergence is shortest in SATD3-SEP because of the independent agents, longer in MATD3 because of the multi-agent setting, and longest in SATD3 (since it fails to reach convergence). For execution, the CPU time of each DRL method is similar and in the order of milliseconds since the policies are directly inferred from the observations by the trained actor networks.

Table 4.2: Mean ± SD of The Community's Average Hourly Cost ($) over Adjacent 100 Episodes for the examined methods

Method	**Episode**			
	1-100	500-600	1000-1100	1500-1600
Rule-based	25.25 ± 0.12	25.25 ± 0.11	25.26 ± 0.12	25.25 ± 0.12
SATD3-SEP	34.44 ± 7.56	24.45 ± 0.16	24.22 ± 0.10	24.04 ± 0.05
SATD3	42.56 ± 5.42	33.87 ± 0.46	29.07 ± 0.35	27.70 ± 0.25
MATD3	64.65 ± 10.41	24.35 ± 0.37	20.94 ± 0.12	20.43 ± 0.13

Method	**Episode**		
	2000-2100	3000-3100	5000-5100
Rule-based	25.22 ± 0.10	25.25 ± 0.13	25.27 ± 0.13
SATD3-SEP	23.97 ± 0.04	23.87 ± 0.04	23.78 ± 0.03
SATD3	27.36 ± 0.22	26.91 ± 0.18	26.22 ± 0.19
MATD3	20.18 ± 0.12	20.06 ± 0.07	19.99 ± 0.05

Table 4.3: Computational Performance of The DRL Methods

Method	**SATD3-SEP**	**SATD3**	**MATD3**
CPU time per episode (s)	1.76	1.54	1.94
Number of episodes	1500	5000[a]	3000
Total CPU time (h)	0.74	2.13[a]	1.62
CPU time at execution (ms)	1.81	1.63	1.75

[a] Failure to converge within 5000 episodes.

4.4.3 Impact of Energy Conversion and P2P Energy Trading

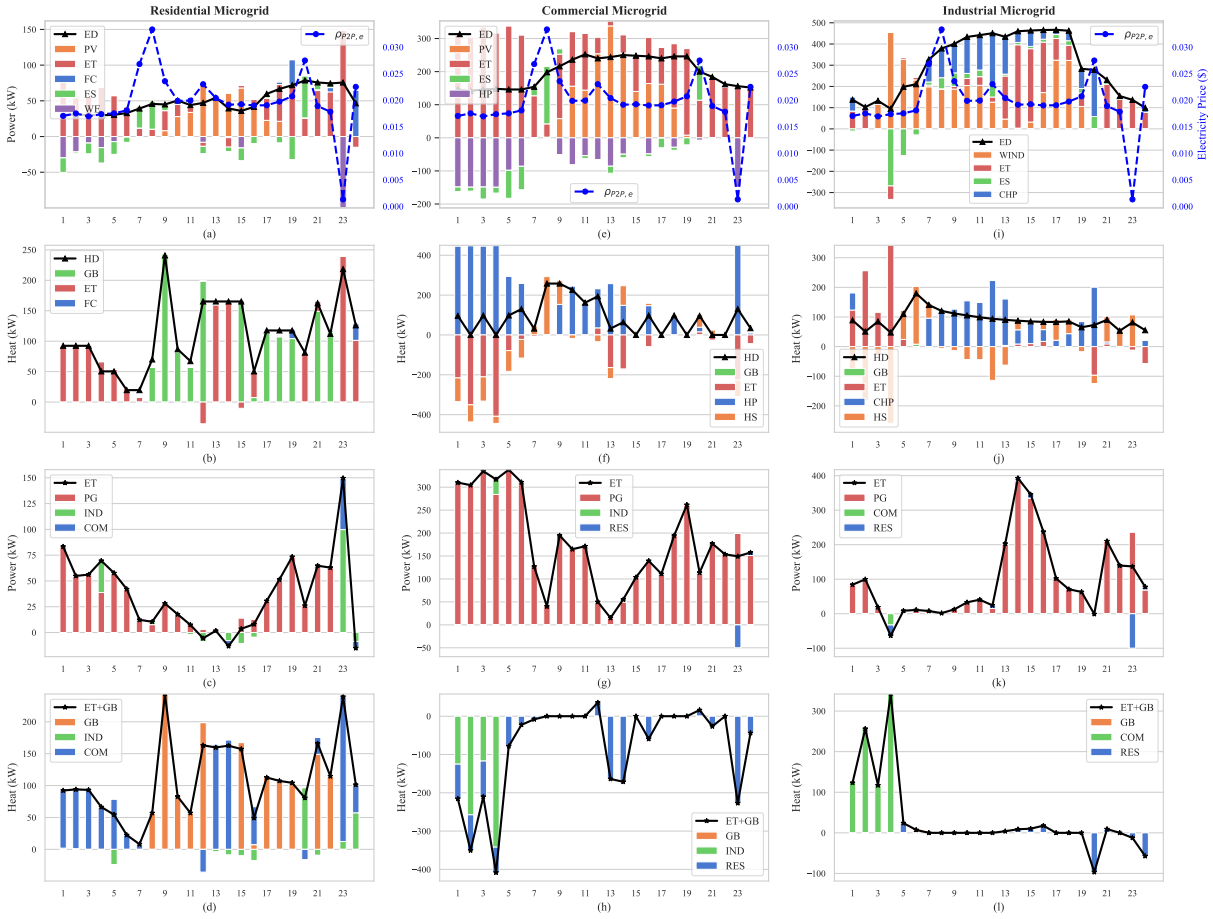


Figure 4.11: P2P energy trading and energy conversion policies for 3 MEMGs under proposed MATD3 method.

Figure 4.11 show the proportion of each MGs' electricity and heat demand that is met in each hour time slot by renewable generation, energy storage, energy trading (including P2P energy trading and trading with the external network), and energy conversion using our proposed MATD3 approach. This figure also shows how our proposed method was able to reduce the average hourly operation cost of each MEMG by revealing the energy trading and energy conversion decisions made at each time slot. The same data was used as in Subsection 4.4.2. For clarification, ED, HD, PV, ES, HS, ET and PG stand for electricity demand, heat demand, solar generation, electrical storage, thermal storage, energy trading and electricity trading with the main grid.

Figure 4.12-4.17 (first two rows in Figure 4.11) show how the electricity demand and heat demand for each MEMG was met using renewable generation, energy storage, energy trading and energy conversion. Figure 4.12, 4.13 and 4.14 reveal that the MGs tend to buy more electricity (labelled in red as ET) when the electricity price is low or the renewable generation is insufficient. The residential MG uses WE to transform purchased surplus electricity to hydro-

gen stored in the hydrogen tank, e.g., Figure 4.12 in hour 22. Later, when needed, FC is used to transform hydrogen to electricity and heat. As shown in Figure 4.16, HP provides a significant amount of heat for the commercial MG and other MGs. For the industrial MG shown in Figure 4.14 and 4.17, CHP is used to provide electricity and heat when wind electricity is insufficient, or the electricity price is high.

Figure 4.18-4.23 (last two rows in Figure 4.11) show the amount of energy traded among three MGs and external networks. P2P energy trading accounts for a considerable proportion of the heat traded. However, the majority of electricity trading is with the main power grid, as all three MGs have insufficient renewable generation to meet their own demand and energy conversion. There is no electricity sold back to the grid in this case, which shows our proposed approach makes appropriate decisions. Figure 4.21-4.23 show the amount of heat that each MEMG trade with others. The commercial MG provides significant heat energy to other MGs via P2P energy trading, which explains why the commercial MG converts much power into heat using HP even when its heat demand is relatively low. These results also show that MGs fulfil their heat demand using P2P heat trading, and only the residential MG needs to buy extra natural gas (Figure 4.21 labelled in orange as GB) from the external network.

These results demonstrate that the proposed MATD3 method can utilise energy conversion to flexibly convert and store the energy when needed. It also allows the community to consume heat energy locally with P2P energy trading, and reduce the surplus electricity sent back to the main grid.

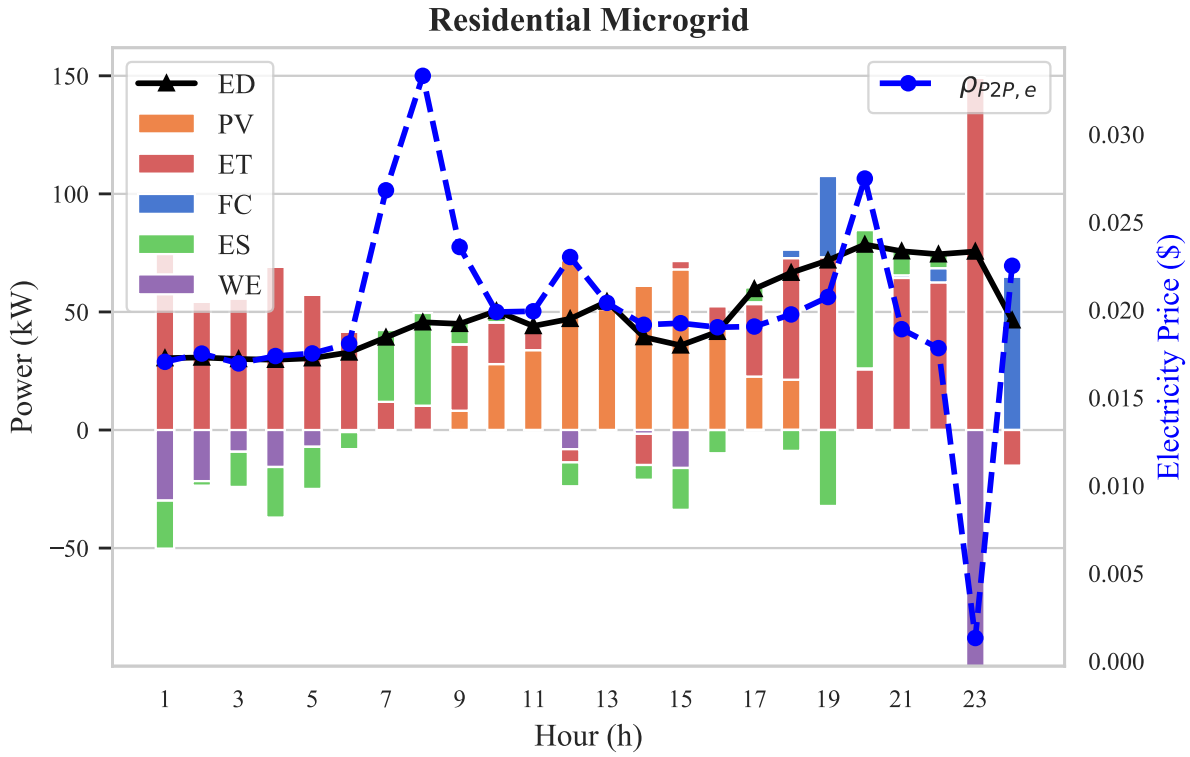


Figure 4.12: The electricity demand and electrical power of the DERs and energy converters for residential MEMG.

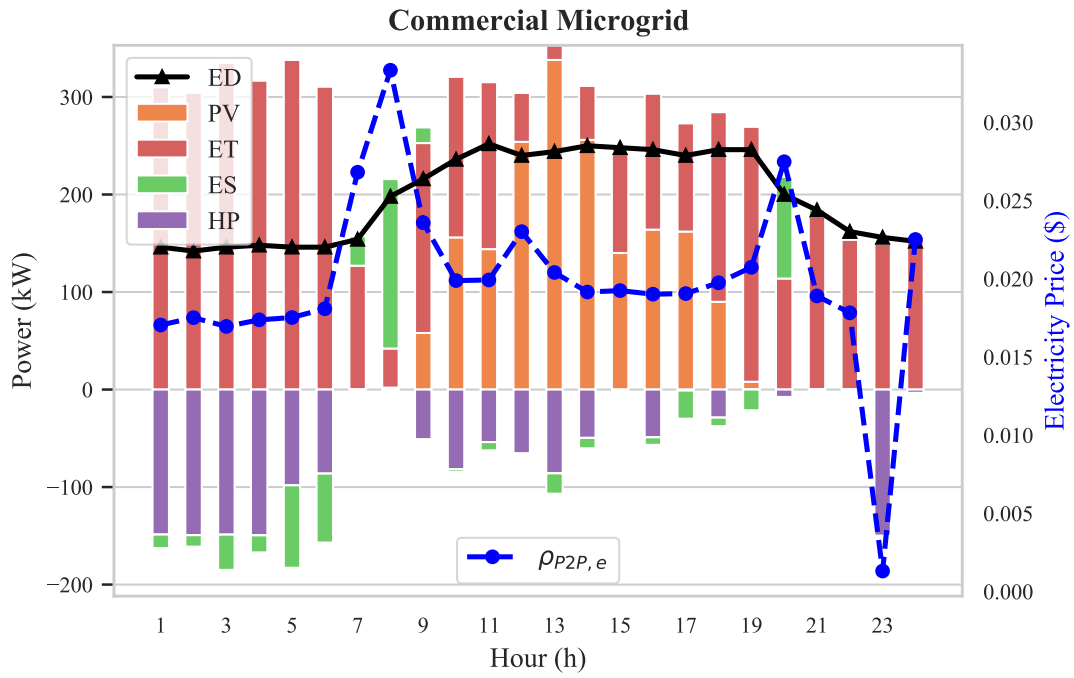


Figure 4.13: The electricity demand and electrical power of the DERs and energy converters for commercial MEMG.

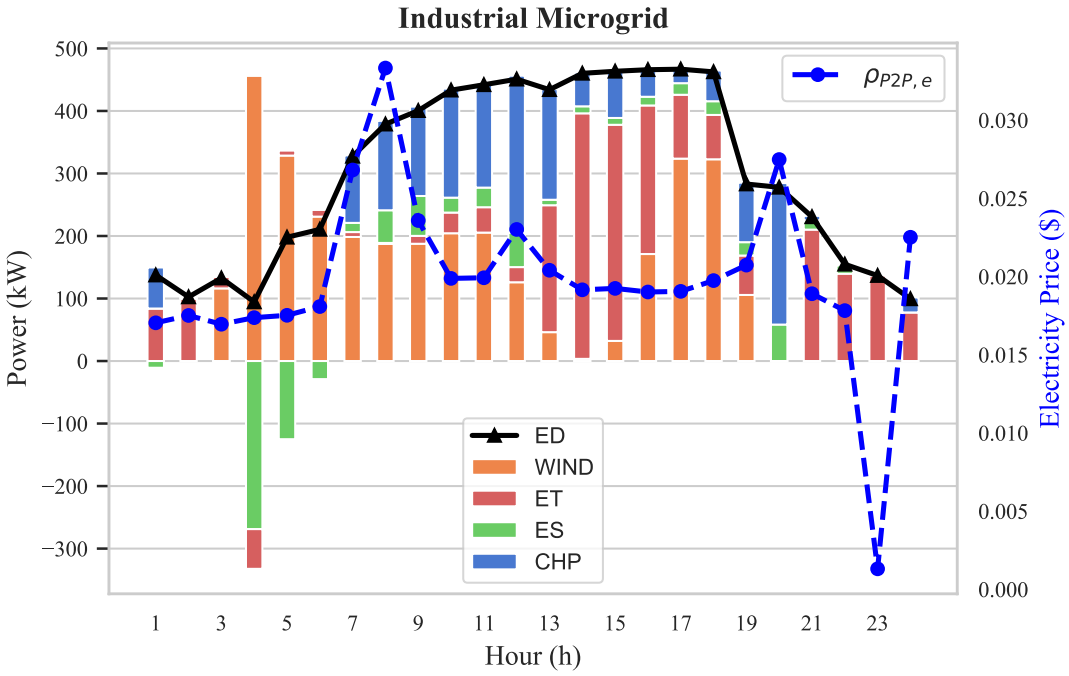


Figure 4.14: The electricity demand and electrical power of the DERs and energy converters for industrial MEMG.

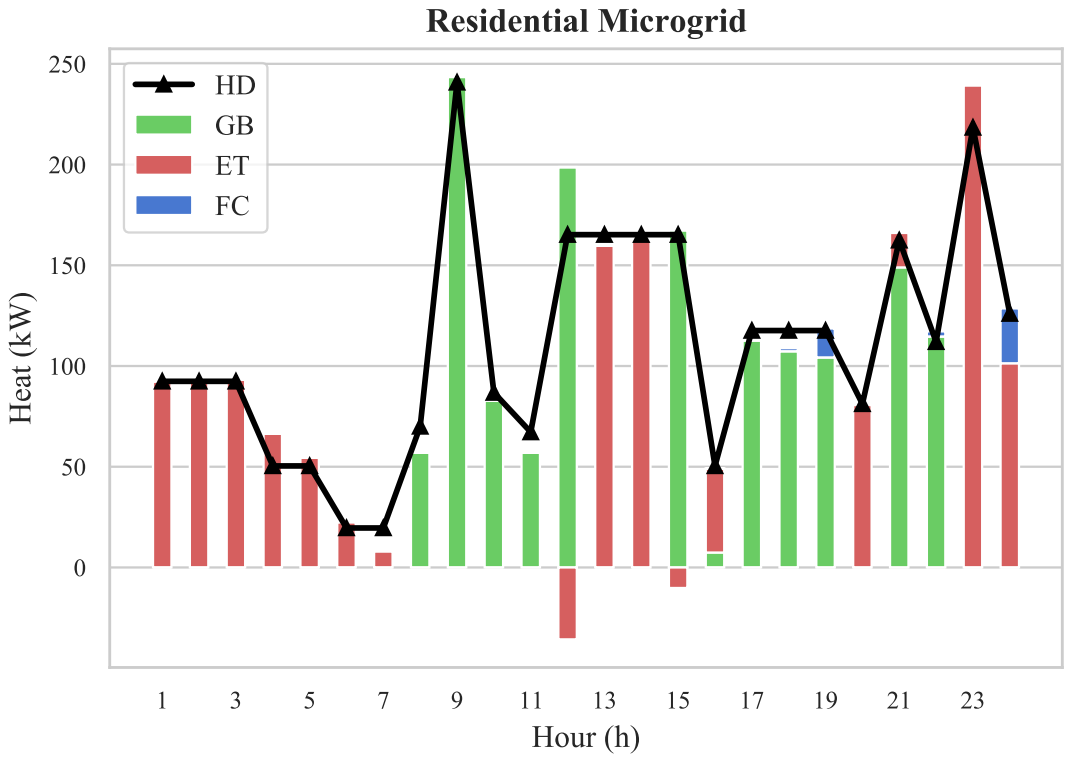


Figure 4.15: The heat demand and heat flow of the DERs and energy converters for residential MEMG.

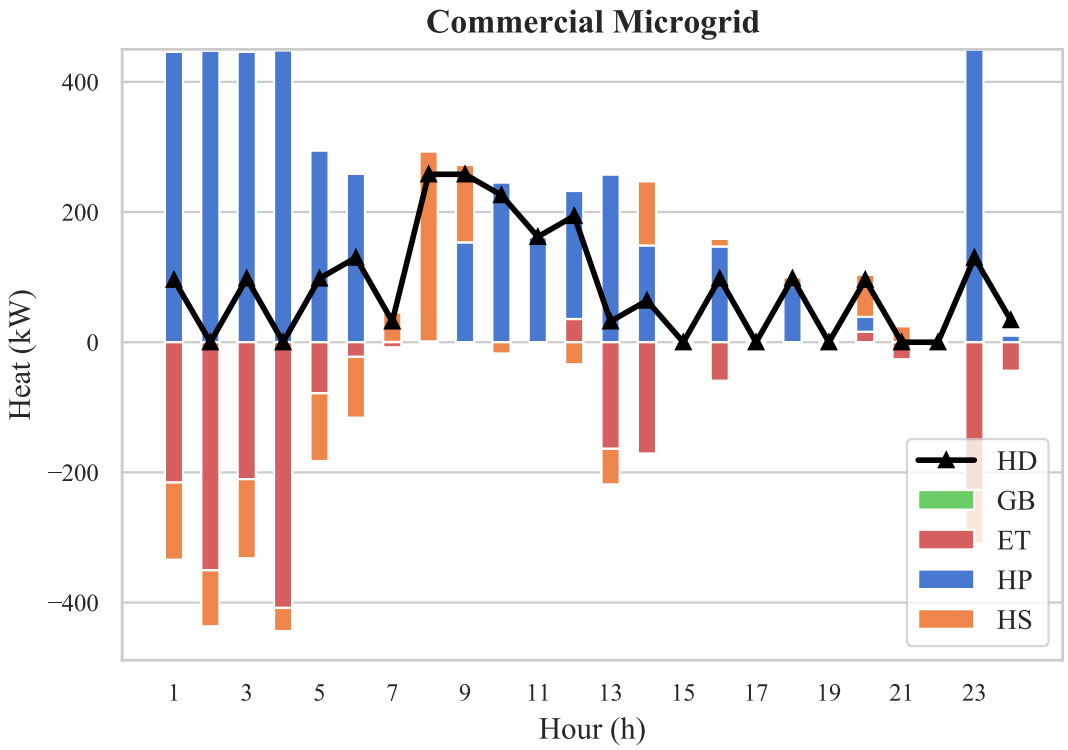


Figure 4.16: The heat demand and heat flow of the DERs and energy converters for commercial MEMG.

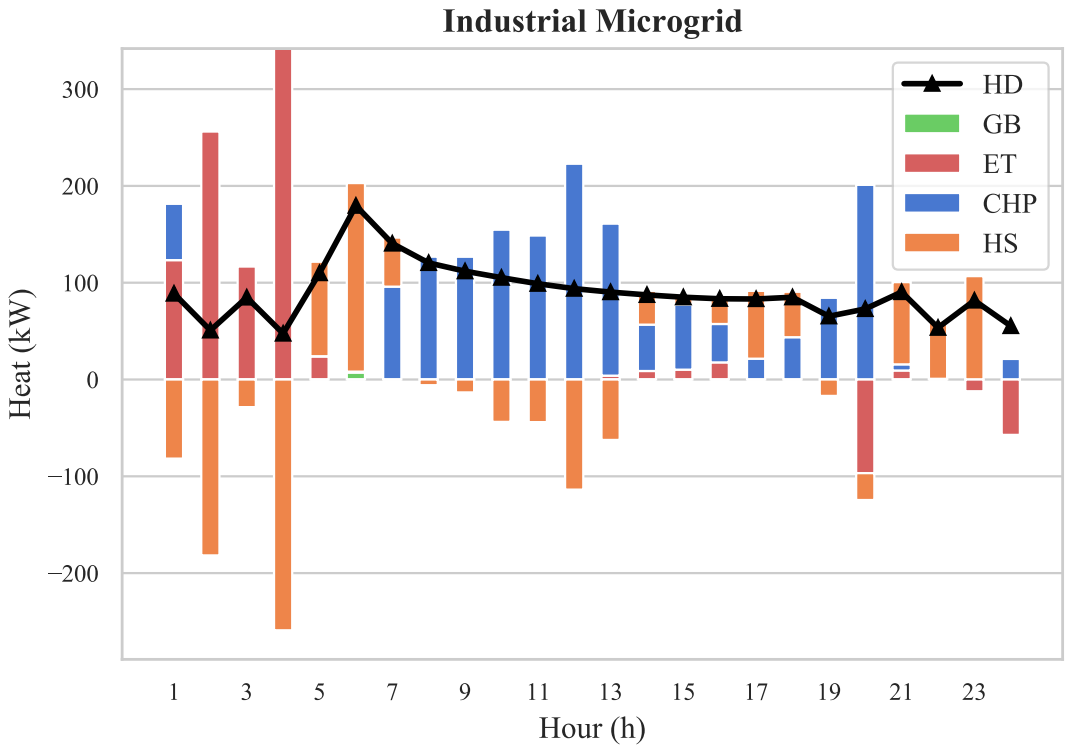


Figure 4.17: The heat demand and heat flow of the DERs and energy converters for industrial MEMG.

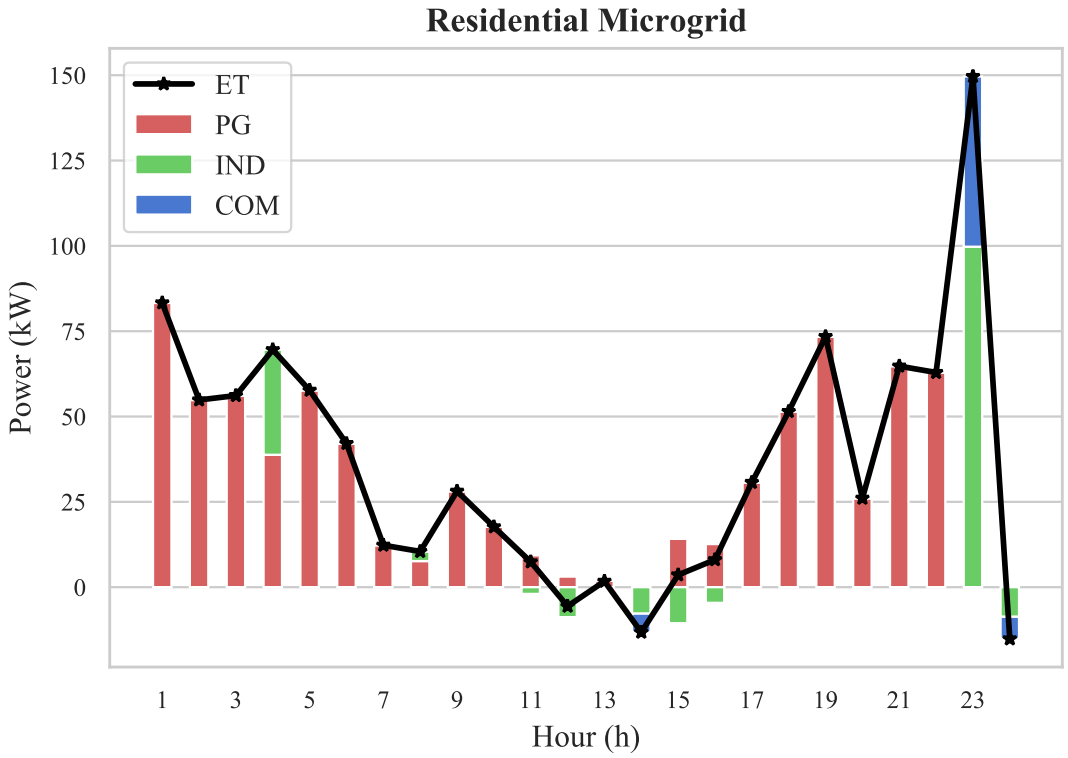


Figure 4.18: The amount of electricity traded with each source for residential MEMG.

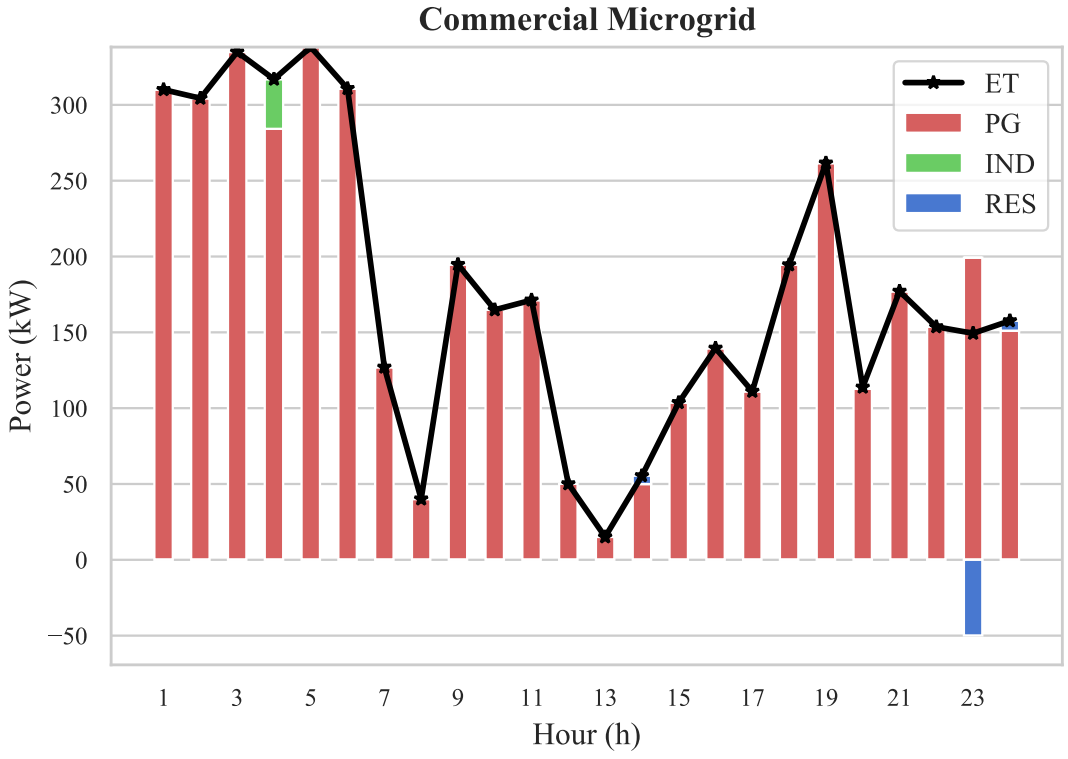


Figure 4.19: The amount of electricity traded with each source for commercial MEMG.

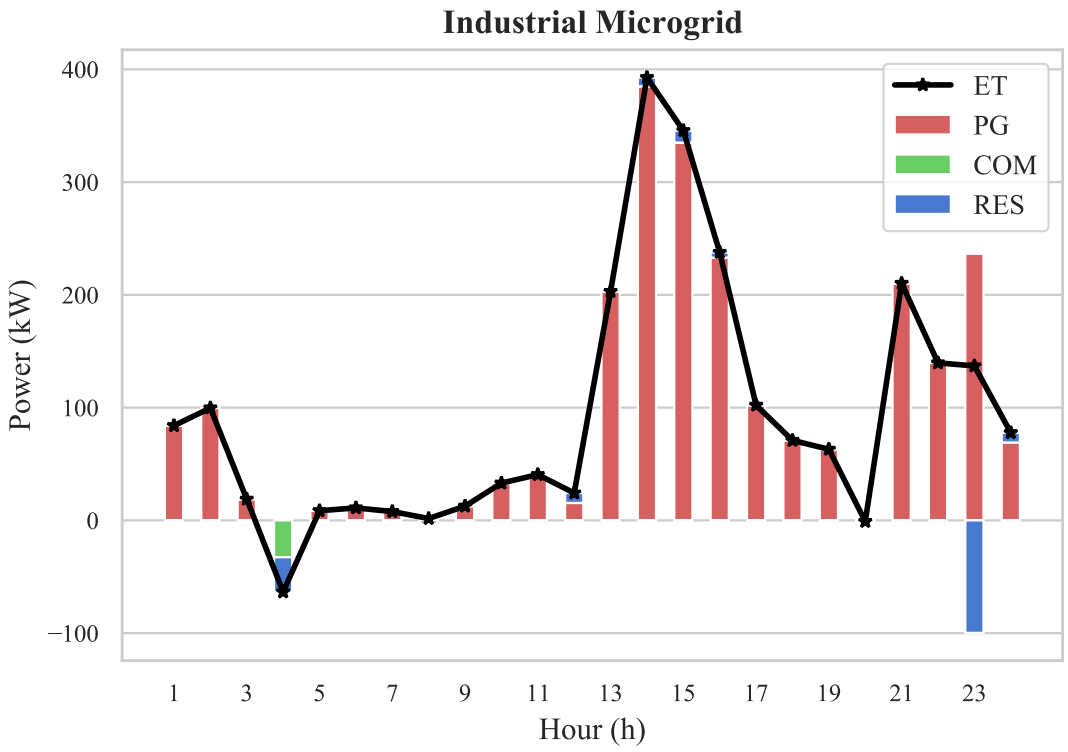


Figure 4.20: The amount of electricity traded with each source for industrial MEMG.

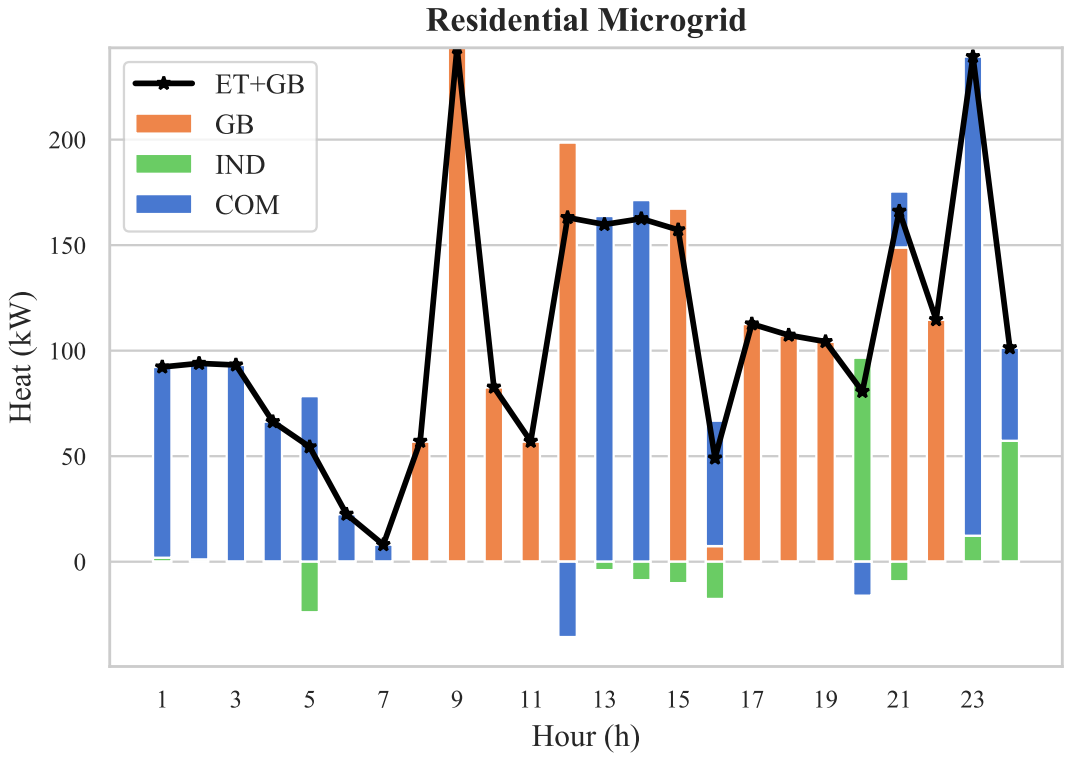


Figure 4.21: The amount of heat traded with each source for residential MEMG.

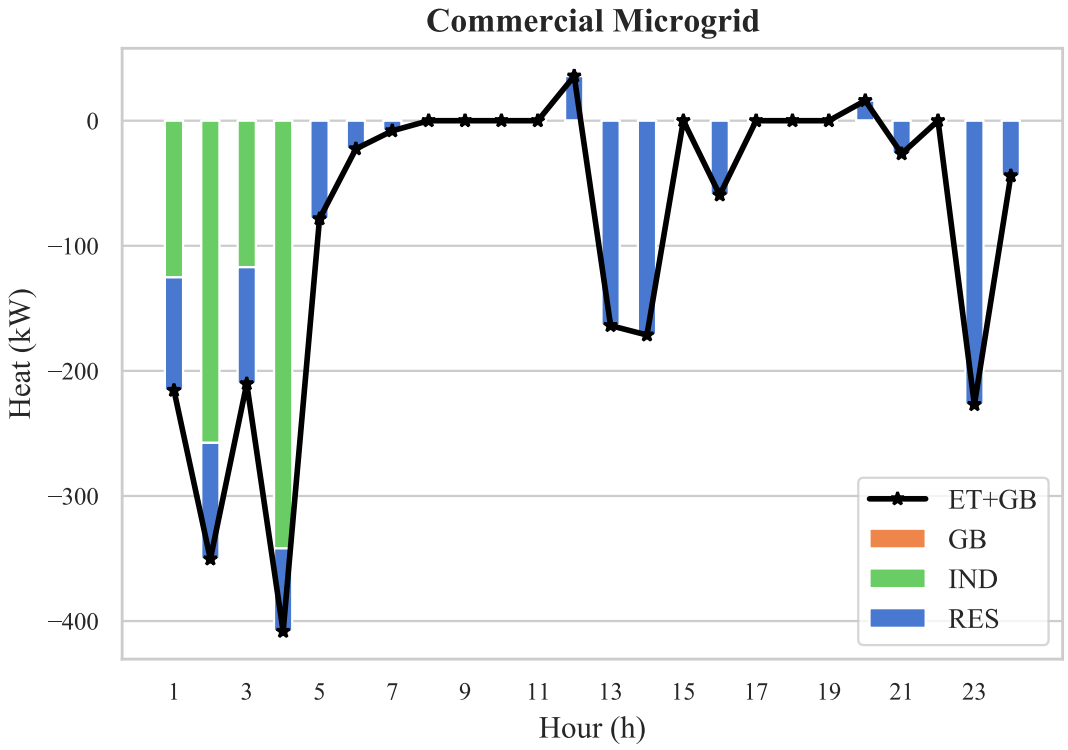


Figure 4.22: The amount of heat traded with each source for commercial MEMG.

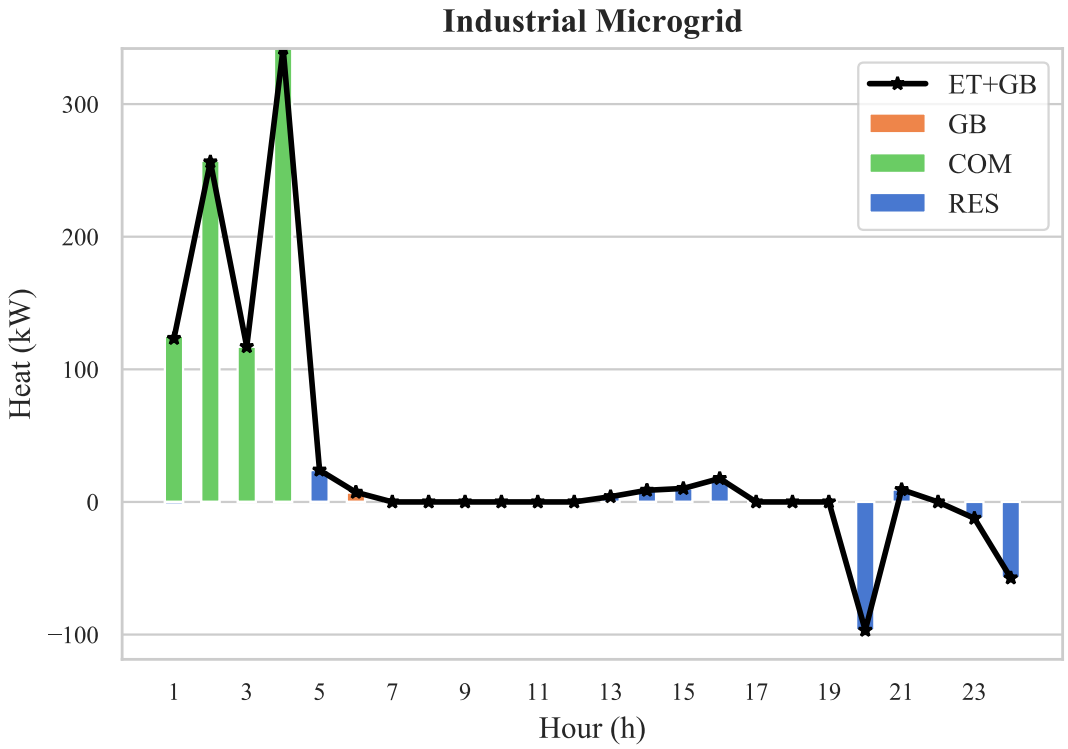


Figure 4.23: The amount of heat traded with each source for industrial MEMG.

4.4.4 Effect of Carbon Tax Price on Costs and CO_2 Emissions

The impact of the carbon tax price on the the average hourly costs for each MEMG using our proposed MATD3 method is shown in Figure 4.24. When the carbon tax price increases, the average hourly costs of each MG increase due to the increase of the environmental cost. The carbon tax price impact is less on the residential MG, because the amount of energy trading with the external network is smaller compared to the commercial and industrial MGs.

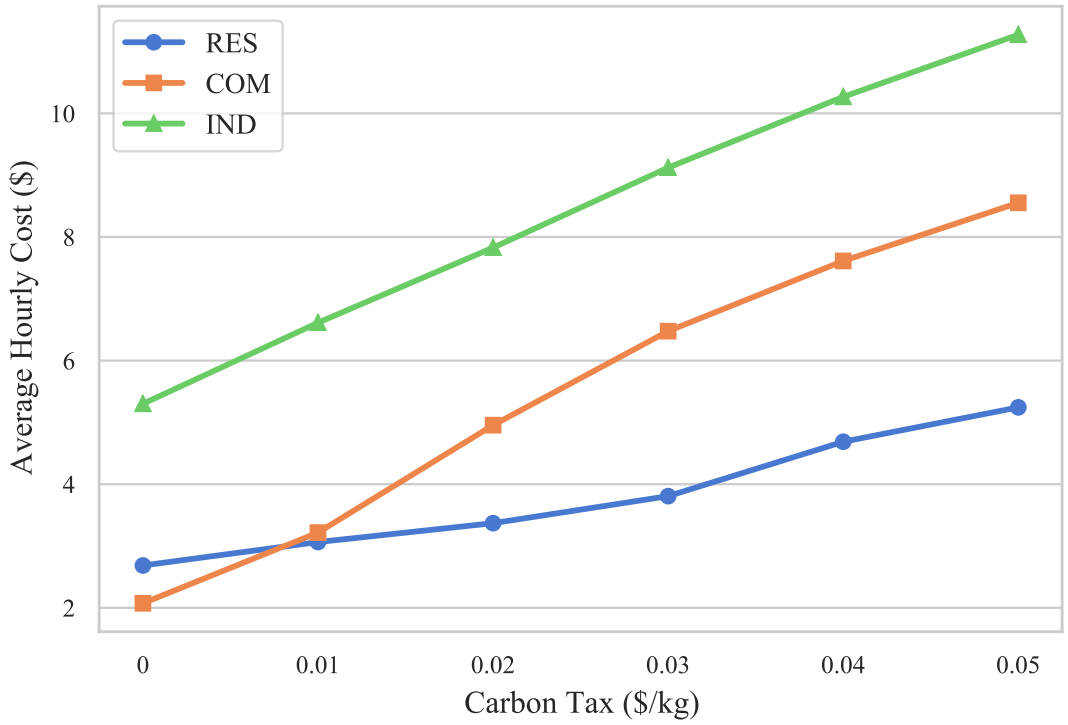


Figure 4.24: Average hourly operation costs at varying carbon tax prices for 3 MEMGs.

The impact of the carbon tax price on the CO_2 emissions for each MEMG using our proposed MATD3 method is shown in 4.25. The CO_2 emissions from the commercial MG are stable when the carbon tax price is lower than 0.02 \$/kg, decrease significantly when the carbon tax increases from 0.02 to 0.04 \$/kg, and stay unchanged when the price increases beyond 0.04 \$/kg. The reason is that the commercial MG converts significant amount of electricity purchased from the main grid to heat and trades the heat with other MGs, since the P2P energy trading profit is higher than the environmental cost when carbon tax price is lower than 0.02 \$/kg. When the carbon tax price is above 0.02 \$/kg, the amount of the heat traded with other MGs is reduced, and the CO_2 emissions are reduced until no more heat trading takes place. The figure shows that there is a sharp drop in CO_2 emissions from the industrial MG when the carbon tax price increases from 0 to 0.02 \$/kg, and the CO_2 emissions remain unchanged when the price increases beyond 0.02 \$/kg. As the carbon tax price increases, the industrial MG uses the CHP to meet a higher proportion of its own electricity demand rather than buying electricity from the main grid, and all

of the electricity demand was met by the CHP when the carbon tax price is 0.02 $/kg. Therefore, the CO_2 emissions are reduced at first, and remain unchanged above the carbon tax price of 0.02 $/kg. The residential MG produces less CO_2 emissions as the carbon tax price increases from 0 to 0.02 $/kg, and then more CO_2 until the carbon tax price reaches 0.04 $/kg, and the same amount of CO_2 above that. This is because the residential MG uses more environmentally friendly approach as the carbon tax price increases at first. However, when the carbon tax price is higher than 0.02 $/kg, it has to use an increasing amount of natural gas to meet its heat demand because the commercial MG begins to reduce selling heat to other MGs. Once the carbon tax price is higher than 0.04 $/kg, the demand is met fully by the natural gas in residential MG.

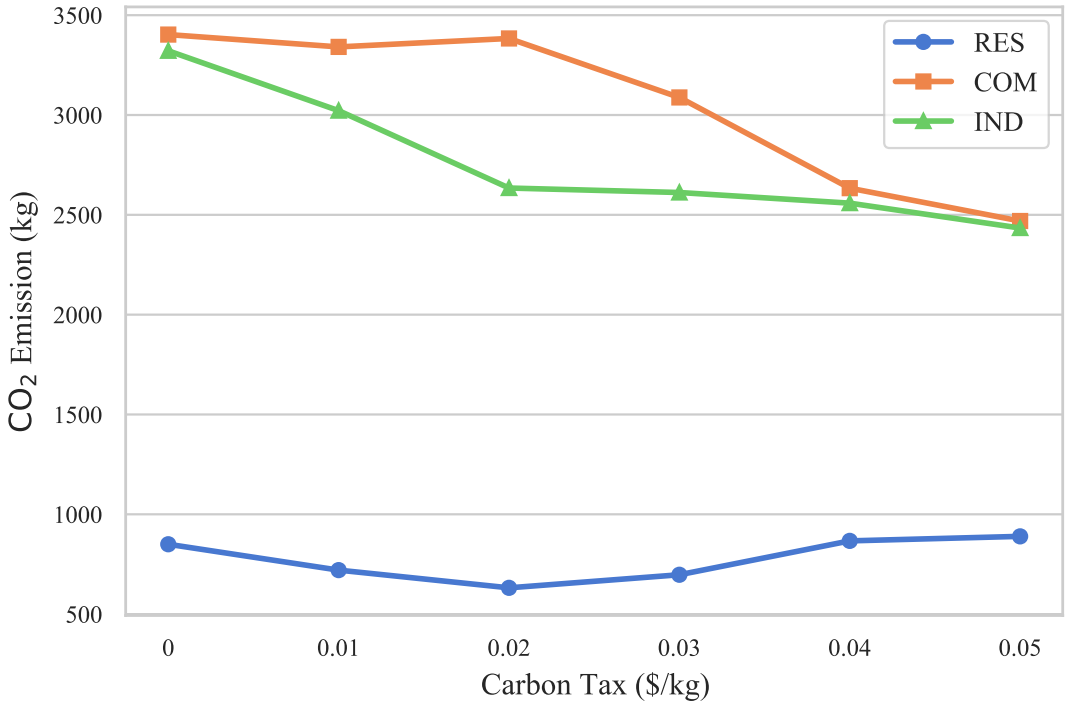


Figure 4.25: CO_2 emissions at varying carbon tax prices for 3 MEMGs.

4.4.5 Scalability of the proposed approach

The proposed approach can be scaled up for a longer period than a day. If the proposed approach is scaled up to a month or shorter, the time horizon in the system problem shown in (4.15) needs to be changed from a day to the new period. This method only needs to train the agents once and the results can be reused. If the extended period is longer, e.g., up to one year, the agents have to be trained periodically using the newest collected data and old data. The computational costs of this method are higher than the first method.

4.5 Conclusion

An external P2P energy trading and internal energy conversion problem was investigated for the interconnected residential, commercial and industrial MEMGs in a local community. The problem was formulated as a POMDP, and a multi-agent deep reinforcement learning approach was proposed to address it. The proposed approach aligns with the nature of P2P energy trading, and can also handle a high-dimensional continuous action space and alleviate overestimation and high variance problems. The case study on three real-world datasets showed that the proposed method significantly reduced all MGs' operation costs. The simulation results also demonstrated that the MATD3 method can utilise energy conversion to flexibly convert and store the energy and allows MGs to consume heat energy locally with P2P energy trading. The simulation results also showed the impact of carbon tax price on the operation cost and CO_2 emissions.

To the best of our knowledge, this work is the first to consider the combined dynamics of energy trading, energy conversion and multi-vector energies (electricity, heat and natural gas) as a whole. More options can be used to match supply with demand, making the system more flexible overall. Increased flexibility provides alternatives to adding additional costly infrastructure to meet demand and supports the inclusion of a higher share of variable renewable energy sources.

Chapter 5

Coordination for Multi-Energy Microgrids in Double Auction Market Using Multi-Agent Deep Reinforcement Learning

5.1 Introduction

Power systems are undergoing a significant transition from fossil fuel resources to the decolonisation of renewable energy resource (RES), promising to address the environmental concerns [154]. However, the less controllable and predictable RES introduce new challenges to power system planning and operation [155]. In this respect, there has been a significant increase in developing multi-energy systems (MES) that interact electricity, gas, heat with each other, constituting a significant opportunity to provide the flexibility of shifting across multiple energy vectors and resulting in a cost-effective and reliable system [131]. Currently, an increasing attention has been made to study MES inside microgrid, forming the multi-energy microgrids (MEMG) [156, 157]. MEMG is composed of various energy loads, generators, storages, and converters under the microgrid concept. Currently, the benefits of using MEMG have been discussed in many studies [157]. Instead of independently scheduling each energy vector, the integrated manner is more efficient to deal with the complementary and synergistic effects of MES, and therefore booting the operation efficiency of MEMG.

Gas and electricity are the two main input energy sources for MEMGs. The gas retail market is normally indifferent to MEMGs, allowing them to buy gas but not sell it back [158]. The electricity retail market under the deregulation is more active and flexible, where MEMGs with RES can sell electricity back to the grid at Feed-in Tariff (FiT) [159]. However, under scenarios where MEMGs need to import energy from the grid, the higher-rated Time-of-Use (ToU) prices, compared with the lower FiT issued by the same utility company, can present a dilemma for MEMGs' net import decision making [160]. Furthermore, when MEMGs participate in the traditional market, they act independently to manage their supply-demand balance. This is,

however, not optimal as the lack of coordination with others leaves untapped the full potential of energy flexibility for achieving overall system supply-demand balance [161]. To this end, an efficient and coordinated P2P energy trading framework for local MEMGs is urgent to maximise the economic benefit and system flexibility.

In last chapter, MADRL is used to find efficient energy trading and energy conversion policies for MEMGs. However, it is assumed that the MGs decide an agreed P2P price for all of the MGs, and then negotiate the amount of electricity traded among themselves. Considering that energy price is a key factor in determining the benefits of local energy trading, the design of pricing mechanism plays a crucial role in P2P energy trading. Furthermore, keeping regulation and safety concerns in mind, MEMGs also desire autonomy and privacy. Therefore, MEMGs may be not willing to disclose their system information which are needed in previous MADRL methods. The following research questions are related to the topic and addressed in this chapter:

1) Is there any efficient P2P energy trading framework other than centralised and decentralised P2P energy trading frameworks and with a fair P2P energy price mechanism? The answer is double auction (DA) market, which has a reasonable trading pricing mechanism and ensure the market efficiency while preserving the private information of the MEMGs.

2) Is it possible to integrate a DA market into the P2P energy trading service since it offers both the benefits of centralised and decentralised markets? The answer is to formulate a Partially Observable Markov Decision Process (POMDP) with discrete time steps for the P2P energy trading and energy conversion problem in the DA market.

3) How will this proposed method ensure that the local MEMGs can keep the privacy of their system information and at the same time achieve maximum social welfare as is achieved by a centralised method? The proposed method, DA-MATD3, integrates the key information of the DA market into the state-of-the-art MATD3 algorithm by connecting the critic networks of the agents with the DA market order books. The details are discussed in the following sections.

The remainder of the paper is organised as follows. Section 5.2 formulates the examined P2P energy trading and energy conversion problem of multiple MEMGs in a DA market. Section 5.3 proposes the DA-MATD3 method. Section 5.4 presents the case studies to evaluate the effectiveness of the proposed method. Section 5.5 draws the conclusions.

5.2 Coordination of MEMGs in Double Auction Market

5.2.1 Problem Setting

We focus on a local energy community, which is composed of a group of MEMGs categorised into residential, commercial, and industrial areas. In detail, the set of components of the proposed MEMGs includes 1) two types of consumption loads: electric load (EL), heat load (HL); 2) two types of RES generators: solar photovoltaic (PV), wind generator (WG); 3) two types

of storage units: electric energy storage (EES), thermal energy storage (TES); and 4) four types of energy converters: combined heat and power (CHP) engine, fuel cell (FC), electric heat pump (EHP), and gas boiler (GB). The MEMGs are categorised into three groups: 1) residential MEMGs with the energy portfolio of EL, HL, PV, EES, TES, FC, GB; 2) commercial MEMGs with the energy portfolio of EL, HL, PV, EES, TES, EHP, GB; and 3) industrial MEMGs with the energy portfolio of EL, HL, WG, EES, TES, CHP, GB.

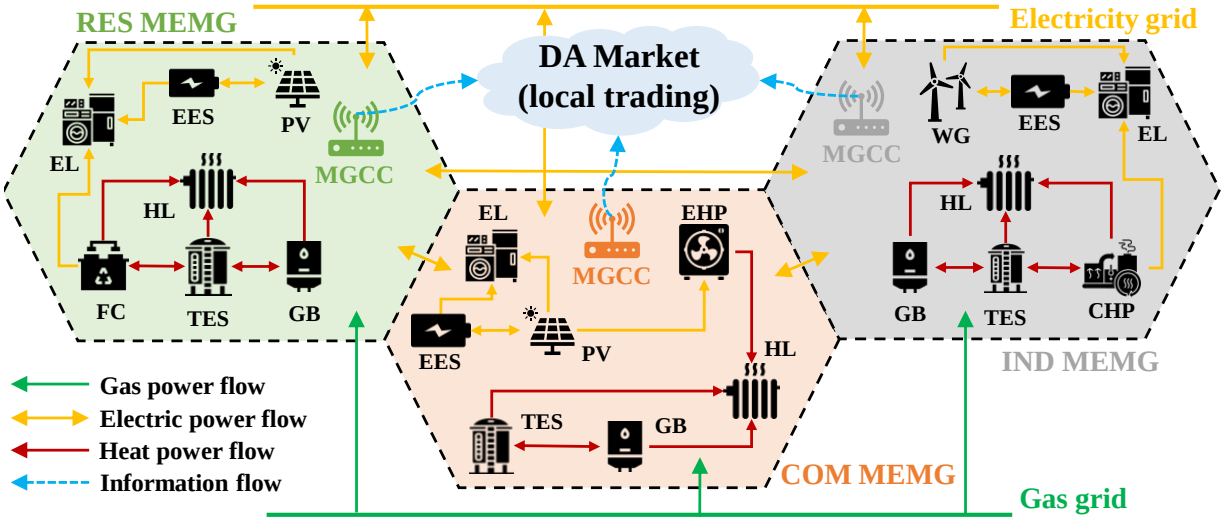


Figure 5.1: P2P energy trading framework and MES of considered MEMGs.

In order to incentive MEMGs to cooperatively participate in local trading, a DA market is introduced in this problem due to its high trading efficiency [120]. As illustrated in Fig. 5.1, the options of each MEMG to supply its consumption loads are diverse. First, MEMGs can manage their own installed energy resources to supply EL and HL. Second, MEMGs can trade their electricity with each other in the DA market. Third, MEMGs are allowed to buy/sell their unbalanced electricity with the utility company at the grid buy/sell prices. Finally, MEMGs can purchase natural gas from the gas grid. The decision-making problem is processed for each hour across a daily horizon, with the objective of minimising energy cost and carbon emission. At each hour, each MEMG equips a microgrid central controller (MGCC) [162] to manage its energy schedules and trading decisions based on: 1) grid information of energy and carbon price signals; 2) local information of its consumption loads, renewable generations, the status of controllable components; and 3) community information of DA market trading prices and quantities.

5.2.2 System Controllable Components

Energy Storage Units

The energy storage units with the high flexibility in MEMGs are characterised by their redistribution ability of off-peak and peak loads and the ability to absorb free RES for the future usage

when energy prices are at the peak. The mathematical models of an EES unit can be formulated as

$$E_{t+1}^{ees} = E_t^{ees} + P_t^{eesc} \cdot \Delta t \cdot \eta^{eesc} + P_t^{eesd} \cdot \Delta t / \eta^{eesd}, \tag{5.1}$$

$$\underline{E}^{ees} \leq E_t^{ees} \leq \overline{E}^{ees}, \tag{5.2}$$

$$0 \leq P_t^{eesc} \leq \overline{P}^{ees} \cdot V_t^{ees}, \tag{5.3}$$

$$\overline{P}^{ees} \cdot (V_t^{ees} - 1) \leq P_t^{eesd} \leq 0, \tag{5.4}$$

where the equality constraint (5.1) corresponds to the storage dynamic transition of battery energy content, taking into account the energy losses caused by the charging and the discharging efficiencies $\eta^{eesc}, \eta^{eesd} \in (0, 1]$. Constraint (5.2) expresses the lower and upper bounds of battery energy content. The following constraints (5.3) and (5.4) ensure that charging and discharging power P_t^{eesc}, P_t^{eesd} are under its power capacity $\overline{P}^{ees}$ and operate mutually exclusive (since the battery cannot charge and discharge simultaneously). As a result, a binary variable $V_t^{ees} \in \{0, 1\}$ is introduced in (5.3)-(5.4) to express the charging status ($V_t^{ees} = 1$) and discharging or idle status ($V_t^{tes} = 0$) of EES unit. Then, the power rate Q_t^{tesc}, Q_t^{tesd} as well as the storage dynamic transition E_{t+1}^{tes} of TES unit can be derived similarly to the EES model (5.1)-(5.4).

Energy Converters

The key to model an MES is to capture the energy conversion relationship between different energy carriers. The studied MEMGs mainly consist of four types of energy converters. CHP engine, a single-input-multi-output converter, is typically characterised by its high energy efficiency compared to independent electricity and heat sources. Therefore, it is considered a critical converter in the MES. The coupled heat and electricity generation of a CHP engine is modelled as

$$P_t^{chp} = \eta^{chpe} \cdot G_t^{chp}, \tag{5.5}$$

$$Q_t^{chp} = \eta^{chpq} \cdot G_t^{chp}, \tag{5.6}$$

$$0 \leq P_t^{chp} \leq \overline{P}^{chp}, \tag{5.7}$$

$$0 \leq Q_t^{chp} \leq \overline{Q}^{chp}, \tag{5.8}$$

where constraints (5.5) and (5.6) respectively indicate the efficiency of CHP engines to convert natural gas into electric and heat power, determined by the conversion efficiency from gas to electricity η^{chpe} and the conversion efficiency from gas to heat η^{chpq}, respectively. The electric and heat output power capacities of CHP are limited by its lower and upper bounds expressed in (5.7) and (5.8), respectively.

Like CHP engines, FC is also a single-input-multi-output converter, characterized by its higher combined efficiency and lower emissions than CHP engines. Given the high thermal efficiency and low operating temperature, FC is more suitable for individual residents with high heat demands. FC can use a variety of fuels. In this work, FC operates on natural gas but could switch to hydrogen if available with minor modification.

$$P_t^{fc} = \eta^{fce} \cdot G_t^{fc}, \tag{5.9}$$

$$Q_t^{fc} = \eta^{fcq} \cdot G_t^{fc}, \tag{5.10}$$

$$0 \leq P_t^{fc} \leq \overline{P}^{fc}, \tag{5.11}$$

$$0 \leq Q_t^{fc} \leq \overline{Q}^{fc}, \tag{5.12}$$

where constraints (5.9) and (5.10) indicate the efficiency of FC to convert natural gas into electric and heat power, respectively. Constraints (5.11) and (5.12) limit the FC output electric and heat power, respectively.

Apart from CHP and FC, the studied MEMGs also include the energy converters of EHP and GB. The EHP produces heat energy by consuming electricity, as presented in equation (5.13), where η^{ehp} represents the energy conversion efficiency from electricity to heat power. The power output of EHP is limited by its capacity $\overline{Q}^{ehp}$ in (5.14).

$$Q_t^{ehp} = \eta^{ehp} \cdot P_t^{ehp}, \tag{5.13}$$

$$Q_t^{ehp} \leq \overline{Q}^{ehp}. \tag{5.14}$$

GB is a vessel converting natural gas to heat energy. The generation of heat from natural gas via GB is given in (5.15) that is affected by the energy conversion efficiency from natural gas to heat power η^{gb}. Furthermore, the boiler has a limitation $\overline{Q}^{gb}$ for outputting heat power, which is considered by (5.16).

$$Q_t^{gb} = \eta^{gb} \cdot G_t^{gb}, \tag{5.15}$$

$$Q_t^{gb} \leq \overline{Q}^{gb}. \tag{5.16}$$

5.2.3 Double Auction Market

The DA market matches multiple buyers (MEMGs with energy deficit) and sellers (MEMGs with energy surplus) who are interested in local trading, and is deemed as a highly efficient mechanism [120]. It is called a double auction because the bids tend to increase while the offers tend to decrease. Therefore, the spread between bids and offers narrows until one of the bids or offers is accepted. It is widely used in the trading of a variety of commodities, including equities and electricity. In this work, we apply the DA market to the local electricity trading, while the heat energy cannot be traded in the community. In general, a DA market lasts for a fixed period of time, known as the auction period (one hour). It allows traders to submit their bids/offers at the beginning of each auction period, then the auctioneer (DA market operator) clears the market and publishes the public market outcomes (trading prices and quantities) at the end of each auction period. More specifically, a DA market comprises:

- A set of buyers $\mathcal{B}$, where each buyer $b \in \mathcal{B}$ defines its trading price p_b and quantity q_b, which means the buyer b would like to buy q_b amount of energy at price p_b.
- A set of sellers $\mathcal{S}$, where each seller $s \in \mathcal{S}$ defines its trading price p_s and quantity q_s, which means the seller s would like to sell q_s amount of energy at price p_s.
- A public order book managed by an auctioneer, where the all accepted bids and offers are listed, respectively. Bids submitted by buyers are sorted by decreasing the submitted buy prices and queue in buy order book $k^b(b, p_b, q_b)$, while offers submitted by sellers are sorted by increasing the submitted sell prices and queue in sell order book $k^s(s, p_s, q_s)$.

The pseudo-code of the clearing process in DA market is given in Algorithm 3. Once an auction period begins, traders submit their order information with a trading price and a corresponding energy quantity to the market, collected by the auctioneer (step 1). All submitted orders are allocated in the order book (step 2). The clearing algorithm iterates down the order books and attempts to match each buy order with sell order (steps 3-12) until the buy price is less than the sell price or no unmatched sell/buy order exists anymore (steps 13-14). Specifically, when two orders get matched, the auctioneer calculates the trading price between the matched buy price and sell price, using the traditional mid-pricing method [120] (step 6), while the trading quantity is equal to the lower value between the two matched orders (step 5). Due to the sorting principle and clearing algorithm, the clearing results promise the social welfare maximization [120]. Finally, at the end of the auction period, the remaining quantity of energy and the unmatched orders are balanced with the utility company at the grid electricity prices. It should be noted that the submitted prices of all traders are bounded between the grid sell (FiT)

and sell (ToU) prices to guarantee the economic benefits in DA market instead of directly trading with the utility company [26].

Algorithm 3: DA market clearing algorithm

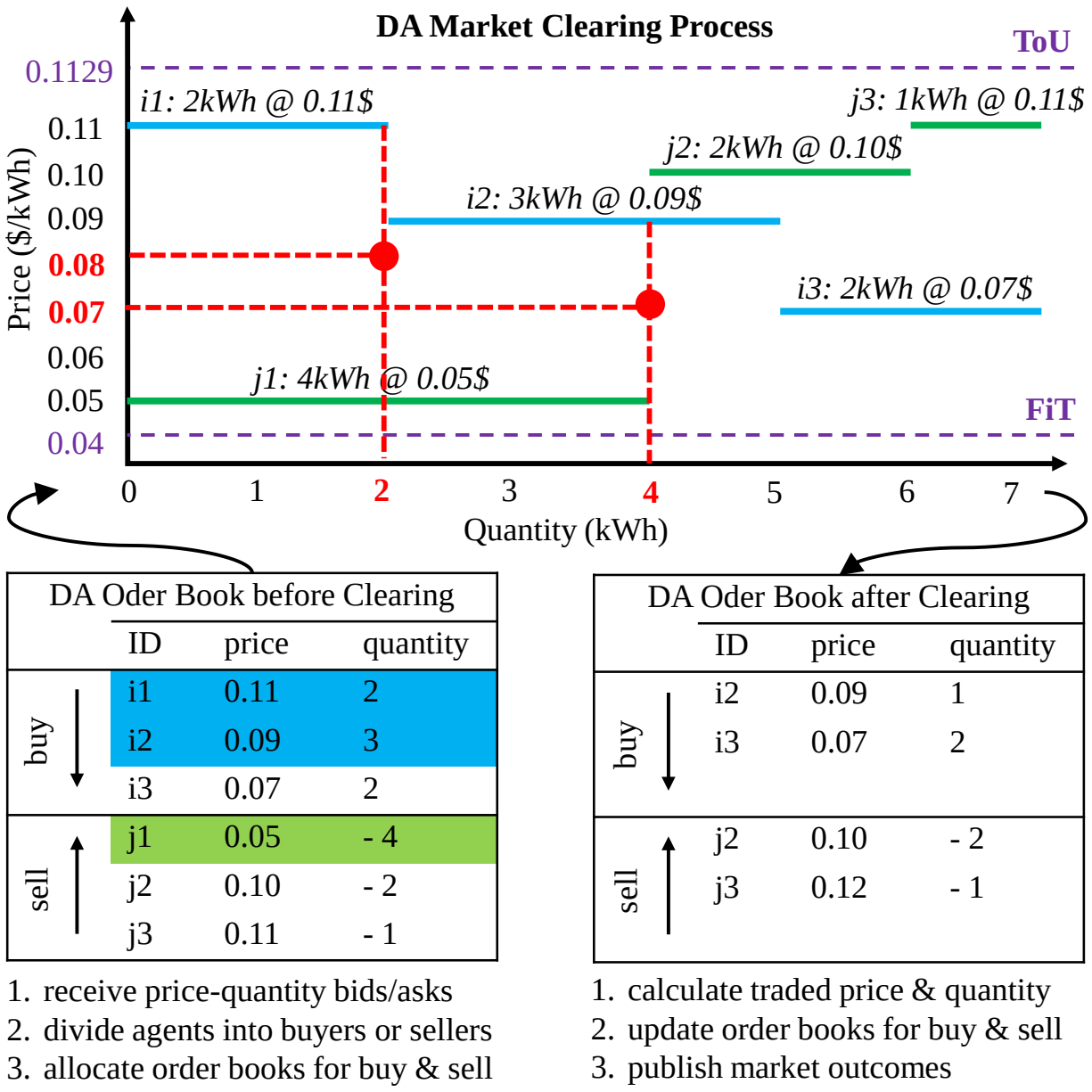
Collect price-quantity bids/offers at auction period t
Allocate order books $k_t^b(b, p_{b,t}, q_{b,t})$ and $k_t^s(s, p_{s,t}, q_{s,t})$ at auction period t
Initialise $b = s = 1$
while $p_{b,t} \geq p_{s,t}$ **do**
    Match the trading energy: q_t^l = min$(q_{b,t}, q_{s,t})$
    Calculate the trading price: $p_t^l = (p_{b,t} + p_{s,t})/2$
    Update buy order book $q_{b,t} \leftarrow q_{b,t} - q_t^l$
    if $q_{b,t} = 0$ **then**
        $b \leftarrow b + 1$
    end
    update sell order book $q_{s,t} \leftarrow q_{s,t} - q_t^l$
    if $q_{s,t} = 0$ **then**
        $s \leftarrow s + 1$
    end
    break if
        $b >$ length of k_t^b or $s >$ length of k_t^s
end
Balance unmatched quantity at FiT (λ_t^s) and ToU (λ_t^b) prices

5.2.4 Double Auction Market Clearing Process

In this section, we provide an numerical example to better illustrate the market clearing principle of a DA market. Taking the scenario in Figure 5.2 as an example. There are six agents participating into DA market.

Order Books

Depending on the sign of the quantity submitted by each agent (positive for buyers, negative for sellers), the six agents are split into three buyers and three sellers. In the left-hand table of Figure 5.2, sellers are arranged according to their order of low-to-high price, while buyers are arranged according to their order of high-to-low price.



DA Oder Book before Clearing

	ID	price	quantity
buy	i1	0.11	2
	i2	0.09	3
	i3	0.07	2
sell	j1	0.05	- 4
	j2	0.10	- 2
	j3	0.11	- 1

1. receive price-quantity bids/asks
2. divide agents into buyers or sellers
3. allocate order books for buy & sell

DA Oder Book after Clearing

	ID	price	quantity
buy	i2	0.09	1
	i3	0.07	2
sell	j2	0.10	- 2
	j3	0.12	- 1

1. calculate traded price & quantity
2. update order books for buy & sell
3. publish market outcomes

Figure 5.2: DA market clearing process.

Market Transactions

We describe and provide all the potential transaction pairs in the provided example like the following subsections because multiple transactions will occur in the market. Each matches one seller and buyer with the same price and quantity.

Transaction 1

The first transaction takes place when the first bid price ($0.11/kWh) is higher than the first ask price ($0.05/kWh). The matching quantity is equal to the minimum amount of buyer $i1$ and seller $j1$ (i.e., 2kWh), and the transaction price is the average of $0.11/kWh and $0.05/kWh (i.e., $0.08/kWh). Buyer $i1$ in this transaction has been completed matched and should be removed,

while buyer $i2$ should be updated at the top of the buy order book. Additionally, seller $j1$'s quantity is reduced to 2kWh, but the sellers' quotes remain unchanged.

Transaction 2

The second transaction takes place since the updated first bid price ($0.09/kWh) is still higher than the first ask price ($0.05/kWh), the matched quantity is equal to the minimum amount of buyer $i2$ and seller $j1$ (i.e., 2kWh) and the transaction price is the average of $0.09/kWh and $0.05/kWh (i.e., $0.07/kWh). Seller $j1$ in transaction 2 is completely matched and should be removed, while seller $j2$ is updated at the top of the sell order book. Furthermore, the quantity of buyer $i2$ is reduced to 1kWh accordingly.

Transaction ends

The transaction in the DA market ends when the current first bid price ($0.09/kWh) is lower than the current first ask price ($0.10/kWh). Finally, the unmatched quantities of traders (right table in Figure 5.2) are balanced by buying/selling from/to the utility company. The unmatched buyers will pay $0.1129/kWh (ToU) and sellers will get $0.04/kWh (FiT) for the remained energy quantity.

Market Outcomes

After all transactions get matched in the DA market, the auctioneer publishes the market clearing outcomes for each participating agents, which are organized in Table 5.1. More specifically, the trading price and quantity for transaction 1 between buyer $i1$ and seller $j1$ are $0.08/kWh and 2kWh, respectively. The cost/revenue for buyer $i1$/seller $j1$ is $0.16. In addition, the trading price and quantity for transaction 2 between buyer $i2$ and seller $j1$ are $0.07/kWh and 2kWh, respectively. The cost/revenue for buyer $i2$/seller $j1$ is $0.14. Finally, the unmatched demand (1kWh for buyer $i2$ & 2kWh for buyer $i3$) are bought at ToU $0.1129/kWh representing the energy cost, while the unmatched generation (2kWh for buyer $j2$ & 1kWh for buyer $j3$) are sold at FiT $0.04 representing the energy revenue.

5.2.5 Energy Coordination as Markov Decision Process

The above-introduced DA market can be formulated as a multi-agent P2P energy trading problem in the form of a finite Partially Observable Markov Decision Process (POMDP) [81] with discrete time steps. The POMDP is then defined with a set of state $\mathcal{S}$ describing the global state of environment $\mathcal{E}$ (DA market), a collection of local observations $\{\mathcal{O}_{1:I}\}$, a collection of action sets $\{\mathcal{A}_{1:I}\}$, a collection of reward functions $\{\mathcal{R}_{1:I}\}$ and a state transition function $\mathcal{T}(s, a_{1:I}, \omega)$, where ω is the environment stochasticity representing uncertain parameters. The time interval between two consecutive time steps is one auction period (Δt = 1 hour). At time step t, each agent i chooses an action $a_{i,t}$ according to its policy $\pi_i(a_{i,t}|o_{i,t})$ conditional on its local observation $o_{i,t}$ and executes this $a_{i,t}$ to the environment $\mathcal{E}$. The environment then moves into the next

Table 5.1: DA Market Outcomes for six agents

	Matched ID	Submitted		Cleared		
		Price	Quantity	Price	Quantity	Cost
Trans.1	$i1$	0.11	2	0.08	2	0.16
	$j1$	0.05	-4	0.08	-2	-0.16
Trans.2	$i2$	0.09	3	0.07	2	0.14
	$j1$	0.05	-2	0.07	-2	-0.14

	Unmatched ID	Price	Quantity	Cost
Trans. end	$i2$	0.1129	1	0.1129
	$i3$	0.1129	2	0.2258
	$j2$	0.04	-2	-0.08
	$j3$	0.04	-1	-0.04

state according to the transition function $\mathcal{T}$. Each agent i obtains the reward $r_{i,t}$ and the next local observation $o_{i,t+1}$. The objective of each agent i is maximising the cumulative discounted reward,

$$R_i = \mathbb{E}_{s\sim\mathcal{T},a_i\sim\pi_i}\left[\sum_{t=0}^{T}\gamma^t r_{i,t}\right] \tag{5.17}$$

where $\gamma \in [0,1)$ is the discount factor and T is the daily horizon of 24 hours. In detail, the components of the POMDP for the proposed P2P energy trading and energy conversion problem are defined as:

Observation

Each MGCC agent i at time step t observes its local observation $o_{i,t}$ that varies for different MEMG categories and can be defined as

$$o_{i,t} = \begin{cases} \left[\lambda_t, L_{i,t}, P^{pv}_{i,t}, E^{es}_{i,t}\right], & \forall i \in \mathcal{I}_{RG}, \forall t \in T \\ \left[\lambda_t, L_{i,t}, P^{pv}_{i,t}, E^{es}_{i,t}\right], & \forall i \in \mathcal{I}_{CG}, \forall t \in T \\ \left[\lambda_t, L_{i,t}, P^{wg}_{i,t}, E^{es}_{i,t}\right], & \forall i \in \mathcal{I}_{IG}, \forall t \in T, \end{cases} \tag{5.18}$$

where the observation $o_{i,t}$ consists of two parts: 1) the exogenous state unaffected by the action includes the sensor data of price signals $\lambda_t = [\lambda^b_t, \lambda^s_t, \lambda^g_t, \lambda^c_t]$ representing the grid electricity buy and sell prices, the gas price, and the carbon price as well as the measured data of consummation loads $L_{i,t} = [P^l_{i,t}, Q^l_{i,t}]$ representing EL and HL, the renewable generation of PV $P^{pv}_{i,t}$ and WG $P^{wg}_{i,t}$; and 2) the endogenous state which serves as the feedback signals of agents' executed action

and represents the system dynamics, including the energy content of EES and TES,

$$E_{i,t}^{es} = \left[E_{i,t}^{ees}, E_{i,t}^{tes}\right]. \tag{5.19}$$

Action

Each MGCC agent i at time step t controls its action $a_{i,t}$ that varies for different MEMG categories and can be defined as

$$a_{i,t} = \begin{cases} \left[a_{i,t}^{p}, a_{i,t}^{ees}, a_{i,t}^{tes}, a_{i,t}^{fc}, a_{i,t}^{gb}\right], & \forall i \in \mathcal{I}_{RG}, \forall t \in T \\ \left[a_{i,t}^{p}, a_{i,t}^{ees}, a_{i,t}^{tes}, a_{i,t}^{ehp}, a_{i,t}^{gb}\right], & \forall i \in \mathcal{I}_{CG}, \forall t \in T \\ \left[a_{i,t}^{p}, a_{i,t}^{ees}, a_{i,t}^{tes}, a_{i,t}^{chp}, a_{i,t}^{gb}\right], & \forall i \in \mathcal{I}_{IG}, \forall t \in T, \end{cases} \tag{5.20}$$

where the action $a_{i,t}$ consists of two parts:

1) the price decision $a_{i,t}^{p} \in [0, 1]$ representing the magnitude of willing price submitted to the DA market as a ratio of FiT and ToU price differentials,

$$p_{i,t} = \lambda_t^s + a_{i,t}^{p} \cdot \left(\lambda_t^b - \lambda_t^s\right). \tag{5.21}$$

2) the energy decisions which comprise of $a_{i,t}^{ees}, a_{i,t}^{tes} \in [-1, 1]$ indicating the mutually exclusive charging (positive) and discharging (negative) power rate of EES and TES as a percentage of their power capacity $[-\overline{P}_i^{ees}, \overline{P}_i^{ees}]$ and $[-\overline{Q}_i^{tes}, \overline{Q}_i^{tes}]$ (EES and TES cannot charge and discharge simultaneously),

$$P_{i,t}^{ees} = P_{i,t}^{eesc} + P_{i,t}^{eesd}, \tag{5.22}$$

$$Q_{i,t}^{tes} = Q_{i,t}^{tesc} + Q_{i,t}^{tesd}. \tag{5.23}$$

The energy decisions also comprise of $a_{i,t}^{fc}, a_t^{gb}, a_t^{ehp}, a_t^{chp} \in [0, 1]$ indicating the magnitude of power schedules as a percentage of their power capacity for FC $G_{i,t}^{fc} \in [0, \overline{G}_i^{fc}]$, GB $G_{i,t}^{gb} \in [0, \overline{G}_i^{gb}]$, EHP $P_{i,t}^{ehp} \in [0, \overline{P}_i^{ehp}]$, and CHP $G_{i,t}^{chp} \in [0, \overline{G}_i^{chp}]$, respectively.

State Transition

The state transition from time step t to $t + 1$ is governed by

$$s_{t+1} = \mathcal{T}(s_t, a_{1:I,t}, \omega_t), \tag{5.24}$$

influenced by the combination of environment state s_t, all agents' actions $a_{1:I,t}$, and environment stochasticity ω_t. In the examined problem, this corresponds to the exogenous states,

$$\omega_t = \left[L_{1:I,t}, P_{1:I,t}^{pv}, P_{1:I,t}^{wg}, \lambda_t\right], \tag{5.25}$$

that are decoupled from the agents' actions and are characterised by inherent variability. In machine learning area, RL translates this problem to a data-driven approach that learns the stochastic characteristics directly from the data sources [81].

By contrast, the state transitions of endogenous states $S_{i,t}^{ees}, S_{i,t}^{tes}$ are determined by actions $a_{i,t}^{ees}, a_{i,t}^{tes}$. Given EES as an example, the mutually quantities $P_{i,t}^{eesc}, P_{t}^{eesd}$ are managed by action $a_{i,t}^{ees}$, and are also restricted by its technical parameters of the minimum/maximum energy level $\underline{E}_i^{ees}, \overline{E}_i^{ees}$, and the charging/discharging efficiency $\eta_i^{eesc}, \eta_i^{eesd}$, which can be expressed as

$$P_{i,t}^{eesc} = [\min(a_{i,t}^{ees}\overline{P}_i^{ees}, (\overline{E}_i^{ees} - E_{i,t}^{ees})/(\eta_i^{eesc}\Delta t)]^{+}, \tag{5.26}$$

$$P_{i,t}^{eesd} = [\max(a_{i,t}^{ees}\overline{P}_i^{ees}, (\underline{E}_i^{ees} - E_{i,t}^{ees})\eta_i^{eesd}/\Delta t]^{-}, \tag{5.27}$$

where $[\cdot]^{+/-} = \max / \min\{\cdot, 0\}$. Given the charging and discharging power $P_{i,t}^{eesc}, P_{i,t}^{eesd}$ and efficiency $\eta_i^{eesc}, \eta_i^{eesd}$, the state transition of $E_{i,t}^{ees}$ from t to $t+1$ can be expressed as

$$E_{i,t+1}^{ees} = E_{i,t}^{ees} + P_{i,t}^{eesc} \cdot \Delta t \cdot \eta_i^{eesc} + P_{i,t}^{eesd} \cdot \Delta t / \eta_i^{eesd}. \tag{5.28}$$

Then, the charging and discharging power $Q_{i,t}^{tesc}, Q_{i,t}^{tesd}$ as well as the state transition $E_{i,t}^{tes}$ of TES can be derived in the similar manner as the EES model (5.26)-(5.28).

To this end, the electricity quantity $q_{i,t}$ submitted to DA market of each agent i at time step t can be expressed as the summation of its individual electric demand and supply power, where the positive value represents the electricity demand to buy while the negative value represents the electricity generation to sell in the DA market

$$q_{i,t} = \begin{cases} \left(P_{i,t}^{l} - P_{i,t}^{pv} - P_{i,t}^{fc} + P_{i,t}^{ees}\right) \cdot \Delta t & \forall i \in \mathcal{I}_{RG}, \forall t \in T \\ \left(P_{i,t}^{l} - P_{i,t}^{pv} + P_{i,t}^{ehp} + P_{i,t}^{ees}\right) \cdot \Delta t & \forall i \in \mathcal{I}_{CG}, \forall t \in T \\ \left(P_{i,t}^{l} - P_{i,t}^{wd} - P_{i,t}^{chp} + P_{i,t}^{ees}\right) \cdot \Delta t & \forall i \in \mathcal{I}_{IG}, \forall t \in T. \end{cases} \tag{5.29}$$

After collecting the price-quantity bids $(p_{i,t}, q_{i,t})$ from all participating agents, the auctioneer allocates the order books $k_t^b(i, p_{i,t}, q_{i,t}), \forall i \in \mathcal{B}$ and $k_t^s(i, p_{i,t}, q_{i,t}), \forall i \in \mathcal{S}$, clears the DA market (Algorithm 3) and publishes the market outcomes $[p_{1:I,t}^l, q_{1:I,t}^l, q_{1:I,t}^g, k_t^b, k_t^s]$, which comprises: 1) the local information of cleared trading price $p_{i,t}^l$, cleared trading quantity $q_{i,t}^l$, the remaining/unmatched quantity balanced with the utility company $q_{i,t}^g$ for each agent i; and 2) the public market information of updated order books k_t^b, k_t^s.

Reward Function

The reward function for each agent i at time step t is designed as two parts: 1) the energy and environment costs; and 2) the penalty imposed to avoid the constraint violations of MES

operation model. Specifically, for these agents who are successfully matched in the DA market will receive the cleared local trading price $p^l_{i,t}$ and quantity $q^l_{i,t}$, then each agent i can calculate its corresponding electricity cost/revenue in the DA market, and the remaining/unmatched quantity $q^g_{i,t}$ will be bought or sold with the utility company at ToU λ^b_t or FiT λ^s_t. For these agents who are unsuccessfully matched in the DA market, their quantity $q^g_{i,t} = q_{i,t}$ (i.e., $q^l_{i,t} = 0$) will be directly traded at λ^b_t or λ^s_t. As a result, the reward term corresponding to the electricity cost for each agent i at time step t can be formulated as

$$r^e_{i,t} = -(p^l_{i,t} q^l_{i,t} \cdot \mathbb{1}_{i,t} + \lambda^b_t [q^g_{i,t}]^+ + \lambda^s_t [q^g_{i,t}]^-), \tag{5.30}$$

where the indicator $\mathbb{1}_{i,t} = 1$ if $i \in \mathcal{B}$ and $\mathbb{1}_{i,t} = -1$ if $i \in \mathcal{S}$. Furthermore, the reward terms corresponding to the gas cost and environment cost out of the DA market for each agent i at time step t can be respectively formulated as

$$r^g_{i,t} = -\lambda^g_t G^g_{i,t} \Delta t, \tag{5.31}$$

$$r^c_{i,t} = -\lambda^c_t E^c_{i,t}, \tag{5.32}$$

where the gas quantity purchased from the natural gas grid varies for three kinds of MEMGs:

$$G^g_{i,t} = G^{gb}_{i,t}, \forall i \in \mathcal{I}_{RG} \cup \mathcal{I}_{CG}, \tag{5.33}$$

$$G^g_{i,t} = G^{chp}_{i,t} + G^{gb}_t, \forall i \in \mathcal{I}_{IG}. \tag{5.34}$$

Note that the electricity demand and supply in each MEMG shown in (5.29) can always be balanced through the internal system together with the external DA market at each time step. However, the heat demand and supply may not be balanced, since extra heat cannot sell back to the grid. More specifically, the power schedules of components (i.e., FC, GB, EHP, CHP, TES) controlled by actions only respect their individual operation models (e.g., power capacity). However, they do not make sure that the heat demand and supply are always balanced. The main factor leading to this issue is that the action selections in the RL algorithm for different dimensions are independent, decoupling the correlation in the optimization-based approach. To adequately account for such operation constraints of heat demand-supply balance, we introduce a penalty term $r^p_{i,t}$ shown in (5.35) for each agent in the reward function, which penalizes the extent of violation of the heat demand-supply balance constraint, with κ denoting a large (negative) penalty weighting factor to ensure its feasibility.

$$r^p_{i,t} = \begin{cases} \kappa \cdot \left| Q^l_{i,t} - Q^{fc}_{i,t} - Q^{gb}_{i,t} + Q^{tes}_{i,t} \right|, & \forall i \in \mathcal{I}_{RG},\ \forall t \in T. \\ \kappa \cdot \left| Q^l_{i,t} - Q^{ehp}_{i,t} - Q^{gb}_{i,t} + Q^{tes}_{i,t} \right|, & \forall i \in \mathcal{I}_{CG},\ \forall t \in T. \\ \kappa \cdot \left| Q^l_{i,t} - Q^{chp}_{i,t} - Q^{gb}_{i,t} + Q^{tes}_{i,t} \right|, & \forall i \in \mathcal{I}_{IG},\ \forall t \in T. \end{cases} \tag{5.35}$$

Thus, the final reward function $r_{i,t}$ of each MGCC agent i at time step t can be expressed as

$$r_{i,t} = r^e_{i,t} + r^g_{i,t} + r^c_{i,t} + r^p_{i,t}, \ \forall i \in I, \ \forall t \in T. \tag{5.36}$$

5.3 Proposed Multi-Agent Reinforcement Learning Method

To solve the POMDP defined above, we propose a novel MADRL method named DA-MATD3 with its training schematic being shown in Figure 5.3. DA-MATD3 derives three concrete implementation details that are insightful and particularly critical to our proposed MEMG P2P energy trading and energy conversion problem: 1) learning an abstracted Q-value function for each agent through the DA market public order books to protect the private information of each MEMG; and 2) forming an actor-critic architecture to handle the high-dimensional continuous state and action spaces of the MEMGs; and 3) taking advantage of double critic networks in Twin delayed DDPG (TD3) algorithm [17] to address the Q-value overestimation problem, thereby stabilising the training performance.

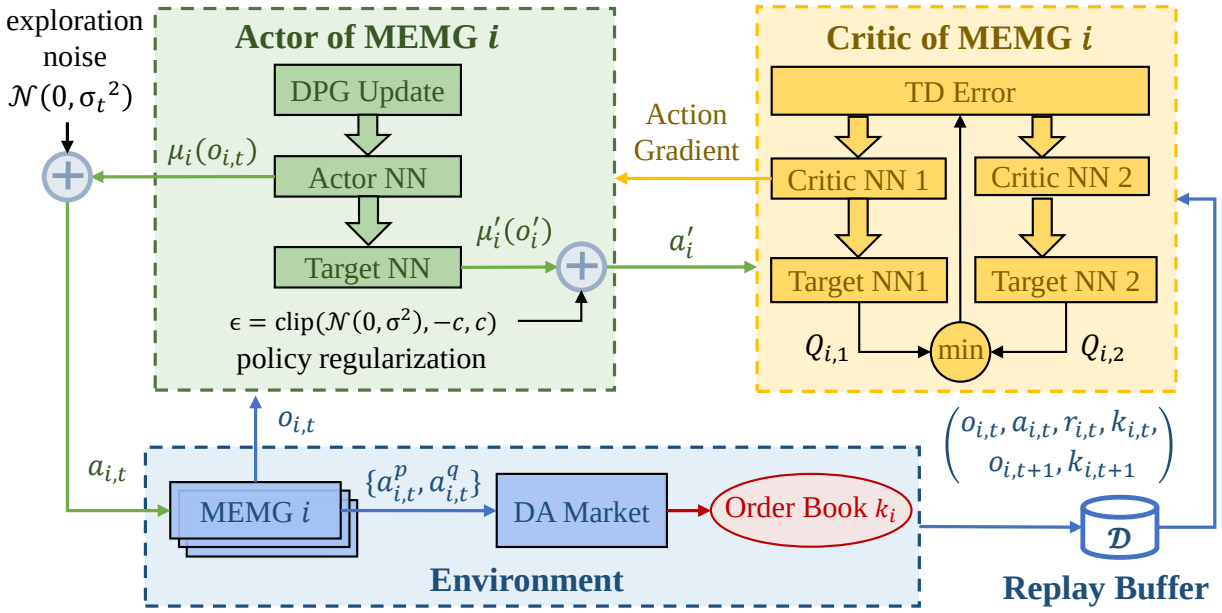


Figure 5.3: Schematic of the proposed DA-MATD3 method for energy coordination.

5.3.1 Multi-Agent Twin Delayed Deep Deterministic Policy Gradient

MATD3 [126], an extension of TD3 to multi-agent setup, is an algorithm that addresses the stability concern that occurred in conventional MADDPG by three key features: 1) using a pair of critics that estimate the current Q-value via a separate target value function; 2) updating the

policy less frequently (delayed update) than the Q-value function; and 3) smoothing the target policy by using a (noise) regularisation technique.

Twin critic networks

The overestimation bias in the conventional MADDPG method has been discussed in [126]. Inspired by the technique in Double Q-learning [146] using a separate target Q-value function to estimate the current Q-value, thus reducing the bias, we introduce for each agent i two separate online critic networks $(Q_{i,1}, Q_{i,2})$ parameterised by $\theta_{i,1}, \theta_{i,2}$, along with two target critic networks $(Q'_{i,1}, Q'_{i,2})$ parameterised by $\theta'_{i,1}, \theta'_{i,2}$. Then the two target values used to update the critic can be written as

$$y_{i,1} = r_i + \gamma Q_{i,1}\big(o'_i, \mu'_i(o'_i), k'_i\big), \tag{5.37}$$

$$y_{i,2} = r_i + \gamma Q_{i,2}\big(o'_i, \mu'_i(o'_i), k'_i\big). \tag{5.38}$$

However, the values of $Q_{i,1}$ and $Q_{i,2}$ cannot be equal, and it is inevitable that the high value may be overestimated. Therefore, we make a slight change on the basis of Double Q-learning, and take the minimum value between these two estimates to get the target Q-value for each agent i.

$$y_i = r_i + \gamma \min_{k=1,2} Q_{i,k}\big(o'_i, \mu'_i(o'_i), k'_i\big). \tag{5.39}$$

With this improvement, MATD3 can simultaneously train two critic networks and pick the minimum value of them, thus alleviating the overestimation phenomenon.

Delayed policy updates

Another potential failure in MADDPG is the variance, which generates noisy gradients during the policy update, thus slowing down the update speed and leading to poor performance [126]. Similar to MADDPG, MATD3 also introduces the target networks to achieve stability in the training process. Apart from this, the algorithm also proposes to delay the actor network update until the critic network is updated after a fixed number of time steps. In this setting, the updates of actor and critic networks are decoupled, i.e., the actor network is updated at a lower frequency than the critic network, to first achieve an accurate Q-value before it is used to update the policy. This less frequent policy update will have a Q-value estimate with lower variance, resulting in better policy performance.

Target policy smoothing regularization

The final technique of MATD3 is smoothing the target policy. Deterministic policies tend to produce high variance of the target when updating the critic, this is caused by overfitting to narrow peaks in the Q-value estimate [126]. MATD3 reduces this variance by adding a clipped

Gaussian noise to the actions of all agents in the critic update:

$$a_i' = \mu_i' \cdot (o_i') + \epsilon, \tag{5.40}$$

where

$$\epsilon = \text{clip}\left(\mathcal{N}\left(0, \sigma\right), -c, c\right). \tag{5.41}$$

This serves as a regularization, such that all actions within this small area have similar Q-values, thereby reducing the variance in the associated estimations. The complete target for the critic resolves to

$$y_i = r_i + \gamma \min_{k=1,2} Q_{i,k}\big(o_i', \mu_i'(o_i') + \epsilon, k_i'\big). \tag{5.42}$$

5.3.2 Abstracted Q-value function

As discussed in Section 5.3.2, it is challenging to directly acquire the local observations and actions by other agents in our proposed problem since the MEMGs are not willing to share their energy portfolios, technical parameters, and energy usage behaviours. This work thus assumes that the agents can use the public order books that epitomise the key information of DA market (thereby abstracting all agents' price-quantity bid information) in the centralised training process. This substantial improvement protects the privacy of all agents. To this effect, we approximate the centralised Q-value as

$$Q_i(o_{1:I}, a_{1:I}) \approx Q_i(o_i, a_i, k_i), \tag{5.43}$$

where

$$k_i = \left\{k_j^b, k_j^s, \ \forall j \in \mathcal{I} \setminus \{i\}\right\} \tag{5.44}$$

denotes combination of buy and sell order books of all agents other than agent i in the DA market. k_i is an embedded function of order books k_j^b, k_j^s that not only abstracts all other agents' observations (e.g., $E_j^l, P_j^{pv}, P_j^{wd}$) as well as actions of the price bids a_j^p and the quantity bids resulting from their energy decisions (e.g., $a_j^{ees}, a_j^{fc}, a_j^{ehp}, a_j^{chp}$), but also displays the DA market dynamics of local trading activities. As a result, this combination provides a good approximation of agents' observations and actions as well as the DA market dynamics. Incorporating k_i into the critic estimation, each agent can make acquainted decisions on the basis of the impact of other agents' actions, albeit not knowing their energy portfolios and usage activities, protecting the privacy of each MEMG.

5.3.3 Nash equilibrium

In our energy coordination problem, the optimal performance of each MEMG is controlled not only by its own energy trading and energy conversion policies, but also the choices of all other

participants in the DA market. It is worthwhile to discuss the rationality of convergence of the proposed MADRL method. The most common solution concept, Nash equilibrium (NE), is defined as follows [163].

A NE of the Markov game $\left(\mathcal{N}, \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, \mathcal{P}, \{R^i\}_{i\in\mathcal{N}}, \gamma\right)$ is a joint policy $\pi^* = \left(\pi^{1,*}, \cdots, \pi^{N,*}\right)$, such that for any $o \in \mathcal{O}$ and $i \in \mathcal{N}$,

$$Q^i_{\pi^{i,*},\pi^{-i,*}}(o,a) \geq Q^i_{\pi^i,\pi^{-i,*}}(o,a), \quad \text{for any } \pi^i. \tag{5.45}$$

As a result of NE, there is no incentive for any agent to deviate from the equilibrium point π^*. Thus, for any agent $i \in \mathcal{N}$, policy $\pi^{i,*}$ is the best response of $\pi^{-i,*}$. As a standard learning objective in MADRL, NE is always present for finite-space infinite-horizon discounted Markov games, but it may not be unique [164]. Therefore, in most MADRL algorithms, this equilibrium point is supposed to be reached if it exists.

5.3.4 Training process

DA-MATD3 is an off-policy MADRL method that requires the past experiences to update the networks. To this end, an experience replay buffer $\mathcal{D}_i$ is employed for each agent i. The buffer is a cache storing the past experiences of agent i acquired from the environment (an experience is a transition tuple $(o_{i,t}, a_{i,t}, r_{i,t}, k_{i,t}, o_{i,t+1}, k_{i,t+1})$. For each time step t, we sample uniformly a minibatch of N experiences from each agent's corresponding replay buffer $\{(o_i^n, a_i^n, r_i^n, k_i^n, o_i^{n+1}, k_i^{n+1})\}_{n=1}^N \sim \mathcal{D}_i$ to compute the mean-squared TD error of two online critic networks as

$$\mathcal{L}(\theta_{i,1}) = \frac{1}{N}\sum_{n=1}^{N}\left[\left(y_i^n - Q_{i,1}(o_i^n, a_i^n, k_i^n)\right)^2\right], \tag{5.46}$$

$$\mathcal{L}(\theta_{i,2}) = \frac{1}{N}\sum_{n=1}^{N}\left[\left(y_i^n - Q_{i,2}(o_i^n, a_i^n, k_i^n)\right)^2\right], \tag{5.47}$$

where

$$y_i^n = r_i^n + \gamma \min_{k=1,2} Q_{i,k}\left(o_i^n, \mu_i'(o_i^{n+1}) + \epsilon, k_i^{n+1}\right). \tag{5.48}$$

The online actor network employs the delayed update after d critic updates, its policy gradient can be expressed as

$$\nabla_{\phi_i} J(\mu_i) = \frac{1}{N}\sum_{n=1}^{N}\left[\nabla_{\phi_i}\mu_i(o_i^n)\nabla_{a_i^n} Q_{i,1}(o_i^n, a_i^n, k_i^n)|_{a_i^n=\mu(o_i^n)}\right]. \tag{5.49}$$

The target networks of two critic and one actor are also employed as the delayed updates after d critic updates.

$$\theta_{i,1}' \leftarrow \tau\cdot\theta_{i,1} + (1-\tau)\cdot\theta_{i,1}', \tag{5.50}$$

Figure 5.4: Flowchart of implementing our proposed DA-MATD3 method for each agent.

$$\theta'_{i,2} \leftarrow \tau \cdot \theta_{i,2} + (1 - \tau) \cdot \theta'_{i,2}, \tag{5.51}$$

$$\phi'_i \leftarrow \tau \cdot \phi_i + (1 - \tau) \cdot \phi'_i, \tag{5.52}$$

where τ is the soft update rate for their target networks. Moreover, in order to help the agents explore the environment and acquire more valuable experiences, we add a random Gaussian

noise $\mathcal{N}(0, \sigma_t^2)$ to the online policy $\mu_i(o_{i,t})$ of each agent i, constructing an exploration policy

$$\hat{\mu}_i(o_{i,t}) = \mu_i(o_{i,t}) + \mathcal{N}(0, \sigma_t^2). \tag{5.53}$$

Finally, the overall training process of the proposed DA-MATD3 is summarised in Algorithm 4. The flowchart of implementing our proposed DA-MATD3 solution method for each agent is shown in Figure 5.4.

Algorithm 4: DA-MATD3 for I agents

Initialise weights $\theta_{i,1}, \theta_{i,2}, \phi_i$ for the online networks and copy them to the target network weights $\theta'_{i,1}, \theta'_{i,2}, \phi'_i$ for each agent i
Initialise replay buffer $\mathcal{D}_i$ for each agent i
for *episode (i.e. trading day) = 1 to* M **do**
    Initialise the environment $\mathcal{E}$ and Gaussian noise $\mathcal{N}(0, \sigma_t^2)$
    for *time step (i.e. 1 hour)* $t = 1$ *to* T **do**
        For agent i, select action $a_{i,t} = \hat{\mu}_i(o_{i,t})$ in (5.53)
        Execute actions $a_{1:N,t}$ to the DA market, then observe reward $r_{i,t}$, next observation $o_{i,t+1}$ and order books $k_{i,t+1}$
        For agent i, store $(o_{i,t}, a_{i,t}, r_{i,t}, k_{i,t}, o_{i,t+1}, k_{i,t+1})$ in $\mathcal{D}_i$
        Update local observations for next time step $o_{i,t} \leftarrow o_{i,t+1}$
        for *agent* $i = 1$ *to* I **do**
            Sample uniformly a mini-batch of N experiences $(o_i^n, a_i^n, r_i^n, k_i^n, o_i^{n+1}, k_i^{n+1})$ from $\mathcal{D}_i$
            Compute critic target value in (5.48)
            Update two online critic networks in (5.46), (5.47)
            if t *mod* $d = 0$ **then**
                Update online actor network in (5.49)
                Update parameters of target networks in (5.50)-(5.52)
            end
        end
    end
end

Table 5.2: Technical Parameters of Controllable Components

Component	Capacity (kW/kWh)	Energy Efficiency
EES	$\underline{E}^{ees}, \overline{E}^{ees}$ = 40, 400 $\overline{P}^{ees}$ = 100	η^{eesc}, η^{eesd} = 0.95
TES	$\underline{E}^{tes}, \overline{E}^{tes}$ = 45, 450 $\overline{Q}^{tes}$ = 90	η^{tesc}, η^{tesd} = 0.9
GB	$\overline{Q}^{gb}$ = 500	η^{gb} = 0.8
EHP	$\overline{Q}^{ehp}$ = 400	η^{ehp} = 3
CHP	$\overline{P}^{chp}$ = 200, $\overline{Q}^{chp}$ = 300	$\eta^{chp,e}, \eta^{chp,q}$ = 0.3, 0.45
FC	$\overline{P}^{fc}$ = 300, $\overline{Q}^{fc}$ = 550	$\eta^{fc,e}, \eta^{fc,q}$ = 0.3, 0.55

5.4 Case Studies

5.4.1 Experimental Setup and Implementation

Experiment Setup

The experiment of this MEMG P2P energy trading and energy conversion problem is evaluated on real-world datasets with an hourly resolution of EL, HL, PV, and WG, which are recorded from Open Energy Data Initiative [149] and RWTH Aachen University [150]. We collect the corresponding electric & heat loads and PV & wind power generations of residential, commercial, and industrial users with hourly resolution for our experiments. Then, these energy users can be classified and aggregated into three MEMGs, where their load and generation profiles are plotted in Figure 5.5, 5.6 and 5.7. To further account for the uncertainties, we add the Gaussian noise (zero mean and 5% std) to the original 1-month collected data as the train set, while using the original collected data as the test set.

The controllable components of three MEMGs are varied and defined above, where their operating parameters are derived from [165] and are presented in Table 5.2. ToU tariff [166] selected as the grid electricity buy price varying for the time is presented in Table 5.3, while FiT as the grid electricity sell price, natural gas price, and carbon price are flat over the day at 0.04 \$/kWh [167], 0.0338 \$/kWh [151] and 0.0316 \$/kg [152], respectively. The averaged carbon emission of using natural gas is 0.245 kg/kWh [152].

Benchmarks

We compare the proposed DA-MATD3 method with the conventional ZI strategy and three state-of-the-art MADRL methods of IDDPG, MADDPG, and MATD3. To further evaluate the benefit

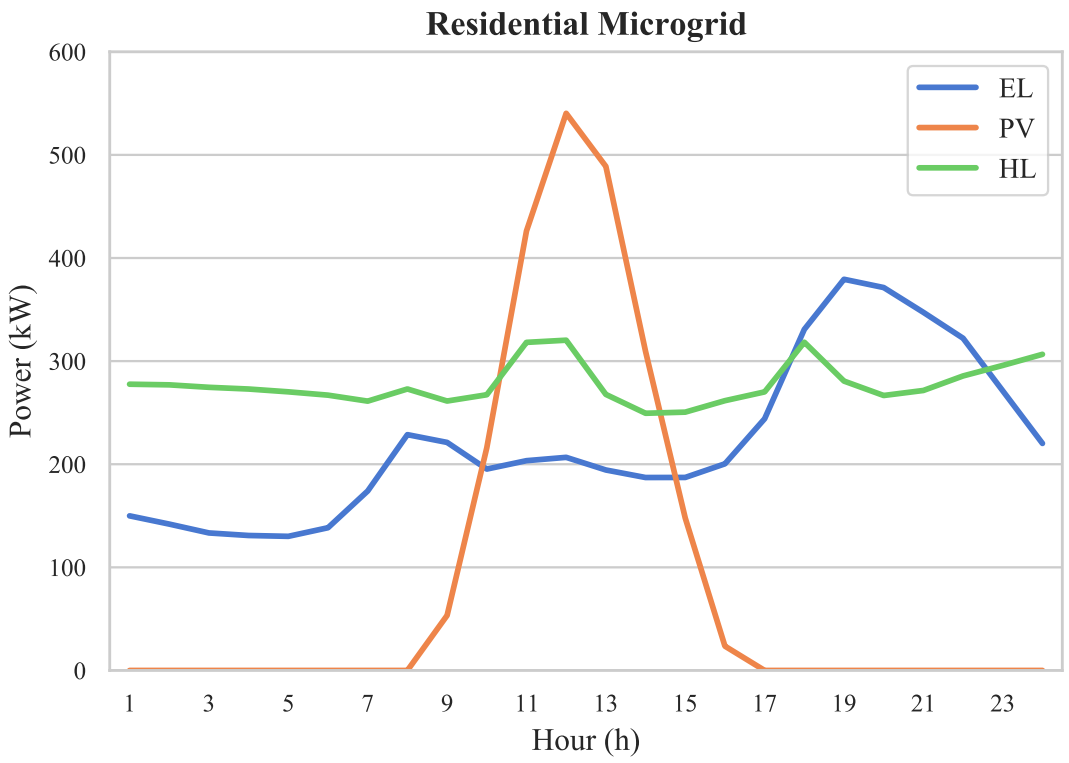


Figure 5.5: Demand and generation profiles Of the residential MEMG.

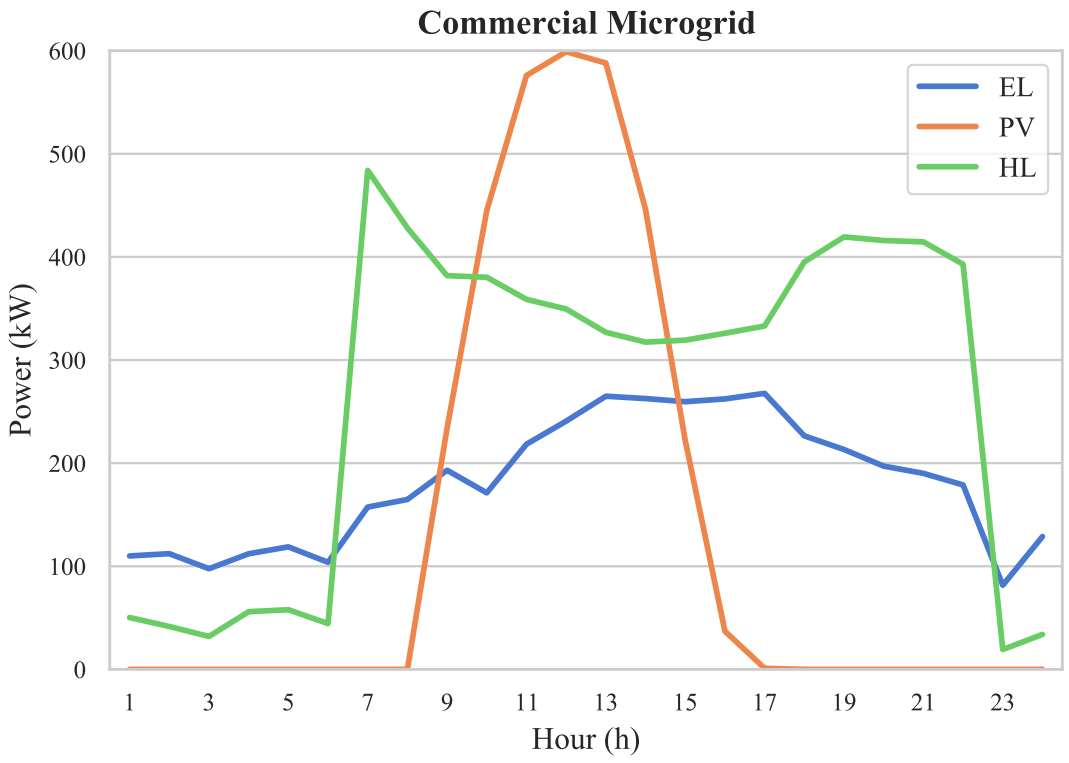


Figure 5.6: Demand and generation profiles Of the commercial MEMG.

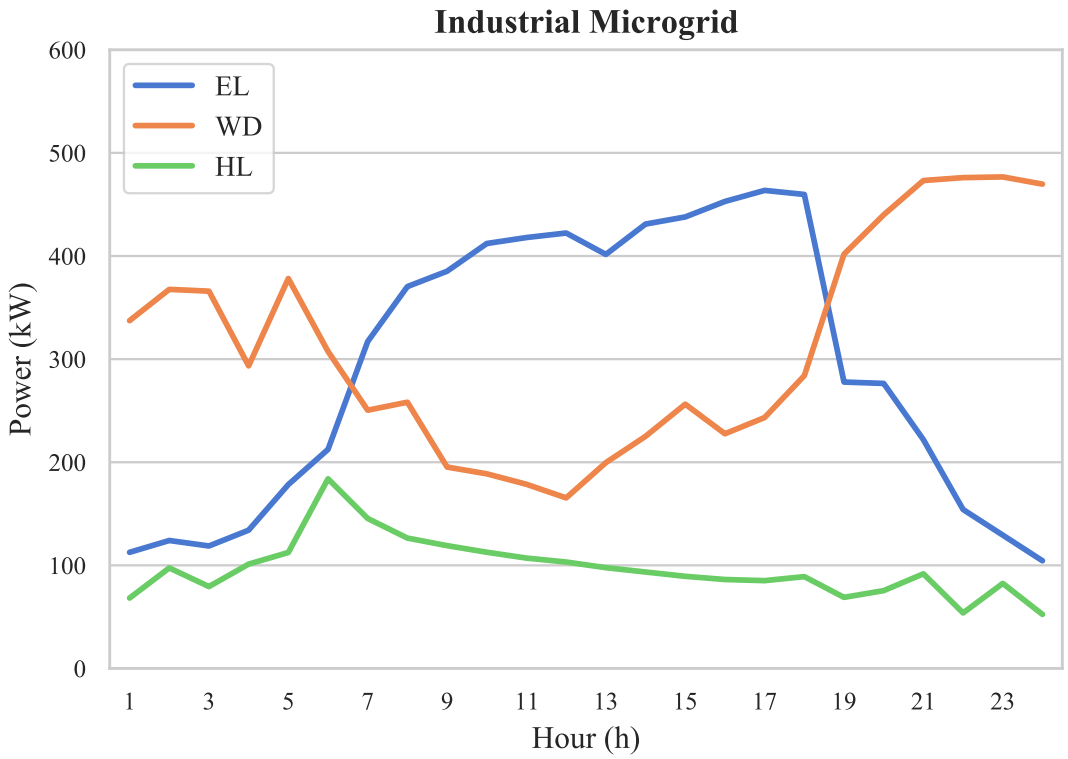


Figure 5.7: Demand and generation profiles Of the industrial MEMG.

Table 5.3: ToU Tariff Structure and Values

Structure	Off-peak	Peak
Time	20:01-17:00 (next day)	17:01-20:00
Price ($/kWh)	0.1129	0.2499

of P2P energy trading in the DA market, we benchmark the performance against one scenario that each MGCC agent trades independently with the utility company using DDPG without P2P energy trading (UDDPG).

Zero Intelligence (ZI) policy is the most common strategy adopted in DA market [121]. For price strategies, ZI agents are subject to a budget constraint (the maximum of grid buy price λ_t^b and the minimum of grid sell price λ_t^s) which forbids the trader to buy or sell at a loss. Then, buyers and sellers of MEMGs select their price bid strategies uniformly at random values between the price limits ($p_{i,t} \sim \mathcal{U}(\lambda_t^s, \lambda_t^b)$). For quantity bid strategies $q_{i,t}$, ZI agents have to run a self-optimization day-ahead problem for their planned energy quantities submitted to the DA market. Specifically, each agent (MGCC) optimizes the energy schedules of all controllable components within its MES by minimizing the cost of daily energy bills and carbon emissions, given the forecasted information of its fixed loads, renewable generations, grid and carbon price signals, as well as the accurate mathematical models and technical parameters of all controllable

components.

Implementations and Hyperparameter Selections

In all of our experiments, we use Adam optimizer for both actor and critic networks with the same learning rate $\alpha = 10^{-3}$. The sizes of replay buffer $\mathcal{D}$ and batch N are 10^5 and 10^2, respectively. We employ $\tau = 10^{-2}$ as the soft update rate and a discount rate $\gamma = 0.9$. The delayed step $d = 2$ for MATD3 and DA-MATD3. For all networks, we use MLPs with two hidden layers with 400 and 300 units, respectively. The sigmoid activation function is used as the actor outputs. The outputs are then scaled linearly to their individual action space. For all examined methods, we run 5×10^3 episodes to evaluate their training performance with the same random seed for the environment and model initialization. The values of the hyperparameters α, τ, d were set based on the original MATD3 [126] paper. The grid search function [168] was used to determine the value of hyperparameter γ to obtain the best performance.

For the RL exploration, we add a Gaussian noise processed to the online policy, the standard deviation decreases exponentially from 2 to 0 within the first 2,000 episodes and stays unchanged until to end in all experiments. The examined DRL methods have been implemented using Tensorflow 2.20 in Python. The case studies have been carried out on a computer with a 8-core 3.80 GHz AMD Ryzen™ 7 5800X CPU and 32 GB of RAM.

5.4.2 Performance Evaluation

This section compares the training performance of five examined MADRL methods and analyses the energy schedules of 3 MEMGs for DA-MATD3 method. Specifically, Figure 5.8, 5.9 and 5.10 illustrate the convergence curve of the episodic reward of 3 MEMGs for different MADRL methods, where the solid lines and the shaded areas respectively depict the moving average over 50 episodes and the oscillations of the reward during the training process. The converged perfor- mance of mean and std of 3 MEMGs' aggregated reward are also compared in Fig. 4. Furthermore, their energy (electricity and gas) costs and carbon emissions (including the community) at convergence are also presented in Table 5.4 for comparison.

Our first observation is that all five methods show an upward trend, and their policies are being improved, even for the UDDPG method converging to the lowest reward (particular high energy cost in Table 5.4), without considering the P2P energy trading benefits. On the other hand, IDDPG, the most straightforward MADRL method, exhibits the highest oscillation and unstable learning behavior, ultimately failing to reach an optimal policy (the highest carbon emission). As discussed in Section 5.3.2, this is because IDDPG focuses on local information while ignoring the others' behaviors, rendering the environment dynamics non-stationary. As such, MADDPG and MATD3 with centralized training can effectively mitigate such non-stationarity issues and exhibit superior training performance. Furthermore, MATD3 owing to its double

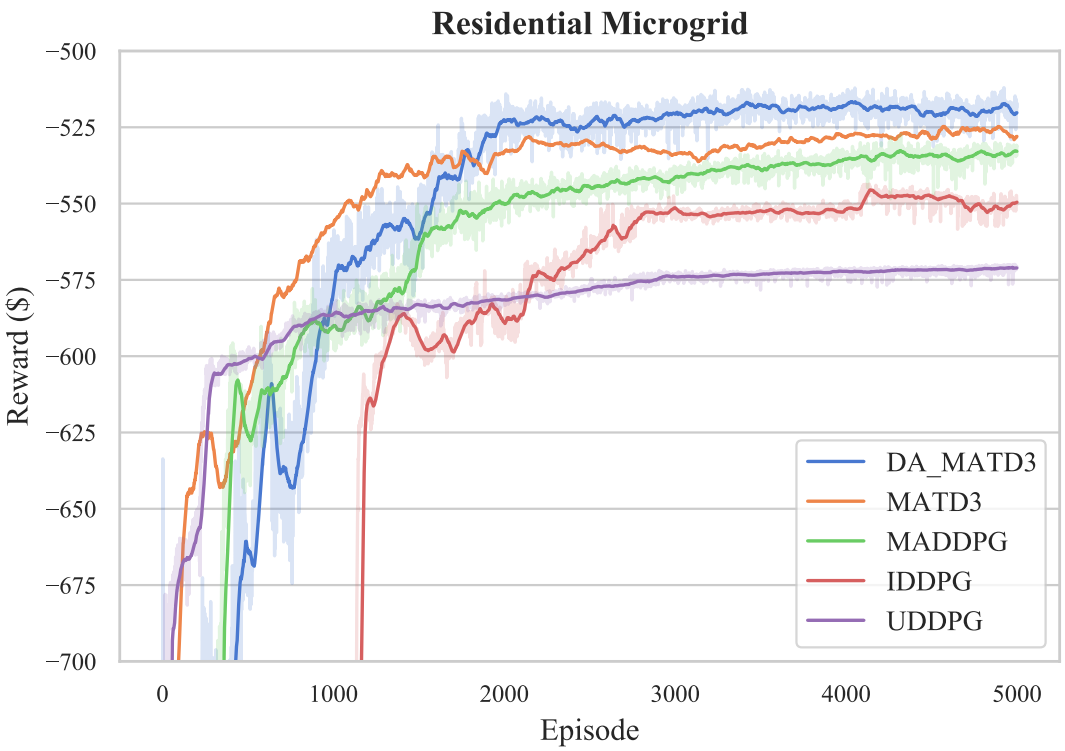


Figure 5.8: Learning curves for the residential MEMG energy coordination under different MADRL methods. Curves are smoothed for visual clarity.

critic networks (more accurate Q-value estimation) can achieve a higher reward with regard to MADDPG. However, both methods suffer from the privacy issue requiring all others' local observations and actions for the centralized critic. To this end, our proposed DA-MATD3 learns the DA market dynamics directly by abstracting the others' observations and actions through the DA market public order books for each agent's critic. The numerical results show that DA-MATD3 learns a very similar performance with MATD3, despite a slight decrease in reward for Industrial MEMG (Fig. 5.10). In relative terms, the proposed DA-MATD3 achieves for the community 3.59% / 1.83% lower energy cost and carbon emission (Table 5.4) over MADDPG / MATD3, respectively.

5.4.3 Analysis of Energy Conversion

To further validate the learned policies of DA-MATD3 for the test set, we provide the energy conversion schedules of 3 MEMGs for both electric and heat supplies in Figure 5.11-5.16. Residential MEMG features abundant PV production during mid-day hours and high EL peaks during night hours as well as a relatively flat HL profile. As its high combined electricity and heating generation efficiencies, FC is learned to supply both EL and HL over the day, apart from the mid-day with PV sources. Furthermore, the MGCC learns to use the storage (EES and TES) flexibility to charge power when energy prices are low or PV is abundant, and discharge

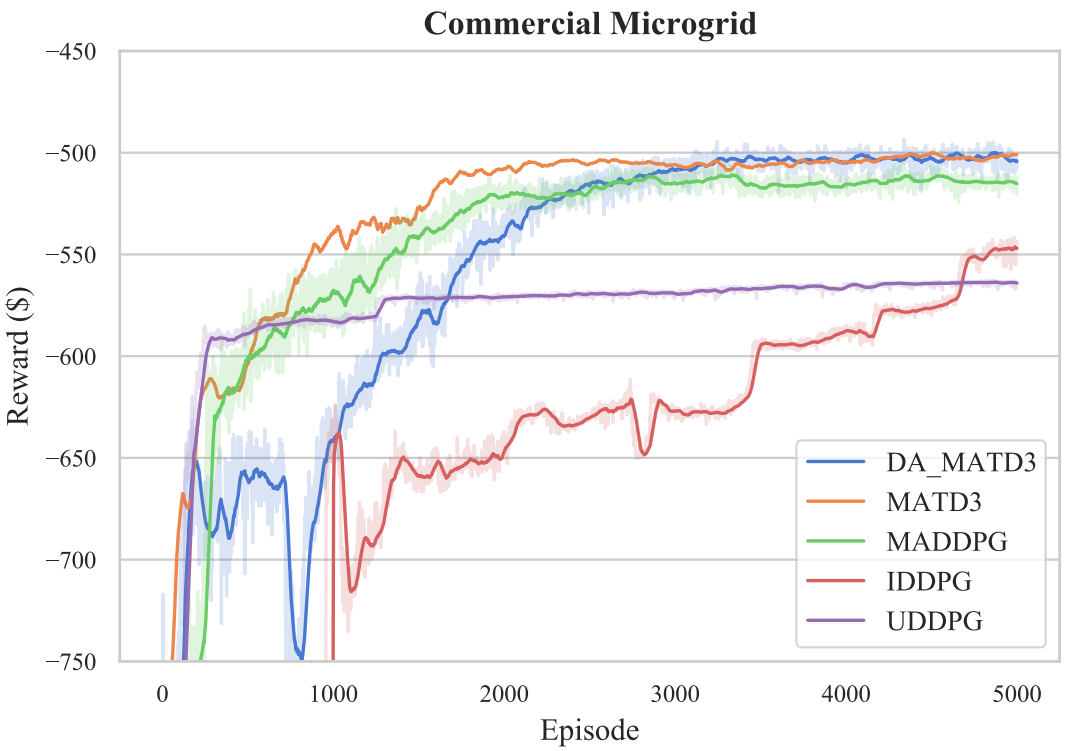


Figure 5.9: Learning curves for the commercial MEMG energy coordination under different MADRL methods. Curves are smoothed for visual clarity.

Table 5.4: Daily Energy Cost and Carbon Emission of 3 MEMGs and Community under Different MADRL Methods

Method	Energy Cost ($)				Carbon Emission (kg)			
MEMGs	RG	CG	IG	Total	RG	CG	IG	Total
UDDPG	477	512	162	1,151	2,806	1,299	1,112	5,217
IDDPG	447	492	40	979	2,914	1,278	1,187	5,379
MADDPG	434	469	-12	891	3,033	1,229	1,064	5,326
MATD3	419	444	-12	851	3,005	1,106	1,144	5,255
DA-MATD3	419	464	-2	881	2,940	951	1,222	5,113

power when the energy price is high or HL is at the peak. Finally, GB is a backup component to supply HL when FC is not in use. Similar to Residential MEMG, Commercial MEMG also features abundant PV, but its HL is concentrated during the daytime. Without the converter from natural gas, the electricity grid and PV are major sources to supply EL. EHP is used to supply HL during the mid-day hours by converting the free PV from electricity to heat power. While EES and TES also exhibit their flexibility to charge cheap and free energy and discharge them to the peak demand hours. Finally, GB in the heat sector is used to supply the left part of HL.

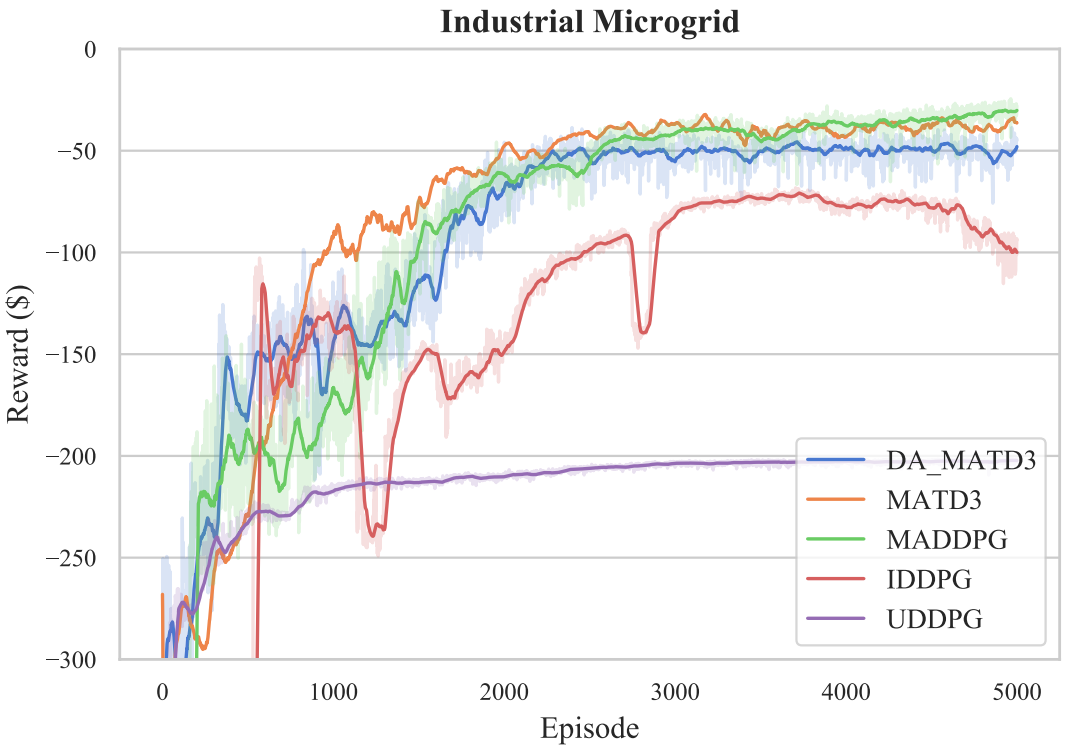


Figure 5.10: Learning curves for the industrial MEMG energy coordination under different MADRL methods. Curves are smoothed for visual clarity.

Unlike Residential and Commercial ones, Industrial MEMG installs WG and its energy usage mainly focus on EL. It can be observed that there is abundant WG production supplying EL and are also used for EES charging power and surplus fed to the grid to obtain additional revenue. The electricity grid partly supplies EL during the mid-day hours with low wind sources. In the heating sector, CHP accounts for the major proportion of HL supply, while TES is learned to discharge to reduce CHP usage when energy prices are high. It can be concluded that the proposed DA-MATD3 is able to learn effective energy conversion policies for all 3 MEMGs to various price signals, demand patterns, and renewable output. In addition, the complementary effect among multi-energy vectors (interaction between electric and heat supplies) can be also verified based on the above analysis.

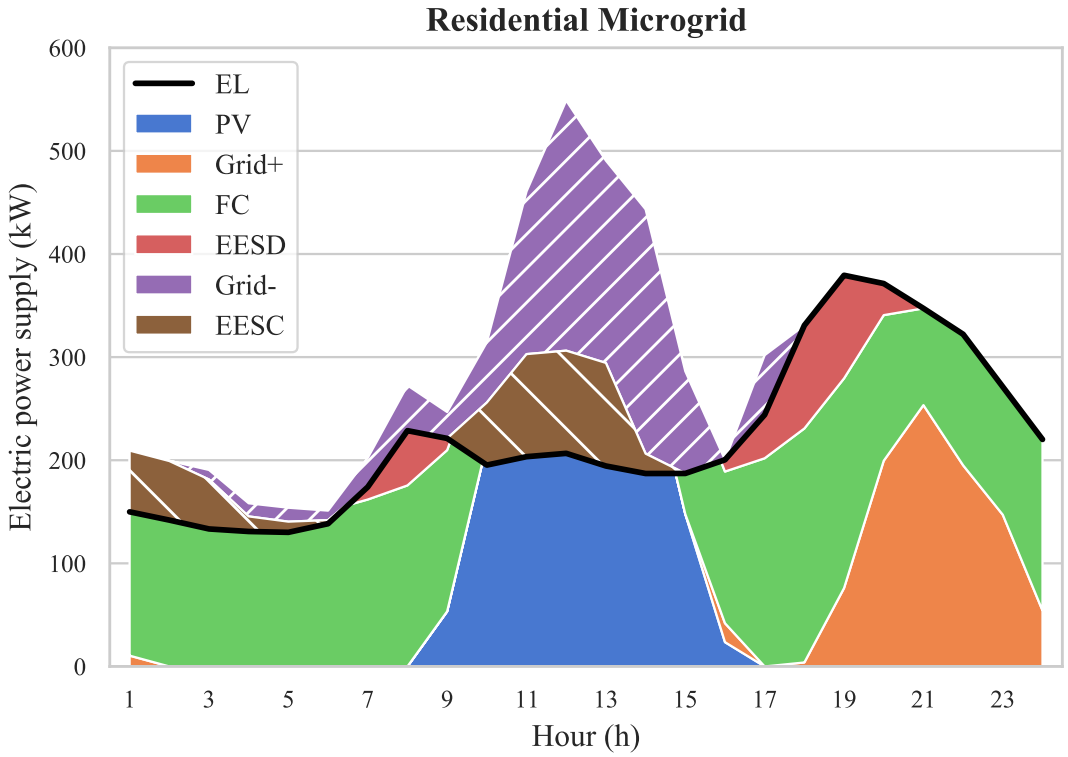


Figure 5.11: Electric power supply and demand for residential MEMG under DA-MATD3 method.

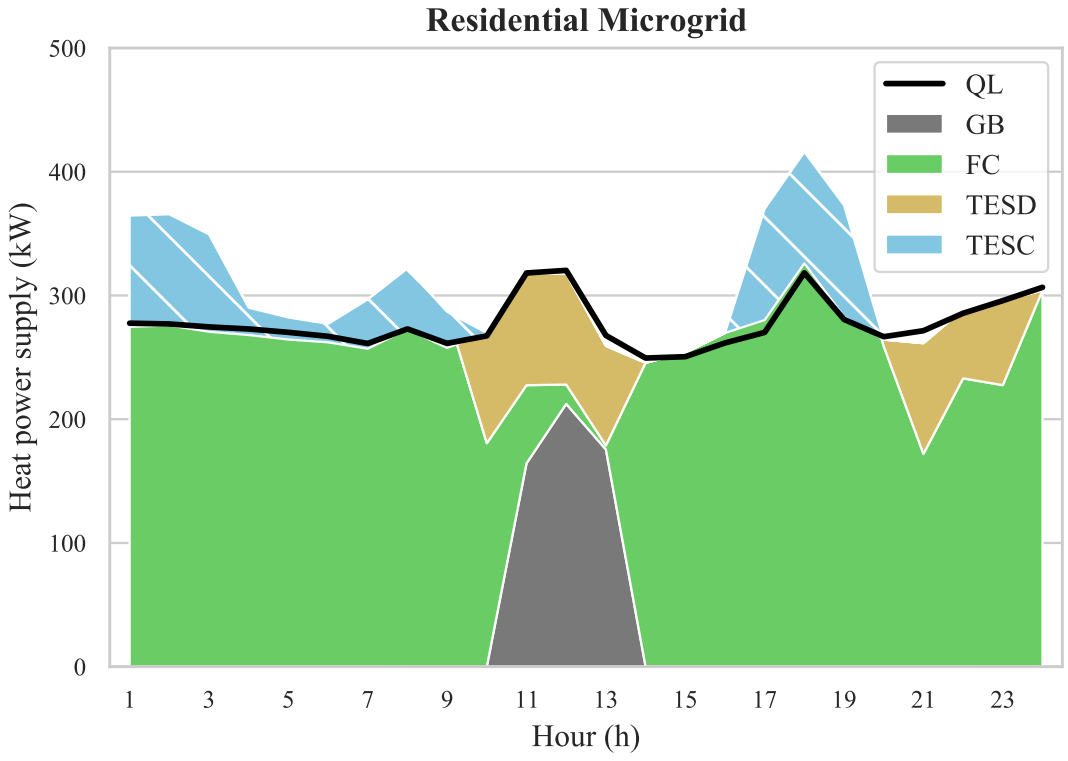


Figure 5.12: Heat power supply and demand for residential MEMG under DA-MATD3 method.

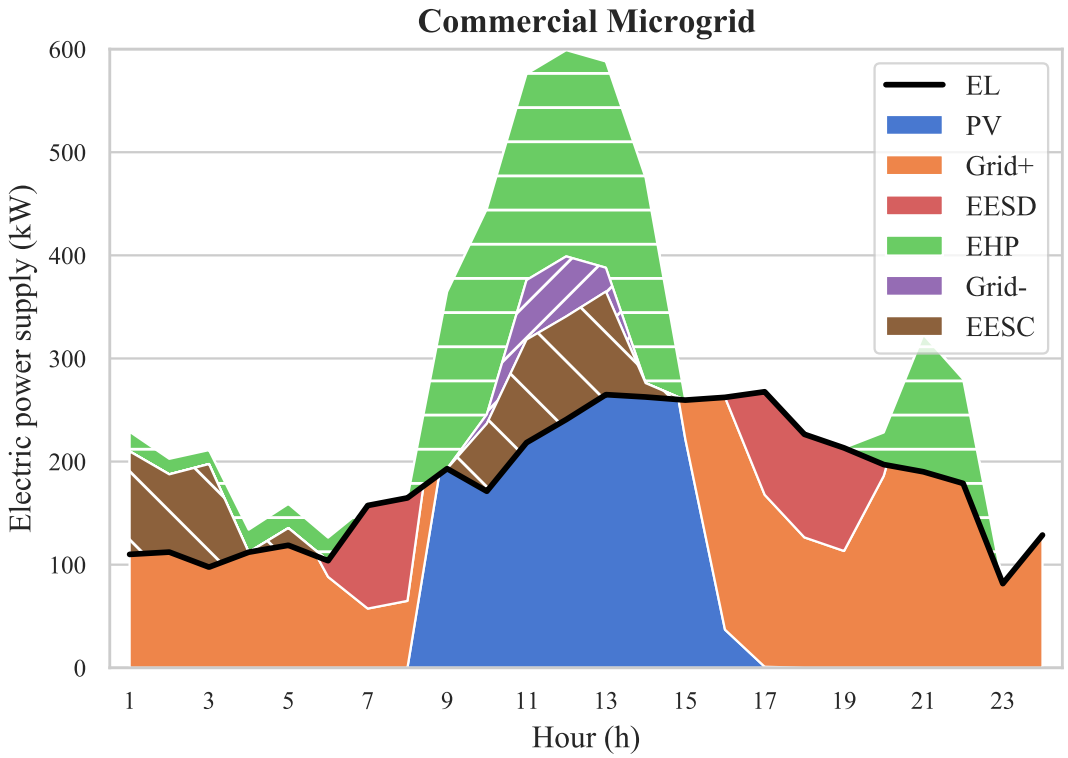


Figure 5.13: Electric power supply and demand for commercial MEMG under DA-MATD3 method.

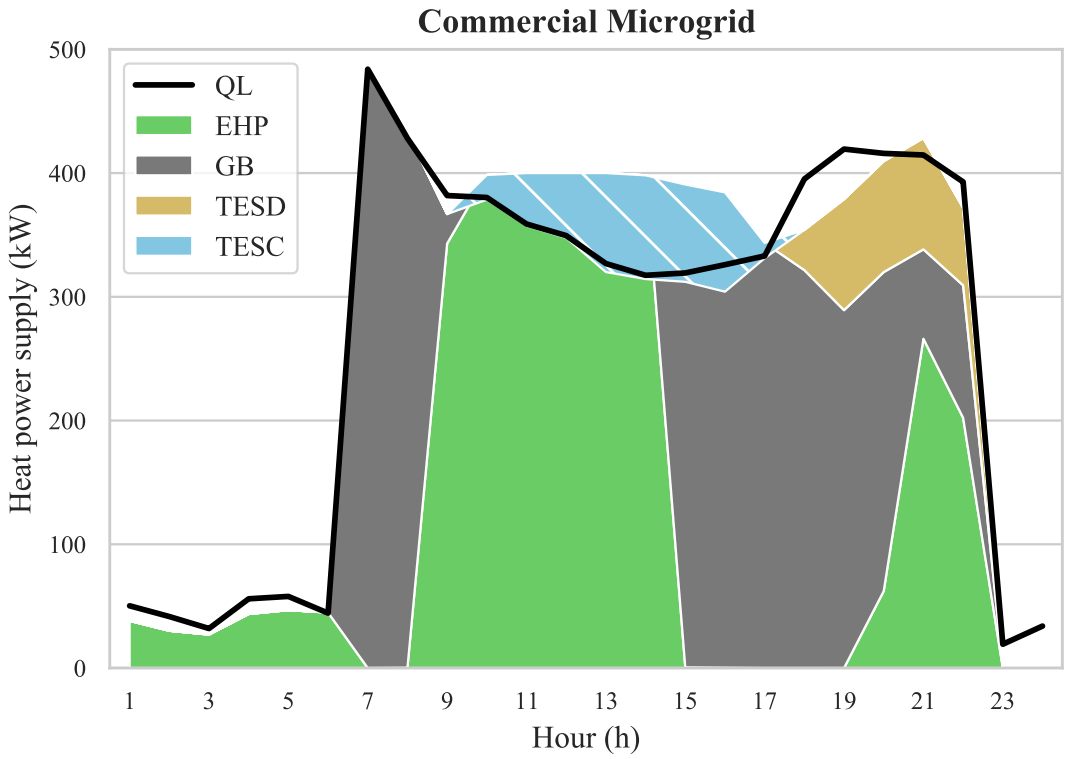


Figure 5.14: Heat power supply and demand for commercial MEMG under DA-MATD3 method.

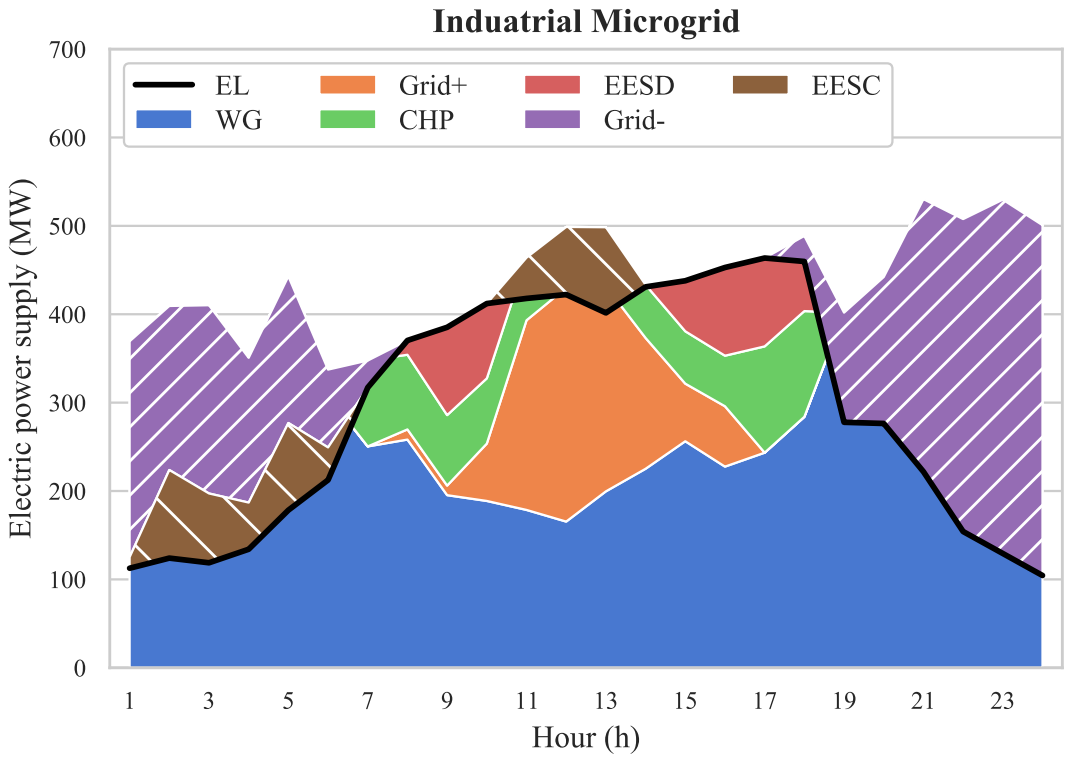


Figure 5.15: Electric power supply and demand for industrial MEMG under DA-MATD3 method.

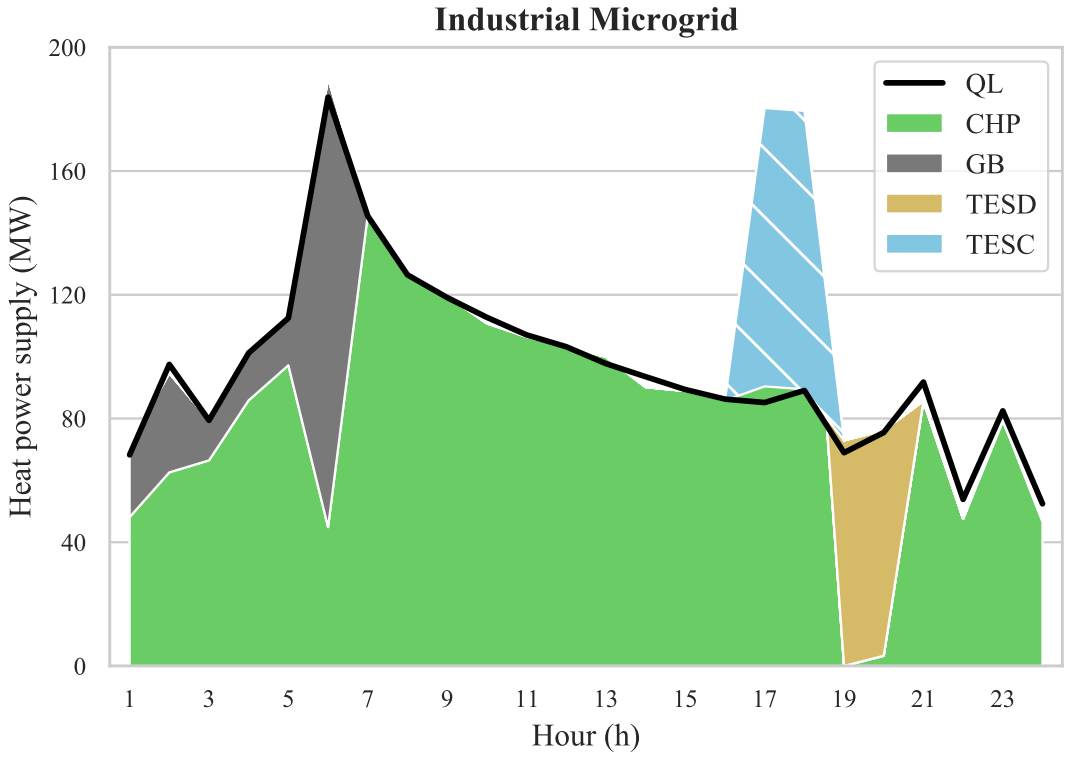


Figure 5.16: Heat power supply and demand for industrial MEMG under DA-MATD3 method.

5.4.4 Benefits of P2P Energy Trading in Double Auction Market

Having demonstrated the superiority of the DA-MATD3 method over the state-of-the-art MADRL methods and analysed the energy schedules of 3 MEMGs, this section aims to compare the P2P energy trading strategies under the dynamic DA-MATD3 method with the statistic ZI policy and quantifying the benefits of P2P energy trading among 3 MEMGs. Figure 5.17, 5.18 and 5.19 show the net load (positive for consumption, negative for generation) of 3 MEMGs under the methods of UDDPG without P2P energy trading and ZI, DA-MATD3 with P2P energy trading but in different trading strategies. Dash lines as the baselines represent the aggregated load of inflexible demand and renewable. Figure 5.20, 5.21 and 5.22 illustrate the local trading quantities and the averaged trading prices under ZI and DA-MATD3 methods.

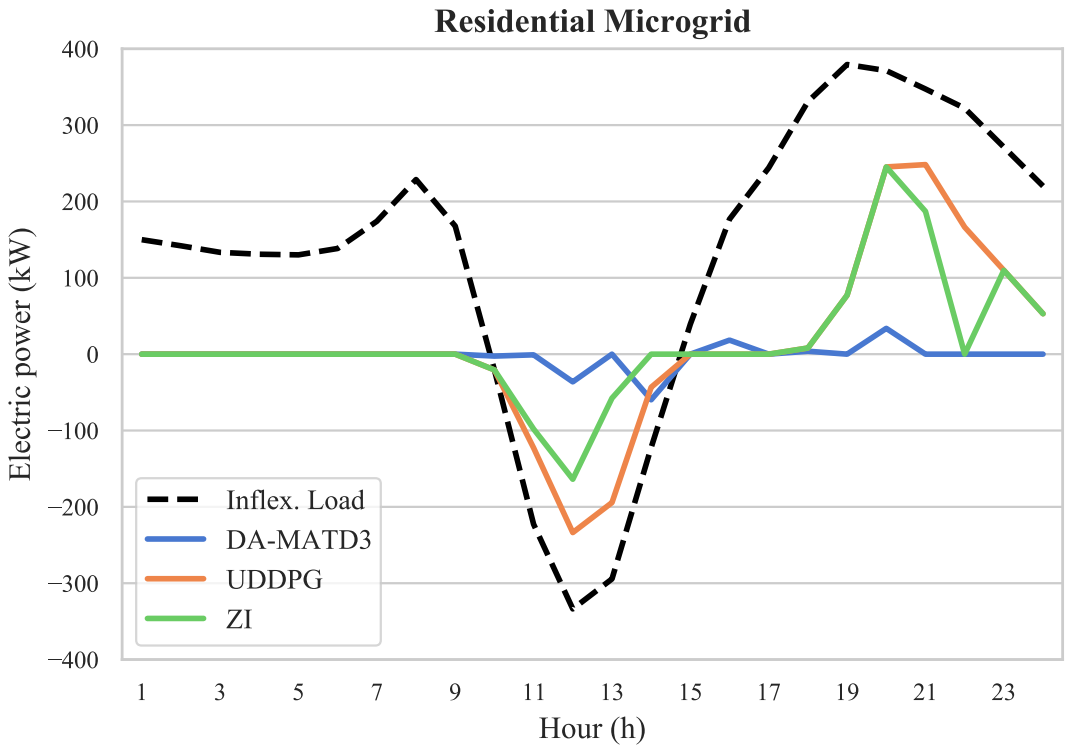


Figure 5.17: Net loads for residential MEMGs under UDDPG, ZI and DA-MATD3 methods.

When P2P energy trading is allowed in DA market, MEMGs with energy surplus/deficiency are incentivised to trade locally. Thereby, we can observe that compared with UDDPG, the generation and demand of 3 MEMGs in Figure 5.17, 5.18 and 5.19 are both reduced under ZI and DA-MATD3, since an amount of energy is balanced locally in DA market, which can also be confirmed in Figure 5.20 and 5.21. The figures show that the DA-MATD3 method trades more frequently and in greater quantities than the ZI method due to the following reasons: 1) For the DA-MATD3 method, the agents are trained to select the suitable trading prices, so that the buyers and the sellers can achieve more trading deals. For the ZI method, the trading prices of the MEMGs are chosen randomly within the range of FiT and ToU, which affects how many

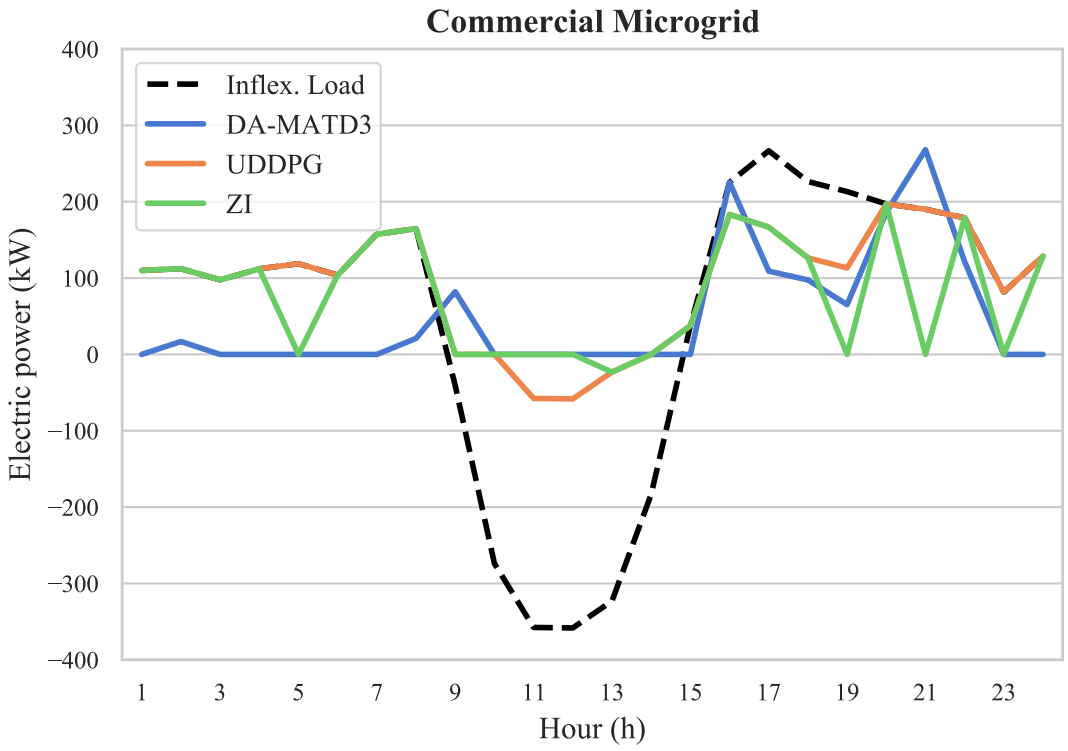


Figure 5.18: Net loads for commercial MEMGs under UDDPG, ZI and DA-MATD3 methods.

times the trading deals are successful. 2) For the DA-MATD3 method, the agents are more likely to trade larger quantities in the DA market to reduce the costs, since each agent considers others' trading strategies. For the ZI method, each MEMG decides the energy trading quantity without considering the trading strategies of the other MEMGs.

More importantly, compared with the non-strategically sampling behaviors in ZI method, MGCC agents under DA-MATD3 learn to trade a large amount of energy locally, thereby reducing their dependence on the upstream utility company. Such results can also be validated in Table 5.5: 1) there is no internal trading under UDDPG, so the net demand and generation (7,382 kWh in total) are all bought at high ToU and sold at low FiT with the utility company; 2) ZI achieves $89 total cost saving by 1,929 kWh internal trading within the DA market; 3) DA-MATD3 achieves the lowest total energy cost by making the highest internal trading at 7,263 kWh. In relative terms, DA-MATD3 achieves 2.82 / 1.76 times lower external trading with the utility company (higher balance of local demand-generation) and 30.65% / 20.54% lower energy cost (more economic benefits of local trading) over UDDPG / ZI methods.

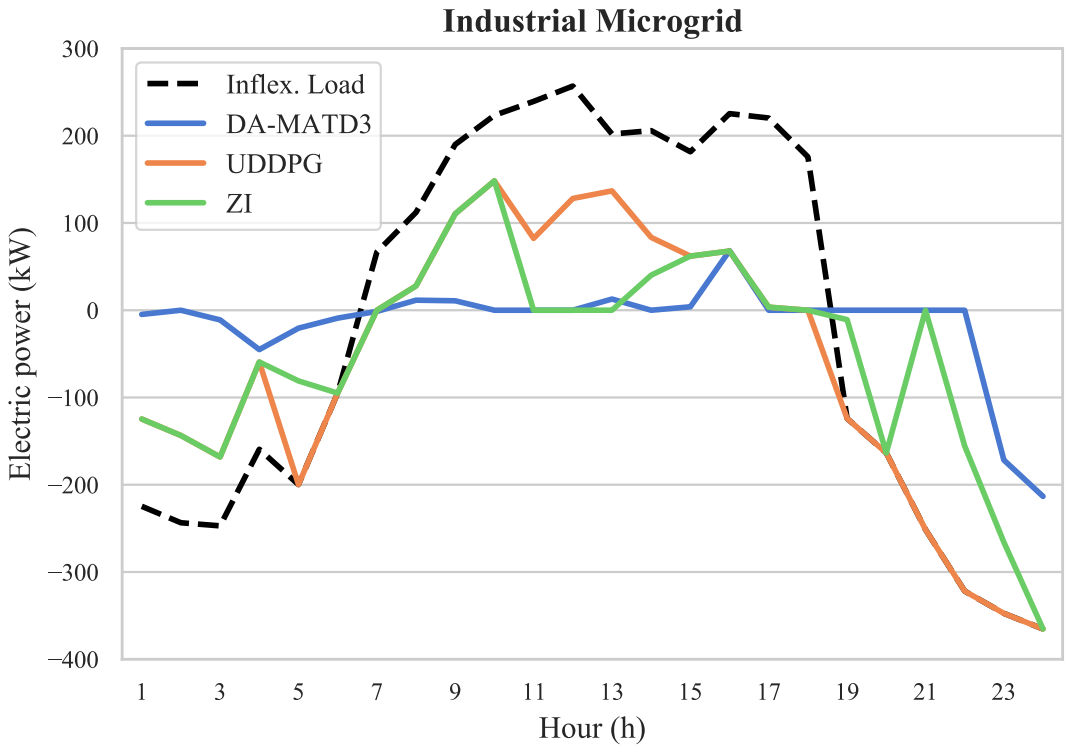


Figure 5.19: Net loads for industrial MEMGs under UDDPG, ZI and DA-MATD3 methods.

Table 5.5: Community Daily Internal, External Trading Quantities, and Energy Costs under UDDPG, ZI, DA-MATD3 Methods

Method	Internal (kWh)	External (kWh)	Energy Cost ($)
UDDPG	-	7,382	1,151
ZI	1,929	5,327	1,062
DA-MATD3	7,263	1,933	881

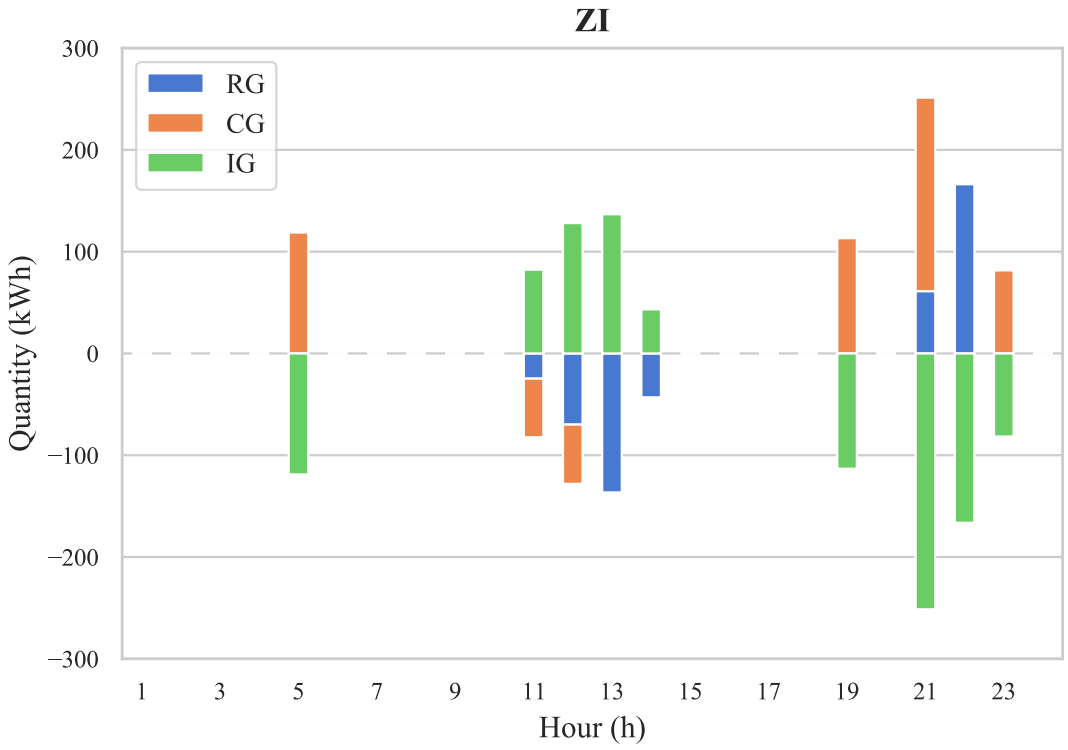


Figure 5.20: Local DA market clearing quantities under ZI method.

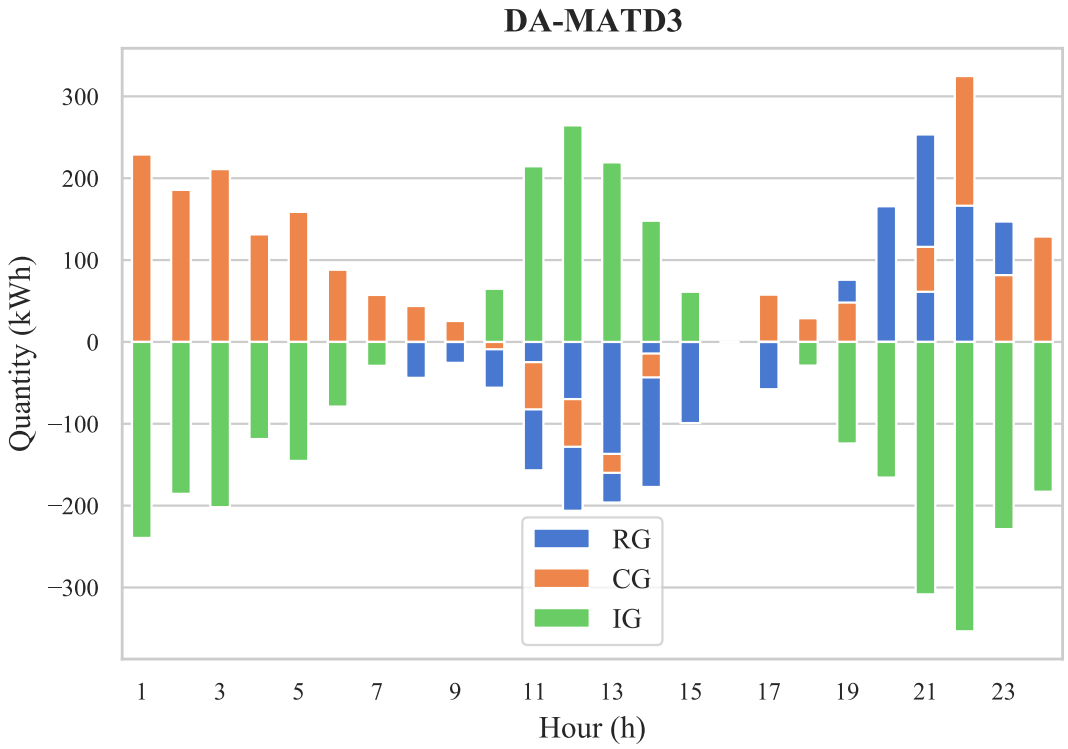


Figure 5.21: Local DA market clearing quantities under proposed DA-MATD3 method.

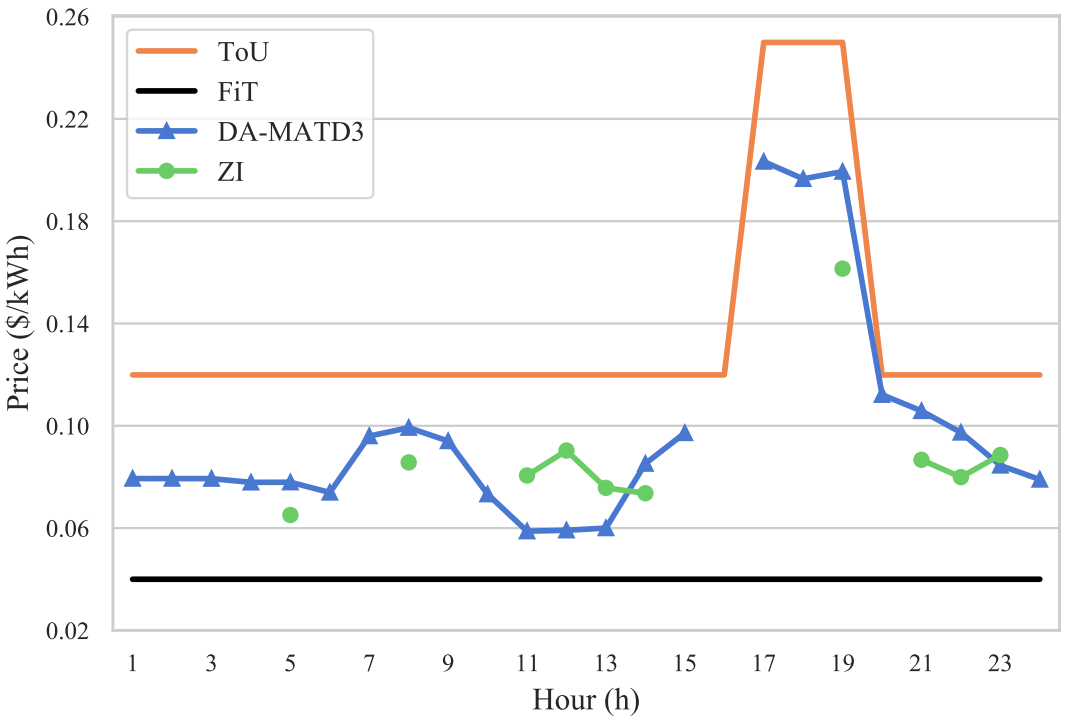


Figure 5.22: Local DA market clearing prices under ZI and DA-MATD3 methods.

5.5 Conclusions

This chapter has proposed a novel MADRL method to address the P2P energy trading and energy conversion problem of multiple MEMGs local trading in a highly efficient DA market, incentivising MEMGs to participate in local trading through a welfare-maximised market clearing algorithm. The examined MEMGs featuring various demand and renewable characteristics, and complex MES operation models are categorised into three heterogeneous residential, commercial, and industrial areas. The proposed MADRL method named DA-MATD3 1) constructs the centralised critic by abstracting the others' observations and actions through the DA market public information, thereby preserving MEMGs' privacy and capturing the market dynamics; 2) uses a pair of critic networks to overcome the Q-value overestimation issue and stabilise the training performance. The effectiveness of the proposed DA-MATD3 method has been evaluated through simulations using a real-world setting. Specifically, the proposed method achieves superior performance in reducing both energy costs and carbon emissions compared to the state-of-the-art ZI and MADRL methods. Finally, the trading strategies and outcomes are also analysed to show the significant economic benefits of the community by more internal energy trading among 3 MEMGs within the DA market.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This thesis investigated the feasibility of using deep reinforcement learning to automatically derive optimal P2P energy trading and energy operation policies for microgrids participating in a local energy trading market. This thesis examined P2P trading in the electricity sector, including developing a realistic model that accounts for continuous action spaces and a set of physical constraints while using deep reinforcement learning. The thesis then extended P2P energy trading to multi-energy systems by analysing the P2P energy trading among multiple multi-energy microgrids using a novel multi-agent DRL algorithm. As a final conclusion, this thesis explored P2P energy trading in the double auction market, investigating how the proposed novel multi-agent deep reinforcement learning algorithm can address privacy concerns while minimising the operation cost of multi-energy microgrids.

Specifically, we studied how to empower P2P energy trading using deep reinforcement learning from the following three aspects:

6.1.1 P2P Energy Trading for Microgrids using Deep Reinforcement Learning

With the development of the DRL algorithm, this chapter investigated the effectiveness and robustness of the DRL algorithms for solving realistic P2P energy trading problems. The P2P energy trading for MGs was formulated as an MDP with a set of physical constraints like transmission loss, the node power limit and battery wear cost.

The chapter proposed a DRL solutions to address the high-dimensional and uncertain data in the P2P energy trading. This method involved discretising the action space into many equal bins and solving the problem using a DQN-based algorithm.

The case study was conducted using 1-year real-world generation and demand data. Simulation results indicated that the DQN-based algorithm can be adapted to the P2P energy trading

problem across different seasons.

Furthermore, this chapter studied the impact of different battery sizes on the operational cost of the MG. Results showed that the proposed method could help MGs choose the most suitable battery to achieve their own social goals.

6.1.2 P2P Energy Trading and Energy Conversion in Interconnected Multi-Energy Microgrids Using Multi-Agent Deep Reinforcement Learning

As the multi-energy system is involved in forming the MEMGs, energy conversion policies are also essential to minimise the operation cost of the MEMGs. Therefore, this chapter proposes the problem with external P2P energy trading and internal energy conversion for three interconnected MEMGs: residential, commercial, and industrial MEMGs in a local community. Due to the intermittent nature of energy production and demand, the problem was formulated as a POMDP.

In order to cope with the non-stationary environment of multi-agent interactions, this chapter proposed a multi-agent TD3 approach. The proposed MATD3 method employs the centralised critic for maintaining Markov property and the decentralised actor for independent local execution. The TD3 algorithm was used to solve the overestimation and high variance problems in the Q-learning based algorithms. The proposed MATD3 approach also includes modifications to the original framework to improve the stability of the P2P energy trading and energy conversion problem.

The case study on three real-world datasets showed that the proposed method significantly reduced all MGs' operation costs. Compared with other DRL algorithms and a rule-based method, the proposed MATD3 method can reduce up to 18.2 %, 27.8% and 23.1% operational cost for residential MG, commercial MG and industrial MG. This chapter also compares the DRL methods' computational performance in terms of training time and execution time, indicating the efficient and stable training performance of the proposed method.

The simulation results also demonstrated the importance of P2P energy trading and energy conversion. Cleverly deciding the P2P energy trading and energy conversion by the proposed method, the MGs can flexibly convert, trade and store the energy when needed. The results showed that no electricity is sold back to the grid, and most of the heat demand is fulfilled by P2P energy trading.

This chapter also investigated the impact of carbon tax price on the operation cost and CO_2 emissions. The carbon tax price impact is less on the residential MG since it is less dependent on the external network than the commercial and industrial MGs. In terms of CO_2 emissions, the increase of the carbon tax price will drive the MGs to reduce emissions. However, a high carbon tax causes the commercial MG to decrease sales of heat to other MGs, increasing heating

costs for others and even increasing greenhouse gas emissions.

Finally, this chapter investigated the scalability of the proposed approach. The proposed approach can be scaled up for a more extended period, including changing the time horizon in the system problem and training the agents periodically with the added new experience.

6.1.3 Coordination for Multi-Energy Microgrids in Double Auction Market Using Multi-Agent Deep Reinforcement Learning

This chapter investigated the market framework of the P2P energy market. The centralised market framework requires the full authority from the MEMGs to collect all the information of their resources and system model. Decentralised markets, however, are not able to guarantee that the solutions offered are optimal. Therefore, this chapter investigated P2P energy trading in a highly efficient double auction market designed to maximise the participated MEMGs.

This chapter proposed a novel MADRL method to address the P2P energy trading and energy conversion problem for multiple MEMGs in the DA market. The MEMGs with different generation and demand patterns are categorised into residential MEMGs, commercial MEMGs and industrial MEMGs. The proposed approach addresses the privacy issues in the original MADRL framework by approximating the centralised critic with the DA market public information. The proposed method also uses a pair of critic networks to overcome the Q-value overestimation issue, resulting in a stable training process.

The effectiveness of proposed DA-MATD3 method has been evaluated through the experiment results on a real-world MES scenario. Specifically, the proposed DA-MATD3 achieves superior performance in reducing energy cost and carbon emission regarding the state-of-the-art MADRL methods and the conventional ZI strategy. The proposed DA-MATD3 abstracted the others' observations and actions through the DA market public order books, alleviating the privacy issue of the centralise critic in MATD3.

Finally, this chapter investigated the value of energy conversion and P2P energy trading in the DA market. The proposed DA-MATD3 is able to learn effective energy conversion policies to manage their energy converters and energy storage systems for all MEMGs with various price signals, demand patterns, and renewable generation. The MEMGs trained by the proposed method also collaboratively trade and consume the energy locally in the DA market most of the time, showing great economic benefits.

6.2 Future Work

Although this thesis fulfils the aims of developing efficient P2P energy trading models for MGs in a local community using DRL techniques, there is still some related work and interesting direction that can be explored in the future.

1. Firstly, the DA market introduced in Chapter 5 focuses on electricity trading only. The limitation is related to the pricing mechanism whereby the grid electricity buy price is higher than the grid electricity sell price. Therefore, the benefits of the DA market will always be guaranteed. In contrast, the benefits of heat trading and its implementation in the DA market are not sufficiently researched. However, trading electricity only may not fully exploit the flexibility of MES among multiple energy sectors. In this context, future work will explore a new market mechanism enabling multi-energy trading in the local MEMGs community.

2. The intermittent nature of renewable energy generation and demand is the main drive to use dynamic programming techniques like DRL over traditional static planning tools like MILP and ADMM. However, the performance of the DRL agent is still affected by the uncertain generation and demand pattern as the actual state of the environment can not be fully observed by the MGs. In this thesis, random Gaussian noise is used to represent the estimated generation and demand (i.e., the observation values), which can effectively improve the DRL algorithm's robustness. However, developing energy forecasting techniques to facilitate the agent's observation could be a promising way to improve the performance of DRL approaches in P2P energy trading. The recurrent neural network, e.g., Long Short-Term Memory (LSTM), could be a suitable algorithm to explore.

3. The complex power system network is also critical in P2P energy trading. Our future work will consider how the network constraints can be integrated into the proposed P2P energy trading optimisation problem. Investigating the line congestion on the market outcomes and the strategic decisions of the MGs with the help of DRL methods is also worth digging into. Take the DA market model of Chapter 5 as an example, future work aims at integrating a set of network constraints into the DA market, where the distribution network operator (DNO) could be involved to help provide the optimal network operations to facilitate the energy trading among MEMGs.

4. This thesis only considered MGs, and a limited number of MGs were simulated to test the performance. Future work will model a more realistic energy community by increasing the population of MEMGs and including small-scale prosumers to form them into a large-scale multi-agent system. However, the number of participants in the P2P energy market will bring significant computational issues. A direction is to look for an abstraction formulation for multi-agent DRL algorithms. Another future direction will be classifying the P2P energy trading participants into several categories depending on the generation and demand pattern to form a hierarchy energy market.

5. The P2P energy trading and energy conversion problem contain many constraints. DRL algorithms often have difficulty learning a good policy for constrained problems. The

constraints defined in the P2P energy problem are essential for the power system's safety and should not be compromised. Recently, a sub-field of RL algorithm, safe RL, was proposed to address this issue. Safe RL is defined to ensure good system performance but also need to obey safety constraints during the learning and execution processes [169]. Our future work will include implementing safe RL techniques into P2P energy trading problems.

6. Finally, the MADRL algorithms and other machine learning techniques can be introduced to solve security issues in P2P energy trading, preventing participating prosumers and MGs from hostile attacks. As part of our future work, we plan to identify abnormal activity occurring in the P2P energy trading market and automatically protect users against cyber-attacks to improve energy security.

Bibliography

[1] C. B. Field, V. Barros, T. F. Stocker, and Q. Dahe, *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press, 2012.

[2] M. Allen, O. Dube, W. Solecki, F. Aragón-Durand, W. Cramer, S. Humphreys, M. Kainuma, J. Kala, N. Mahowald, Y. Mulugetta *et al.*, "Global warming of 1.5° c. an ipcc special report on the impacts of global warming of 1.5° c above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty," *Sustainable Development, and Efforts to Eradicate Poverty*, 2018.

[3] Scottish Government. (2021) Policy: Renewable and low carbon energy. [Online]. Available: https://www.gov.scot/policies/renewable-and-low-carbon-energy

[4] Department for Business, Energy & Industrial Strategy. (2022) Energy Trends: UK renewables. [Online]. Available: https://www.gov.uk/government/statistics/energy-trends-section-6-renewables

[5] ——. (2020) Energy and emissions projections. [Online]. Available: https://www.gov.uk/government/collections/energy-and-emissions-projections

[6] W. Tushar, C. Yuen, H. Mohsenian-Rad, T. Saha, H. V. Poor, and K. L. Wood, "Transforming energy networks via peer-to-peer energy trading: The potential of game-theoretic approaches," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 90–111, Jul. 2018.

[7] E. Mengelkamp, J. Gärttner, K. Rock, S. Kessler, L. Orsini, and C. Weinhardt, "Designing microgrid energy markets: A case study: The brooklyn microgrid," *Appl. Energy*, vol. 210, pp. 870–880, Jan. 2018.

[8] W. Tushar, T. K. Saha, C. Yuen, P. Liddell, R. Bean, and H. V. Poor, "Peer-to-peer energy trading with sustainable user participation: A game theoretic approach," *IEEE Access*, vol. 6, pp. 62 932–62 943, Oct. 2018.

[9] J. Guerrero, A. C. Chapman, and G. Verbič, "Decentralized p2p energy trading under network constraints in a low-voltage network," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5163–5173, Sep. 2019.

[10] T. Baroche, P. Pinson, R. L. G. Latimier, and H. B. Ahmed, "Exogenous cost allocation in peer-to-peer electricity markets," *IEEE Trans. Power Syst.*, vol. 34, no. 4, pp. 2553–2564, Jul. 2019.

[11] C. A. Floudas and X. Lin, "Mixed integer linear programming in process scheduling: Modeling, algorithms, and applications," *Ann. Oper. Res.*, vol. 139, no. 1, pp. 131–162, Oct. 2005.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2010, vol. 3, no. 1.

[13] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," *arXiv preprint arXiv:1708.05866*, Sep. 2017.

[14] P. Mancarella, G. Andersson, J. A. Peças-Lopes, and K. R. Bell, "Modelling of integrated multi-energy systems: Drivers, requirements, and opportunities," in *Proc. 19th Power Syst. Comput. Conf. (PSCC'2016)*, Genoa, Italy, pp. 1–22.

[15] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR'2016)*, San Juan, Puerto Rico, pp. 1–14.

[16] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS'2017)*, Long Beach, CA, USA, pp. 6379–6390.

[17] S. Fujimoto, H. V. Hoof, and D. Meger, "Addressing Function Approximation Error in Actor-Critic Methods," in *Proc. 35th Int. Conf. Mach. Learn. (ICML'2018)*, Stockholm, Sweden, pp. 2587–2601.

[18] C. Zhang, J. Wu, Y. Zhou, M. Cheng, and C. Long, "Peer-to-Peer energy trading in a Microgrid," *Appl. Energy*, vol. 220, pp. 1–12, Jun. 2018.

[19] W. Tushar, T. K. Saha, C. Yuen, D. Smith, and H. V. Poor, "Peer-to-Peer Trading in Electricity Networks: An Overview," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3185–3200, Jul. 2020.

[20] W. Tushar, B. Chai, C. Yuen, S. Huang, D. B. Smith, H. V. Poor, and Z. Yang, "Energy storage sharing in smart grid: A modified auction-based approach," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1462–1475, May 2016.

[21] D. Papadaskalopoulos and G. Strbac, "Decentralized participation of flexible demand in electricity markets—part i: Market mechanism," *IEEE Trans. Power Syst.*, vol. 28, no. 4, pp. 3658–3666, Apr. 2013.

[22] Y. Zhou, J. Wu, C. Long, and W. Ming, "State-of-the-art analysis and perspectives for peer-to-peer energy trading," *Engineering*, vol. 6, no. 7, pp. 739–753, Jul. 2020.

[23] A. Lüth, J. M. Zepter, P. Crespo del Granado, and R. Egging, "Local electricity market designs for peer-to-peer trading: The role of battery flexibility," *Appl. Energy*, vol. 229, pp. 1233–1243, Nov. 2018.

[24] J. M. Zepter, A. Lüth, P. Crespo del Granado, and R. Egging, "Prosumer integration in wholesale electricity markets: Synergies of peer-to-peer trade and residential storage," *Energy Build.*, vol. 184, pp. 163–176, Feb. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378778818330378

[25] M. R. Alam, M. St-Hilaire, and T. Kunz, "Peer-to-peer energy trading among smart homes," *Appl. Energy*, vol. 238, pp. 1434–1443, Mar. 2019.

[26] J. Guerrero, A. C. Chapman, and G. Verbič, "Decentralized P2P energy trading under network constraints in a low-voltage network," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5163–5173, Sep. 2019.

[27] T. Morstyn, A. Teytelboym, and M. D. McCulloch, "Bilateral contract networks for peer-to-peer energy trading," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2026–2035, Mar. 2019.

[28] E. Sorin, L. Bobo, and P. Pinson, "Consensus-based approach to peer-to-peer electricity markets with product differentiation," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 994–1004, Mar. 2019.

[29] M. T. Devine and P. Cuffe, "Blockchain electricity trading under demurrage," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2323–2325, Mar. 2019.

[30] P. Scarabaggio, S. Grammatico, R. Carli, and M. Dotoli, "Distributed demand side management with stochastic wind power forecasting," *IEEE Trans. Control Syst. Technol.*, vol. 30, no. 1, pp. 97–112, Jan. 2022.

[31] T. Sousa, T. Soares, P. Pinson, F. Moret, T. Baroche, and E. Sorin, "Peer-to-peer and community-based markets: A comprehensive review," *Renew. Sust. Energ. Rev.*, vol. 104, pp. 367–378, Apr. 2019.

[32] A. Paudel, K. Chaudhari, C. Long, and H. B. Gooi, "Peer-to-peer energy trading in a prosumer-based community microgrid: A game-theoretic model," *IEEE Trans. Ind. Electron.*, vol. 66, no. 8, pp. 6087–6097, Aug. 2019.

[33] F. Moret and P. Pinson, "Energy collectives: A community and fairness based approach to future electricity markets," *IEEE Trans. Power Syst.*, vol. 34, no. 5, pp. 3994–4004, Sep. 2019.

[34] J. Kang, R. Yu, X. Huang, S. Maharjan, Y. Zhang, and E. Hossain, "Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3154–3164, Dec. 2017.

[35] C. Long, J. Wu, C. Zhang, L. Thomas, M. Cheng, and N. Jenkins, "Peer-to-peer energy trading in a community microgrid," in *Proc. IEEE Power Energy Soc. Gen. Meet. (GM'2017)*, Chicago, IL, USA, pp. 1–5.

[36] T. Chen and W. Su, "Indirect Customer-to-Customer Energy Trading with Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4338–4348, Jul. 2019.

[37] U.S. Energy Information Administration. (2022) Solar explained: Photovoltaics and electricity. [Online]. Available: https://www.eia.gov/energyexplained/solar/photovoltaics-and-electricity.php

[38] H. L. Tsai, C. S. Tu, Y. J. Su *et al.*, "Development of generalized photovoltaic model using MATLAB/SIMULINK," in *Proc. World Congress Eng. Comput. Sci. (WCECS'2008)*, vol. 2008, San Francisco, USA, pp. 1–6.

[39] J. An, M. Lee, S. Yeom, and T. Hong, "Determining the peer-to-peer electricity trading price and strategy for energy prosumers and consumers within a microgrid," *Appl. Energy*, vol. 261, p. 114335, Mar. 2020.

[40] S. Y. Lee, Y. Hwang, and S. Lee, "Frequency-driven convolutional neural network for enhancing noise-robustness of bearing fault detection," in *INTER-NOISE NOISE-CON Congress Conf. Proc. (NOISE-CON'2020)*, vol. 261, no. 1, Institute of Noise Control Engineering, pp. 5635–5645.

[41] B. Ren and Q.-C. Zhong, "Ude-based robust control of variable-speed wind turbines," in *39th Ann. Conf. IEEE Ind. Electron. Soc. (IECON'2013)*, 2013, pp. 3818–3823.

[42] M. M. Arsoon and S. M. Moghaddas-Tafreshi, "Peer-to-peer energy bartering for the resilience response enhancement of networked microgrids," *Appl. Energy*, vol. 261, p. 114413, Mar. 2020.

[43] S. Baros and M. D. Ilić, "Distributed torque control of deloaded wind dfigs for wind farm power output regulation," *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4590–4599, Nov. 2017.

[44] M. Vahedipour-Dahraie, H. Rashidizadeh-Kermani, M. Shafie-Khah, and P. Siano, "Peer-to-peer energy trading between wind power producer and demand response aggregators for scheduling joint energy and reserve," *IEEE Syst. J.*, vol. 15, no. 1, pp. 705–714, Mar. 2021.

[45] J. Ruotsalainen, J. Karjalainen, M. Child, and S. Heinonen, "Culture, values, lifestyles, and power in energy futures: A critical peer-to-peer vision for renewable energy," *Energy Res. Soc. Sci.*, vol. 34, pp. 231–239, Dec. 2017.

[46] Y. Xiao, X. Wang, P. Pinson, and X. Wang, "A local energy market for electricity and hydrogen," *IEEE Trans. Power Syst.*, vol. 33, no. 4, pp. 3898–3908, Jul. 2018.

[47] D. Zhu, B. Yang, Q. Liu, K. Ma, S. Zhu, C. Ma, and X. Guan, "Energy trading in microgrids for synergies among electricity, hydrogen and heat networks," *Appl. Energy*, vol. 272, p. 115225, Aug. 2020.

[48] H. Mehrjerdi, "Peer-to-peer home energy management incorporating hydrogen storage system and solar generating units," *Renew. Energy*, vol. 156, pp. 183–192, Aug. 2020.

[49] Y. Ye, D. Qiu, X. Wu, G. Strbac, and J. Ward, "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3068–3082, Jul. 2020.

[50] L. Cabeza, I. Martorell, L. Miró, A. Fernández, and C. Barreneche, "Introduction to thermal energy storage (tes) systems," in *Advances in thermal energy storage systems*. Elsevier, 2015, pp. 1–28.

[51] C. Zhang, Y. Xu, Z. Li, and Z. Y. Dong, "Robustly coordinated operation of a multi-energy microgrid with flexible electric and thermal loads," *IEEE Trans. Smart Grid*, vol. 10, no. 3, pp. 2765–2775, May 2019.

[52] M. S. Sadaghiani and M. Mehrpooya, "Introducing and energy analysis of a novel cryogenic hydrogen liquefaction process configuration," *International journal of hydrogen energy*, vol. 42, no. 9, pp. 6033–6050, Mar. 2017.

[53] V. Oldenbroek, L. A. Verhoef, and A. J. van Wijk, "Fuel cell electric vehicle as a power plant: Fully renewable integrated transport and energy system design and analysis for smart city areas," *Int. J. Hydrogen Energy*, vol. 42, no. 12, pp. 8166–8196, mar 2017.

[54] O. Z. Sharaf and M. F. Orhan, "An overview of fuel cell technology: Fundamentals and applications," *Renew. Sustain. Energy Rev.*, vol. 32, pp. 810–853, Apr. 2014.

[55] U.S. Energy Information Administration. (2022) Hydrogen explained: Use of hydrogen. [Online]. Available: https://www.eia.gov/energyexplained/hydrogen/use-of-hydrogen.php

[56] J. Baschuk and X. Li, "A general formulation for a mathematical pem fuel cell model," *Journal of power sources*, vol. 142, no. 1-2, pp. 134–153, Mar. 2005.

[57] A. Keçebaş, M. Kayfeci, and M. Bayat, "Chapter 9 - electrochemical hydrogen generation," in *Solar Hydrogen Production*, F. Calise, M. D. D'Accadia, M. Santarelli, A. Lanzini, and D. Ferrero, Eds. Academic Press, 2019, pp. 299–317. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128148532000096

[58] Wright Renewable Heating. (2022) A quick guide to explain how an Air Source Heat Pump works. [Online]. Available: https://www.wrightrenewableheating.co.uk/how-heat-pumps-work

[59] A. H. Nosrat, L. G. Swan, and J. M. Pearce, "Simulations of greenhouse gas emission reductions from low-cost hybrid solar photovoltaic and cogeneration systems for new communities," *Sustainable Energy Technologies and Assessments*, vol. 8, pp. 34–41, Dec. 2014.

[60] U.S. Environmental Protection Agency. (2022) What Is CHP? [Online]. Available: https://www.epa.gov/chp/what-chp

[61] Department for Business, Energy & Industrial Strategy. (2022) Smart Meter Statistics in Great Britain: Quarterly Report to end September 2021. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1035290/Q3_2021_Smart_Meters_Statistics_Report.pdf

[62] R. E. Edwards, J. New, and L. E. Parker, "Predicting future hourly residential electrical consumption: A machine learning case study," *Energy Build.*, vol. 49, pp. 591–603, Jun. 2012.

[63] R. Granell, C. J. Axon, and D. C. Wallom, "Clustering disaggregated load profiles using a dirichlet process mixture model," *Energy Convers. Manag.*, vol. 92, pp. 507–516, Mar. 2015.

[64] F. McLoughlin, A. Duffy, and M. Conlon, "A clustering approach to domestic electricity load profile characterisation using smart metering data," *Appl. energy*, vol. 141, pp. 190–199, Mar. 2015.

[65] O. Jogunola, A. Ikpehai, K. Anoh, B. Adebisi, M. Hammoudeh, H. Gacanin, and G. Harris, "Comparative analysis of p2p architectures for energy trading and sharing," *Energies*, vol. 11, no. 1, p. 62, Dec. 2017.

[66] C. Zhang, "Peer-to-peer energy trading in electrical distribution networks," Ph.D. dissertation, Cardiff University, 2017.

[67] H. Kanchev, D. Lu, F. Colas, V. Lazarov, and B. Francois, "Energy management and operational planning of a microgrid with a pv-based active generator for smart grid applications," *IEEE Trans. Ind. Electron.*, vol. 58, no. 10, pp. 4583–4592, Oct. 2011.

[68] D. Webb, J. Kneifel, C. O'Fallon *et al.*, "Developing cost functions for estimating solar photovoltaic system installed and life cycle costs using historical quote data," 2020.

[69] B. S. Hartono, Y. Budiyanto, and R. Setiabudy, "Review of microgrid technology," in *13th Int. Conf. Qual. Res. (QiR'2013)*, Yogyakarta, Indonesia, pp. 127–132.

[70] W. Saad, Z. Han, H. V. Poor, and T. Basar, "Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 86–105, Sep. 2012.

[71] Y. Wang, W. Saad, Z. Han, H. V. Poor, and T. Başar, "A game-theoretic approach to energy trading in the smart grid," *IEEE Trans. Smart Grid*, vol. 5, no. 3, pp. 1439–1450, May 2014.

[72] J. Björnerstedt and J. W. Weibull, "Nash equilibrium and evolution by imitation," IUI Working Paper, Tech. Rep., 1994.

[73] W. Tushar, T. K. Saha, C. Yuen, T. Morstyn, N. A. Masood, H. Vincent Poor, and R. Bean, "Grid influenced peer-to-peer energy trading," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1407–1418, Mar. 2020.

[74] P. Huang, A. Scheller-Wolf, and K. Sycara, "Design of a multi–unit double auction e–market," *Computational Intelligence*, vol. 18, no. 4, pp. 596–617, Dec. 2002.

[75] W. Saad, Z. Han, H. V. Poor, and T. Başar, "A noncooperative game for double auction-based energy trading between phevs and distribution grids," in *IEEE int. conf. smart grid com. (SmartGridComm'2011)*. IEEE, pp. 267–272.

[76] G. Sierksma and Y. Zwols, *Linear and integer optimization: theory and practice*. CRC Press, 2015.

[77] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decent. Bus. Rev.*, p. 21260, 2008.

[78] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *IEEE Access*, vol. 4, pp. 2292–2303, May 2016.

[79] W. Zou, D. Lo, P. S. Kochhar, X.-B. D. Le, X. Xia, Y. Feng, Z. Chen, and B. Xu, "Smart contract development: Challenges and opportunities," *IEEE Trans. Softw. Eng.*, vol. 47, no. 10, pp. 2084–2106, Oct. 2021.

[80] L. Thomas, Y. Zhou, C. Long, J. Wu, and N. Jenkins, "A general form of smart contract for decentralized energy systems management," *Nature Energy*, vol. 4, no. 2, pp. 140–149, Jan. 2019.

[81] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[82] M. v. Otterlo and M. Wiering, "Reinforcement learning and markov decision processes," in *Reinforcement learning*. Springer, 2012, pp. 3–42.

[83] B. W. Balleine, N. D. Daw, and J. P. O'Doherty, "Multiple forms of value learning and the function of dopamine," in *Neuroeconomics*. Academic Press, jan 2009, pp. 367–387.

[84] C. J. Watkins and P. Dayan, "Technical Note: Q-Learning," *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992.

[85] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–33, Feb. 2015.

[86] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, no. 4, pp. 838–855, 1992.

[87] S. Park, J. Lee, G. Hwang, and J. K. Choi, "Event-Driven Energy Trading System in Microgrids: Aperiodic Market Model Analysis with a Game Theoretic Approach," *IEEE Access*, vol. 5, pp. 26 291–26 302, Nov. 2017.

[88] G. El Rahi, S. R. Etesami, W. Saad, N. B. Mandayam, and H. V. Poor, "Managing Price Uncertainty in Prosumer-Centric Energy Trading: A Prospect-Theoretic Stackelberg Game Approach," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 702–713, Jan. 2019.

[89] W. Kou and S. Y. Park, "Game-theoretic approach for smartgrid energy trading with microgrids during restoration," in *Proc. IEEE Power Energy Soc. Gen. Meet. (GM'2017)*, Chicago, IL, USA, pp. 1–5.

[90] R. Li, W. Wei, S. Mei, Q. Hu, and Q. Wu, "Participation of an Energy Hub in Electricity and Heat Distribution Markets: An MPEC Approach," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3641–3653, Jul. 2019.

[91] V. François-lavet, R. Fonteneau, and D. Ernst, "Deep Reinforcement Learning Solutions for Energy Microgrids Management," in *Eur. Work. Reinf. Learn.*, 2016, pp. 1–7.

[92] G. K. Venayagamoorthy, R. K. Sharma, P. K. Gautam, and A. Ahmadi, "Dynamic Energy Management System for a Smart Microgrid," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 8, pp. 1643–1656, aug 2016.

[93] B. V. Mbuwir, F. Ruelens, F. Spiessens, and G. Deconinck, "Battery energy management in a microgrid using batch reinforcement learning," *Energies*, vol. 10, no. 11, pp. 1–19, 2017.

[94] Y. Chen, W. Wei, F. Liu, E. E. Sauma, and S. Mei, "Energy Trading and Market Equilibrium in Integrated Heat-Power Distribution Systems," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 4080–4094, Jul. 2019.

[95] L. Xiao, X. Xiao, C. Dai, M. Pengy, L. Wang, and H. V. Poor, "Reinforcement Learning-based Energy Trading for Microgrids," 2018.

[96] K. Zhang, S. Troitzsch, S. Hanif, and T. Hamacher, "Coordinated Market Design for Peer-to-Peer Energy Trade and Ancillary Services in Distribution Grids," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 2929–2941, Jul. 2020.

[97] M. Yan, M. Shahidehpour, A. Paaso, L. Zhang, A. Alabdulwahab, and A. Abusorrah, "Distribution Network-Constrained Optimization of Peer-to-Peer Transactive Energy Trading among Multi-Microgrids," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1033–1047, Mar. 2021.

[98] W. Zhong, S. Xie, K. Xie, Q. Yang, and L. Xie, "Cooperative P2P Energy Trading in Active Distribution Networks: An MILP-Based Nash Bargaining Solution," *IEEE Trans. Smart Grid*, vol. 12, no. 2, pp. 1264–1276, Mar. 2021.

[99] K. Anoh, S. Maharjan, A. Ikpehai, Y. Zhang, and B. Adebisi, "Energy Peer-to-Peer Trading in Virtual Microgrids in Smart Grids: A Game-Theoretic Approach," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1264–1275, Mar. 2020.

[100] S. Cui, Y. W. Wang, Y. Shi, and J. W. Xiao, "A New and Fair Peer-to-Peer Energy Sharing Framework for Energy Buildings," *IEEE Trans. Smart Grid*, vol. 11, no. 5, pp. 3817–3826, Sep. 2020.

[101] Y. Chen, W. Wei, F. Liu, and S. Mei, "A multi-lateral trading model for coupled gas-heat-power energy networks," *Appl. Energy*, vol. 200, pp. 180–191, Aug. 2017.

[102] C. Wang, W. Wei, J. Wang, L. Wu, and Y. Liang, "Equilibrium of Interdependent Gas and Electricity Markets with Marginal Price Based Bilateral Energy Trading," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4854–4867, Sep. 2018.

[103] P. Jiang, S. Lu, W. Gu, S. Yao, R. Bo, C. Wu, and Z. Wu, "A Two-Stage Game Model for Combined Heat and Power Trading Market," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 506–517, Jan. 2019.

[104] D. Xu, B. Zhou, N. Liu, Q. Wu, N. Voropai, C. Li, and E. Barakhtenko, "Peer-to-Peer Multienergy and Communication Resource Trading for Interconnected Microgrids," *IEEE Trans. Ind. Informatics*, vol. 17, no. 4, pp. 2522–2533, Apr. 2021.

[105] T. Morstyn, A. Teytelboym, C. Hepburn, and M. D. McCulloch, "Integrating P2P Energy Trading with Probabilistic Distribution Locational Marginal Pricing," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3095–3106, Jul. 2020.

[106] Z. Li and T. Ma, "Peer-to-peer electricity trading in grid-connected residential communities with household distributed photovoltaic," *Appl. Energy*, vol. 278, pp. 1–13, Nov. 2020.

[107] F. Zeng, Z. Bie, S. Liu, C. Yan, and G. Li, "Trading Model Combining Electricity, Heating, and Cooling Under Multi-energy Demand Response," *J. Mod. Power Syst. Clean Energy*, vol. 8, no. 1, pp. 133–141, Jan. 2020.

[108] T. Morstyn and M. D. McCulloch, "Multiclass energy management for peer-to-peer energy trading driven by prosumer preferences," *IEEE Trans. Power Syst.*, vol. 34, no. 5, pp. 4005–4014, Sep. 2019.

[109] M. Khorasany, Y. Mishra, and G. Ledwich, "A decentralized bilateral energy trading system for peer-to-peer electricity markets," *IEEE Trans. Ind. Electron.*, vol. 67, no. 6, pp. 4646–4657, Jun. 2020.

[110] Z. Guo, P. Pinson, S. Chen, Q. Yang, and Z. Yang, "Chance-constrained peer-to-peer joint energy and reserve market considering renewable generation uncertainty," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 798–809, Jan. 2021.

[111] T. Chen and S. Bu, "Realistic Peer-to-Peer Energy Trading Model for Microgrids using Deep Reinforcement Learning," in *Proc. IEEE PES Innov. Smart Grid Technol. Eur. (ISGT-Europe'2019)*, Bucharest, Romania, pp. 1–5.

[112] X. Lu, X. Xiao, L. Xiao, C. Dai, M. Peng, and H. V. Poor, "Reinforcement Learning-Based Microgrid Energy Trading With a Reduced Power Plant Schedule," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 728–10 737, Dec. 2019.

[113] S. Thrun and A. Schwartz, "Issues in Using Function Approximation for Reinforcement Learning," in *Proc. 4th Connect. Model. Summer. 1993*, Hillsdale, New Jersey, pp. 255–263.

[114] E. A. M. Cesena, N. Good, A. L. Syrri, and P. Mancarella, "Techno-economic and business case assessment of multi-energy microgrids with co-optimization of energy, reserve and reliability services," *Appl. energy*, vol. 210, pp. 896–913, Jan. 2018.

[115] C. Li, Y. Xu, X. Yu, C. Ryan, and T. Huang, "Risk-averse energy trading in multi-energy microgrids: A two-stage stochastic game approach," *IEEE Trans. Industr. Inform.*, vol. 13, no. 5, pp. 2620–2630, Oct. 2017.

[116] N. Liu, J. Wang, and L. Wang, "Hybrid energy sharing for multiple microgrids in an integrated heat–electricity energy system," *IEEE Trans. Sustain. Energy*, vol. 10, no. 3, pp. 1139–1151, Jul. 2019.

[117] D. Xu, B. Zhou, K. W. Chan, C. Li, Q. Wu, B. Chen, and S. Xia, "Distributed multi-energy coordination of multimicrogrids with biogas-solar-wind renewables," *IEEE Trans. Industr. Informat.*, vol. 15, no. 6, pp. 3254–3266, Jun. 2019.

[118] D. Xu, Q. Wu, B. Zhou, C. Li, L. Bai, and S. Huang, "Distributed multi-energy operation of coupled electricity, heating, and natural gas networks," *IEEE Trans. Sustain. Energy*, vol. 11, no. 4, pp. 2457–2469, Oct. 2020.

[119] W. Gu, S. Lu, Z. Wu, X. Zhang, J. Zhou, B. Zhao, and J. Wang, "Residential cchp microgrid with load aggregator: Operation mode, pricing strategy, and optimal dispatch," *Appl. Energy*, vol. 205, pp. 173–186, Nov. 2017.

[120] D. Friedman, *The double auction market: institutions, theories, and evidence*. Routledge, March 2018.

[121] P. Vytelingum, D. Cliff, and N. R. Jennings, "Strategic bidding in continuous double auctions," *Artif. Intell.*, vol. 172, no. 14, pp. 1700–1729, Sep. 2008.

[122] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[123] D. Qiu, Z. Dong, X. Zhang, Y. Wang, and G. Strbac, "Safe reinforcement learning for real-time automatic control in a smart energy-hub," *Appl. Energy*, vol. 309, p. 118403, Mar. 2022.

[124] T. Zhang, D. Yue, and N. Zhao, "Energy optimization management of multi-microgrid using deep reinforcement learning," in *Proc. Chin. Autom. Congr. (CAC)*. IEEE, 2020, pp. 4049–4053.

[125] Y. Xu, L. Yu, G. Bi, M. Zhang, and C. Shen, "Deep reinforcement learning and blockchain for peer-to-peer energy trading among microgrids," in *Proc. Int. Conf. Internet of Things (iThings)*. IEEE, 2020, pp. 360–365.

[126] J. Ackermann, V. Gabler, T. Osa, and M. Sugiyama, "Reducing overestimation bias in multi-agent domains using double centralized critics," *arXiv preprint arXiv:1910.01465*, 2019.

[127] S. Han, S. Han, and H. Aki, "A practical battery wear model for electric vehicle charging applications," *Appl. Energy*, vol. 113, pp. 1100–1108, 2014.

[128] V. Francois-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, *An Introduction to Deep Reinforcement Learning*, 2018.

[129] Pecan Street Inc. (2018) Dataport. [Online]. Available: https://www.pecanstreet.org/dataport/

[130] ISO New England. (2018) Electricity price. [Online]. Available: https://www.iso-ne.com/

[131] P. Mancarella, "MES (multi-energy systems): An overview of concepts and evaluation models," *Energy*, vol. 65, pp. 1–17, Feb. 2014.

[132] R. Khurmi and R. Sedha, *Materials Science*, 5th ed. New Delhi: S. Chand & Company Ltd, 2014.

[133] G. Glenk and S. Reichelstein, "Economics of converting renewable power to hydrogen," *Nat. Energy*, vol. 4, no. 3, pp. 216–222, Feb. 2019.

[134] R. Dixon, J. Li, and M. Wang, "Progress in hydrogen energy infrastructure development—addressing technical and institutional barriers," in *Compendium of Hydrogen Energy*. Elsevier, 2016, pp. 323–343.

[135] GOV.UK. (2021) £166 million cash injection for green technology and 60,000 UK jobs. [Online]. Available: https://www.gov.uk/government/news/166-million-cash-injection-for-green-technology-and-60000-uk-jobs

[136] Energy Digital. (2020) Ofgem approves SGN hydrogen homes project in Scotland. [Online]. Available: https://energydigital.com/oil-and-gas/ofgem-approves-sgn-hydrogen-homes-project-scotland

[137] A. Züttel, "Hydrogen storage methods," *Naturwissenschaften*, vol. 91, no. 4, pp. 157–172, Mar. 2004.

[138] S. M. Saba, M. Müller, M. Robinius, and D. Stolten, "The investment costs of electrolysis–a comparison of cost studies from the past 30 years," *Int. J. Hydrog. Energy*, vol. 43, no. 3, pp. 1209–1223, 2018.

[139] M. Schalenbach, G. Tjarks, M. Carmo, W. Lueke, M. Mueller, and D. Stolten, "Acidic or alkaline? towards a new perspective on the efficiency of water electrolysis," *J. Electrochem. Soc.*, vol. 163, no. 11, p. F3197, Aug. 2016.

[140] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. NY: Springer, 2012.

[141] H. Demsetz, "The cost of transacting," *Q. J. Econ.*, vol. 82, no. 1, pp. 33–53, Feb. 1968.

[142] Y. Wu, X. Tan, L. Qian, D. H. K. Tsang, W.-Z. Song, and L. Yu, "Optimal pricing and energy scheduling for hybrid energy trading market in future smart grid," *IEEE Trans Industr Inform*, vol. 11, no. 6, pp. 1585–1596, Dec. 2015.

[143] S. Sharma, A. Verma, Y. Xu, and B. K. Panigrahi, "Robustly Coordinated Bi-level Energy Management of a Multi-Energy Building under Multiple Uncertainties," *IEEE Trans. Sustain. Energy*, vol. 12, no. 1, pp. 3–13, Jan. 2021.

[144] H. Li, Z. Wan, and H. He, "Constrained EV Charging Scheduling Based on Safe Deep Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, May 2020.

[145] Y. Jiang, C. Wan, C. Chen, M. Shahidehpour, and Y. Song, "A Hybrid Stochastic-Interval Operation Strategy for Multi-Energy Microgrids," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 440–456, Jan. 2020.

[146] H. V. Hasselt, "Double Q-learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS'2010)*, Vancouver, Canada, pp. 2613–2621.

[147] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[148] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," in *Proc. 33th Int. Conf. Mach. Learn. (ICML'2016)*, New York City, NY, USA, pp. 1329–1338.

[149] Office of Energy Efficiency & Renewable Energy. (2011) Commercial and Residential Hourly Load Profiles for all TMY3 Locations in the United States. [Online]. Available: https://openei.org/datasets/dataset/

[150] RWTH Aachen University. (2014) Smart Energy Data: Aachen/ Cologne Virtual Power Plant. [Online]. Available: https://data.lab.fiware.org/organization/rwth-aachen-university

[151] U.S. Energy Information Administration. (2018) United States Natural Gas Industrial Price. [Online]. Available: https://www.eia.gov/dnav/ng/hist/n3035us3m.htm

[152] ——. (2018) U.S. Energy-Related Carbon Dioxide Emissions. [Online]. Available: https://www.eia.gov/environment/emissions/carbon/

[153] H. R. Ellamla, I. Staffell, P. Bujlo, B. G. Pollet, and S. Pasupathi, "Current status of fuel cell based combined heat and power systems for residential sector," *J. Power Sources*, vol. 293, pp. 312–328, Oct. 2015.

[154] G. Strbac, *et al.*, "Cost-effective decarbonization in a decentralized market: The benefits of using flexible technologies and resources," *IEEE Power Energy Mag.*, vol. 17, no. 2, pp. 25–36, Mar.-Apr. 2019.

[155] MIT Energy Initiative, *et al.*, "Managing large-scale penetration of intermittent renewables," Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, Tech. Rep., Apr. 2011.

[156] N. Lidula and A. Rajapakse, "Microgrids research: A review of experimental microgrids and test systems," *Renew. Sustain. Energy Rev.*, vol. 15, no. 1, pp. 186–202, Jan. 2011.

[157] Z. Li and Y. Xu, "Optimal coordinated energy dispatch of a multi-energy microgrid in grid-connected and islanded modes," *Appl. Energy*, vol. 210, pp. 974–986, Jan. 2018.

[158] P. W. MacAvoy, *The natural gas market*. Yale University Press, 2008.

[159] T. Couture and Y. Gagnon, "An analysis of feed-in tariff remuneration models: Implications for renewable energy investment," *Energy policy*, vol. 38, no. 2, pp. 955–965, Feb. 2010.

[160] J. Yang, J. Zhao, F. Luo, F. Wen, and Z. Y. Dong, "Decision-making for electricity retailers: A brief survey," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4140–4153, Sept. 2017.

[161] T. Morstyn, N. Farrell, S. J. Darby, and M. D. McCulloch, "Using peer-to-peer energy-trading platforms to incentivize prosumers to form federated power plants," *Nat. Energy*, vol. 3, no. 2, pp. 94–101, Feb. 2018.

[162] A. Kaur, J. Kaushal, and P. Basak, "A review on microgrid central controller," *Renew. Sustain. Energy Rev.*, vol. 55, pp. 338–345, Mar. 2016.

[163] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[164] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[165] W. Huang, N. Zhang, J. Yang, Y. Wang, and C. Kang, "Optimal configuration planning of multi-energy systems considering distributed renewable energy," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1452–1464, Mar. 2019.

[166] Salt River Project, "Srp time-of-use price plan." [Online]. Available: https://srpnet.com/prices/home/tou.aspx

[167] Organisation for Economic Co-operation and Development, "Renewable energy feed-in tariffs." [Online]. Available: https://stats.oecd.org/Index.aspx?DataSetCode=RE_FIT#

[168] S. M. LaValle, M. S. Branicky, and S. R. Lindemann, "On the relationship between classical grid search and probabilistic roadmaps," *Int. J. Rob. Res.*, vol. 23, no. 7-8, pp. 673–692, Aug. 2004.

[169] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.