Cost and Power Loss Aware Coalitions under Uncertainty in Transactive Energy Systems

by

Mohammad Sadeghi

Thesis submitted to the University of Ottawa
in partial fulfillment of the thesis requirement for the degree of

Doctor of Philosophy in Electrical and Computer Engineering

Abstract

The need to cope with the rapid transformation of the conventional electrical grid into the future smart grid, with multiple connected microgrids, has led to the investigation of optimal smart grid architectures. The main components of the future smart grids such as generators, substations, controllers, smart meters and collector nodes are evolving; however, truly effective integration of these elements into the microgrid context to guarantee intelligent and dynamic functionality across the whole smart grid remains an open issue. Energy trading is a significant part of this integration.

In microgrids, energy trading refers to the use of surplus energy in one microgrid to satisfy the demand of another microgrid or a group of microgrids that form a microgrid community. Different techniques are employed to manage the energy trading process such as optimization-based and conventional game-theoretical methods, which bring about several challenges including complexity, scalability and ability to learn dynamic environments. A common challenge among all of these methods is adapting to changing circumstances. Optimization methods, for example, show promising performance in static scenarios where the optimal solution is achieved for a specific snapshot of the system. However, to use such a technique in a dynamic environment, finding the optimal solutions for all the time slots is needed, which imposes a significant complexity. Challenges such as this can be best addressed using game theory techniques empowered with machine learning methods across grid infrastructure and microgrid communities.

In this thesis, novel Bayesian coalitional game theory-based and Bayesian reinforcement learning-based coalition formation algorithms are proposed, which allow the microgrids to exchange energy with their coalition members while minimizing the associated cost and power loss. In addition, a deep reinforcement learning scheme is developed to address the problem of large convergence time resulting from the sizeable state-action space of the methods mentioned above. The proposed algorithms can ideally overcome the uncertainty in the system. The advantages of the proposed methods are highlighted by comparing them with the conventional coalitional game theory-based techniques, Q-learning-based technique, random coalition formation, as well as with the case with no coalitions. The results show the superiority of the proposed methods in terms of power loss and cost minimization in dynamic environments.

To
my beloved parents

Acknowledgements

First and foremost, I am incredibly grateful to my supervisor, Professor Melike Erol-Kantarci, for her invaluable advice, continuous support and patience during my Ph.D. study in Canada. Her immense knowledge and great experience guided me through this research. This thesis would not have been possible without her help, support and guidance.

I would like to express gratitude to Dr. Shahram Mollahasani for his insightful comments during our research collaboration, which is an essential contribution to this thesis. I would also like to thank the committee members for providing precious feedback.

I would like to thank my colleagues in the NETCORE lab for a cherished time spent together in the lab and in social settings. My appreciation also goes out to my family and friends for their encouragement and support throughout my studies.

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the university's regulations concerning plagiarism, including those concerning consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Table of Contents

List of Tables

List of Figures

Abbreviations

AI Artificial Intelligence 34

BCG Bayesian Coalitional Game Theory 2

BCRL Bayesian Coalitional Reinforcement Learning 80

BEB Bayesian Exploration Bonus 73

BRL Bayesian Reinforcement Learning 2

CG Coalitional Game Theory 1

CVaR Conditional Value at Risk 38

DQN Deep Q-Network 27

DQN-CF DQN-Based Coalition Formation 104

DRL Deep Reinforcement Learning 2

EVs Electric Vehicles 2

FME Fully Myopic Estimation 90

G2V Grid to vehicle 18

GPS Global Positioning System 49

ICT Information and Communication Technologies 6

LSTM Long Short Term Memory 31

MAPE Maximum a Posterior Estimation 90

MDP Markov Decision Process 23

MGs Microgrids 1

PDF Probability Density Function 38

POMDP Partially Observable MDP 24

SBC Strong Bayesian Core 22

SP Solar Panel 8

V2G Vehicle to Grid 18

WBC Weak Bayesian Core 22

WT Wind Turbine 9

List of Symbols

C Random variable for coalition

v Coalition value

$\mathbf{M}$ Set of players

$\mathbf{S}$ Vector of all possible states

$\mathbf{A}$ Vector of all possible actions

B Belief function

$\mathbf{T}$ Types vector of all the players

$\overline{\mathbf{u}}$ Vector of the expected payoffs of all players

$\mathbb{T}$ Set of all possible players types vectors

$\mathbf{T}^C$ Types vector of members of coalition C

$\mathbb{T}^C$ Set of all possible types vectors of members of coalition C

u^i Immediate payoff of the i-th agent

u^i_j Immediate payoff of player j realized by player i

H Transition function

a Action

s State

r Reward function

τ Iteration step

γ Discount factor

σ Policy of agent

$\mathbf{\Omega}$ Observation space of POMDP

O Observation function

o Observation

σ^* Optimal policy

B_a^o Updated belief

$\mathbf{x}$ Input feature vector of a neural network

$\mathbf{y}$ Output vector of a neural network

$\mathbf{w}$ Weight vector of a neural network

$(z(\mathbf{x}))$ Pre-activation function

$\psi(\mathbf{x})$ Neuron activation function

b Bias value

F Output of the neuron

M Number of agents

M_{MG} Number of MGs

M_{EV} Number of EVs

M_{CS} Number of charging stations

e^g Generate power of MG

e^d Power demand of MG/EV

e^b Battery capacity of EV

e^q Equivalent surplus/demand of MG/EV

PL Power loss

E_{ij} Power flow from MG i to j

$\mathbb{R}_{ij}$ Resistance of the line per km

U_i Voltage

D_{ij} Distance between MG i and j

ρ Transformer loss coefficient

ϵ_i Probability of traveling to the second station

p_e Probability of observation error when a fixed behavior of an agent observed as moving

p_c Probability of observation error when a moving behavior of an agent observed as fixed

T_{fixed} Agent of type fixed

T_{moving} Agent of type moving

$\mathbf{T}_k^C$ k-th possible types vector of members of coalition C

$\wp$ Prior probability

ζ_i The relative ratio of their contribution

ϖ Adjustable constant

χ Counting parameter

S_{ij} Cost of transferring power

ω Weighting coefficient of virtual cost

ω_s Weighting coefficient of virtual cost in short distance energy trading

ω_l Weighting coefficient of virtual cost in long distance energy trading

ω_0 Weighting coefficient of virtual cost in energy trading with macrogrid

δ Scaling factor

D_{tr} Threshold distance

T_C Player type committed

T_{UC} Player type uncommitted

η Belief parameter of uncommitted player

κ Belief parameter of uncommitted player

ν Belief parameter of uncommitted player

O_C Committed observation

O_{UC} Uncommitted observation

p_{e_1} Probability of observation error when a full commitment behaviour is observed as half commitment

p_{e_2} Probability of observation error when a full commitment behaviour is observed as zero commitment

p_{c_1} Probability of observation error when a half commitment behaviour is observed as full commitment

p_{c_2} Probability of observation error when a zero commitment behaviour is observed as full commitment

ϕ Battery level

Φ Set of possible battery levels

ε Power level

Ξ Set of possible power levels

C_k k-th coalition

β Bonus weight in BEB

$\mathcal{T}$ Time upper bound

$\boldsymbol{\lambda}_t$ Counting parameter

T^m Agent type

$\mathbf{T}^{-m}$ Belief vector about the type of other agents

B^m Agent belief

$\mathbf{A}^{C_k}$ Set of all possible coalitional action

$B^m(T^{-m})$ m-th MG's beliefs about the types of other players

θ Demand from value

$\boldsymbol{\theta}^{C_k}$ Demand vector of coalition C_k

π_m^k Proposal by prosper m

λ^m Counting parameter

Chapter 1

Introduction

1.1 Motivation

Recently, centralized-unidirectional systems of electric power transmission, and distribution and demand-driven control systems have been gradually evolving into a massive heterogeneous mix of utility grids and Microgrids (MGs), along with renewables being integrated into energy generation mix and residential demand-response for smart homes [1, 2]. These groundbreaking technologies call for higher quality and reliability in the electricity distribution system.

Power grids have undergone significant transformations since the mid-2000s, and they continue to evolve [1]. Transactive energy systems are receiving particular emphasis as part of the future smart grid. A transactive distribution system is composed of several MGs (e.g., buildings, homes, solar farms) with a generation capacity that can satisfy part of the demand of other MGs by allowing peer-to-peer energy trading through a transactive energy market and using an underlying communication technology [3]. An MG is a small-scale electricity distribution system with loads, generation capacity, storage, and islanding capability.

The early form of MGs emerged in the military [4]. A network of MGs has been initially proposed in [5], and more recently, community MGs, which focus more on collaboration rather than market-based interactions, have emerged to serve communities during disasters and to increase the reliability of the smart grid [6]. Meanwhile, MG communities using Coalitional Game Theory (CG), with the objective of minimizing power loss, have been first explored in [7].

Despite the advances towards a smart power grid, the emergence of Electric Vehicles (EVs) and their ability to charge from and discharge to the distribution system impose challenges on the smart grid end. Furthermore, automotive companies are extensively competing to bring self-driving or autonomous cars to the market, implying that autonomous EVs will be a dominant element of future smart cities. Consequently, it is essential to investigate the effects of EVs on the power grid and specifically on transactive energy systems and MGs. High penetration of EVs can affect the operation of MGs in different aspects due to their uncertain behaviors, such as imposing unexpected load which may accumulate in peak hours or uncertainty about which MG they will be charging/discharging from when using public charging stations. Also, the uncertainty of an entity type, whether an MG (fixed) or an EV (mobile), can impact decisions. Therefore, it is crucial to consider the uncertainty inflicted by EVs.

While the main elements of the future smart grid, such as power generators, electric power substations, controllers, smart meters, and collector nodes are well defined, their truly harmonic integration requires intelligent functionality at the distribution system level. To effectively deal with the swift transition from the legacy grid to the future transactive energy systems, with MG communities and peer-to-peer energy trading, intelligence from machine learning and game theory needs to be incorporated.

1.2 Contribution and Summary of Chapters

This thesis aims to study the potential of machine learning, in particular reinforcement learning methods combined with CG-based methods, to address energy trading problems among MGs under uncertainty. We proposed Bayesian Coalitional Game Theory (BCG)-based, Bayesian Reinforcement Learning (BRL)-based and Deep Reinforcement Learning (DRL)-based methods to address the coalition formation challenge in energy trading problems considering the uncertainties in the system. The contribution of this thesis can be highlighted in the following lines.

- **Power Loss and Cost Efficiency:** Among the many advantages of energy trading between MGs, power loss minimization and cost-efficiency of energy transfer are two of the most significant benefits. Power is lost in the transmission and distribution system due to heating dissipation in power lines. The amount of loss and the total cost of energy transfer are proportional to distance and the amount of transmitted power. Any power loss results in less revenue and imposes further costs on the utility side or buyer MGs. Therefore, minimizing the distance between interconnections reduces

power loss and consequently the cost, which can be realized by interconnecting several MGs and allowing peer-to-peer energy trading between them. In this thesis, the problem of energy trading among MGs is addressed with the aim to minimize the cost in chapter 4, chapter 6 and chapter 7, and minimize power loss over the line during energy transactions in chapter 3 and chapter 5.

- **Modeling the Uncertainties in the System:** Most of the studies in the energy trading domain consider a static environment, which can be perfectly investigated using optimization and conventional game-theoretical methods. However, considering uncertainties in such dynamic environments is crucial for designing real-world scenarios. In this thesis, we consider uncertainty in different aspects of the system. Particularly, we consider dynamic demand and generation for each of the MG units in the system. We consider future generation and demand misprediction as a source of uncertainty in chapter 4. In chapter chapter 3, penetration of EVs and their mobility is modeled as a source of uncertainty. We also consider communication errors in chapter 3 and chapter 4.

- **BCG-Based and BRL-Based Methods:** While optimization methods and conventional coalition formation methods can perfectly perform in small-scale static scenarios, using these methods in a large-scale dynamic environment imposes several difficulties such as complexity, scalability and inability to adapt with dynamic environments. To cope with these problems, we propose BCG-based and BRL-based methods in chapter 3 to chapter 6 in order to overcome the uncertainty and reach an autonomous distributed system.

- **Convergence and Complexity:** To address the problem of convergence and complexity of action-state space, we introduced a DRL-based method in chapter 7.

1.2.1 Publications

[**B01**] **M. Sadeghi**, M. Erol-Kantarci and H. T. Mouftah, "Connected and Autonomous Electric Vehicle Charging Infrastructure Integration to Microgrids in Future Smart Cities," Chapter in *Connected and Autonomous Vehicles in Smart Cities*, pp. 1-17. CRC Press, 2020.

[**J02**] **M. Sadeghi**, S. Mollahasani and M. Erol-Kantarci, "Cost-Optimized Microgrid Coalitions Using Bayesian Reinforcement Learning," *Energies* 14, no. 22 (2021): 7481.

[**J01**] **M. Sadeghi**, S. Mollahasani and M. Erol-Kantarci, "Power Loss-Aware Transactive Microgrid Coalitions under Uncertainty," *Energies* 13, no. 21 (2020): 5782.

[**C03**] **M. Sadeghi** and M. Erol-Kantarci, "Deep Reinforcement Learning Based Coalition Formation for Energy Trading in Smart Grid," *2021 IEEE 4th 5G World Forum (5GWF)*, 2021, pp. 200-205.

[**C02**] **M. Sadeghi**, S. Mollahasani and M. Erol-Kantarci, "Cost-Aware Dynamic Bayesian Coalitional Game for Energy Trading among Microgrids," *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021, pp. 1-6.

[**C01**] **M. Sadeghi** and M. Erol-Kantarci, "Power Loss Minimization in Microgrids Using Bayesian Reinforcement Learning with Coalition Formation," *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2019, pp. 1-6.

1.2.2 Organization of Thesis

The rest of this thesis is organized as follows.

Chapter 2- Background and Related Works: We provide comprehensive background information on MGs. Then we explain the algorithms and techniques used in this thesis. Finally, we provide a literature review on the energy trading problem among MGs under uncertainty.

Chapter 3- Power Loss-Aware Transactive MG Coalitions under Uncertainty: We propose a BCG approach to form coalitions and effectively address uncertainty in the varying charging/discharging locations of EVs. In our approach, each MG and EV agent assumes a prior belief function over the type of other agents (either fixed MG or moving EV). As the agents interact through the iterations of BCG, they update their belief estimations, which eventually result in a coalition formation that minimizes power loss. The contribution of this chapter is formulating the uncertainty arising from agents' location and including observation error in BCG.

Chapter 4- Cost-Aware Dynamic BCG for Energy Trading among MGs: We consider effects of uncertainty resulting from the misprediction in future generation and demand, which makes MGs refuse to deliver or accept the energy they committed to provide in the coalition. A BCG-based method is proposed to establish coalitions while addressing the commitment uncertainty effectively. The contribution of this chapter can be summarized as formulating uncertainty based on MGs type and considering observation error. The results show improvement in terms of cost minimization compared to the conventional CG-based and Q-learning-based approaches.

Chapter 5- Power Loss Minimization in MGs Using BRL with Coalition Formation: We propose a BRL-based approach to form coalitions and effectively address uncertainty in generation and demand. We assume that MGs are equipped with a battery to store energy. Our results demonstrate a significant reduction in power loss compared to a case with no coalitions and improvement over other CG-based and learning techniques.

Chapter 6- Cost-Optimized MG Coalitions Using BRL: Chapter 6 investigates the energy trading problem among MGs, where each MG has different levels of energy surplus or demand in each epoch. The dynamic nature of energy levels causes uncertainty in our system. In this chapter, we improve BCG-based and BRL-based methods presented in previous chapters by proposing a BRL-based coalition formation scheme for energy trading among MGs. Through the course of interactions with the environment, the proposed method help agents gradually update their belief about the type of other agents and learn from past experiences to maximize the long-term expected reward. We compared the proposed method with two BRL-based models, and Q-learning, BCG and conventional CG-based approaches.

Chapter 7- DRL-Based Coalition Formation for Energy Trading in Smart Grid: We propose a DRL-based approach to address the problem of long convergence time and scalability in conventional reinforcement learning methods. The proposed algorithm helps form coalitions that effectively overcome uncertainty in generation and demand in energy trading problems among MGs. Our results show improvement in cost minimization compared with an existing Q-learning-based scheme and a conventional CG-based approach.

Chapter 8- Future Directions and Conclusion: Finally, we provide the future directions and we conclude the thesis.

Chapter 2

Background and Related Works

2.1 Introduction

In this chapter, we provide an overview of the concept of MG and explain the analytical methods used in this thesis. In section 2.2, MG infrastructure, renewable energy generation resources of MGs, effects of integrating EVs on the operation and management of MGs, and impacts of shifting to more distributed control systems, are explained. In Section 2.3, BCG-based, BRL-based and DRL-based methods that are adopted in this thesis are explained. In Section 2.4, the chapter focuses on the energy trading problem among MGs under uncertainty and investigates studies that address such problems. Finally, the chapter concludes by identifying research gaps.

2.2 Background

Current cities are experiencing drastic transformation due to the quick incorporation of technology in the daily life of residents. Therefore, establishing a framework to unify all aspects of daily life in a city, as well as digitalizing services and embedding intelligence into their functions, are crucial. Smart cities are urban areas that widely employ Information and Communication Technologies (ICT) such as different types of sensors which collect data in order to manage resources and improve the quality of life in the city [8–11].

Following developments in smart cities, the power grid, as an essential infrastructure supporting cities, has been undergoing major changes since the mid-2000s. The main

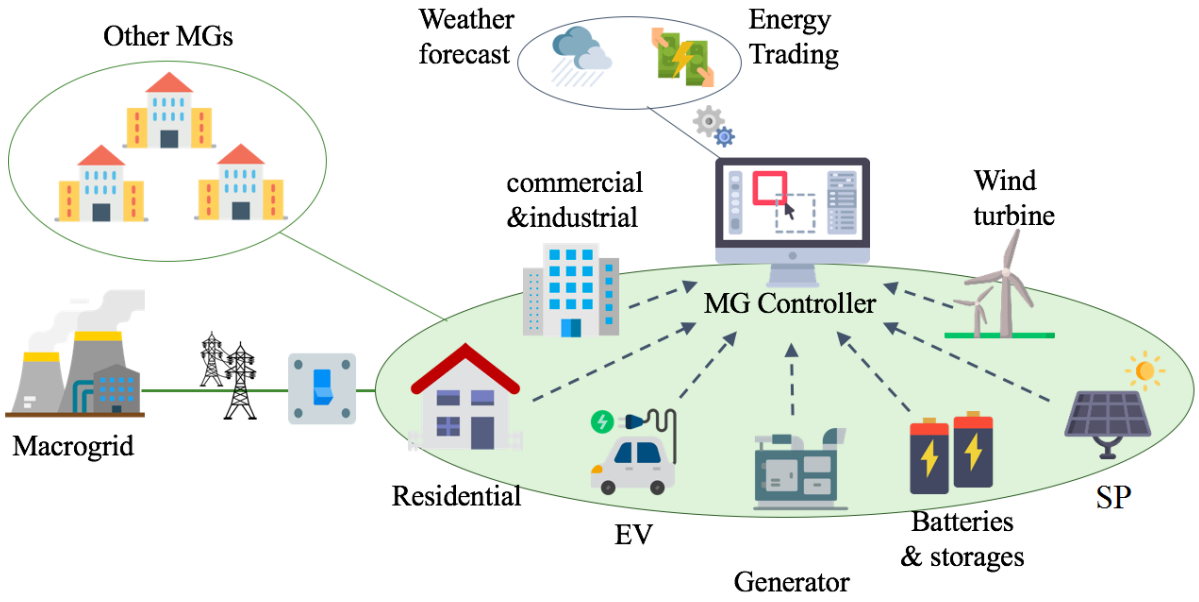

Figure 2.1: MG block diagram.

drive for the changes in the power grid has been the desire to make power generation less dependent on fossil fuels. This requires increased use of renewables, which are intermittent and hence call for innovations in storage technologies. On the other hand, lower consumption or demand response became another important area contributing to lowered peak-hour electricity consumption. MGs can play an essential role in bringing the energy resources (primarily renewable energy) to where it is needed. Implementing MG structure in smart cities may result in efficient generation of renewable energy, less power loss in the grid and optimized load regulation, which all result in less consumption of fossil fuels and reduced gas emissions [12, 13]. Although military MGs existed for several decades [14], their co-existence within the commercial distribution system has only recently become a feasible idea. In the next section, we provide a detailed overview of the state-of-the-art developments in MGs research.

2.2.1 Microgrids

There are various definitions of an MG in the literature [15, 16]; one of the most cited ones is provided by the U.S. Department of Energy as *"a group of interconnected loads and distributed energy resources within clearly defined electrical boundaries that acts as a single controllable entity. An MG can connect and disconnect from the grid to enable it to*

operate in both grid-connected or island mode" [17]. According to this definition, an MG needs to be a) distinguishable from the rest of the grid system as an independent unit; b) should include resources of energy that help to not rely on distant resources and sustain as a single unit; c) should be able to run effectively in case of losing connection to the main grid. The general block diagram of an MG system is demonstrated in Fig. 2.1.

Several MG implementations already exist around the world [18]. Santa Rita Jail is a real implementation example of campus/institutional MG [19]. In [20], the implementation of a military-based MG is presented.

Although it is possible to construct an MG without renewable energy resources, the true potential of opportunities for future MGs arises from integrating such resources. Therefore, renewable energy resources are important components of MGs. In the next section, we overview the widely used renewable energy generators.

2.2.2 Renewable Energy Generation Resources

There are no strict criteria in the design of MGs, and the design process depends on the specific requirements of the project and economic concerns. There is a broad selection of energy generation resources and storage that can be implemented in the design of MGs. Diesel engines, microturbines, fuel cells such as solid oxide and alkaline, and renewable generation resources are among generator candidates. In contrast, sodium-sulfur and lithium-ion batteries, flow batteries such as zinc-bromine and polysulphide bromide batteries, hydrogen from hydrolysis and kinetic energy storage can be suitable storage choices.

Among generation options, renewable resources have recently gained remarkable attention due to the following reasons:

1. Low carbon emission;
2. Less dependence on fossil fuels;
3. Longer lifespan compared to conventional resources;
4. Noise-free in the case of Solar Panel (SP);
5. Low operational cost.

Although these sources can be employed effectively in the MG, the deployment of these resources has some disadvantages, which are summarized in the following:

1. Currently, the installation cost of renewable generators is significantly higher than conventional resources. For example, the installation cost of an SP can be up to 10 times higher than that of a diesel generator.
2. The implementation of renewable resources is geographically limited.
3. Renewable resources are less energy efficient compared to diesel generators.

Despite the downsides mentioned above, renewable resources are in the stage of development, and research shows a promising future in the advancement of these resources. Different types of renewable resources can be deployed in an MG; these may be categorized as the following major types:

- SP
- Wind Turbine (WT)
- Mini-hydro

In the following, these renewable generator categories are discussed in more detail.

SP

The idea of SP generation is to generate electrical energy from free and limitless solar energy. The efficiency of SPs is impacted not only by environmental parameters such as geographical location, the intensity of solar radiation, temperature and cloud obstruction, but also system parameters such as the performance of employed SP modules, the efficiency of converters, inverter and the adopted control scheme.

Variation in irradiance and cloud obstruction patterns can crucially impose voltage disturbances, resulting in the disconnection of the inverters from the grid and consequently loss of energy. Today's SPs may demonstrate low efficiency in the long term due to cloud coverage and fluctuation in solar irradiance intensity. Studies show that SP systems perform 5 to 10 times less efficiently compared to diesel generators [21]. Instead, SPs have a very long lifetime, which sometimes expands to about 20 to 25 years, and the efficiency only drops to 80 percent after this period.

Numerous MG testbeds have employed SPs [14, 22–25]. The Nice Grid project in Nice, France, is an example of a successful smart SP-based MG [25] since solar power is highly

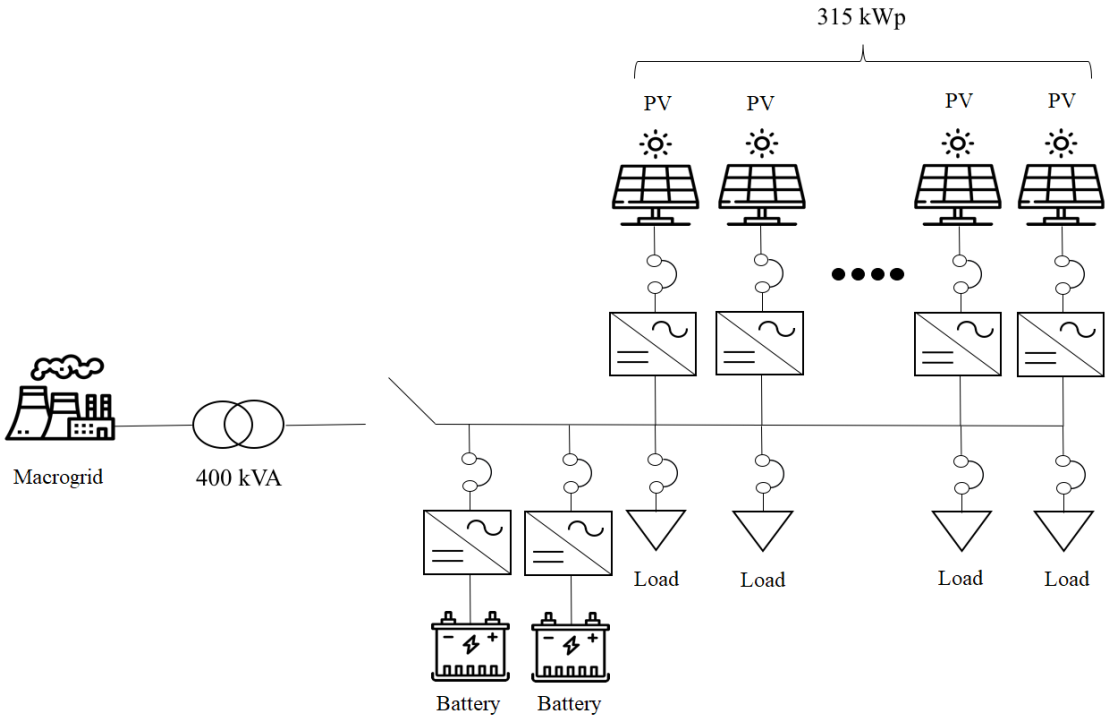

Figure 2.2: Continuon project structure.

available where it is located. The MG features 2.5 MW of SP generation capacity with an energy storage capacity of 1.5 MW, while the interruptible demand is 3.5 MW. One of the project's objectives has been to optimize the distribution grid management, considering the high deployment of SPs. Full functionality in the islanded mode, which makes MG only rely on the SPs, has been another objective of this project. Different finance models have also been studied to explore the feasibility of involving the MG in the energy market. The project is co-funded by Grid4EU and the French government. In [25], an MG project in Bronsbergen, Netherlands, known as Continuon is discussed. The project includes 109 cottages with SP-covered roofs, a battery energy storage and a control system developed to manage the islanding mode. The block diagram of this MG is depicted in Fig. 2.2.

WT

A WT generates electrical energy by converting wind energy into electricity. The WT structure consists of two essential parts: electrical and mechanical. The motion energy of wind is captured in the form of rotational energy in the mechanical part, which is converted

into electrical energy in the electrical part. The WT includes three main components: the tower, the rotor and the nacelle. The nacelle is equipped with an electrical generator and mechanical power transmission parts. The rotor normally includes more than two blades that extract motion energy from wind. A gearbox is used to transfer the captured rotational energy to the shaft of an electrical generator to generate electric power. WTs can be categorized into two categories: vertical axis and horizontal axis. Vertical axis WTs are common for small units where kW ranges up to 100 kW, while larger units mostly have horizontal axis WTs, which support the order of MW.

The implementation of WTs has been increasing exponentially around the world. More than two-thirds of the installed small WTs are only in the United States and China. The main benefit of employing a WT is its CO_2 free nature. However, the implementation cost is still a significant shortcoming. Inconsistency of wind speed, inability to predict wind speed accurately, occupying vast areas, aesthetics concerns and bird strikes are among other drawbacks.

Several MGs around the world have adopted WTs as a source of energy generation in combination with other sources [14, 26–28]. The MG project Atenea started in 2013, and its main objective has been to support the energy demand for lighting the facility and testbed types of equipment. This MG features 20 kW of WT generation working parallel with 25 kW of SP generation and a 55-kVA diesel generator.

Mini-hydro

Hydropower generation technology takes advantage of the kinetic energy of water which is converted into electrical energy. Hydropower can be categorized as storage, run-of-river, pumped storage and ocean hydropower. Mini-hydro is employed in MG projects which are equipped with a generator with a capacity of up to 10MW. The potential of mini-hydro is estimated to be 173 GW globally, 75 GW of which was achieved in 2012. China, Brazil and the United States are the top countries that deploy hydropower. CO_2-free emission is the main advantage of employing this future-generation technology. However, the installation of dams and water reservoirs occupies an area in the order of thousands of kilometers, which imposes adverse effects on the surrounding environment. Therefore, a comprehensive study on geographical, environmental and hydrological characteristics is crucial before any construction and installation.

Several MG projects have included hydropower generation as the renewable resource in their system [14, 29–31]. Kodiak Island, the second-largest island in the state of Alaska, is a community MG that has been utilizing hydropower since the 1980s [30]. Several WT

farms and a hydro turbine are installed on the island. Currently, the Kodiak Island MG includes 500 kW of hydroelectric, 9 MW of wind generation, and 35 MW of gas/diesel. Additionally, the island is equipped with 2.5 MW/2 MWh of battery storage, which helps to guarantee the stability of the MG system.

2.2.3 EVs and the Effects of Integrating EVs into MGs

With an increasing adoption rate, EVs have recently started to be a part of smart cities, as the convenience of charging facilities is improving. Automotive manufacturers are also making significant advancements towards self-driving or autonomous vehicles, which implies that these will be a part of the future smart cities. Naturally, connectivity will be an essential property of many aspects of such vehicles. In this section, we will focus on the integration of EVs into MGs and their co-existence in future smart cities [32].

A high level of penetration of EVs in the electricity distribution system may affect the operation of MGs due to their uncertain behaviors. Therefore, considering uncertainty as a problem in the design stage is crucial. A summary of different studies on the integration of EVs into MGs and their effects on the operation of MGs is presented in Table 2.1 and discussed with more details in the following.

In [35], the effect of EVs' charging demand on the optimized operation and management of MGs are examined. The authors propose a novel charging scheme and a smart stochastic framework to manage the charging demand of EVs in the residential communities and public charging stations. This framework is evaluated on an MG that includes various renewable energy sources such as microturbine, WT, photovoltaic and fuel cells, equipped with a battery. The authors conclude that with the cost reduction achieved by the presented framework, higher integration of EVs into the MGs system can be supported.

In [36] and [37], the operation problem of integrating EVs to MGs is investigated. The authors present a scheme for co-optimizing energy and power management of EVs with multiple sources. The management scheme is designed at two levels. Dynamically limiting search space and utilizing a meta-heuristic technique are adopted at these two levels, respectively. The results demonstrate a reduction in power loss.

In [38], a dynamic optimal power flow formulation is proposed, which satisfies the security of MGs and industrial constraints while considering EVs' energy-related constraints. The proposed scheme is implemented in an industrial MG that includes 12 factories with combined heat and power systems, an SP generation system equipped with SP storage, and various types of EVs operating in connected and stand-alone modes. The loads and

Table 2.1: The effect of integrating EVs on the operation management of MGs.

Reference	Objective	Model	Renewable resources	Storage
[33]	Optimizing operation management of MGs	stochastic	SP, WT	battery
[34]	Co-optimizing energy and power management of EVs	stochastic	–	battery
[35]	Managing the charging demand	stochastic	SP, WT	–
[36]	Co-optimizing energy and power management of EVs	deterministic	–	battery
[37]	Co-optimizing energy and power management of EVs	deterministic	–	battery
[38]	Minimizing charging cost and regulate voltage profile	deterministic	SP	–
[39]	Optimizing the size of building equipment	deterministic	SP	battery

charging rates of the EVs are optimized to minimize charging costs and regulate voltage profiles.

In [33], the authors evaluate the impact of EV charging demand on the optimized operation management of MGs with different renewable energy sources and battery storage. The authors compare their smart charging strategy with conventional methods through simulation on two test systems, and the results show the superiority of their modified symbiotic organisms search method in solving the optimal operation management problem. Furthermore, they conclude that although the charging demand of EVs results in higher costs for the MG, with the proposed smart charging, the cost impact of charging demand is significantly reduced.

In [40], the optimized operation management of smart MGs and EVs charging infrastructure development is studied. This study shows that although EVs are alluring for smart cities and highly accepted due to their green environmental impacts, the standard to connect EVs to variable-source MGs has not been fully adopted in the industry.

In [39], authors study a case in which interconnected cooperative buildings form an MG exchanging energy among themselves to enhance their independence from the grid and also reduce power loss. In this setting, the proposed solution aims to optimize the power management as well as the capacities of the building equipment. Each building is equipped with SPs, energy storage units and EVs. The load patterns of the buildings, electricity prices and carbon emission taxes, and efficient charging and discharging of the EVs are taken into account in the design of MG configuration. The authors proposed a cooperative game theoretical approach that optimally determines the size of building equipment and significantly reduces the dependency on the main grid.

In [34], a novel scheme for robust optimization of MG power management is proposed, which encourages the buildings in the MG to transfer energy among themselves. The uncertain behavior of EVs is considered, which may impact the MG operation. The results show that the proposed solution achieves significant carbon and cost efficiency.

2.2.4 MG Control Methods in the Presence of EVs

Integrating and controlling EVs in MGs network can be achieved in centralized (Fig. 2.4. a), decentralized (Fig. 2.4. b) or distributed (Fig. 2.4. c) manners. These three control typologies are further described below.

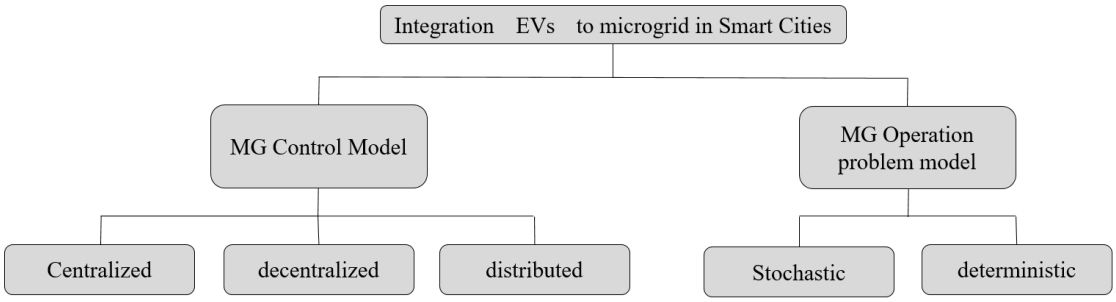

Figure 2.3: Various aspect of integration EVs to MG in smart cities.

MG Centralized Control

In the centralized approach, all the EVs or EVs' operators only interact with central MG management. This center is responsible for collecting and analyzing the received data and responding with the appropriate control signal. The central approach results in improved controllability and reduced scalability.

In [41], the authors proposed a central structure to integrate EVs into the MGs network. A scheme called optimal power setpoints calculator is designed, which aims to minimize active power variance with low and medium voltage substations. This algorithm finds optimized load profiles of EVs by using evolutionary particle swarm optimization as well as considering the data collected from the battery, EV operator's behavior and routing pattern.

In [42], a central scheme is presented to balance generation and load and then regularize the voltage in an integrated system. A central control device is implemented that monitors the exchange of active power between the three phases and computes the power setpoint of EVs that may connect to MG in different phases. This method results in phase balance in the MG.

For decades, the central control schemes have been well-established and widely implemented in practice. However, the increase in consumers' distributed energy generation and storage, and the integration of renewable energy sources make the central control approaches less effective. The central approaches are unable to operate and control future MG systems for several reasons:

1. The increase in the number of users, including the number of EVs integrated with MGs, imposes a heavy computational load that cannot be handled with the central approach.

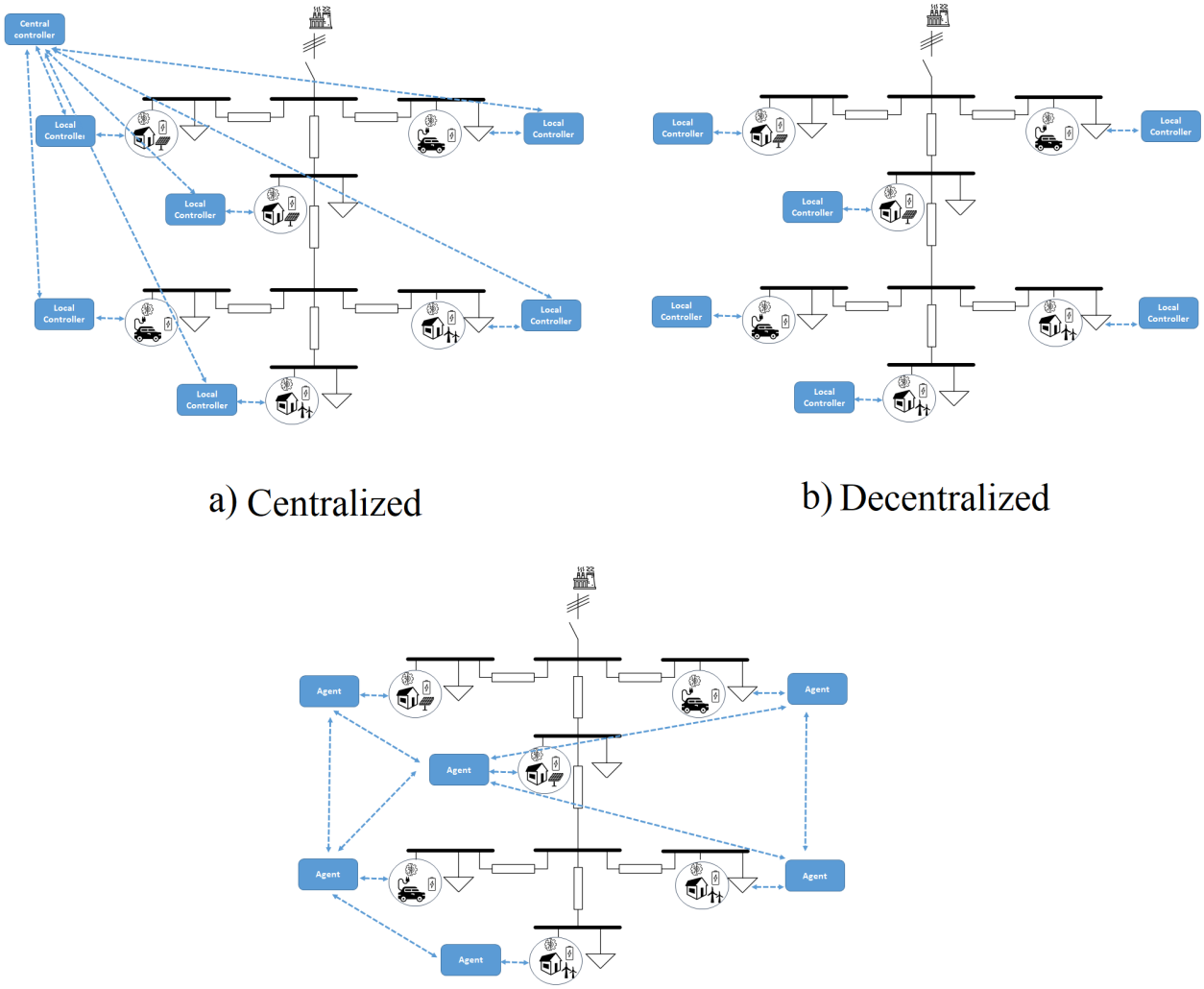

Figure 2.4: MG control typologies.

2. It is challenging to expand the central control systems, although MGs need to evolve and expand very fast.

3. A single point of failure results in the failure of the whole system. Therefore, these control methods are more suitable for small-scale systems.

4. High level of connectivity is necessary for central approaches since each agent should be connected to the central unit individually, which imposes enormous cost as the number of agents grows.

These reasons have resulted in a shift from the central to more distributed structures.

MG Decentralized Control

Decentralized and distributed are terms often used interchangeably; however, there is a difference between these two. In the decentralized control scheme, each user is controlled by either itself or a higher control level. Most of the decisions are made considering the local measurements, and the number of local connections is limited, resulting in a lower level of connectivity than the central control approach. These systems are relatively more robust against a single point failure in the system. If a higher control level or an agent under its control fails, the whole system is still functional. However, due to the low level of connectivity and lack of information exchange between users, global optimization and reliability of the whole network cannot be guaranteed.

MG Distributed Control

Distributed control schemes have the advantages of both centralized and decentralized control, and also overcome the difficulties of both. In the distributed scheme, neighbor agents can share information among themselves. This sharing of information plus the availability of local measurements in distributed control leads to global optimization and robustness of the whole system. The distributed control has a variety of applications.

In [43], a distributed control scheme for coordination of EVs is proposed to optimize energy sharing and regularization of voltage in a commercial MG, which will be able to work as an independent electric unit. The authors designed an EV storage controller that can operate in a distributed or decentralized mode. The controller monitors the reference power of the EV aggregator. The proposed framework not only results in the efficiency of voltage or frequency regulation and system robustness but also it is economically beneficial to EV and aggregator operators.

In conclusion, distributed MG control approaches are significant revolutions in the MG industry, making them very effective tools for coordinating and integrating EVs into the MG network. The advantages are summarized in the following:

1. Distributed MG control is easy to modify and expand and it supports scalability.
2. Distributed MG control is computationally cost-efficient since the computational load is distributed among all agents.
3. A single point of failure would not affect the operation of the whole system.
4. It is highly adaptable for the expansion and integration of EVs into the system.

Challenges and Future Directions in MG Control

The integration of EVs into the MG communities has evolved significantly over the last decade. However, there are still major challenges which require proper solutions. In the following, some of these challenges are outlined.

- *Load Management:* MG communities are designed to transfer energy among themselves to reduce the cost and power loss. However, introducing EVs in these communities adds additional load, which affects the stability of MG communities and imposes uncertainty in the energy demand. Therefore, load management schemes should be re-investigated to fulfill the challenge of adding EVs' loads to the MGs.
- *Charging and Discharging:* Grid to vehicle (G2V) and Vehicle to Grid (V2G) are vastly studied and implemented. However, these technologies are still in the development stages. The two-way power transfer infrastructure is in the design stages, and wide implementation of these infrastructures relies on the industrial models, standards and economic justifications.
- *Distributed Control:* As the number of MGs and EVs increases, the complexity of controlling those entities will also increase. This will consequently impose a heavy computational load that cannot be handled with only one single centralized control unit. Therefore, careful consideration and shifting to distributed control will be essential.
- *Big Data Analytics:* As the number of users with a dynamic nature increases in the system, the amount of data that EVs and MGs generate increases significantly. The analysis of this massive amount of data is necessary to develop charging policies, to design smart charging schemes, to study energy efficiency and the capacity of the system as well as to evaluate the financial aspect of energy transactions. Big data analytics is a promising approach to analyzing this vast amount of data. However,

applying big data on an MG system with integrated EVs faces various challenges. First, real-world data are limited. Furthermore, the analysis time frame needs to be reduced to seconds to satisfy MGs and EVs applications requirements. Therefore, fast and effective big data analytic tools are needed to be developed for the MG applications [44].

- *Security and Privacy:* As the power system moves toward a more distributed architecture, the urge to share more information about each element of the system increases. Although sharing information is crucial for proper data analysis, the privacy of consumers should be taken into account [45, 46]. A study on the trade-off between user privacy and data sharing is needed to find the right amount of data sharing which achieves the required data analysis accuracy. Other than that, customers may become more prone to cyber attacks as EVs become more widespread. Therefore, cyber attack protection should be addressed in future studies.

2.2.5 Energy Trading among MGs

One interesting aspect of MG technology is the possibility of energy trading among distribution system entities. Energy trading, or in other words, using surplus energy from one MG to supply loads at another MG, has been a challenging research question for the past few years. Generally, the problem of energy trading among MGs can be described as a set of connected MGs that can exchange energy in a specific region. Each MG is equipped with energy generation and storing units (see Fig. 2.1). MGs are also connected to the macrogrid, and energy trading can be between one MG and the macrogrid or between two or more MGs. In the literature, it is assumed that for a specific time interval, some MGs have surplus energy and prefer to sell energy, while others suffer from a lack of energy and wish to buy energy. This condition can perfectly be modeled as a trading game with distributed players. Different methods are developed to address energy trading problems among MGs. On the other hand, the distributed and independent nature of MGs results in various uncertainties in the system, such as uncertainties in generation, demand and energy pricing.

In this thesis, we investigate the problem of energy trading among MGs under uncertainty and we use CG-based methods and machine learning methods, particularly reinforcement learning-based methods to address this problem. In the next section, we provide an overview of the methods used in the thesis to solve energy trading problems and in Section 2.4, the research attempts on energy trading and uncertainty in MG systems are reviewed.

2.3 Concepts and Methods

Game theory-based and reinforcement learning-based techniques are becoming prevalent tools to answer engineering problems in various applications. In particular, with the emergence of MGs and an urge for decentralized, self-organizing and independent systems, it has become critical to design proper intelligent models that help effectively investigate and analyze the behavior and interactions of the agents in the future smart grid systems. To this end, in the following, we present the concepts of BCG, reinforcement learning, BRL and DRL, which are used in this thesis to address the energy trading problem in MG systems under uncertainty.

2.3.1 Bayesian Coalitional Game Theory

Generally, game-theoretical methods can be divided into two categories: non-cooperative game theory [47] and cooperative game theory [48]. In non-cooperative game theory, each player decides its strategy independently, intending to maximize its own benefit (utility) or reduce its own costs. Non-cooperative games can be solved by employing equilibrium concepts such as the famous Nash equilibrium [47]. While non-cooperative games analyze competitive strategies, cooperative game theory studies the behavior of rational players in collaborative scenarios. CG methods fall under the cooperative games category [47]. CG methods have been widely employed in various domains such as economics and political science. In typical cooperative scenarios, players have limited ability and information. To this end, players have to collaborate with one another to accomplish mutual tasks. Considering that rational players are ultimately motivated by their own share of payoff, having a payoff division strategy is crucial for establishing a stable collaboration. CG enables cooperative players to attain stability and fairness by providing a strategy for dividing the payoffs.

CG games branch into two subdivisions: CG with transferable utility [49] and CG with non-transferable utility [50]. In transferable utility games, it is assumed that a divisible commodity such as money exists that coalition members can distribute it among themselves based on a fairness mechanism. This dividable value is known as coalition value which shows the total utility of the coalition. Although we can model a wide range of cooperative scenarios as a CG with transferable utility, defining the coalition value with a single number is impossible in some scenarios. Instead, these scenarios can be modeled using CG with non-transferable utility.

In this thesis, the objective is two minimize the total power loss and cost associated

with energy trading among coalition members. Since both quantities are dividable, our considered scenarios lie in the realm of transferable utility games. Therefore, from now on, when we refer to CG, we refer to CG games with transferable utility. In CG, a coalition can be defined with a pair (C, v) which, C denotes the members of a coalition that cooperate to achieve a higher coalition value v [51].

In CG, two crucial assumptions are made. First, it is assumed that the value of each coalition is fixed and deterministic. Second, it is assumed that coalition values are common knowledge among coalition members. These assumptions fail in practical systems under uncertainty. To address the problem of uncertainty in cooperative scenarios, BCG is proposed in the literature [51–53]. A BCG game $G = \langle \mathbf{M}, \mathbf{A}, \mathbf{T}, B, \overline{\mathbf{u}} \rangle$ consists of [51]:

- $\mathbf{M}$ is a set of players.
- $\mathbf{A}$ is a set of coalitional actions. In each coalition, coalition members agree on a unified coalitional action.
- $\mathbf{T} \in \mathbb{T}$ is the types vector of all the players and $\mathbb{T}$ is a set of all possible types vectors of all players. The uncertainty in the system can be modeled with type. In other words, type reflects the ability and influence of players on a specific task and the beliefs of a player about the types of others capture its uncertainty about others influences on the coalitional task. Each player only knows its own type and has no information regarding other players' types.
- B is the belief function. $B(\mathbf{T})$ is the joint belief probability of a player over other players' types. In the beginning, each player considers a prior probability over other players' types. After a considerable number of epochs, the effect of the considered prior will be negligible [51].
- $\overline{\mathbf{u}}$ is the vector of the expected payoff of all players, and $\overline{u}^i$ shows the expected payoff realized by player i. In the BCG, we use expected payoff (utility[1]) realized by the players rather than immediate payoff since the system is dynamic and the payoff of coalition changes based on the types of coalition members. This will help to overcome the uncertainty and consequently form a stable coalition. Since we consider a discrete space, we can denote the expected payoff realized by i-th agent as follows:

$$\overline{u}^i\left(C, \mathbb{T}^C\right) = E\left[u^i\left(C, \mathbb{T}^C\right)\right] = \sum_{\mathbf{T}^C \in \mathbb{T}^C} B^i\left(\mathbf{T}^C\right) u^i\left(C|\mathbf{T}^C\right), \tag{2.1}$$

[1] In this thesis, we will use the terms payoff and utility interchangeably

where $\mathbf{T}^C$ denotes a possible types vector of members of coalition C and $\mathbb{T}^C$ denotes a set of all possible types vectors of coalition C. $B^i\left(\mathbf{T}^C\right)$ is joint belief probability of agent i for a given types vector $\mathbf{T}^C$. $u^i(C|\mathbf{T}_C)$ shows the immediate payoff can be achieved by agent i given types vector $\mathbf{T}^C$.

As previously mentioned, in non-cooperative scenarios, equilibrium concepts such as Nash equilibrium are employed to find a stable solution for the games. In BCG, the concept of the Bayesian core is proposed to achieve stable coalition formation. Two forms of Bayesian core that are adopted in this thesis can be defined as follows [51]:

1. *Weak Bayesian Core (WBC):* We assume that a coalitional structure is in the WBC of a BCG if none of the players believe that a better coalition exists (in terms of expected payoff) than the one they are in now. This definition can be formulated for i-th player in coalition C as follows:

$$\overline{u}^i\left(C, \mathbb{T}^C\right) \geq \overline{u}^i\left(C', \mathbb{T}^{C'}\right). \tag{2.2}$$

2. *Strong Bayesian Core (SBC):* We assume that a coalitional structure is in the SBC of a BCG if a) none of the players believe that a better coalition exists (in terms of expected payoff) than the one they are in now and b) each player, based on its realization about the expected payoff of other players, believes that other cannot achieve higher payoffs in any other coalition formation. This definition can be formulated for i-th and j-th players in coalition C as follows:

$$\begin{aligned} &\overline{u}^i\left(C, \mathbb{T}^C\right) \geq \overline{u}^i\left(C', \mathbb{T}^{C'}\right) \\ &\overline{u}^i_j\left(C, \mathbb{T}^C\right) \geq \overline{u}^i_j\left(C', \mathbb{T}^{C'}\right), \end{aligned} \tag{2.3}$$

where $\overline{u}^i_j\left(C, \mathbb{T}^C\right)$ shows the expected payoff of player j realized by player i.

In Chapter 3 and Chapter 4, we employ WBC, while in Chapter 6, SBC is used. In each iteration of the game, players update their beliefs about other players' types based on the observation until reaching a stable coalition formation. The belief update mechanisms are explained separately in each chapter based on the considered system model and constraints.

2.3.2 Reinforcement Learning

We first need to define the Markov Decision Process (MDP) as an essential part of the reinforcement learning [54]. An MDP consists of four elements $(\mathbf{S}, \mathbf{A}, H, r)$, where $\mathbf{S}$ is a vector of all possible states s. $\mathbf{A}$ is a set of all possible actions. $H(s', s, a) = \Pr\{s'|s, a\}$ shows the chance of transition from state s to state s' while taking action a. $r(s, a)$ expresses the reward that the agent receives by taking action a in the state s. The reinforcement learning problem can be defined as the problem of finding the optimal mapping strategy from actions to states ($\sigma : \mathbf{S} \rightarrow \mathbf{A}$) for the MDP with the known or unknown transition probabilities. Essentially, in an reinforcement learning setup, the goal of agent is to learn the best policy that results in the maximum total expected reward over the time as follows:

$$\max_{\sigma(s)} \mathrm{E}\left[r_{\tau+1} + \gamma r_{\tau+2} + \gamma^2 r_{\tau+3} + ... \,|s_\tau = s, a_\tau = a\right], \tag{2.4}$$

where $\sigma(s)$ is the policy that selects optimal action a at state s at iteration step τ. $0 < \gamma < 1$ represents the discount factor that control the effect of future rewards. To find a policy that maximize the total expected rewards as in Eq. (2.4), the action-value function is employed to determine numerically how good is a policy. The action-value function correspond with selecting action a at the state s is given by:

$$Q_\sigma(s, a) = \mathrm{E}_\sigma\left[r_{\tau+1} + \gamma r_{\tau+2} + \gamma^2 r_{\tau+3} + ... \,|a_\tau = s, a_\tau = a\right]. \tag{2.5}$$

The optimal value function can be achieved using brute-force methods. However, for the large actions-states space, it becomes intractable. Temporal difference methods such as Q-learning use an estimation of the value function to find the optimal policy. Q-learning is a type of reinforcement learning algorithm where the objective is to obtain a sub-optimal policy by choosing actions that maximize the expected current and future rewards. In Q-learning, the Q-values should be updated considering Bellman's equation as follows [54]:

$$q(s, a) = (1 - \alpha)\, q_{t-1}(s, a) + \alpha\left[r(s, a) + \gamma \max q_{t-1}(s, a)\right], \tag{2.6}$$

where α is the learning rate and γ is discount factor. In reinforcement learning approaches the design of state and reward functions is a critical stage, since it highly impacts the outcome of the learning system.

2.3.3 Bayesian Reinforcement Learning

In MDP, an agent computes and uses the optimal policy to take action a in state s and then receives reward r and ends up in state s'. However, in practical systems, it is common

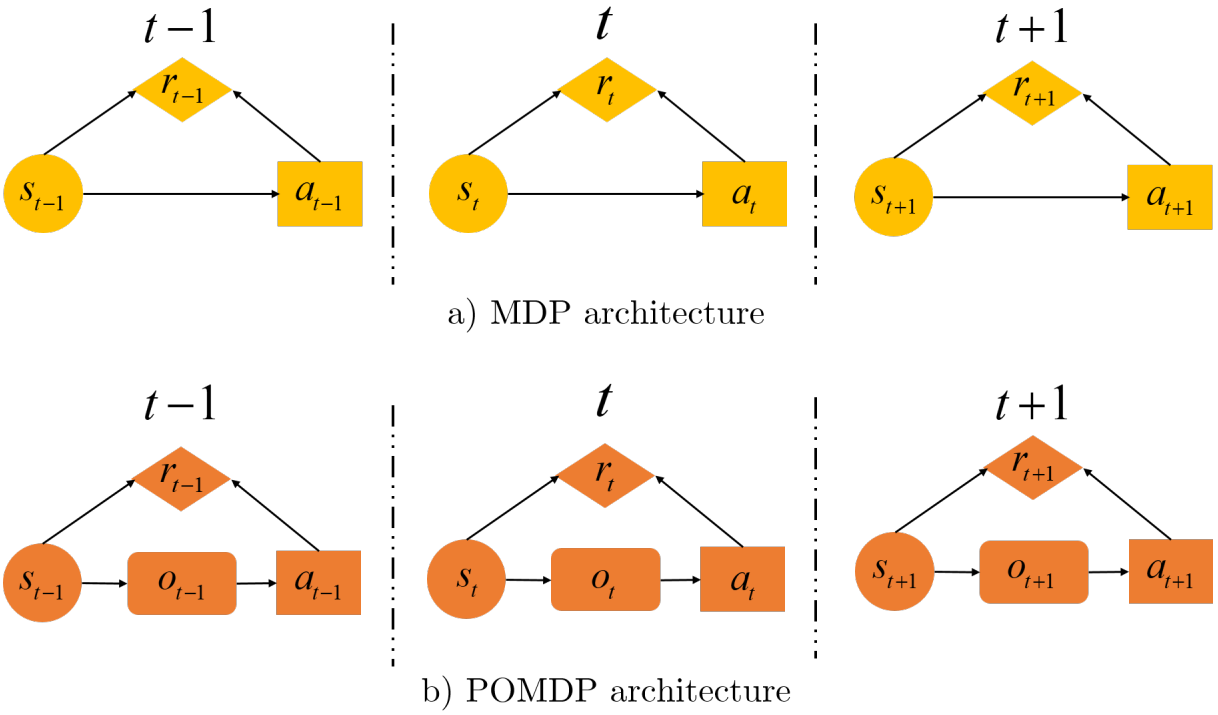

Figure 2.5: MDP and POMDP architectures.

that agents cannot determine the current state with complete reliability. Therefore, to effectively act in scenarios with a partially observable state, the history of previous actions and observations must be considered, which helps agents clarify the states of the world. The Partially Observable MDP (POMDP) [55–57] framework provides a systematic method to address this problem. A POMDP can be represented with tuple $\langle \mathbf{S}, \mathbf{A}, H, r, \mathbf{\Omega}, O \rangle$ where [55]:

- $\mathbf{S}$ is set of all possible states in the system.
- $\mathbf{A}$ represents set of actions.
- $H(s', s, a) = \Pr(s'|s, a) : \mathbf{S} \times \mathbf{A} \to \Pi(\mathbf{S})$ is the state-transition function and it is a probability distribution over all possible ending states s' when the agent takes action a in state s.
- $r(s, a)$ shows reward function when the agent takes action a in state s.

- $\mathbf{\Omega}$ is set of all possible observation that the agent can experience.
- $O(s', a, o) = \Pr(o|s', a) : \mathbf{S} \times \mathbf{A} \rightarrow \Pi(\mathbf{\Omega})$ is the observation function and it is a probability distribution over all possible observation o when the agent takes action a and ends in state s'.

A POMDP is a special form of MDP in which the agents cannot observe the current state. Instead, in POMDP, agents rely on observation based on the action and resulting state (see Fig 2.5). Nevertheless, agents still aim to maximize their expected discounted future rewards.

Fig. 2.6 shows a simplified block diagram of the BRL agent, which consists of two parts: a state estimator and a policy. State estimator updates belief state based on the current observation, last action, and last belief state. Different definitions can be considered for belief state. In the following, belief state B is defined as the probability distribution over all possible states in $\mathbf{S}$. $B(s)$ shows agents belief probability of being in state s where $0 \leq B(s) \leq 1$ for all $s \in \mathbf{S}$ and $\sum_{s \in \mathbf{S}} B(s) = 1$. In each iteration, receiving a new observation o after taking action a and considering a prior belief state, the agent can calculate posterior belief state $B'(s')$ for some state s' based on Bayes' theorem as follows [55]:

$$\begin{aligned} B'(s') &= \Pr(s'|o, a, B) \\ &= \frac{\Pr(o|s', a, B) \Pr(s'|a, B)}{\Pr(o|a, B)} \\ &= \frac{\Pr(o|s', a) \sum_{s \in \mathbf{S}} \Pr(s'|a, B, s) \Pr(s|a, B)}{\Pr(o|a, B)} \\ &= \frac{O(s', a, o) \sum_{s \in \mathbf{S}} H(s', a, s) B(s)}{\Pr(o|a, B)}. \end{aligned} \tag{2.7}$$

$\Pr(o|a, B)$ can be regarded as a normalizing coefficient. Finally, the output of the state estimator is $B'(s')$. In the BRL algorithm, first, a prior distribution is assigned to the initial beliefs of the agent about the values of the unknowns in the system. Then, this belief will be updated continuously as the agent observes the unknown parameters. To demonstrate the belief update mechanism with an example, we consider a simple scenario adopted from [55] as shown in Fig. 2.7.

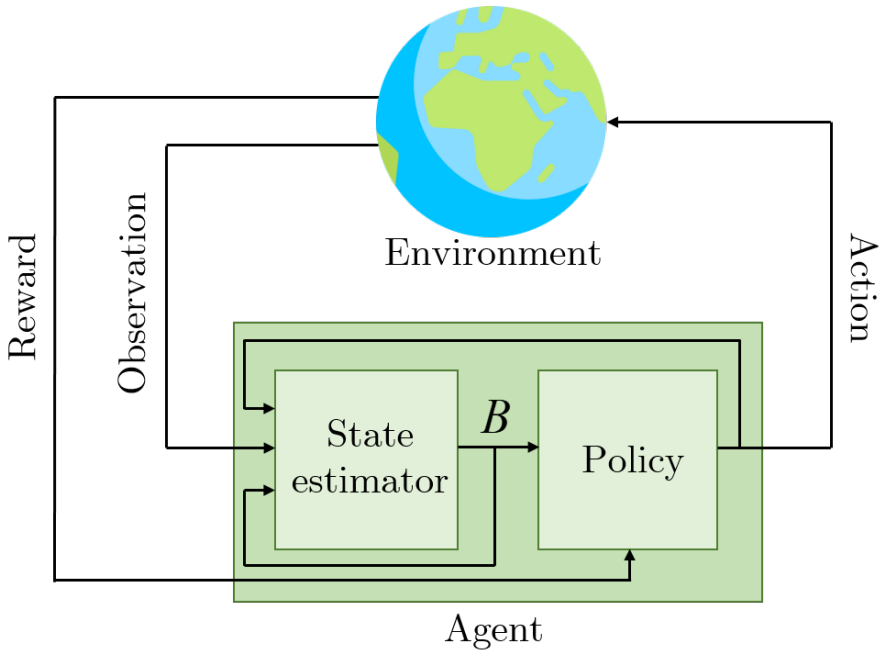

Figure 2.6: Block diagram of a BRL agent.

In this example, each of the four blocks corresponds to one of the states, and the final state is denoted by the green flag. The agent has no information regarding its state at the beginning and only relies on two observations. Agent receives first observation when ends in states 1, 2, and 4 and receives second observation when lands in final state 3. The agent can either move one block to the right or one block to the left as the set of possible actions. However, the system is faulty and the agent might move in the opposite direction with the probability of 0.1. When movement in some directions is not feasible (moving right in state 4 and moving left in state 1), the agent stays in the same block. The agent continues to interact with the environment until landing in the final state 3. It is logical to assume that in the initial step, the agent can be in states 1,2, or 4 with the same chance. Therefore, the initial belief state is equal to [0.333 0.333 0.000 0.333]. If the agent chooses action right and does not receive final state observation, the belief state can be updated as [0.100 0.450 0.000 0.450]. If the agent again decides to move in the right direction and still does not receive final state observation, then the chance of being in state 4 increases, and the belief state can be updated as [0.100 0.164 0.000 0.736]. The agent continues interacting with the environment and evolving its belief about its location (state) until reaching the final state.

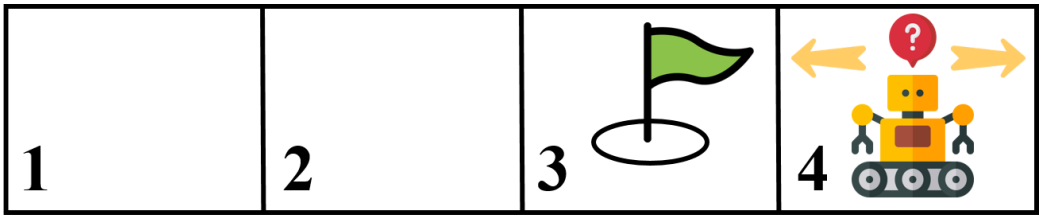

Figure 2.7: Example of a BRL agent interacting with environment.

The second component of a BRL agent is policy. In BRL, the strategy is to map from belief function to actions as $\sigma : B \rightarrow \mathbf{A}$. We can calculate the value of a specific policy σ as the expected sum of discounted reward over infinite time in the future, which is given by [56]:

$$V^{\sigma}(B) = \sum_{\tau=0}^{+\infty} \gamma^{\tau} r(\sigma(B_{\tau})), \tag{2.8}$$

where γ and B_{τ} express the discount factor and belief at time τ, respectively. We are interested in finding the optimal policy σ^*. The optimal policy has the highest value for all the belief states, i.e., $V^{\sigma^*}(B) > V^{\sigma}(B)$ and the corresponding value function of the optimal policy satisfies the Bellman's equation as follows:

$$\begin{aligned} V^* &= \max_{a \in \mathbf{A}} Q(B, a) \\ &= \max_{a \in \mathbf{A}} \sum_{o \in \mathbf{\Omega}} \Pr\{o|B, a\}[r(B, a) + \gamma V^*(B_a^o)]. \end{aligned} \tag{2.9}$$

$Q(B, a)$ represents the action-value function in the case of taking action a. B is the prior belief state and B_a^o is the updated posterior belief state.

2.3.4 Deep Reinforcement Learning

Long convergence in tabular reinforcement learning methods such as Q-learning is a significant drawback. Q-table is used to store Q-values in conventional reinforcement learning. Q-table is updated each episode which results in considerable convergence time and high complexity. Therefore, to overcome the problem of long convergence time, a deep neural network is employed to estimate the Q-values and learn the correlation between input sequences. This method is known as Deep Q-Network (DQN) [58] which is a specific category of the DRL methods. To discuss DQN, we first briefly introduce the neural network and deep learning [59].

The structure of biological neurons motivated the design of neural networks. In Fig. 2.8, a general block diagram of a neural network is depicted. Neural networks consist of an input layer, an output layer and hidden layers in between. Each layer is built from a set of artificial neurons that works as a mathematical function known as activation function. In the first layer, pre-processing is done on the input feature $\mathbf{x}$. At the output layer, neurons produce the outcome vector $\mathbf{y}$. Neurons in a specific layer are connected to the neurons in the next layer, and each connection has a weight that should be adjusted in the training phase.

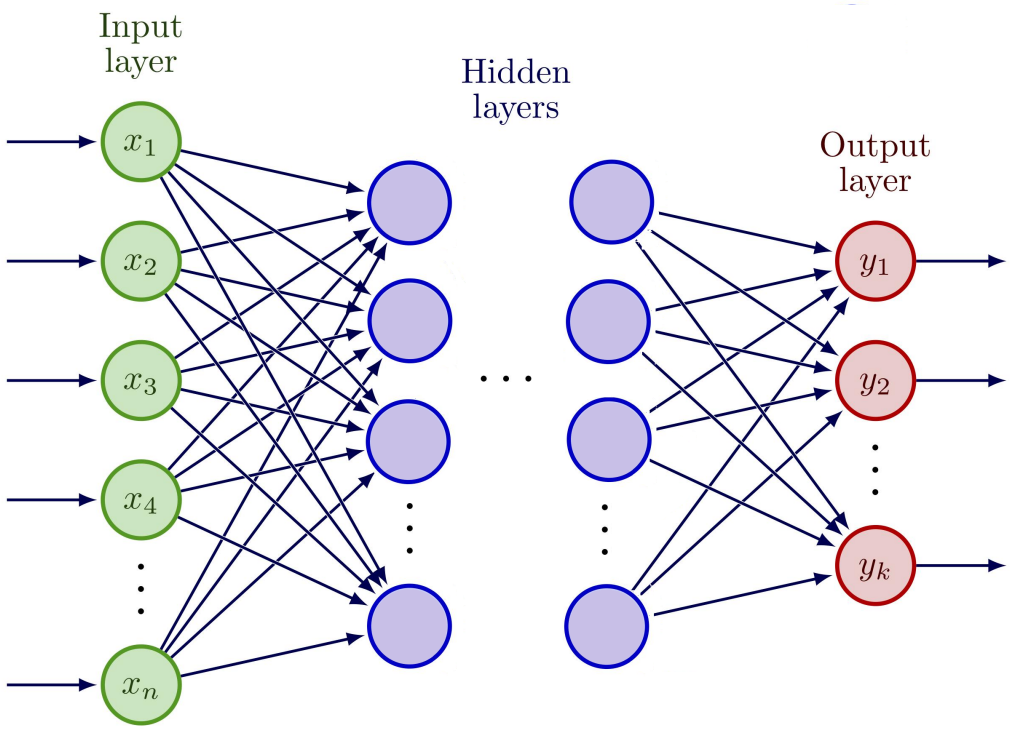

Figure 2.8: Block diagram of the neural network.

Fig. 2.9 shows the structure of a neuron in which $\mathbf{w}$ represent the weight vector in the pre-activation function $z(\mathbf{x})$. $\psi(\mathbf{x})$ and b denote neuron activation function and bias value, respectively. $z(\mathbf{x})$ can be defined as:

$$z(\mathbf{x}) = b + \sum_{i=1}^{N} w_i x_i. \tag{2.10}$$

Therefore, the output $F(\mathbf{x}; b, \mathbf{w})$ can be calculated as following:

$$F(\mathbf{x}; b, \mathbf{w}) = \psi(z(\mathbf{x})) = \psi\left(b + \sum_{i=1}^{N} w_i x_i\right). \tag{2.11}$$

Different forms of activation functions are used in deep learning methods such as linear,

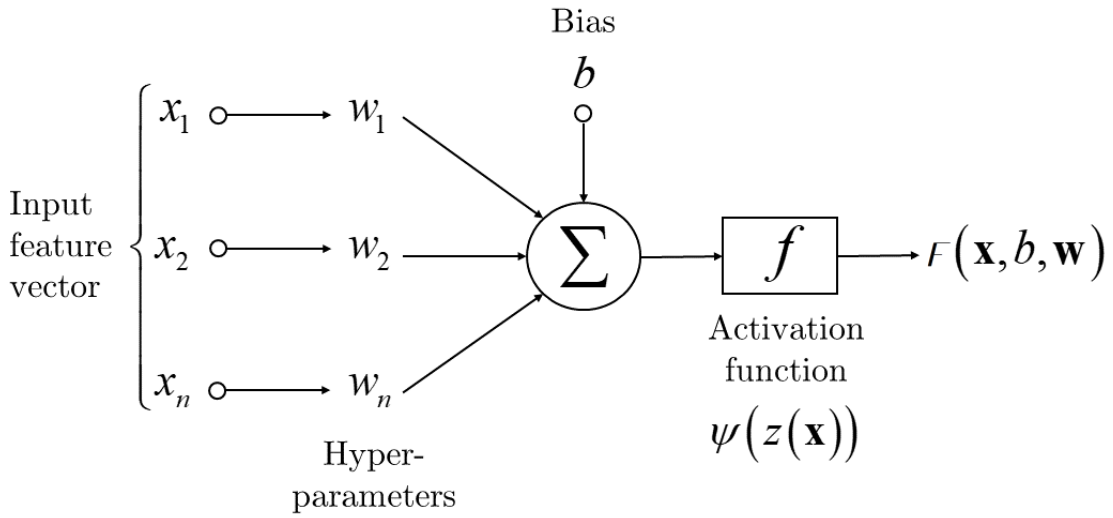

Figure 2.9: Structure of a neuron.

hyperbolic tangent, sigmoid and rectified linear activation functions. In Fig. 2.10, these activation functions are plotted.

Neural networks can be implemented in various architecture such as feed-forward, convolutional, or recurrent neural network. A neural network that consist of one layer known as shallow neural network while a network with more than one layer is called deep neural network. Fig. 2.8 shows a deep feed-forward neural network in which information flows through multiple layers in only one direction. On the contrary, Fig. 2.11 shows a deep recurrent neural network in which feedback connections are integrated between layers. In recurrent neural network, state of neurons depends on the input feature and the current state which helps the neural network to learn the long sequential information. Therefore, we can obtain the activation function in recurrent neural network as following:

$$F_\tau(\mathbf{x}) = f(F_{\tau-1}(\mathbf{x}), x_\tau; b, \mathbf{w}), \tag{2.12}$$

where $F_\tau(\mathbf{x})$ is the output at τ-th iteration and $F_{\tau-1}(\mathbf{x})$ is the output at $\tau-1$-th iteration.

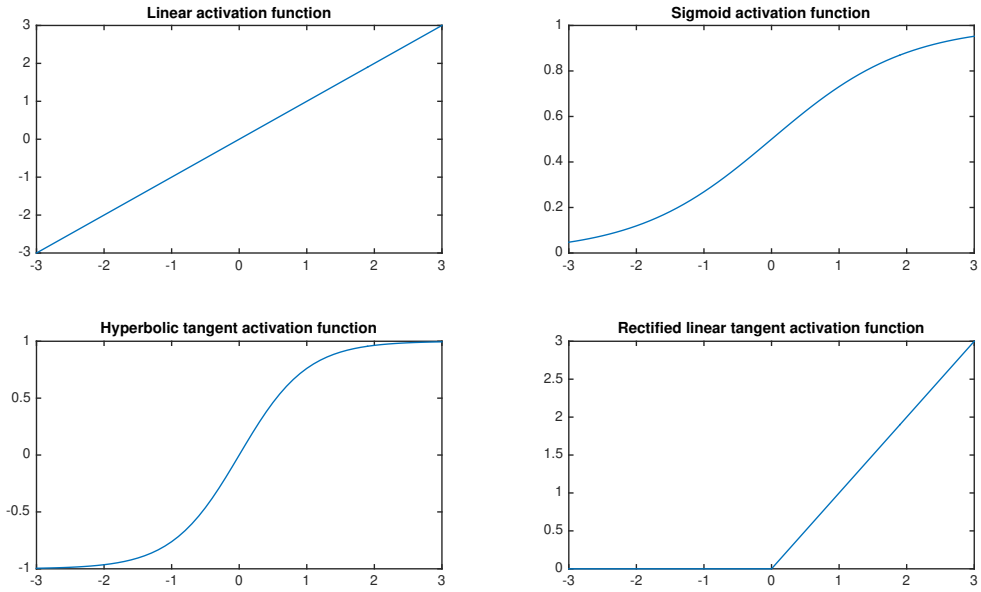

Figure 2.10: Examples of activation functions of a neuron.

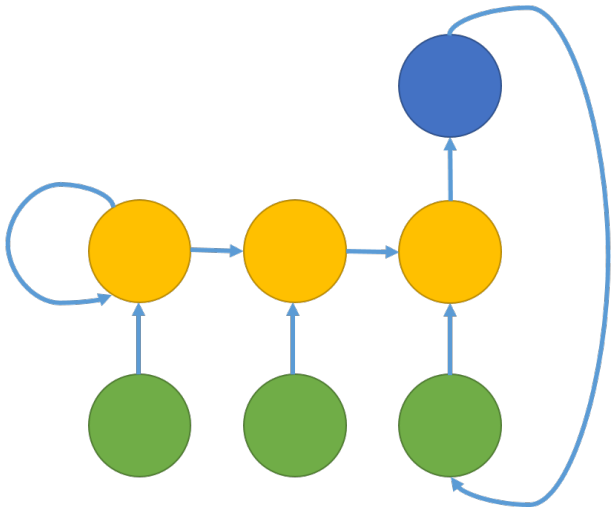

Figure 2.11: Feedback connections in a recurrent neural network.

In DQN, a neural network is used to estimate the Q-values [60]. Q-values vary dynamically, and consequently, the target values will change, which may result in an unstable out-

put. Also, the training data should be discontinuous, while the state and action transitions are consecutive in Q-learning. To address these issues and improve the DQN, experience replay and target network have been proposed as two DQN improvement solutions [58]. The loss function in DQN is defined as follows:

$$L(w) = E(r(s,a) + \gamma \max q_{\tau+1}(s', a'; \mathbf{w}) - q_{\tau}(s, a; \mathbf{w})), \tag{2.13}$$

where $\mathbf{w}$, s and s' are the weight of the neural network, the current state and the next state, respectively. $r(s,a) + \gamma \max q_{\tau+1}(s', a'; \mathbf{w})$ is the training target and $q_{\tau}(s, a; \mathbf{w})$ is the predicted result.

In this thesis,in Chapter 7, the Long Short Term Memory (LSTM) which is a subset of the recurrent neural network, is used to estimate the Q-values [61]. LSTM is capable of capturing the long-term dependencies compared to the conventional recurrent neural networks and, at the same time, addresses the problem of vanishing gradient. Considering the block diagram of the LSTM network in Fig. 2.12, x_t and h_{t-1} are the input value at time t and the output value at time $t-1$ respectively. c_{t-1} represents cell state at time $t-1$. The current output is shown by h_t and c_t denotes the current cell state.

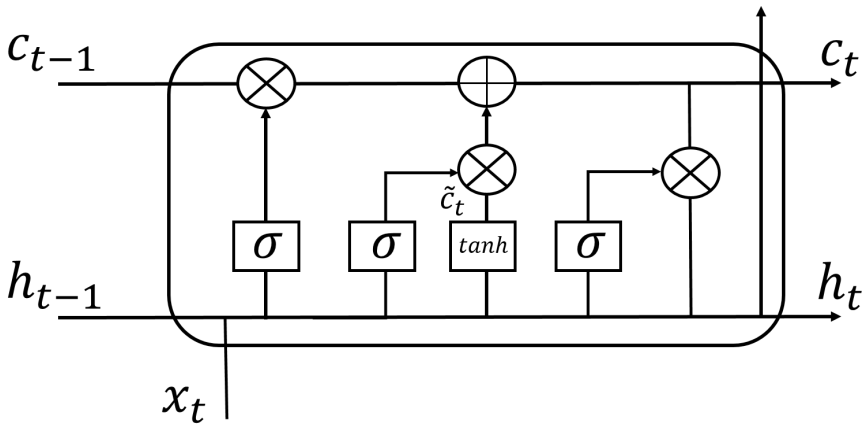

Figure 2.12: Block diagram of the LSTM network.

2.4 Related Works

In this section, we provide a survey on the energy trading problems among MGs in the literature and then we review the research works that considered uncertainty in analyzing

energy trading and management of MGs.

2.4.1 Energy Trading among MGs

Based on the techniques used in the literature, energy trading studies can be divided into four categories: optimization method, game-theoretical methods, CG methods, and machine learning methods, in particular reinforcement learning-based methods. In the following, we have reviewed each category.

Optimization-Based Energy Trading Problems

The energy trading problem is examined using constrained optimization under different system constraints [62–67]. A novel energy management strategy is developed in [62], using linear programming to study the optimal scheduling problems of consumers with multiple energy demands. In [63], the authors proposed a day-ahead forecasting energy market scheme that could be used by system operators at the distribution level to improve the efficiency of the applications of distributed energy resources. In [64], non-linear programming is used to design a peer-to-peer energy sharing system in an MG community, using aggregated battery control.

Game-Theoretical-Based Energy Trading Problems

Game theory has been widely used to address energy trading problems in MG communities [68–78]. In [68], a game-theoretical approach is proposed for distributed energy trading between MGs. In this study, a set of interconnected MGs aims to exchange energy with each other and also with the macrogrid. In this market, those MGs with surplus energy can choose to sell part of their energy and store the rest for the future. Likewise, those MGs that suffer from a shortage of energy or wish to store energy for the future can buy energy. In [68], a two-level continuous kernel Stackelberg game is employed, in which seller and buyer MGs are classified as leaders and follower players, respectively. To find the Stackelberg equilibrium of the formulated problem, the authors use a backward induction technique, where first the best response of buyers are found in the followers-level game, and then, the results are plugged into the utility function of each seller to solve the problem at the leaders-level game. The numerical evaluations demonstrate that sellers can achieve higher utilities since the sellers act as the leaders and have the advantage of choosing their

strategies first. The results show that the Nash equilibrium of the Stackelberg game is lower bounded by half of the optimal cooperative centralized solution.

In [69], a priority-based energy trading game is proposed. First, the amount of contribution in the past, in terms of energy provided by the buyer to the grid, is calculated; then, the amount of energy requested by that buyer is prioritized. This prioritization model is considered in the paper to eliminate the difficulty of energy pricing. An efficient method for energy distribution is proposed in this paper. The aim is to maximize the sum of satisfaction of all the buyers, known as social welfare. To do so, an iterative algorithm is used to optimally assign energy to buyers using the water filling approach. At the buyer level, the traditional game theory is considered, and the existence of Nash equilibrium is proved. The buyer and seller utilities are selected according to the research in [70].

In [71], the authors presented a single-leader, multiple-follower Stackelberg game in which the central power station is considered as a leader (buyer) and followers are the sellers who decide to sell their surplus energy to the central unit. The goal of this paper is to maximize the sum of all followers' utilities while satisfying the minimum cost for the central power station. The approach is decentralized since the main station does not have control over the sellers.

In [76–78], double auction-based methods are used to tackle energy trading problems. In [76], energy trading between prosumers and consumers is facilitated by a continuous double auction algorithm. A prediction model is developed, which helps prosumers to analyze market prices. The proposed method enhances both the prosumer operation and market profit at the same time. In [77], authors proposed a double auction-based method to address the energy trading problem among residential MGs. The results show the effectiveness of the proposed method in managing renewable generations and demand uncertainties. In [78], the peer-to-peer energy trading problem among EVs is solved using a double auction-based method which maximizes social welfare while addressing security and privacy issues.

CG-Based Energy Trading Problems

CG is a subset of game theory in which players cooperate to maximize a shared payoff (utility) and then distribute the received payoff among themselves. The energy trading problem among MGs can efficiently be modeled as a CG problem [7, 79–82]. As such, MGs can form coalitions for a specific time interval where some MGs have surplus energy and are willing to supply energy to others, while others generate less than their demands and are willing to buy energy.

In [7], an energy trading problem in the MGs community is investigated to minimize the energy transfer power loss over the line. In this study, a conventional CG-based approach is proposed that enables MGs to exchange energy with nearby MGs through the coalition formation process. Furthermore, in [79], the authors propose a CG scheme to solve the problem of energy management in local energy communities. It is demonstrated that the achieved coalition scheme results in a higher total payoff, and the allocation scheme effectively flattens the peak and minimum. However, the resulting coalition cannot guarantee the maximum payoff for all agents, and the agent with renewable energy resources may leave the coalition as they get dissatisfied with their received expected payoff. The problem of energy management of the community MGs has been addressed in [80] using CG. Despite [7], which only focuses on energy loss, the authors expanded the objective functions to maximize the expected profit of MGs and usage of renewable energy while minimizing power loss and consumer discomfort. Although uncertainty is proposed in problem formulation, the designed scheme does not effectively address it.

In [81], the authors propose a CG-based energy trading, where in each coalition, an auction-based matching is employed first to calculate the utility of the coalition, then the concept of Shapely Value [83] is used to divide coalition utility among coalition members fairly. In [82] a two-stage algorithm is proposed for the energy trading problem of MGs. In the first stage, a coalition formation algorithm is used, and then in the second stage, a matching game is employed to manage the energy exchange inside each coalition.

Reinforcement Learning-Based Energy Trading Problems

Machine learning algorithms are critical tools that enable many applications ranging from computer vision to sentiment analysis, self-organized systems and robotics. However, an Artificial Intelligence (AI)-enabled smart grid calls for machine learning techniques different from those developed for conventional application areas. Thus, novel machine learning techniques that are exclusively designed to meet the unique challenges of the smart grid and MGs are needed.

In [84], authors propose two learning automata-based methods for optimal power management in smart grids. These methods aim to control the power consumption by different grid users and identify the required energy for various distribution substations. Learning automata enables dynamic power analysis and finds the corresponding optimal usage. In [85], authors propose a dynamic demand response and distributed generation management method for a residential MG community. The distributed generation management works with distributed generation and monitors the stochastic load and wind power. It aims to decrease the overall cost of energy consumption for the residential community while

keeping users highly satisfied. In [86], a fully distributed learning approach is proposed for optimal reactive power dispatch. In this method, a multi-agent Q-learning algorithm is employed that minimizes the active power loss while guaranteeing bus voltages range constraints and reactive power generation constraints. The implemented reinforcement learning is model-free and can learn the near-optimal solution without prior knowledge of the system. However, having prior knowledge of the system will boost the learning process. In [87], the temporal difference reinforcement learning approach is used to achieve the optimal control policy for residential energy storage. The proposed method does not rely on precise predictions of power generation/consumption and only requires partial system information. Temporal difference reinforcement learning is implemented to gain a shorter convergence time and higher performance. The state, action and reward function are chosen so that the objective function achieves costumers' bill minimization.

The problem of dynamic pricing in smart grid with reinforcement learning methods is visited in [88]. In this paper, the service provider sets the electricity price in the retail market. In order to overcome the challenges in implementing dynamic pricing, a reinforcement learning algorithm is developed. Also, to overcome the problem of convergence time and high computational complexity in reinforcement learning algorithms, an approximate state is proposed and virtual experience is adopted. The numerical results guarantee that the proposed method can effectively set the price without prior knowledge of the system. Another attempt for dynamic pricing is made in [89]. In this paper, the authors propose a scenario in which the service provider is considered as a broker between the utility grid and customers by buying energy from the main grid and reselling it to energy consumers. Due to the lack of information from energy consumers and also the problem ahead of consumers for scheduling, employing dynamic pricing is complicated. Thereby, to ease this issue, authors propose a reinforcement-based dynamic pricing and energy consumption scheduling to help energy providers and consumers to learn their best strategies. Another aspect of the smart grid problem studied with the aid of machine learning is the communication network for smart grid applications. In [90, 91], reinforcement learning-based algorithms are presented for the problem of distributed energy management with wireless networks in MGs.

Energy trading is also among the problems that can be tackled with machine learning approaches, specifically reinforcement learning approaches such as Q-learning, BRL and DRL. Several studies deployed reinforcement learning methods to address energy trading problems. In [72], the problem of energy trading between MGs is visited while protecting their private strategies. The system model is adopted from [68]. In this problem, a market operator is introduced which collects MGs actions, and each MG chooses its action randomly and individually. To overcome the problem of incomplete information, a new

scheme is proposed, which combines non-cooperative repetitive Stackelberg game and reinforcement learning algorithms. It is proved that learning techniques help each MG to reach the best response achieved by the Stackelberg game. This has been done by connecting the average utility maximization and the best strategy. In the first learning technique, the action set is considered finite, and the epsilon-optimality is guaranteed. In the second learning technique, the action set is continuous, and the numerical results show that in this method, by finding the optimal mean, the best response of each player can be achieved.

In [92], a set of connected MGs is considered, which can transfer energy to each other and to the macrogrid. Each MG is equipped with a battery to store energy and local energy generators such as WTs and SPs. The level of these renewable energy resources is not constant and varies by time and thereby should be estimated based on the generation history. Energy trading between MG is modeled as a non-cooperative distributed game, in which each MG aims to maximize its own utility function. A hot-booting Q-learning-based approach is implemented to achieve the Nash equilibrium of the dynamic repeated game. Since the MG interactions in the next state depend on the current battery level, energy generation and energy transfers, each MG can employ Q-learning to learn the best trading strategy. The conventional Q-learning approach starts the learning process with zero information about the game condition. However, in this paper, Hot-booting Q-learning is employed in which the initial values are derived based on the training data achieved during the actual experiments and then fed into the Q-learning system to obtain the initial values. The results show that the hot-booting method gains significant efficiency in the convergence time and also increases the overall profit of the players.

In [93–95], DQN is used to address peer to peer energy trading problems. In [93], the authors of [92] have improved their work by designing a DQN-based approach. The scenario of the problem and game model is the same as [92]. DQN estimates the values of the Q-table and therefore improves the convergence rate and system performance. The epsilon-greedy algorithm is also considered, which helps the system avoid staying in the local optimum and search for other possibilities. The simulation results show that this approach can be more efficient in terms of players' utility than the hot-booting method. In these works, maximizing MGs' revenue in trading energy is considered as the objective of the system. In this thesis, we consider power loss and the cost associated with transferring power as the system's objectives.

2.4.2 Uncertainty Analysis

The idea of developing MG communities is to move from a centralized grid system to a more distributed heterogeneous system that can operate independently from the main grid

and at the same time be able to exchange energy with nearby MGs, which makes MGs environments so uncertain that practical implementation of these systems without careful analysis of uncertainty sources is impossible. In the following, major sources of uncertainty in MGs are discussed.

- *Variable Loads:* Flexible load in MG is a major source of uncertainty. Flexible load depends on various parameters such as weather conditions, variation in energy pricing and consumers' decisions. While it is straightforward to forecast fixed load, forecasting variable load imposes uncertainty on the system.

- *Energy Generation:* MGs highly rely on renewable resources, which generate variable energy. The generation of renewables such as SP and WT depends on the weather condition, which is not varying periodically. Therefore accurate prediction of variable generation is challenging.

- *Energy Pricing:* In energy trading among MGs, fixed energy pricing is not available. Different factors such as available energy in the MGs market, total energy demand, consumer response to the pricing, congestion of transmission lines, and power losses affect energy market pricing and impose crucial uncertainty on the design of MG systems.

- *EVs:* High penetration of EVs in MG systems is another issue that introduces uncertainty to MG systems. EVs have variable loads, which depend on several parameters such as drivers' decisions, drivers' response to the energy market, traffic situation. Additionally, due to their moving nature, EVs can charge or discharge from different places, which imposes location-wise uncertainty on the system.

- *Islanding:* MGs' islanding capability is another factor in imposing uncertainty. MGs switch to islanding mode when the main grid distribution network suffers from a disturbance. MGs switch back to normal mode when the disturbance is eliminated. However, the duration and location of disturbances are not known to MGs. Although major outages are not common, their impact in case of happening is so significant that it makes islanding analysis so vital to the system.

Different methods are proposed in the literature to solve MG's energy trading and management problem while addressing uncertainty. Different optimization techniques are used to deal with uncertainty in energy trading management in MG communities, such as stochastic, interval and chance-constraints optimizations. Among optimization methods, stochastic optimization is the most applied method in energy trading management

problems of the MGs under uncertainties [96–101]. In this method, a Probability Density Function (PDF) is assigned to the uncertain parameter in the system and several scenarios are generated based on the considered PDF to model the uncertainty in the system. Increasing the number of scenarios can improve the precision of the considered model while imposing computational complexity. Therefore, the scenario generation and reduction technique is employed to overcome the computational complexity burden.

In [96], the authors used particle swarm optimization to economically optimize the energy trading management in an incentive-based demand response MG system. The proposed method helps MG operators with decision-making that considers unprecedented scenarios. In [97], stochastic mixed-integer linear programming is employed to solve the MG scheduling problem. The authors considered the variation in the generation of SP and WT, and market-clearing price as the sources of uncertainties in objective function to enhance the system's reliability. Despite the implementation of the scenario generation and reduction technique, the method cannot fully satisfy the security and economic constraints, which can impact the robustness of this method in case of significant forecasting errors.

To address the limitations of stochastic optimization, Conditional Value at Risk (CVaR) is used in [98–101]. CVaR is a risk assessment measure to estimate the market risk or credit risk. In [98], the authors investigated the problem of energy trading among MGs, integrating the uncertainties in day-ahead and real-time market prices. A CVaR-based stochastic optimization method is proposed to address the optimization problem. In [99] and [100], the uncertainties in wind speed and market price are considered in the MG energy management problem. The authors used CVaR-based stochastic optimization to handle the risk of overall cost increase. In [101], a CVaR-based energy management approach is proposed to minimize operational cost while optimizing resilience in commercial building MGs. The CVaR is used to overcome the power generation and electricity market price uncertainties in this work.

Robust optimization is another optimization method used to address uncertainty in energy trading and management in MGs [102–104]. A mixed-integer non-linear robust optimization method is proposed in [102] to solve the day-ahead scheduling problem in MGs with SP, WT and micro-turbine generators and batteries. In this paper, variation in power generation of renewable resources, power demand and electricity price are considered the origin of uncertainty in the system. In [103], the uncertainties of renewable generation and load are expressed as an uncertain set generated by interval prediction. A mixed-integer robust optimization method is used for energy management which maximizes the total exchange cost while addressing the uncertainty. The authors of [104] proposed a robust optimization-based solution for multi-stage MG energy trading management to improve energy utilization efficiency. In this study, uncertainties in the generation of renewables

and demand challenge the operation of multi-stage MGs. A two-stage robust optimization is employed to formulate the mathematical problem, which is solved by the column-and-constraint generation algorithm.

In [105, 106], authors applied chance-constrained optimization to address the energy trading problem in MGs under uncertainties, while in [107, 108] interval optimization method is used to solve the energy trading problem considering the uncertainties. Despite all the advantages of the optimization-based approaches in dealing with uncertainty in MG systems, these methods suffer from various limitations. Optimization-based approaches heavily rely on the PDF of the uncertain parameters of the system. Assigning an accurate PDF to the unknown parameter is not feasible in most scenarios. Also, the behavior of unknown parameters might change over time, which necessitates finding a new PDF to model uncertainty in the system. High computational complexity is another burden for the optimization methods. While methods such as robust optimization reduce the complexity, these methods optimize worst-case scenarios, which is overly pessimistic. Furthermore, most optimization-based methods require a central controller to perform the optimization process, which is not a favorable option as the MGs evolve toward more independent and distributed architectures.

While game theory-based methods are promising approaches to address the problems in distributed systems, applications of these methods in energy trading and management of MGs under uncertainties are limited. In [109], a distributed Bayesian game theory-based method is proposed for the demand-side energy management of MGs. Each player's goal in the modeled Bayesian game is to optimize the expected payoff through a repeated two-player game of incomplete information. In this paper, players aim to overcome the uncertainty in the energy consumption of themselves and other players. In [110], the authors proposed a distributed strategy for energy-sharing among MGs. Bayesian Nash equilibrium is used to optimize the pricing model and the call-auction approach is adopted as the energy trading mechanism.

Machine learning techniques and specifically reinforcement learning techniques attract significant attention when dealing with MG systems under uncertainty due to their ability to adapt to an unknown environment and perform with minimum knowledge and supervision in a distributed manner. Therefore, reinforcement learning techniques can be a prominent solution to MG systems' energy management and trading problems under uncertainties. Various studies employed different reinforcement learning techniques in MGs systems. In [91], a multi-agent Bayesian DRL is proposed for energy management in MGs under power supply uncertainty and communication failure uncertainty. In this paper, the authors modeled the interaction of agents with the environment as a POMDP. The Bayes' theorem is employed to develop a belief update method in which agents can up-

date their belief about the actions of other agents in the case of communication failures. A double-DQN architecture is used in Bayesian DRL to approximate Q-values and avoid over-estimation.

In [111], the authors model the battery bank energy management problem as an MDP. It is assumed that uncertainty in environmental parameters can cause power flow to deviate from the anticipated values. Therefore, battery bank agents trained with a BRL-based method to balance any real-time power mismatch optimally in the integrated energy system.

A Bayesian regularized deep neural network is presented in [112] to guarantee robust and reliable power flow for MGs, particularly in the islanding mode. The uncertainty in renewable energy generation imposes a mismatch in the generation and demand of MGs. To this end, Battery energy storages are used to compensate for the inflicted mismatch. MGs employ the Bayesian regularized deep neural network to effectively manage the charge and discharge of batteries such that energy shortage is avoided.

In [113], the Bayesian ensemble method is used to train DRL agents for the energy management system of an extended range EVs for package delivery. Uncertainty estimation is conducted for each action during the package delivery, guaranteeing robust travel in unfamiliar new conditions.

In [114], the authors proposed a DRL-based energy scheduling of MGs under the uncertainty in the renewable energy generation, demand, and market energy price. The energy management problem in MGs is formulated as an MDP with the objective of minimizing the overall cost. The proposed DRL-based approach solves the MDP while overcoming the uncertainties.

The authors of [115] model integrated energy systems dynamic dispatch problem as an MDP and a DRL-based method is proposed as the solution to the designed MDP. The uncertainties in renewable generation and energy demand are considered in formulating the MDP.

2.4.3 Research Gap

In Table 2.2, the research attempts investigating the energy trading problem and uncertainty in MGs that were previously discussed are summarized. As demonstrated in the literature, game-theoretical methods such as CG-based methods can perfectly model energy trading problems in distributed MGs. Also, it is shown that BCG-based and reinforcement learning techniques can best address the uncertainty in the system without relying on a

central controller or having a stochastic model for the source of uncertainty as it is required in optimization methods. However, most coalition formation studies investigate energy trading in static scenarios. In this thesis, we solve energy trading problems among MGs considering various sources of uncertainty. We employ BCG, BRL and DRL to address uncertainty challenges while minimizing cost and power loss in energy trading among MGs.

Table 2.2: Summary of research works on energy trading and uncertainty problems.

Category	Work	Method	Elements	Objective
Optimization	[62]	linear programming	MG, macrogrid, battery, renewable resources	optimal energy scheduling
	[63]	alternating direction method of multipliers	MG, macrogrid, battery, renewable resources	distributed multiclass energy management
	[64]	mixed integer linear programming	MG, macrogrid, PV	minimize net energy cost
	[65]	epsilon constraint	MG, macrogrid	optimal energy scheduling
	[66]	primal-dual gradient method	MG, macrogrid, renewable resources	maximize overall welfare
	[67]	bi-level optimization, mixed-integer linear programming	MG, macrogrid, EVs, battery, renewable resources	minimizing total dispatch cost, energy trading costs and costs of batteries degradation
	[96]	particle swarm optimization	MG, macrogrid, renewable resources	energy trading management considering uncertainty
	[97]	stochastic mixed-integer linear programming	MG, macrogrid, SP, WT	scheduling problem in MGs considering uncertainty

	[98]	CVaR-based stochastic optimization	MG, macrogrid, renewable resources	energy trading among MGs considering uncertainty
	[99]	CVaR-based stochastic optimization	MG, macrogrid, WT	MGs energy management considering uncertainty
	[100]	CVaR-based stochastic optimization	MG, macrogrid, WT	MGs energy management considering uncertainty
	[101]	CVaR-based stochastic optimization	MG, macrogrid, EVs, SP	MGs energy management considering uncertainty
	[102]	mixed-integer non-linear robust optimization	MG, macrogrid, battery, renewable resources	day-ahead scheduling in MGs considering uncertainty
	[103]	mixed-integer robust optimization	MG, macrogrid, renewable resources	MGs energy management considering uncertainty
	[104]	robust optimization	MG, macrogrid, renewable resources	energy trading among MGs considering uncertainty
	[105]	chance-constrained optimization	MG, macrogrid,renewable resources	energy trading among MGs considering uncertainty
	[106]	chance-constrained optimization	MG, macrogrid,renewable resources	energy trading among MGs considering uncertainty
	[107]	interval optimization	MG, macrogrid, renewable resources	energy trading among MGs considering uncertainty
	[108]	interval optimization	MG, macrogrid, renewable resources	energy trading among MGs considering uncertainty
Game-Theory	[68]	Stackelberg game	MG, macrogrid, battery	minimize net energy cost
	[69]	Stackelberg game	MG, macrogrid	maximize the sum of social welfares

	[70]	Stackelberg game	MG, macrogrid	minimize net energy cost
	[71]	Stackelberg game	MG, macrogrid	minimize net energy cost
	[73]	Stackelberg game	MG, macrogrid	minimze total electricity demand
	[74]	non-cooperative game	MG, macrogrid, battery, renewable resources	minimize net energy cost
	[75]	non-cooperative evolutionary game, Stackelberg game	MG, macrogrid, battery, PV	maximize the sum of social welfares
	[76]	double auction	MG, macrogrid, battery, renewable resources	overall cost
	[77]	double auction	MG, battery, renewable resources	maximize overall welfare
	[78]	double auction	EVs, macrogrid	social welfare maximization
	[109]	Bayesian game theory	MG, macrogrid, renewable resources	demand-side energy management of MGs considering uncertainty
	[110]	Bayesian game theory	MG, macrogrid, renewable resources	energy trading among MGs considering uncertainty
Coalition Formation	[7]	CG	MG, macrogrid	minimize power loss
	[79]	CG	MG, macrogrid	energy management
	[80]	CG	MG, macrogrid	energy management
	[81]	CG	MG, macrogrid	minimize net energy cost
	[82]	CG	MG, macrogrid	minimize net energy cost

Reinforcement Learning	[72]	learning automaton	MG, macrogrid	average utility maximization
	[92]	Q-learning	MG, macrogrid, battery	average utility maximization
	[93]	DQN	MG, macrogrid, battery, renewable resources	average utility maximization
	[94]	DQN	MG, macrogrid, battery, renewable resources	average utility maximization
	[95]	Q-learning	MG, macrogrid	maximize overall
	[91]	multi-agent Bayesian DRL	MG, macrogrid, renewable resources	MGs energy management considering uncertainty
	[111]	BRL	MG, macrogrid	MGs energy management considering uncertainty
	[112]	Bayesian regularized deep neural network	MG, macrogrid, battery, renewable resources	MGs energy management considering uncertainty
	[113]	DRL	MG, macrogrid, EVs	EVs energy management considering uncertainty
	[114]	DRL	MG, macrogrid, renewable resources	energy scheduling of MGs considering uncertainty
	[115]	DRL	MG, macrogrid, renewable resources	dynamic dispatch considering uncertainty

Chapter 3

Power Loss-Aware Transactive MG Coalitions under Uncertainty

3.1 Introduction

Peer-to-peer energy trading within MG communities emerges as a key enabler of the future transactive distribution systems and the transactive electricity market. In many modern MGs, EVs have been considered a viable storage option due to their ease of use (plug-and-play) and their growing adoption rates by drivers. On the other hand, the dynamic nature of EVs escalates the uncertainty in the transactive distribution system. The present chapter investigates the problem of energy trading among MGs and EVs with the aim of power loss minimization, where the location of EVs imposes uncertainty. A novel BCG-based algorithm [51] is proposed, which allows the MGs and EVs to reduce the overall power loss by allowing them to form coalitions intelligently [116]. The proposed scheme is compared with a conventional CG-based approach and a Q-learning-based approach.

The rest of this chapter is organized as follows. In Section 3.2, the system model is described. In Section 3.3, the BCG scheme is explained. Numerical results are provided in Section 3.4, and finally, the conclusion remarks are presented in Section 3.5.

3.2 System Model

We consider a network of interconnected MGs and EVs as illustrated in Fig. 3.1. The network under consideration includes M_{MG} MGs, M_{EV} EVs and M_{CS} charging stations.

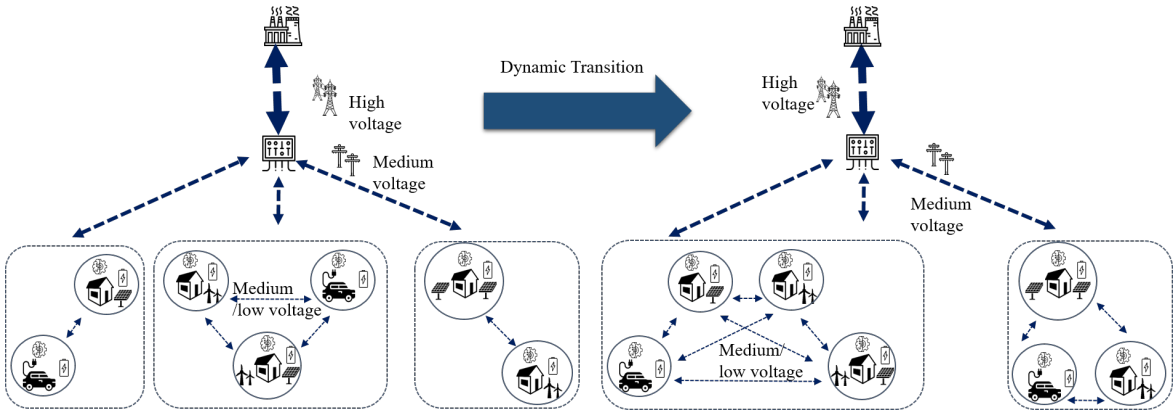

Figure 3.1: Block diagram of a system of MGs. Due to the dynamicity of the system, different coalitions can be formed during each epoch.

Each MG is individually connected to a utility grid (macrogrid), and EVs are connected to the network through charging stations. The MGs and EVs trade energy among themselves and/or with the macrogrid. For a given time slot, each MG $i \in M_{MG}$ may generate power e_i^g and may have demand denoted by e_i^d. EVs are equipped with a battery with a capacity of b_i^g and may have demand denoted by e_i^d. Therefore, the surplus energy (or shortage if the load exceeds generation) of the MGs can be defined as $e_i^q = e_i^g - e_i^d$, which denotes the amount of energy that the MG would potentially export to/import from the network. Similarly, for EVs $e_i^q = e_i^b - e_i^d$ represents surplus (or demanded) energy. It is considered that, in a given period, some MGs and EVs have a surplus (or shortage) of energy and desire to enter energy trading with the transactive distribution system. EVs need to drive to a charging station to participate in a coalition. It is assumed that EVs would rationally drive to the closest charging station to have minimum driving cost. Considering the generation and demand levels, at each epoch, an MG or an EV may move from the seller group to the buyer group or vice versa. Energy trading among the MGs and EVs (through the charging stations) results in power loss over the distribution lines. Given the line resistance $\mathbb{R}_{ij}$ between agent i and agent j per km and voltage U_i, power loss can be calculated as: [7, 117, 118]:

$$PL(E_{ij}) = \mathbb{R}_{ij} D_{ij} \frac{{E_{ij}}^2}{U_i^2} + \rho E_{ij}, \tag{3.1}$$

where ρ expresses the fraction of power loss that happens in the transformer at the interconnection point between the MGs and the macrogrid (macro station) and D_{ij} denotes distance in the energy transfer between i-th and j-th MGs and index 0 represent transac-

tion with macrogrid. It should be mentioned that to satisfy the total demand e_i^q of agent j in transferring power E_{ij} from MG i the following constraint should be satisfied since a portion of power will be lost in the line:

$$E_{ij}(e_j^q) = e_j^q + PL(E_{ij}(e_j^q)) \tag{3.2}$$

This equation may have zero, one, or two solutions for a set of given parameters. In the case of two positive solutions, the smallest root is considered. Whenever Eq. (3.2) does not have solution, we consider that the energy of $\frac{(1-\rho)U_i^2}{2\mathbb{R}_{ij}}$ transmitted from macrogrid [7].

3.3 BCG with Transferable Utility

To reduce transmission power loss from a distant macrogrid, MGs and EVs can participate in energy trading by joining a coalition. Forming cooperative groups and trading energy among the close by MGs (coalitions), i.e., peer-to-peer energy trading, is a promising approach to reduce the transmission power loss since the line loss depends on the distance and the power. Note that in most of the previous game-theoretical frameworks implemented in the energy trading problems, each player's goal is to increase its revenue through efficient selling and buying of energy. Unlike those studies, in this chapter, the goal of each rational agent is to minimize its objective function, which is its share of power loss and consequently decrease the total coalitional power loss. A coalition is determined with a pair (C, v) where C denotes a set of agents that agree to form a coalition, and v represents the value function which calculates the total payoff of the coalition. In this scheme, a coalition can have several different coalition values, v, according to the internal energy trading policy. Setting energy trading within a coalition in an optimized way can result in the optimal coalition value which is minimum power loss. The optimal coalition value is defined as:

$$v_{max}(C) = max\left[-\left(\sum PL(E_{ij}) + \sum PL(E_{i0}) + \sum PL(E_{0j})\right)\right], \tag{3.3}$$

where $PL(E_{ij})$ shows the loss due to energy trading among seller i and buyer j while $PL(E_{i0})$ and $PL(E_{0j})$ represent the situation that the coalition has an overall surplus generation (agent i transfer energy to macrogrid) and demand (agent j receive energy from macrogrid), respectively. Algorithm 3.1 is implemented to ensure the least possible power loss occurs in a coalition transaction in each time slot. The output of this algorithm will be used in Algorithm 3.2 when calculating the coalition value later in Eq. (3.4).

Algorithm 3.1 Greedy energy transaction to achieve $v_{max}(C)$

Initialization: Divide the members of coalition C to two groups of buyers and sellers.
Main loop:
while $\sum_{i \in C} |e_i^q| \neq 0$ **do**
    Find the shortest distance seller i and buyer j with $e_i^q \& e_j^q \neq 0$
    if $\sum_{i \in C} e_i^q > 0 \,\&\, PL(E_{ij}(min\{e_i^q, e_j^q\}) > PL(E_{i0}(min\{e_i^q, e_j^q\})$ **then**
        Do the energy transaction of amount $\min\{e_i^q, e_j^q\}$ between seller i and macrogrid.
    else if $\sum_{i \in C} e_i^q < 0 \,\&\, PL(E_{ij}(min\{e_i^q, e_j^q\}) > PL(E_{0j}(min\{e_i^q, e_j^q\})$ **then**
        Do the energy transaction of amount $\min\{e_i^q, e_j^q\}$ between buyer j and macrogrid.
    else
        Do the transaction of amount $\min\{e_i^q, e_j^q\}$ between seller i and buyer j.
    end if
    Update e_i^q, e_j^q and total power loss.
end while
$v_{max}(C)$ is equal to negative of total power loss.

In this chapter, to preserve agents' privacy, it is assumed that coalition members do not have any information regarding the location and type of other agents (EVs and MGs), and a coalition leader is responsible for computation and sharing utilities. Therefore, the agents should be equipped with a technique to overcome this lack of information, which imposes uncertainties on the coalition values, which in turn causes uncertainty in the system. The uncertainty can be explained at two levels. First, the type of coalition members is unknown to the other members. Second, the locations of the agents are unknown. Therefore, computing the payoff (utility) is not straightforward. This introduces uncertainty as EVs in a specific coalition charge and discharge in different charging stations, which varies the power loss in the coalition. Therefore, it is critical to employ a method that allows agents to learn about others agents to overcome uncertainties and refine the coalition formation process. In this chapter, we propose a BCG-based method that aids agents in overcoming the uncertainties about other agents' types. BCG method is explained in Section 2.3.1. We assume that two types of agents are involved in the game: fixed and moving agents.

- Fixed: MGs are considered as fixed agents. Power loss resulting from selling or buying unit power from a specific MG is always constant for these agents.
- Moving: Since the EVs are driving, they might not always charge or discharge at the

same charging station, which results in a varying power loss when transferring power to a specific destination. EVs are assumed to utilize the closest charging station when they participate in coalitions.

Without loss of generality and for the sake of simplicity, we assume that there are two charging stations in the network. Also, it is assumed that EV i travels to station 1 with probability $1-\epsilon_i$ and the other station with probability ϵ_i. MGs and EVs have no information about the type of the other agents. Two observation errors with Bernoulli distribution [119] are considered. p_e is the probability of observation error when a fixed behavior of an agent is observed as moving, while the chance of having correct observation of a fixed behavior is $1-p_e$. Also, p_c represents the chance of observation error when a moving behavior of an agent is observed as fixed while the chance of having correct observation of a moving behavior is $1-p_c$. p_e and p_c are assumed close to zero and happen due to several reasons such as communication error or Global Positioning System (GPS) inaccuracy. The BCG is formulated as follows:

- *Agents*: The set of M_{MG} interconnected rational MGs and M_{EV} rational EVs.
- *Types*: It is assumed that system includes two types of agents, fixed type T_{fixed} and moving type T_{moving}. Agents have full information regarding their own type, while the types of others can not be observed at the beginning. T_{fixed} transfer energy from the same location with probability 1. On the other hand, T_{moving} agents may use default station '1' or station '2' with the probabilities $1-\epsilon_i$ and ϵ_i, respectively. When T_{moving} agents use the station '1', their behavior will be observed as fixed behavior by other agents. It is assumed that agents' types will be constant for a long period.
- *Expected Payoff*: Eq. (3.3) formulates the total utility of the coalition. However, each user in the coalition should have its own share of this utility. To share the total utility of a coalition, we use the proportional fair division algorithm [120] which divides the total utility among the members of a coalition proportional to their power loss when trading energy only with the macrogrid (known as transfer function [120]). Therefore, the immediate payoff of the i-th agent, which is the share of utility, can be expressed as:

$$u^i = \zeta_i \left(v_{max}(C) - \sum_{j \in C} v(\{j\}) \right) + v(\{i\}) \tag{3.4}$$

where ζ_i, the relative ratio of their contribution, equals to $\frac{v(\{i\})}{\sum_{j \in C} v(\{j\})}$ and $v(\{j\})$ represents the power loss or value of a coalition with single member agent j.

In the BCG, we use expected payoff (utility) realized by the agents rather than immediate payoff since the system is dynamic and the payoff of coalition changes with respect to the type of coalition members. This will help to overcome the uncertainty and consequently form a stable coalition. Since we consider a discrete space, we can denote the expected payoff realized by i-th agent as follows:

$$\overline{u}^i\left(C, \mathbb{T}^C\right) = E\left[u^i\left(C, \mathbb{T}^C\right)\right] = \sum_{\mathbf{T}^C \in \mathbb{T}^C} B^i\left(\mathbf{T}^C\right) u^i\left(C|\mathbf{T}^C\right) = \sum_{k=1}^{2^{|C|-1}} B^i\left(\mathbf{T}_k^C\right) u^i\left(C|\mathbf{T}_k^C\right) \tag{3.5}$$

where $\mathbf{T}^C$ denotes a possible types vector of members of coalition C, $\mathbb{T}^C$ expresses a set of all possible types vectors of coalition C and $\mathbf{T}_k^C$ shows the the k-th possible types vector of members of coalition C. Since in our scenario there are two possible types for each agent, the total number of possible combinations of other agents' types believed by agent i is equal to $2^{|C|-1}$. $B^i\left(\mathbf{T}_k^C\right)$ is joint belief probability of agent i about other agents with respect to index k which represents the k-th possible combination of agents' types believed by agent i. $u^i\left(C|\mathbf{T}_k^C\right)$ shows the immediate payoff which can be achieved by agent i from k-th possible combination of agents' types. $B^i\left(\mathbf{T}_k^C\right)$ can be found as:

$$B^i\left(\mathbf{T}_k^C\right) = \prod_{j \in C \setminus \{i\}} p_{ij}\left(T_x\right), \tag{3.6}$$

where $T_x \in \{T_{fixed}, T_{moving}\}$ and $p_{ij}\left(T_x\right)$ denotes the belief probability of agent i that the type of agent j is T_x. For the sake of brevity we assume $p_{ij}(T_{fixed}) = b_{ij}$ and consequently $p_{ij}(T_{moving}) = 1 - b_{ij}$. We elaborate on the definition of expected utility with an example in Appendix A.1.

- *Action*: All coalition members agree on a coalitional action of accepting or rejecting a joining proposal.
- *Stability*: The concept of WBC is adopted to guarantee the stability of BCG.

The users build their preferences according to their expected payoff. The expected payoff captures the beliefs about the type of other agents. However, these beliefs should change over time with the observation to represent the correct type for other agents. Therefore, an algorithm to update beliefs is essential. The concept of the Bayes' theorem [119] is employed to derive an equation for belief update. Considering an agent j, which sold (or purchased) energy, then the type of agent j for the traded energy can be either

of the two observations, which are denoted as O_{fixed} and O_{moving}. Consequently, we are interested in calculation of $p_{ij}(T_{fixed}|O_{fixed})$ and $p_{ij}(T_{fixed}|O_{moving})$. According to Bayes' theorem [119, 121], agent i can update its belief about agent j as follows (see Appendix A.2 for extended details):

$$p_{ij}^{\tau_{ij}}(T_{fixed}|O_{fixed}) = \frac{b_{ij}^{\tau_{ij}}(1-p_e)}{b_{ij}^{\tau_{ij}}(1-p_e) + (1-b_{ij}^{\tau_{ij}})((1-\epsilon_{ij}^{\tau_{ij}+1})(1-p_e) + \epsilon_{ij}^{\tau_{ij}+1}p_c)}, \tag{3.7}$$

$$p_{ij}^{\tau_{ij}}(T_{fixed}|O_{moving}) = \frac{b_{ij}^{\tau_{ij}}p_e}{b_{ij}^{\tau_{ij}}p_e + (1-b_{ij}^{\tau_{ij}})(\epsilon_{ij}^{\tau_{ij}+1}(1-p_c) + (1-\epsilon_{ij}^{\tau_{ij}+1})p_e)}, \tag{3.8}$$

where τ_{ij} denotes the number of iteration that agent i observed agent j. If the agent j is a moving agent, then the belief probability of agent i that agent j will transfer energy from charge station 2 is denoted by ϵ_{ij}. $\epsilon_{ij}^{\tau_{ij}+1}$ can be updated as the weighted sum of previous values $\epsilon_{ij}^{\tau_{ij}}$ and current value $\epsilon_{ij}^{\tau_{ij}^*}$ using exponential moving average [122]:

$$\epsilon_{ij}^{\tau_{ij}+1} = \varpi\epsilon_{ij}^{\tau_{ij}^*} + (1-\varpi)\epsilon_{ij}^{\tau_{ij}} \tag{3.9}$$

where ϖ is adjustable constant and $\epsilon_{ij}^{\tau_{ij}^*}$ can be found using Bayes' theorem as follows:

$$\frac{\left|\chi_{ij}^{\tau_{ij}+1}(O_{fixed})\right|}{\left|\chi_{ij}^{\tau_{ij}+1}\right|} = b_{ij}^{\tau_{ij}}(1-p_e) + (1-b_{ij}^{\tau_{ij}})((1-\epsilon_{ij}^{\tau_{ij}^*})(1-p_e) + \epsilon_{ij}^{\tau_{ij}^*}p_c), \tag{3.10}$$

where $\left|\chi_{ij}^{\tau_{ij}+1}(O_{fixed})\right|$ and $\left|\chi_{ij}^{\tau_{ij}+1}\right|$ denote the number of actual fixed and total observations, respectively. In Eq. (3.10), the right hand side expression shows theoretical equivalent of $p_{ij}(O_{fixed})$ while the left hand side expression is the probability calculated based on the the actual observations. Therefore we can obtain $\epsilon_{ij}^{\tau_{ij}*}$ as:

$$\epsilon_{ij}^{\tau_{ij}*} = \frac{1-p_e-\frac{\left|\chi_{ij}^{\tau_{ij}+1}(O_{fixed})\right|}{\left|\chi_{ij}^{\tau_{ij}+1}\right|}}{(1-b_{ij}^{\tau_{ij}})(1-p_e-p_c)}. \tag{3.11}$$

We then use Eq. (3.9) and Eq. (3.10) to find the posterior in Eq. (3.7) or Eq. (3.8). We use these updated probabilities in Algorithm 3.2 to form coalitions. In each iteration,

we assume one agent is randomly chosen, and it proposes to join a new coalition. The new coalition will accept a new member if, all existing members will be able to maintain their current expected payoff or achieve a higher expected payoff value (known as Pareto order [123]) in the new coalition, while the expected payoff values for other agents remain intact (or improved). After forming new coalitions, coalition members use the observation information to update their beliefs about other members of their coalition according to Eq. (3.7)- Eq. (3.11). At the end of each iteration, Algorithm 3.1 is executed in each coalition to plan energy trading among coalition members in a way that minimizes power loss inside the coalition. A Consecutive merge and split iterations happen until the system of agents reach a coalition formation, from whereon no agent has any incentive to further merge to a new coalition based on its realized expected utility.

Algorithm 3.2 BCG formation for energy trading among MGs and EVs

1.6

Initialization: Randomly assign all agents to the coalitions
Main loop:
for Each time slot $t = 1$ to $\mathcal{T}$ **do**
    for Agent $i = 1$ to $M_{MG} + M_{EV}$ **do**
        Update current payoff $u^i(t)$
    end for
    Coalition formation
With probability of $1/(M_{MG} + M_{EV})$ choose agent i as the proposer;
    For agent i:
    Send the joining proposal to all agents m, $m \in C_k / \{i\}$
    If $\overline{u}\{i \in C_k\} \geq \overline{u}\{i \notin C_k\}$ for all $m \in C_k / \{i\}$ then set $i \in C_k$ and update the u for $m \in C_k / \{i\}$.
    Update Eq. (3.7)- Eq. (3.11) for all the members of coalition C_k
end for

Baseline I - Q-Learning-based Coalition Formation: To compare the proposed technique to a well-known machine learning based solution, we employ Q-learning based approach as it is explained in Section 2.3.2. The immediate reward of agent i can be calculated in the same way as immediate payoff in Eq. (3.4) as follows:

$$r_i(s,a) = \zeta_i \left(v_m ax(C) - \sum_{j \in C} v(\{j\}) \right) + v(\{i\}). \tag{3.12}$$

MGs and EVs are assumed as the agents, and the agent's action is to refuse or accept the joining proposition to their coalition from the proposer agent while the state is the vector of coalition memberships. Epsilon-greedy method [124] is employed in order to consider action exploration.

Baseline II - CG-Based method: We implement the CG-Based method proposed in [7]. Besides Q-learning and CG-based approaches, we also compare the proposed approach with a case where there are no coalitions.

3.4 Performance Evaluation

3.4.1 Simulation Parameters and Setup

For the numerical evaluation, we consider that the number of MGs and EVs ($M_{MG}+M_{EV}$) varies between 4 and 10, which is a realistic assumption considering the actual cases of community MGs. Interconnected MGs and macrogrid are located at random locations in a 10km by 10km area [7]. A day is divided into 24 time slots. Each time slot is considered as one iteration in the simulation. In each iteration, one snapshot of the Algorithm. 3.2 is executed. EVs have varying levels of energy where surplus or demand patterns are generated randomly based on a Gaussian random variable and periodically repeated after a day with slight variations as in [7]. The mean and variance values of generation and demand are adopted from [125]. Observation error values (p_e and p_c) are chosen analogous to [121]. We assume the resistance, $\mathbb{R}_{ij}$, is the same in all the lines. Energy transfer with the macrogrid happens in medium voltage U_0, energy transfer among MG happens in low voltage U_i, and the losses inside the MG are not considered. The simulation parameters are summarized in Table 3.1. Physical parameters ($\mathbb{R}_{ij}$, U_0, U_i) are selected based on [126]. The proposed BCG method is compared with a CG-based method and a Q-learning-based method, as well as when there are no coalitions. The results are obtained from 15 runs, and each run includes at least 5000 iterations. 95% confidence interval is considered in the demonstrated results.

3.4.2 Simulation Results

In Fig. 3.2, the average power loss per user versus the number of varying MGs and EVs is presented. Average power loss per user (MGs or EVs) can be computed by dividing the total power loss (calculated with Eq. (3.1)) by the number of agents (MGs and EVs).

Table 3.1: Summary of simulation parameters.

parameters	**value**
Line resistance (R_{ij})	0.2
Medium voltage (U_0)	50 kv
Low voltage (U_i)	22 kv
Transformer loss fraction (ρ)	0.02
Moving average parameter (ϖ)	0.6
Observation error (p_e)	0.1
Observation error (p_c)	0.01
Learning rate of Q-learning (α)	0.5
Discount factor of Q-learning (γ)	0.8

The number of MGs ranges from 4 to 10, and there are 2 EVs. The proposed scheme is compared with the CG-based, Q-learning, and no coalition schemes, showing that the BCG scheme has approximately %15 and %40 less power loss than Q-learning and CG-based schemes, respectively. Furthermore, it provides a significant advantage over not having coalitions. As expected, when the number of MGs increases, the power loss is less for both schemes since the distance between agents is shorter.

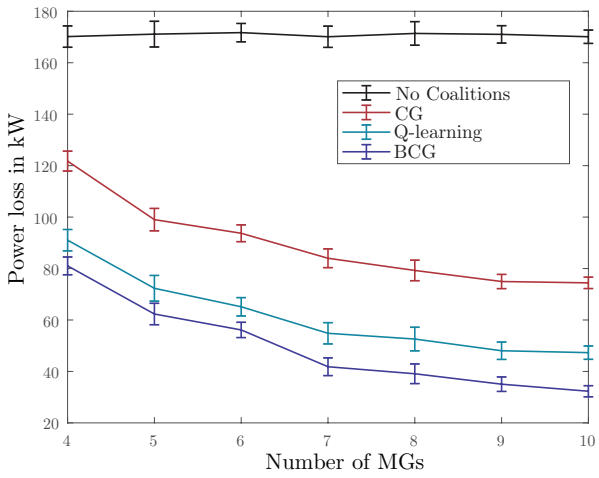

Figure 3.2: Average loss per user versus the number of MGs.

In Fig. 3.3, it is demonstrated that the average power loss per user versus the percentage of EVs ranges from 0 to 75 percent. In these simulations, the total number of agents is set

to 10, where the number of MGs changes from 2 to 8. Furthermore, it is shown that when the percentage of EVs increases, the average power loss increases as expected. This is due to the increased uncertainty introduced by the EVs. Besides, it can be observed that the BCG method outperforms other methods as the uncertainty in the system increases.

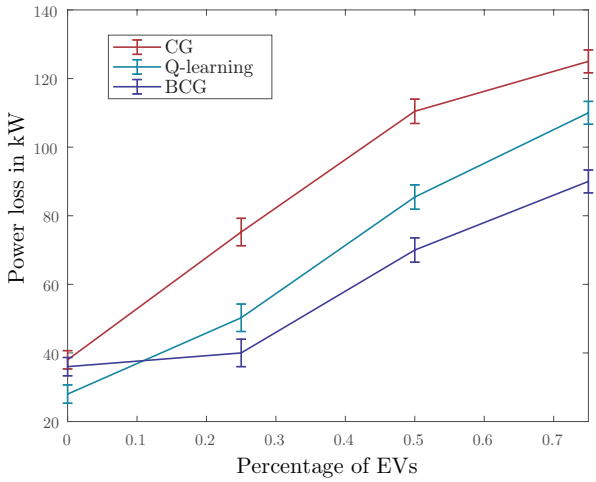

Figure 3.3: Average power loss per user versus the percentage of EVs.

In Fig. 3.4, the convergence of the belief of the MG agent over the behavior of EVs is shown. To demonstrate this, a system with one MG and two EVs is considered. Values of $\epsilon_{12}^{\tau_{12}}$ and $\epsilon_{13}^{\tau_{13}}$ are set to be arbitrary values 0.6 and 0.3 respectively. The aim of this evaluation and setting a predefined value for $\epsilon_{ij}^{\tau_{ij}}$ is to confirm the accurate convergence of the algorithm. As seen in the figure, estimated parameters converge to the actual values over time.

In real-life scenarios, multiple charging stations could be available in a specific area. In Fig. 3.5, The number of charging stations that EVs can visit is increased. We show the power loss for BCG, Q-learning and CG-based schemes when we have two and four charging stations. In this scenario, we have 2 EVs out of 8 agents. The results demonstrate that as the number of charging stations increases, power loss increases as well. This is because, with the increase in the number of charging stations (uncertainty about which charging station will be used increases), the agents have more difficulty reaching optimal coalitions, resulting in more power loss. Nevertheless, BCG is able to incur less loss than other schemes. To better overcome the uncertainty in BCG, we need to modify the belief probabilities Eq. (3.7)- Eq. (3.11) based on the number of charging stations. To this end,

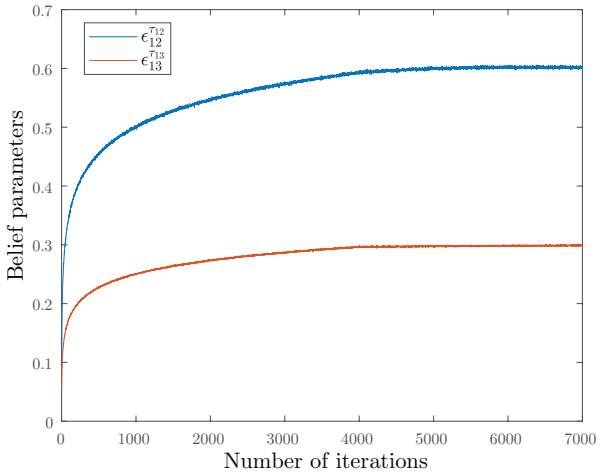

Figure 3.4: The convergence of belief parameters $\epsilon_{ij}^{\tau_{ij}}$ of MG 1 about EVs 2 and 3 versus number of iteration.

in Appendix A.3, generalized belief probabilities are derived for any number of charging stations.

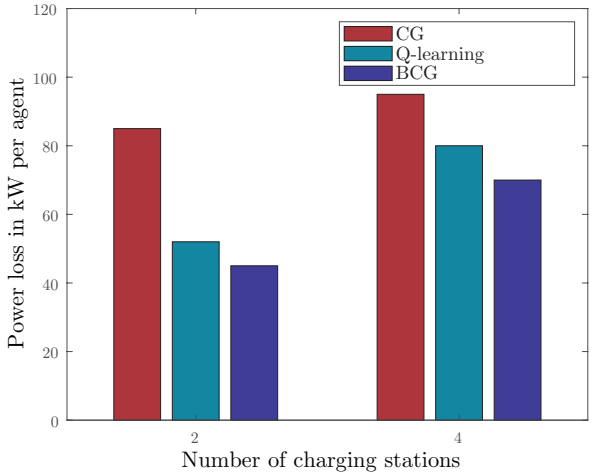

Figure 3.5: Average power loss versus the degree of freedom.

3.5 Conclusion

The future transactive energy systems will be built on peer-to-peer energy trading and MG communities. As EV penetration grows dramatically, EVs will become fundamental elements in the energy system. However, the mobility of EVs and their high penetration impose various uncertainties in the system, such as unexpected load, which may accumulate in peak hours or hot spots, or uncertainty about which charging station to charge/discharge from when using public charging stations. In this chapter, the problem of energy trading was visited with the aim of power loss minimization considering the uncertainties introduced by the existence of EVs as dynamic elements in the system. A novel BCG approach was proposed to form optimized coalitions of MGs and EVs, which results in less energy transfer from macrogrid or distant MGs while overcoming the uncertainty introduced by EVs. In this method, players/agents (MGs and EVs) refine their belief about the type of other players with iterative observations of the environment until converging to their final belief. Whereas CG aims to maximize instantaneous payoff, in this method, the goal of each agent is to optimize its expected payoff through the various iterations. Comparing the proposed approach with a CG-based and a Q-learning-based approaches, we achieved a reduction of 16% in power loss compared to the Q-learning-based method. BCG can be used to overcome different sources of uncertainties. While in this chapter we employed BCG to address uncertainty imposed by the location of EVs, in the next chapter BCG is used to overcome the uncertainty resulting from the generation and demand predictions.

Chapter 4

Cost-Aware Dynamic BCG for Energy Trading among MGs

4.1 Introduction

The variations in demand and generation and the dynamic nature of MG communities sometimes result in a false prediction of MGs' generation and demand in future epochs. Consequently, such MGs will fail to satisfy their trading commitment which imposes uncertainties in the system. In this chapter, the problem of energy trading among MGs is revisited to minimize the cost, considering the uncertainties imposed by the lack of information about whether the MGs can fulfill their commitment. A BCG-based method is proposed, which helps the MGs to minimize the overall cost by forming stable coalitions. In our approach, each MG considers a prior belief over the commitment (known as type) of other MGs. MGs can update and enhance the precision of their belief estimations by interacting with the environment through the iterations of BCG, which finally leads to a coalition formation that minimizes cost [127].

The rest of this chapter is organized as follows. The system model is demonstrated in Section 4.2. In Section 4.3, the BCG is formulated and explained. Numerical results are presented in Section 4.4, and finally, the chapter is concluded in Section 4.5.

4.2 System Model

We consider a network of M interconnected MGs where each MG is also connected to the macrogrid (utility grid). We assume that each MG i $(1 < i < M)$, generates power e_i^g and has demand e_i^d. Consequently, the total surplus energy (or demand) of the MGs can be denoted as $e_i^q = e_i^g - e_i^d$ and shows the amount of energy that the MG is expected to export (or import) from the other MGs. Therefore MGs enter an energy trading process with each other and with the macrogrid. At each iteration, each MGs may move from buyers' position to seller position or vice versa according to their generation and demand levels.

Energy trading among the MGs (and with the macrogrid) is associated with various kinds of costs. We consider two sets of costs in our model. First, the *operational cost* that includes the power loss in the line, transformer loss from high to medium or low voltages, maintenance cost etc. However, this physical cost is not the only cost in the system. For example, imagine two distant MGs want to trade energy. Practically, having a direct distribution line between all the MGs is not feasible, and most of the time, only close by MGs may be interconnected. Therefore we assume that when there is no distribution line between MGs, the seller MG exports energy to the macrogrid (or an external storage operator), and the buyer MG imports from the macrogrid or the storage operator. Besides, one aspect of energy trading among MGs is islanding capability. Any reliance on energy trading with the macrogrid will put the chance of islanding at risk, which could impose further costs when a coalition of MGs is in islanding mode. All of these third parties involved in energy trading bring about extra unforeseen costs, which are termed in this chapter as *virtual costs*. Therefore, we can define the total cost of transferring power E_{ij} from i-th MG to j-th MG as follows:

$$S_{ij} = \omega D_{ij} E_{ij} + \delta PL(E_{ij}), \tag{4.1}$$

where δ and D_{ij} denote the scaling factor and the distance in the energy transfer between i-th and j-th MGs. The second term reflects the cost introduced with power loss in which $PL(E_{ij})$ shows the amount of power loss in transferring energy between i-th and j-th and can be obtained as in Eq. 3.1 which is scaled by the coefficient δ. ω is a weighting coefficient of virtual cost, which can be found as:

$$\omega = \begin{cases} \omega_s & i, j \neq 0 \ and \ D_{ij} \leq D_{tr} \\ \omega_l & i, j \neq 0 \ and \ D_{ij} > D_{tr}. \\ \omega_0 & i = 0 \ or \ j = 0 \end{cases} \tag{4.2}$$

We assume that the virtual cost is a linear function of distance and transferred energy that is weighted by the parameter ω which is considered to be a small value ω_s for MGs closer than a threshold D_{tr} and larger value ω_l for distant MGs further than a threshold D_{tr}. We assume that there is no direct line between distant MGs, and as a result, the weight associated with virtual cost increases compared to close by MGs. Note that index 0 stands for transferring energy with macrogrid, and ω_0 denotes the weight of transferring energy with macrogrid, which is larger than ω_s and ω_l. When ω_0 increases, the chance of islanding and less energy transaction with macrogrid increases.

Finally, the goal of the defined systems is to minimize the total cost. Therefore, the objective function is formulated as follows:

$$\begin{aligned} &\min \sum_{i=0}^{M} \sum_{j=0}^{M} S_{ij} \\ &s.t: \sum_{i=1}^{M} e_i^q + \sum_{i=0}^{M} \sum_{j=0}^{M} P(E_{ij}) = \sum_{j=0}^{M} E_{0j} - \sum_{i=0}^{M} E_{i0}. \end{aligned} \tag{4.3}$$

The constraint refers to the fact that the total extra surplus/demand should be equal to the total energy transferred or received from the macrogrid. To reduce the total cost, we consider that MGs prefer to trade energy with nearby MGs, resulting in less cost than energy transfer with distant MGs and macrogrid. Therefore, forming cooperative energy trading groups among the close by MGs, known as a coalition of MGs, is a promising method to decrease the total cost. Furthermore, there will be zero transformer loss in the medium or low voltage, which is the operating range of energy transfer among MGs. A coalition can be defined with a pair (C, v) which, C denotes the members of a coalition that cooperate to achieve a higher coalition value v [128]. In our problem, we define the coalition value as the amount of cost imposed during energy transfer among coalition members plus the cost of extra surplus power or extra demand that should be exported to or imported from the main grid. Since we are interested in minimizing the cost, we can define the coalition value as the negative form of cost as follows:

$$v_C = -\sum_{i=0}^{|C|} \sum_{j=0}^{|C|} S_{ij}, \tag{4.4}$$

where $|C|$ is the number of members of coalition C. It should be noted that index 0 stands for importing energy from (exporting to) the macrogrid. After forming the coalition, we have to schedule the energy transfer in a way that minimizes the cost in the coalition. Therefore, we define the coalition utility (payoff) in the proposed system as the maximum

achievable coalition value $v_{max}(C)$ which can be obtained as

$$v_{max}(C) = \max \left\{ -\sum_{i=0}^{|C|} \sum_{j=0}^{|C|} S_{ij} \right\}. \tag{4.5}$$

In order to manage energy transfer inside the coalition and achieve $v_{max}(C)$, an algorithm similar to the one in Algorithm 3.1 is used. Each MG estimates surplus energy or demand for the next hour (next iteration). However, this estimation is not always perfect and may result in the coalition exporting or importing energy from a distant macrogrid. Therefore, it is necessary to equip MGs with a tool to overcome uncertainty regarding MGs' commitments and reach a sub-optimal coalition formation. In the following section, we present a BCG approach to address the uncertainty in the system.

4.3 Cost-Aware BCG

In this section, we propose a BCG game that aims to apprehend the uncertainty of MGs' "type" (type of their commitment to a coalition). We assume that MGs predict their future generation and demand. However, not all of the MGs have an accurate prediction. Therefore, they may not be able to deliver their commitments to their coalitions. We assume that there are two sets of MGs in the game as follows:

- Committed MGs which always deliver or accept the promised amount of energy in the coalition
- Uncommitted MGs according to their false prediction or their local benefits sometimes can not satisfy their energy commitment to the coalition and either export/import half or none of the promised amount of energy[1].

MGs have no information about the type of the other MGs at the beginning. So, first, we formulate our scenario as a BCG which is explained in Section 2.3.1. In our systems, players are a group of M interconnected and rational MGs. Each coalition member chooses to accept or reject a joining proposal from non-member MG based on its own expected payoff, which will be explained in more detail later. In this system, MGs are assumed to have two types, committed, T_C, and uncommitted, T_{UC}, types. MGs have

[1]Various energy levels can be considered however for simplicity, here, we chose to develop the model with three levels.

Table 4.1: Summary of observation errors

Type of error	Probability	Description
Error Set 1	p_{e_1}	Full commitment behaviour is observed as half commitment
	p_{e_2}	Full commitment behaviour is observed as zero commitment
	$1 - p_{e_1} - p_{e_2}$	Error free observation of full commitment behaviour
Error Set 2	p_{c_1}	Half commitment behaviour is observed as full commitment
	p_{c_2}	Zero commitment behaviour is observed as full commitment
	$1 - p_{c_1} - p_{c_2}$	Error free observation of uncommitted behaviours

zero information regarding the types of others in the initialization stage. MGs of type T_C import from or export to the coalition the promised energy with probability 1. While MGs of type T_{UC} are assumed to have three levels of commitment according to their situation. Uncommitted MGs may completely commit with the probability η or deliver half of the energy commitment with the probability κ or finally zero commitment with the probability ν where $\eta+\kappa+\nu = 1$. For simplicity, we can assume MGs do not change their type during the decision horizon. The MGs decide their preferences based on their beliefs about the type of other MGs where the beliefs are updated over time with the observations. The concept of Bayes' theorem is used to derive equations for belief update. Considering an MG i, which initially commits to export (or import) a specific amount of energy, the actual commitment can be either of the two observations, which are denoted as O_C and O_{UC}.

In practical systems, some observation errors may happen, for instance, due to imperfect communications. In this chapter, we consider two sets of observation errors. First, the observation errors that happen when a full commitment behavior of an MG is observed as uncommitted behavior known as false-positive error. The second set of observation errors happens when an uncommitted behavior of an MG is observed as full commitment behavior known as false-negative errors. These errors are modeled with categorical distribution [119], and the probabilities of these considered errors are presented in Table 4.1. All the error probabilities are assumed to have small values (less than 0.1).

Considering the prior $p_{ij}(T_C) = b_{ij}^{\tau_{ij}}$ and the fact that likelihood $p_{ij}(O_C|T_C) = 1 - p_{e_1} - p_{e_2}$, we are interested in the calculation of the posterior probabilities $p_{ij}(T_C|O_C)$ or $p_{ij}(T_C|O_{UC})$. $p_{ij}(T_C|O_C)$ denotes the belief probability of i-th MG that MG j is of the type committed when a committed behavior is observed while $p_{ij}(T_C|O_{UC})$ shows the probability that that MG j is of the type committed when an uncommitted behavior is observed by MG i. Depending on the new observation, an MG i can update its belief about the type of MG j using to Bayes' theorem as (similar to A.6):

$$p_{ij}^{\tau_{ij}}\left(T_C\,|O_C\,\right)=\frac{b_{ij}^{\tau_{ij}}\left(1-p_{e_1}-p_{e_2}\right)}{b_{ij}^{\tau_{ij}}\left(1-p_{e_1}-p_{e_2}\right)+\left(1-b_{ij}^{\tau_{ij}}\right)\left[\eta_{ij}^{\tau_{ij}+1}\left(1-p_{e_1}-p_{e_2}\right)+\kappa_{ij}^{\tau_{ij}+1}p_{c_1}+\nu_{ij}^{\tau_{ij}+1}p_{c_2}\right]}, \tag{4.6}$$

or

$$p_{ij}^{\tau_{ij}}\left(T_C\,|O_{UC}\,\right)=\frac{b_{ij}^{\tau_{ij}}\left(p_{e_1}+p_{e_2}\right)}{b_{ij}^{\tau_{ij}}\left(p_{e_1}+p_{e_2}\right)+\left(1-b_{ij}^{\tau_{ij}}\right)\left[\eta_{ij}^{\tau_{ij}+1}\left(p_{e_1}+p_{e_2}\right)+\kappa_{ij}^{\tau_{ij}+1}\left(1-p_{c_1}\right)+\nu_{ij}^{\tau_{ij}+1}\left(1-p_{c_2}\right)\right]}, \tag{4.7}$$

where τ_{ij} denotes the number of iteration that MG i observed MG j. If the MG j is an uncommitted MG, then the expected probabilities by MG i that MG j will either import or export fully, half or none of the promised energy is denoted by $\eta_{ij}^{\tau_{ij}+1}$, $\kappa_{ij}^{\tau_{ij}+1}$ or $\nu_{ij}^{\tau_{ij}+1}$, respectively. Using Bayes' theorem, we can derive:

$$\frac{\left|\chi_{ij}^{\tau_{ij}+1}\left(O_C\right)\right|}{\left|\chi_{ij}^{\tau_{ij}+1}\right|}=b_{ij}^{\tau_{ij}}\left(1-p_{e_1}-p_{e_2}\right)+\left(1-b_{ij}^{\tau_{ij}}\right)\left[\eta_{ij}^{\tau_{ij}+1}\left(1-p_{e_1}-p_{e_2}\right)+\kappa_{ij}^{\tau_{ij}+1}p_{c_1}+\nu_{ij}^{\tau_{ij}+1}p_{c_2}\right], \tag{4.8}$$

and

$$\frac{\left|\chi_{ij}^{\tau_{ij}+1}\left(O_{1/2C}\right)\right|}{\left|\chi_{ij}^{\tau_{ij}+1}\right|}=b_{ij}^{\tau_{ij}}p_{e_1}+\left(1-b_{ij}^{\tau_{ij}}\right)\left[\eta_{ij}^{\tau_{ij}+1}p_{e_1}+\kappa_{ij}^{\tau_{ij}+1}\left(1-p_{c_1}\right)\right]. \tag{4.9}$$

$|\chi^{\tau_{ij}+1}\left(O_C\right)|$, $\left|\chi^{\tau_{ij}+1}\left(O_{1/2C}\right)\right|$ and $|\chi^{\tau_{ij}+1}|$ denote the number of actual full, half and no commitment observations, respectively. Considering Eq. (4.8) and Eq. (4.9) and the fact that:

$$\eta_{ij}^{\tau_{ij}+1}+\kappa_{ij}^{\tau_{ij}+1}+\nu_{ij}^{\tau_{ij}+1}=1, \tag{4.10}$$

we have three equations and three unknowns, and we can use linear solver methods such as Gauss-Seidel [129] to update $\eta_{ij}^{\tau_{ij}+1}$, $\kappa_{ij}^{\tau_{ij}+1}$ and $\nu_{ij}^{\tau_{ij}+1}$. These updated parameters then are used in Eq. (4.6) and Eq. (4.7) to update the posterior beliefs. Each member of a coalition should have its own share of payoff. We employ the proportional fair division algorithm, which fairly divides the total payoff among the MGs in a given coalition relative to the cost each MG experiences in individual energy exchange with the macrogrid and can be computed using the transfer function as follows:

$$u^i=\zeta_i\left(v_{max}(C)-\sum_{j\in C}v(\{j\})\right)+v(\{i\}), \tag{4.11}$$

where ζ_i is equal to $\frac{v(\{i\})}{\sum_{j\in C} v(\{j\})}$ which represents the relative ratio of each MGs contribution and $v(\{j\})$ represent the value of a singleton coalition with single member MG j.

Since the system is dynamic and payoff is changing based on the behaviour of MGs, to find an stable coalition formation in the BCG, we are interested in the expected payoff instead of the immediate payoff. The expected payoff can be found using Eq. (3.5) as:

$$\overline{u}^i\left(C, \mathbb{T}^C\right) = E\left[u^i\left(C, \mathbb{T}^C\right)\right] = \sum_{k=1}^{2^{|C|-1}} B^i\left(\mathbf{T}_k^C\right) u^i\left(C|\mathbf{T}_k^C\right). \tag{4.12}$$

MGs update their belief probabilities based on Algorithm 4.1 in order to build a precise belief system about other MGs. In every epoch, one MG is randomly chosen as a proposer and decides to join a new coalition or stay in the current coalition. The new coalition members will accept the joining proposal if and only if at least one of the existing members gains a higher expected payoff value in the new coalition formation, while the expected payoff values for other MGs are not decreasing. At the end of each iteration, coalition members update their belief probabilities about the type of other MGs according to the observation information using Eq. (4.6)-(4.12). Consecutive merges and splits happen until the system reaches a stable coalition formation, in which no MGs have any incentive to join a new coalition. The concept of WBC is employed to ensure the stability of BCG

Algorithm 4.1 BCG scheme in MGs energy trading.

Initialization: Distribute MGs randomly among coalitions
for Each time slot $t = 1$ to $\mathcal{T}$ **do**
    for MG $i = 1$ to M **do**
        Update current payoff $u^i(t)$
    end for
    Select the proposer MG with probability of $1/M$ (MG_i);
    MG_i sends the joining proposal to m MGs , $m \in C_k/\{i\}$
    if $\overline{u}\{i \in C_k\} \geq \overline{u}\{i \notin C_k\}$ for all $m \in C_k/\{i\}$ then set $i \in C_k$ and update the u for $m \in C_k/\{i\}$.
    All the members of coalition C_k update probabilities using Eq. (4.6)-(4.12)
end for

4.4 Performance Evaluation

4.4.1 Simulation Parameters and Setup

In our simulations, we consider a region of 20km by 20km [7]. We assume that there is not any direct link between two MGs farther than 5 km (D_{tr} = 5 km). Within that region, we assume the MGs are distributed randomly. The number of MGs varies between 4 and 10 [7]. We consider a 24 hours period. Each hour is considered as one iteration in the simulation. In each iteration, one snapshot of Algorithm. 4.1 is executed. We generate load and generation patterns randomly based on a Gaussian random variable and it is periodically repeated after a day with slight variations as in [7]. The mean and variance values of generation and demand are adopted from [125]. Physical parameters ($\mathbb{R}_{ij}$, U_0, U_i, ρ) are selected based on [126]. Observation error values (p_{e_1}, p_{e_2}, p_{c_1} and p_{c_2}) are chosen analogous to [121]. We compare the proposed BCG method with Q-learning and CG-based methods. For the sake of brevity, we refer the readers to Section 3.3 for the CG and Q-learning implementations. The results are obtained over 15 runs with 1500 iterations, and the average results are plotted. 95% confidence interval is considered in the demonstrated results. The considered values of simulation parameters are summarized in Table 4.2.

Table 4.2: Summary of simulation parameters

parameters	**value**
Line resistance (R_{ij})	0.2
Medium voltage (U_0)	50 kv
Low voltage (U_i)	22 kv
Transformer loss fraction (ρ)	0.02
Threshold distance (D_{tr})	5 km
Virtual cost parameter (w_s)	0.02
Virtual cost parameter (w_l)	0.04
Virtual cost parameter (w_0)	0.08
Scaling parameter (δ)	0.95
Observation errors (p_{e_1}, p_{e_2})	0.1
Observation errors (p_{c_1}, p_{c_2})	0.01
Learning rate of Q-learning (α)	0.5
Discount factor of Q-learning (γ)	0.8

4.4.2 Simulation Results

Fig. 4.1 presents the average cost per MG where we have 25% of the MGs are uncommitted. We have compared the proposed BCG algorithm with Q-learning and the CG-based methods. As expected, by increasing the number of MGs, the cost is reduced. Moreover, since BCG is designed to overcome uncertainty, it demonstrates better performance in terms of cost compared to Q-learning and the CG-based methods.

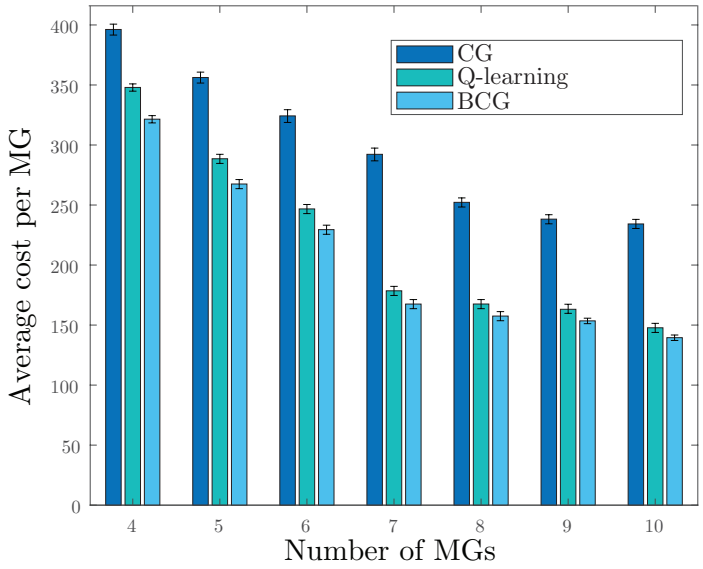

Figure 4.1: Average cost versus number of MGs.

In Fig. 4.2 to evaluate the effect of increasing uncommitted MGs, we show the average cost per MG. In this case, we vary the percentage of uncommitted MGs between 0 to 75 %. As expected, the average cost is also higher for a higher percentage of uncommitted MGs. The reason for this is that as more uncertainty is introduced to the system, coalitions start to export or import more energy from resources outside their coalitions. Besides, it can be observed that BCG outperforms Q-learning and CG-based method by almost 15% and 25% in terms of cost minimization due to uncertainty awareness of the proposed model.

In Fig. 4.3 to evaluate the effect of increasing the cost of transferring energy with the macrogrid, we demonstrate the average energy transfer with macrogrid versus virtual cost weighting parameter ω_0 ranging from 0.06 to 0.2. It is shown that when ω_0 increases, the average energy transfer with the macrogrid decreases, which gives a chance to the coalition of MGs to operate in islanding mode.

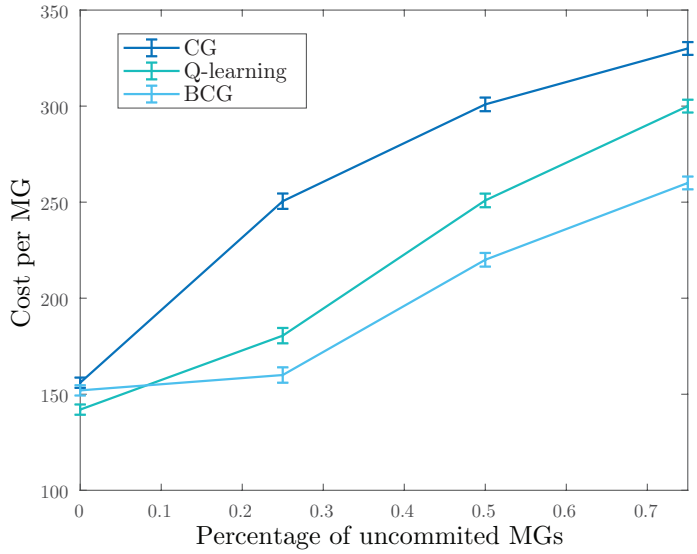

Figure 4.2: Average cost per MG versus the percentage of uncommitted MGs.

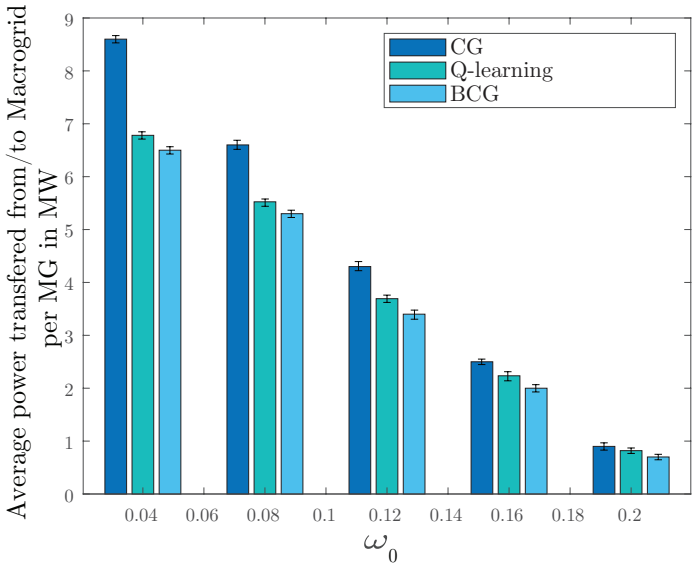

Figure 4.3: Average energy transfer with macrogrid versus virtual cost weighting parameter ω_0.

In Fig. 4.4, we show the belief parameters of a specific MG i, η_{ij}, κ_{ij} and ν_{ij}, about MG j. As seen in the figure, this simulation is evaluated for 1500 iterations, and the parameters converge after approximately 1000 iterations. $\eta_{ij} + \kappa_{ij} + \nu_{ij} = 1$, which guarantees the validity of the algorithm. Besides, the result shows that the algorithm converges in a relatively acceptable number of iterations and has high adaptability, making it a suitable candidate for a dynamic environment.

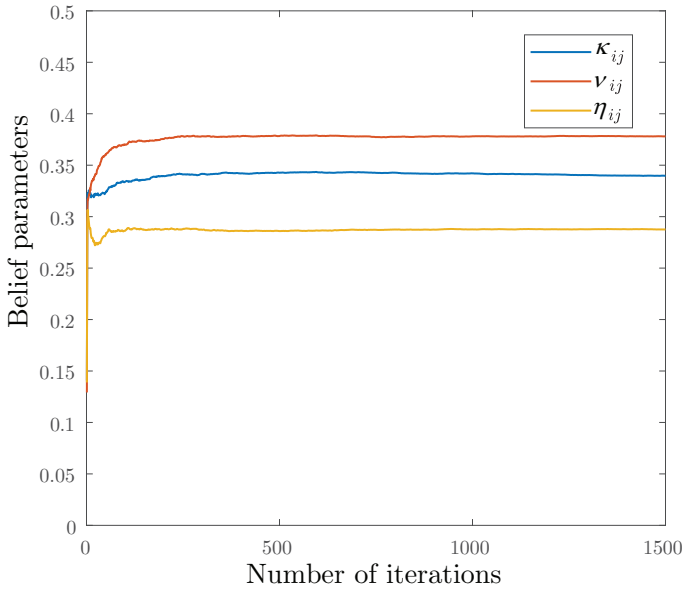

Figure 4.4: Parameters convergence versus number of iterations.

4.5 Conclusion

In this chapter, we investigated the coalitional energy trading problem with the aim of cost minimization in a system with uncertainties and observation errors. We proposed a BCG approach to overcome the uncertainties that arise from the uncommitted members in the coalitions to form optimized coalitions of MGs, resulting in less energy transfer from macrogrid or distant MGs. We compared the proposed approach with Q-learning and CG-based methods, and significant reductions of almost 15% and 25% in cost have been achieved. In Chapter 3 and Chapter 4, BCG helped agents to overcome their uncertainties about the type of other agents. In the next chapter, we propose a BRL-based method

that allows the agents to learn the best coalition formation strategies from the experiences achieved through interactions with the environment.

Chapter 5

Power Loss Minimization in MGs Using Bayesian Reinforcement Learning with Coalition Formation

5.1 Introduction

In this chapter, we focus on energy trading for the purpose of power loss minimization. We assume MGs form coalitions to avoid exporting energy from the utility grid or a distant MGs which might cause higher line losses due to increased distance. We propose a novel BRL-based algorithm, which allows the MGs to reduce the overall power loss. In our approach, each MG agent considers a prior density function over the states of the system. As the agents interact through the iterations of the BRL-based scheme, they update their estimations, finally reaching coalitions that minimize power loss [130]. We compare this scheme with a CG-based approach, Q-learning-based approach, random coalition formation approach, and a case with no coalitions. We assume that MGs have energy storing capability, and we show that power loss can be further reduced by proper sizing of the storage unit (battery).

The rest of this chapter is organized as follows. In Section 5.2, the system model is described. In Section 5.3, the BRL scheme is explained. Numerical results are provided in Section 5.4 and finally, the conclusion is presented in Section 5.5.

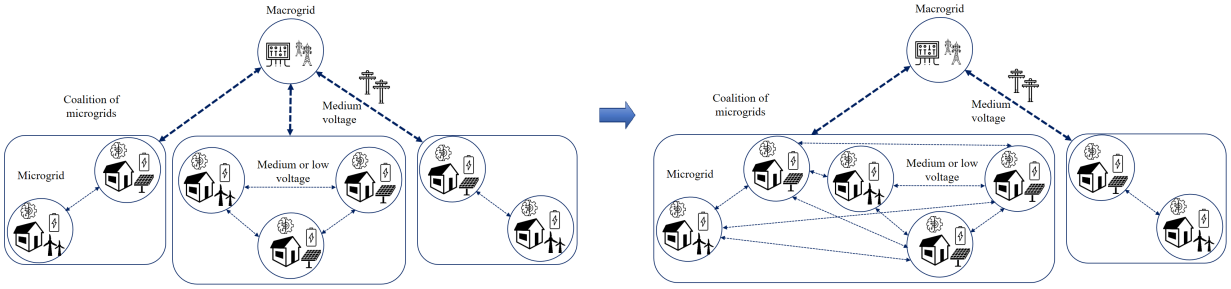

Figure 5.1: Block diagram of a system of MGs.

5.2 System Model

In this chapter, we consider a network of M interconnected MG which are also connected to the utility grid or the macrogrid. The MGs can trade energy with each other or with the utility grid. The simple block diagram of the system is demonstrated in Fig. 5.1. In a specific time slot, each MG $i \in M$, generates power e_i^g and has demand denoted by e_i^d. We also consider that the MGs are equipped with a battery having a capacity of e_i^b, and they can store some of their generated energy in the battery. The surplus energy is defined as $e_i^q = e_i^g - e_i^d - e_i^b$ and it corresponds to the amount of energy that the MG is willing to sell (or buy). We assume that for a given interval, some MGs have surplus energy and can export energy (seller MGs) while others need to import energy (buyer MGs). At each iteration, an MG can move from the sellers group to the buyers group, depending on the interplay between its generation and demand. Power trading among the MGs, as well as selling power to the utility grid, results in power loss over the distribution lines. The amount of this power loss can be expressed as in Eq. (3.1):

$$PL(E_{ij}) = \mathbb{R}_{ij} D_{ij} \frac{{E_{ij}}^2}{U_i^2} + \rho E_{ij}. \tag{5.1}$$

5.3 Power Loss Minimization using BRL

In this chapter, we propose a BRL-based scheme that aims to facilitate energy trading among MGs with minimum power loss. To reduce the power loss due to energy transfer from the distant macrogrid, we consider that MGs can participate in energy trading by joining a coalition. Forming cooperative groups and trading energy among the close by

MGs (coalitions) is a promising approach to reduce the transmission power loss since the line loss depends on the distance and the amount of traded power. A coalition is determined with a pair (C, v) where C denotes a set of players that agree to form a coalition and v is a function that defines the total utility achieved by that coalition. Optimal scheduling of internal energy trading can result in an optimal coalition value. The maximum coalition value can be obtained as follows, which is defined in Eq. (3.3) :

$$v_{max}(C) = max\left[-\left(\sum PL(E_{ij}) + \sum PL(E_{i0}) + \sum PL(E_{0j})\right)\right]. \tag{5.2}$$

We use Algorithm 3.1 to achieve $v_{max}(C)$. We assume each MG is equipped with a battery and stores the surplus generation partially. The amount of energy to be stored is associated with the defined utility function. The proposed BRL approach aims to optimize the decision to join a coalition and the amount of energy to be stored at the battery for each MG in a distributed way. We assume that MGs have variation in the amount of generation and demand, which triggers new coalition formation.

Energy trading among MGs is a dynamic process that can be formulated as an MDP in which each agent decides on the battery level as well as the desirable coalition to join. The MDP is modeled with the following elements. We denote $s = \{\varepsilon, \phi\}$ as the state of the system which is limited to possible states set $\mathbf{S} = \mathbf{\Xi} \times \mathbf{\Phi}$ for the MG i which includes power level $\varepsilon \in \mathbf{\Xi}$ (which corresponds to the quantified state of surplus/demand) and the battery level $\phi \in \mathbf{\Phi}$ (which corresponds to the state of charge). An action $a = \{\phi, C_k\}$ chosen from possible actions set $\mathbf{A} = \mathbf{\Phi} \times \mathbf{C}$ includes the decisions about the coalition to join and the level of energy to be stored in the battery. The state transition function $H(s, a, s')$ is selected from $\mathbf{H} : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \Rightarrow [0, 1]$ and defined as the transition probability from current state s to next state s' taking action a by the MG. The reward function $r(s, a)$ is expressed as the immediate reward r when action a is chosen at the state s and computed in Eq. (5.3) . The goal of each MGs is to obtain the optimal policy $\pi : \mathbf{S} \times \mathbf{H} \Rightarrow \mathbf{A}$ which result in maximized long term expected reward.

Eq. (5.2) formulates the total utility (payoff) of the coalition. Each user in the coalition should have its own share from this utility. We use the proportional fair division algorithm which is given by Eq. (3.4) to define the immediate reward of the i-th MG as:

$$r_i = \zeta_i \left(v(C) - \sum_{j \in C} v(\{j\}) \right) + v(\{i\}), \tag{5.3}$$

where ζ_i is the relative ratio of their contribution equals to $\frac{v(\{i\})}{\sum_{j \in C} v(\{j\})}$ and $v(\{j\})$ represents the value of a coalition with single member MG j.

The agents use BRL to form the coalitions since the lack of information regarding the state of MGs in other coalitions imposes uncertainties on the system and coalition values. Therefore, it is critical to employ a method that allows MGs to learn about each coalition to remove uncertainties and refine the coalition formation process. BRL is a method that aids agents in overcoming uncertainties through repeated interactions with other agents. In the proposed power loss minimizing BRL, each MG considers a prior density function over the state transitions of the system. MGs keep updating their information about the system's state using Bayes' rule. Note that assigning prior density function will not affect the long-term learning process of each MG since, after a large number of iterations, the effect of the assigned prior density function will be insignificant. Therefore, we assume that the state transition function H follows the widely used Dirichlet distribution. Modeling state transition probabilities as a Dirichlet distribution is explained in detail in Appendix B. Considering the method proposed in Appendix B and adopting the Bayesian Exploration Bonus (BEB) method [131] we can derive optimal value function as follows:

$$V_t(s) = \max_a \left\{ r(s,a) + \frac{\beta}{1 + \boldsymbol{\lambda}_t(s,a)} + \sum_{s'} \Pr(s' \,|s,a) V_{t-1}(s') \right\} \tag{5.4}$$

β denotes a fixed value that determines the effect of the additional bonus. $\boldsymbol{\lambda}_t(s,a)$ expresses the number of times that action a has been taken in state s and can be calculated as $\boldsymbol{\lambda}_t(s,a) = \sum_{s'} \boldsymbol{\lambda}_t(s,a,s')$. $\Pr(s' \,|s,a)$ can be computed using Eq. (B.6). BEB algorithm chooses the actions greedily taking into account both state transitions mean estimate and an additional bonus for state and action pairs which have been comparatively less experienced. It means that, at each state, this scheme solves an MDP using the mean of the current belief state for the probability of transition, and an extra exploration bonus equal to $\beta/(1 + \boldsymbol{\lambda}_t(s,a))$. The summary of this approach is provided in Algorithm 5.1. The algorithm is divided into an initialization step and the main loop. In the initialization, each MGs randomly is being assigned to a coalition and the state pair $\{\varepsilon, \phi\}$ is set equal to initial surplus/needed energy and zero, respectively. Afterward, the current coalition formation, C, is broadcasted to all MGs. The main loop includes two parts: learning and forming the coalition. In the learning stage, MGs update current rewards, estimate transition probability and bonus gain, and finally update the value function $V_t(s)$. In the coalition formation step, a proposer MG i is selected from the set of all MGs, M, randomly. Then the chosen proposer takes the optimum action considering Eq. (5.4). The proposer MG which is randomly selected proposes to join a coalition. As MGs decide on the action that results in the maximized sum of current and future expected utility, Eq. (5.4) maximizes long-term expected reward of the considered energy trading system.

Algorithm 5.1 Coalition formation with BRL for distributed energy trading among MGs

Initialization: Initialize a discount rate and β .
At time $t = 0$:
for MG $i = 1$ to M **do**
    Randomly select coalition C_k, set the state $\{\varepsilon_i, \phi_i\} = \{\varepsilon_i, 0\}$.
    Broadcast C to all MGs and set the $V = 0$
end for
Main loop:
for Each time slot $t = 1$ to $\mathcal{T}$ **do**
    for MG $i = 1$ to M **do**
        Update current reward $u_i(t)$
        Estimating the probability of transition and
        Update update value function (Eq. (5.4))
    end for
    BR Coalition formation with the probability of $1/M$ the proposer MG_i is selected from the set M;
    For MG_i:
    Take a action $a = \{C_k, \phi\}$ that maximise $V_t(s)$
    Sends a to all MG_m, $m \in C_k / \{i\}$
    If $u\{i \in C_k\} \geq u\{i \notin C_k\}$ for all $m \in C_k / \{i\}$ then set $i \in C_k$ and update the u for $m \in C_k / \{i\}$.
end for

5.3.1 Q-learning Approach

We implement a Q-learning-based method as it is discussed in Section 2.3.2. Agents, reward function, and states are considered the same as in the previous section. The Q-learning policy is to choose actions maximizing the Q-value. In order to consider action exploration, Q-learning uses the epsilon-greedy method.

5.3.2 Game-Theoretical Approach

In order to compare our proposed method with other techniques, in this section, we implement a CG-based method similar to what is proposed in [7] and implemented in previous chapters. An MGs will be randomly chosen and propose to join a new coalition. The new coalition will accept new members if there is an old member who will be able to achieve a higher payoff within the new formation without hurting any of the other MGs payoff (known as Pareto order [123]). Consecutive merge and split iterations happen until the system of MGs reaches a coalition formation, from whereon no MG has any incentive to further merge to a new coalition. The summary of this scheme is demonstrated in Algorithm 5.2.

Algorithm 5.2 CG for distributed energy trading among MGs

Initialization:
for MG $i = 1$ to M **do** randomly select coalition C_k
end for
Main loop:
for Each time slot $t = 1$ to $\mathcal{T}$ **do**
    for MG $i = 1$ to M **do**
        Update current utility $u_i(t)$
    end for
    Game theoretical coalition formation With the probability of $1/M$ the proposer MG_i is selected from the set M;
    For MG_i**:**
    Sends a to all MG_m, $m \in C_k/\{i\}$
    If $u\{i \in C_k\} \geq u\{i \notin C_k\}$ for all $m \in C_k/\{i\}$ then set $i \in C_k$ and update the u for $m \in C_k/\{i\}$.
end for

5.4 Performance Evaluation

5.4.1 Simulation Parameters and Setup

For the numerical evaluation, we set up a network of MGs where M is between 4 and 10 within an area of 10 km by 10 km. Macrogrid is located in the middle of the considered area, and MGs are located at random [7]. A full day is divided into 24-time slots, where each time slot is considered as one iteration in the simulation. In each iteration, one snapshot of Algorithm. 5.2 is executed. Load and generation patterns are generated randomly based on a Gaussian random variable and periodically repeated after a day with slight variations as in [7]. The mean and variance values of generation and demand are adopted from [125]. Initial values of BRL parameters are selected based on [131] and then tuned for the considered scenario. The simulation parameters are summarized in Table 5.1. We compare the proposed BRL method with the non-cooperative method (no coalition formation), random coalition formation scheme, CG-based method and Q-learning-based algorithm. The results are obtained over 15 runs, and the average results are plotted.

Table 5.1: Summary of parameters.

parameter	**value**
Number of MGs	4 to 10
Number of coalitions	4
Number of battery levels	4
Considered area	10km*10km
$\mathbb{R}_{ij}$	0.2
U_0	50 kv
ρ	0.02

5.4.2 Simulation Results

In Fig. 5.2, we present the average loss per user versus the number of MGs ranging from 4 to 10. We have compared five different schemes in this evaluation. In the no coalition scheme, we considered that each MG just transfers energy with a macrogrid, and there is no coalition formation included. In the random scheme, we considered a random scenario for coalition formation, in which MGs are assigned to coalitions randomly. The following three schemes, CG-based, Q-learning based, and proposed BRL schemes, were introduced

in Section 5.3. As expected, when the number of MGs increases, it results in less power loss. As the system has more adaptation capability with the variations in the environment, there is less power loss. BRL scheme has less power loss compared to Q-learning-based and CG-based schemes.

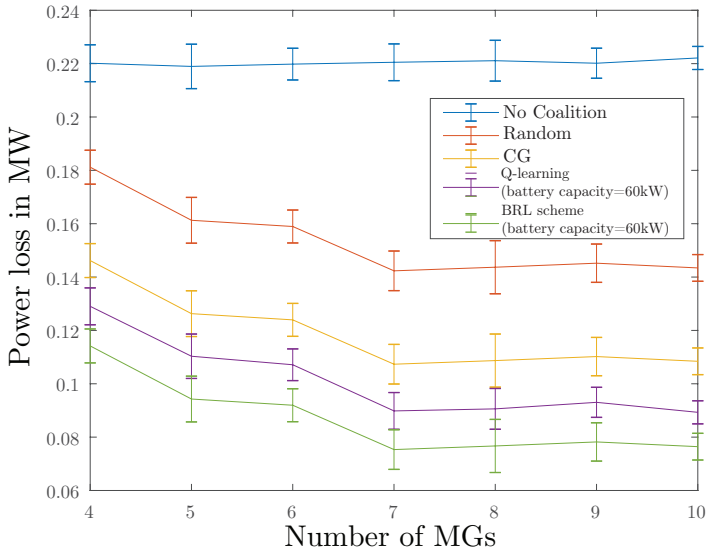

Figure 5.2: Average power loss per user versus the number of MGs.

In Fig. 5.3 to evaluate the effect of battery capacity, we demonstrate the average loss per user versus the number of MGs ranging from 4 to 10 for three different scenarios with different battery capacities. It is shown that when the battery capacity of MGs increases, the average power loss decreases as expected. Also, when we have more MGs, we will have less power loss in the system. This is expected since more energy can be stored for future use, and power is not lost during transmission.

In Fig 5.4, we show the convergence of the average power loss per user versus the number of iterations. This figure demonstrates the accumulative average power loss in time for the BRL method with different battery capacities. As seen in the figure, power loss (which also indicates the negative of the average utility of the MG) converges over time.

To demonstrate the complexity of our method in comparison with the CG-based method, in Fig. 5.5 we show the simulation run time for both schemes. The results demonstrate that the proposed scheme completes in less time than the CG-based approach. The results are obtained with a PC with Intel(R) Core(TM)15-6500 CPU and 8192MB RAM.

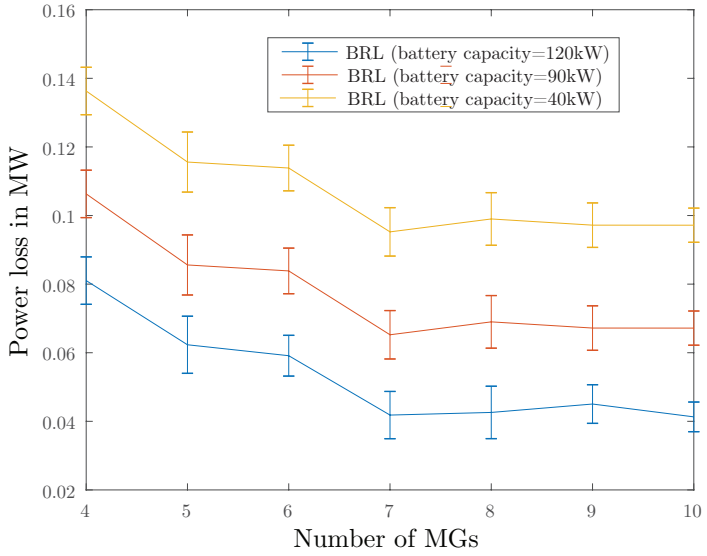

Figure 5.3: Average power loss per user versus the number of MGs.

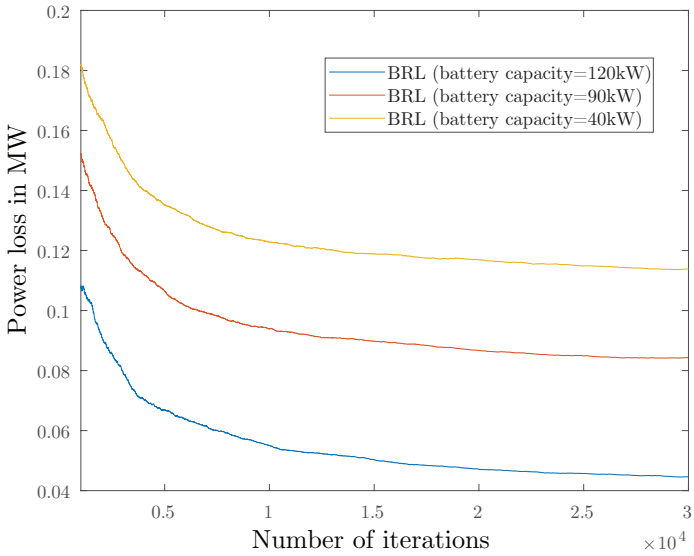

Figure 5.4: Average power loss per user versus number of iteration.

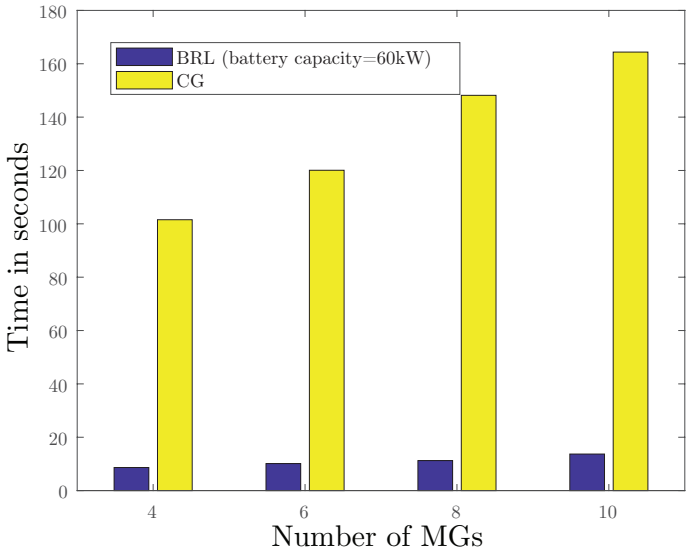

Figure 5.5: Run time of proposed learning method and CG based scheme.

5.5 Conclusion

In this chapter, we visited the problem of energy trading with the aim of power loss minimization. BRL has been used to form coalitions of MGs to avoid exporting energy from the utility grid or distant MGs which might cause higher line losses. Moreover, BRL allowed MGs to learn the energy generation pattern and take optimal coalition decisions. Comparing our proposed scheme with the Q-learning and CG-based methods, a significant reduction in power loss has been achieved. In the next chapter, we combine BCG and BRL to develop a method that helps agents update their beliefs about the type of other agents while enabling them to learn from past experiences.

Chapter 6

Cost-Optimized MG Coalitions Using Bayesian Coalitional Reinforcement Learning

6.1 Introduction

MGs need to optimize the scheduling of their demands and energy levels while trading their surplus with others to minimize the overall cost. However, various factors can affect this, such as uncertainty caused by the generation/consumption of renewable energy and the complexity of interconnected MGs and their interplay. Thus, reaching optimal scheduling is challenging. This chapter addresses the energy trading problem among MGs by minimizing the cost while uncertainty exists in MG generation and demand.

In Chapter 5, we aimed to minimize power loss while addressing the uncertainties from the energy level of agents using BRL. Although we use BRL to track the transition probabilities, the provided learning algorithm suffers from a lack of a system of beliefs about other involved agents in the system. Different than Chapter 5, in Chapter 3 and Chapter 4, we address the uncertainty resulting from the type of agents (moving/fixed and committed/uncommitted, respectively) and make a system of belief using the BCG method in the presence of observation error. However, these BCG schemes cannot learn from past experiences the same as the method presented in Chapter 5. To this end, we propose a comprehensive Bayesian Coalitional Reinforcement Learning (BCRL) method for energy trading problems, which includes belief systems and gradually learns the best action through interaction with the environment to minimize the energy trading cost among

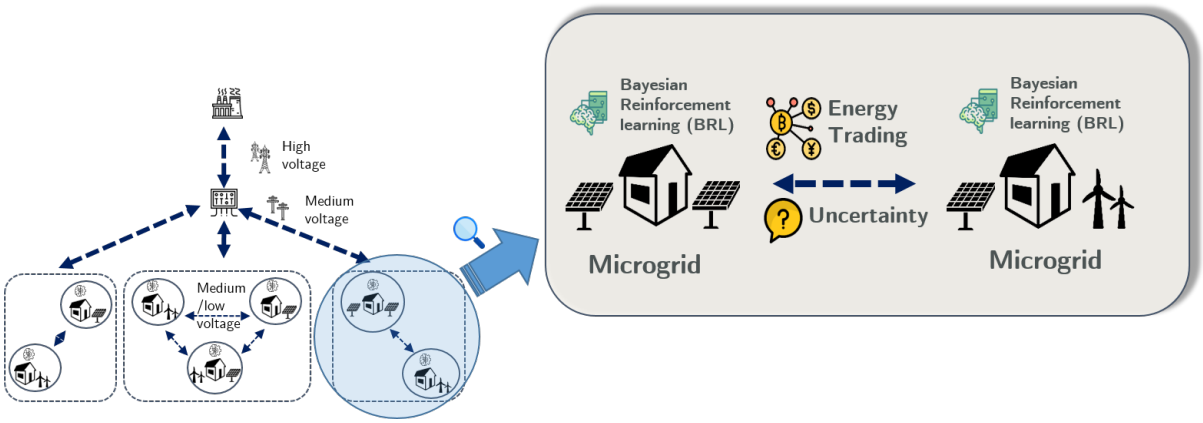

Figure 6.1: Block diagram of a system of MGs. The figure illustrates that due to the dynamicity of the system, different coalitions can be formed during each epoch.

MGs by forming stable coalitions [132]. This algorithm was first introduced in [51, 52] and an application of this model was also developed for device-to-device communications in wireless networks [133].

The rest of this chapter is organized as follows. In Section 6.2, the system model is demonstrated. In Section 6.3, the Bayesian coalition formation scheme is illustrated. In Section 6.4, BCRL based scheme is proposed. In Section 6.5, the numerical results are evaluated, and finally, the conclusions are presented in Section 6.6.

6.2 System Model

In this chapter, we consider a network of M interconnected MGs while each MG is also connected to the main utility grid known as the macrogrid as shown in Fig. 6.1. The amount of generated energy by MG $m \in M$ and its demand are presented by e_m^g and e_m^d, respectively. Therefore, we can find the total surplus or shortage energy of each MG as $e_m^q = e_m^g - e_m^d$, which represents the energy that each MG is required to export to or import from the network. As a result, MGs initiate an energy trading process among each other and with the macrogrid to satisfy their export/import requirements. Due to the dynamic nature of the system, each MG can be either a seller or buyer of energy during each epoch.

The process of energy trading, either among MGs or with the macrogrid, imposes a variety of costs. We define the total cost of power transaction E_{mn} from m-th MG to n-th MG as following which is explained in Section 4.2:

$$S_{mn} = \omega D_{mn} E_{mn} + \delta PL(E_{mn}). \tag{6.1}$$

Energy trading among nearby MGs is less costly than trade energy between distant MGs and macrogrid. Therefore, forming coalitions of close-by MGs to trade energy in their groups is a promising approach that reduces the overall cost. In addition, there is no transformer loss in the operating range of energy trading among MGs (low or medium voltage).

We can formulate a coalition with a pair (C, v_C). C expresses the coalition in which coalition members cooperate to gain a higher coalitional value $v(C)$. In this chapter, we consider the total cost of energy trading among coalition members plus the cost of trading extra energy with the macrogrid as the coalition value. The objective is to minimize the cost. Therefore, the coalition value as the negative form of cost can be formulated as follows:

$$v(C) = -\sum_{m=0}^{|C|}\sum_{n=0}^{|C|} S_{mn}, \tag{6.2}$$

where $|C|$ shows the number of members of coalition C. Index 0 expresses any transactions with the main grid. When the coalition is formed, energy transfer among the coalition members needs to be scheduled in a way that minimize the total cost in the coalition. Therefore, the coalition payoff in our considered system is defined as the maximum achievable coalition value $v_{max}(C)$ which is given by:

$$v_{max}(C) = \max\{-\sum_{m=0}^{|C|}\sum_{n=0}^{|C|} S_{mn}\}. \tag{6.3}$$

6.3 Bayesian Coalition Formation Game

In this section, we propose a Bayesian coalition formation scheme that tackles the uncertainty in the power level of the MGs. The proposed BCG in this section is adopted from [51] and differs slightly from the proposed BCG in Chapter 3 an Chapter 4.

6.3.1 Game Formulation

The Bayesian coalition formation game can be characterized by a set of agents ($\mathbf{M}$), a set of agent types ($T^m \in \mathbb{T}$), agent belief function (B^m), a set of coalition actions ($\mathbf{A}^{C_k}$), a set of outcomes known as states ($s^m \in \mathbf{S}$), and the reward functions (r^m).

To employ the Bayesian coalition formation game in our problem, we can describe the Bayesian coalition formation game as a cost minimization model in an MG community where a set of M rational MG agents is involved in the coalition formation game. A coalition C_k represents a set of MGs that allows them to trade energy among themselves. The m-th MG's type T^m stands for the MG's power level. Each MG is only aware of its type (T^m) but not the types of other MGs. The m-th MG's beliefs about the types of other players is denoted by $B^m(\mathbf{T}^{-m})$ that consists of a joint distribution over $\mathbf{T}^{-m}$, which is the probability assigned to other agents about their types. We assume that any coalition of MGs has a restricted set of coalition actions $\mathbf{A}^{C_k}$. A collective coalition action a^{C_k} is an action that is approved by all coalition members of C_k about a new member to join their coalition. The coalitional action is only observable for the coalition members and hidden from agents in other coalitions.

We consider the coalition tag of each MG as its state in each iteration of the game. Therefore, agents' state profile vector can be defined as $\mathbf{s} = (s^1, ..., s^M)$. Any state profile vector $\mathbf{s}$ results in a joint reward $R^{C_k}(\mathbf{s}, \mathbf{T}^{C_k})$ for the members of types $\mathbf{T}^{C_k}$ in coalition C_k which is calculated as:

$$R^{C_k}(\mathbf{s}, \mathbf{T}^{C_k}) = v_{max}(C_k) = \sum_{m \in C_k} r^m(\mathbf{s}, \mathbf{T}^{C_k}). \tag{6.4}$$

We use the proportional fair division method to distribute the coalition reward among coalition members, allocating each member a share of the coalition reward proportionate to their non-cooperative cost. Therefore, $r^m(\mathbf{s}, \mathbf{T}^{C_k})$ is defined as:

$$r^m(\mathbf{s}, \mathbf{T}^{C_k}) = \zeta_m \left(v_{max}(C_k) - \sum_{n \in C_k} v(\{n\}) \right) + v(\{m\}). \tag{6.5}$$

where ζ_m is equal to $\frac{v(\{i\})}{\sum_{j \in C} v(\{j\})}$ and demonstrates the relative contribution of each MG. $v(\{j\})$ show a single member coalition.

6.3.2 Stability Notation

Like all cooperative games, in the BCG, players with a common interest or members of a specific coalition maximize their joint reward, known as the Bayesian coalitional value function. We compute the Bayesian value of coalition C_k with the members of type $\mathbf{T}^{C_k}$ as follows:

$$V(C_k|\mathbf{T}^{C_k}) = \max_{a^{C_k} \in \mathbf{A}^{C_k}} \sum_{\mathbf{s} \in \mathbf{S}} \Pr\{\mathbf{s}|C_k, a^{C_k}, \mathbf{T}^{C_k}\} r^{C_k}(\mathbf{s}, \mathbf{T}^{C_k}) = \max_{a^{C_k} \in \mathbf{A}^{C_k}} Q(C_k, a^{C_k}|\mathbf{T}^{C_k}) \tag{6.6}$$

where $\Pr\{\mathbf{s}|C_k, a^{C_k}, \mathbf{T}^{C_k}\}$ represents the probability of transitioning to state $\mathbf{s}$ in coalition C_k with members of types $\mathbf{T}^{C_k}$ when taking coalitional action a^{C_k}. $Q(C_k, a^{C_k}|\mathbf{T}^{C_k})$ shows the action value function given the coalitional type $\mathbf{T}^{C_k}$. It can be seen that $V(C_k|\mathbf{T}^{C_k})$ is a function of the actual "type" of coalition members while the "type" of a MG is not known by the other MGs inside the coalition. It is worth clarifying here that in Eq. (6.3), for a given coalition form, energy trading among coalition members should be planned in way that results in the minimum possible power loss, which is defined as maximum coalition value $v_{max}(c)$. This should be distinguished from the Bayesian coalitional value function, where coalition members should take the coalitional action resulting in the optimum coalition formation. Therefore, to estimate the Bayesian coalition value, coalition members need to rely on their beliefs about the "type" of other players. We call this estimation the expected value of coalition C_k and coalition member m can compute its expected coalition value according to its beliefs B^m as follows:

$$V(C_k, B^m) = \max_{a^{C_k} \in \mathbf{A}^{C_k}} \sum_{\mathbf{T}^{C_k} \in \mathbb{T}^{C_k}} B^m(\mathbf{T}^{C_k}) Q(C_k, a^{C_k}|\mathbf{T}^{C_k}) = \max_{a^{C_k} \in \mathbf{A}^{C_k}} Q(C_k, a^{C_k}, B^m) \tag{6.7}$$

where $Q(C_k, a^{C_k}, B^m)$ demonstrates the expected value of coalition C_k when action a^{C_k} is taken while the system's belief is equal to B^m. $\mathbb{T}^{C_k}$ denotes set of all possible type vectors for the members of coalition C_k. Since all the MGs have their specific systems of beliefs, it is common for MGs to end up with different estimations of $Q(C_k, a^{C_k}, B^m)$ and consequently different estimations of $V(C_k, B^m)$. Therefore, none of the MGs can reach the accurate estimations about the coalitional reward R^{C_k} and their share of reward r^m. To this end, players need a system to estimate their achievable rewards for cooperating in coalitional activity. We define value-demand (different from energy demand) θ^m as the share of the coalitional value that MG m believes in receiving in the coalition. Having the coalition structure C_k with the demand vector $\boldsymbol{\theta}^{C_k} = (\theta^1, \theta^2, ..., \theta^{|C_k|})$, MG m's belief about the expected reward of MG $n \in C_k$ by taking coalitional action a^{C_k} can be estimated by:

$$Q_n^m(C_k, a^{C_k}, \boldsymbol{\theta}^{C_k}) = \frac{\theta^n Q(C_k, a^{C_k}, B^m)}{\sum\limits_{i \in C_k} \theta^i}. \tag{6.8}$$

Belief of MG m about its personal expected reward is defined by $Q^m(C_k, a^{C_k}, \boldsymbol{\theta}^{C_k})$.

Considering the above-mentioned definitions, the concept of a SBC can be defined as follows [52].

Definition 1: *We assume that a tuple of a specific coalitional structure and a specific demand vector* $(C_{k'}, \boldsymbol{\theta}^{C_{k'}})$ *are in the SBC of a BCG if*:

- No player believes there exists a better tuple than $(C_{k'}, \boldsymbol{\theta}^{C'_k})$.
- All of the coalition members accept it based on their beliefs about the expected rewards of other players.

This definition can be formulated as follows:

$$Q^m(C_{k'}, a^{C_{k'}}, \boldsymbol{\theta}^{C_{k'}}) > Q^m(C_k, a^{C_k}, \boldsymbol{\theta}^{C_k}) \tag{6.9}$$

and

$$Q_n^m(C_{k'}, a^{C_{k'}}, \boldsymbol{\theta}^{C_{k'}}) > Q_n^m(C_k, a^{C_k}, \boldsymbol{\theta}^{C_k}). \tag{6.10}$$

Eq. (6.9) demonstrates the preference of MG m for itself and Eq. (6.10) shows the preference of MG n believed by MG m.

6.3.3 Coalition Formation

In this section, we define the Bayesian coalition formation process that we present in this chapter. We assume that negotiations among the MGs to merge and split from coalitions happen over an infinite number of iterations. At every iteration, there is a pair of coalitional structure and demand vector, which is named the coalitional agreement $(C, \boldsymbol{\theta})$, on which all players agree. All the MGs have the chance to modify this coalition agreement (a rational player changes the agreement to improve its utility). We call the MG m who attempt to change the coalitional agreement a proposer since it proposes to change the agreement in either one of the following ways:

- A proposer can stay in its current coalition C_k and propose a new demand θ^m.
- A proposer can decide to split from its current coalition and propose merging to other coalition $C_{k'}$ with new demand θ^m.

The MGs have the following finite set of actions (or negotiations options): (1) if an MG is a proposer the action is to make proposal $\pi_m^k = (C_k, \{\theta^i\}_{i \in C_k}{}^{\circ}\theta^m)$ which means joining (or staying in) coalition C_k with the new demand θ^m. (2) If an MG is a responder to a proposal, it has the following action options: (i) either accept ($\Upsilon_k^m = 1$) or (ii) reject ($\Upsilon_k^m = 0$), in response to the presented proposal. We can summarize the proposition procedure as follows. At the beginning of every iteration, a proposer m is chosen randomly from all the MGs with an equal probability of $1/M$. Then, the proposer presents the proposal π_m^k to join or stay in the coalition C_k with demand θ^m. After that, members of

the coalition C_k independently accept or reject the offered proposal without having any information regarding the action of other members. To respond to the proposal, all the individual responder actions need to be unified in a single coalitional action. To this end, we introduce function f, which maps all the responding actions into the coalitional action a^{C_k}. We define this coalitional action as follows:

$$a^{C_k} = \begin{cases} f(\pi_k^m), & if \prod_{i \in C_k \setminus \{m\}} \Upsilon_k^m = 1 \\ f(C_k, \boldsymbol{\theta}^{C_k}), & otherwise \end{cases}. \tag{6.11}$$

This means that coalition members accept a proposal if all coalition members approve it; otherwise, the proposal will be rejected, and the existing coalitional agreement will be in effect. We assume that all the players are rational, which means that the proposer submits a proposal that maximizes its expected reward. Meanwhile, because the other players are also rational, they only accept a proposal that does not degrade their expected reward. Therefore, a rational proposal is to offer the maximum possible demand θ_{max}^m that does not degrade the expected reward of other players according to the beliefs of the proposer about other players. This particular proposal is achievable for the proposer MG if:

$$\frac{\theta^n Q(C_k \cup \{m\}, a^{C_k}, B^m)}{\sum_{i \in C_k \cup \{m\}} \theta^i} \geq Q_n^m, \ \forall n \in C_k, \tag{6.12}$$

where $a^{C_k} = f(\pi_k^m)$ and Q_n^m is the expected reward of MG n believed by MG m. If proposer MG m finds π_k^m to be feasible, it expects all the responders to accept the proposal according to its system of beliefs about others. It should be noted that this feasibility is just an expectation, and the proposer is not sure that the proposal will be accepted or rejected since it does not know what is best for the responder. Considering Eq. (6.12), the proposer can estimate θ_{max}^m as follows:

$$\theta_{\max}^m(C_k) = \min_{n \in C_k} \frac{\theta^n Q(C_k \cup \{m\}, a^{C_k}, B^m) - Q_n^m \sum_{i \in C_k \cup \{m\}} \theta^i}{Q_n^m}. \tag{6.13}$$

The requested demand by the proposer is restricted to the interval $[0, \theta_{max}^m(C_k)]$. To simplify the search for a proper demand, we define a unit Δ, making the proposer to propose a demand as integral multiplies of Δ. Therefore we can define the possible demand vector as $[0, \Delta, 2\Delta, ..., \lfloor \theta_{\max}^m(C_k)/\Delta \rfloor \Delta, \theta_{\max}^m(C_k)]$.

6.4 BCRL method for energy trading among MGs

Types of players in a coalition dynamically change since MGs' generation and demand vary in time. As a consequence, the coalition values change, which imposes uncertainty on the system. Combining BRL with BCG gives the players the chance to learn from previous experiences and overcome uncertainties about the type of other agents through interactions in the coalition formation process. In this section, we present the cooperative multi-user BRL suitable for the coalition formation process in MGs, which is called BCRL [51].

6.4.1 BCRL

In the following, we extend the previously discussed conventional BRL in Section 2.3.3 to the case of multiple agents in a coalition formation game. Our goal is to find the optimal coalition formation in the Bayesian coalition formation game.

Let us assume that the initial belief of the MG m is denoted by $B^m = B^m(\mathbf{T}^{C_k})$ where $\mathbf{T}^{C_k}$ shows the types of players in the coalition C_k, similar to the unknown in the conventional BRL. Each MG m in the coalition C_k with the coalition action a^{C_k} can compute a long-term expected action value based on its belief B^m at each time slot t as follows [51]:

$$\begin{aligned} &Q_t^m(C_k, a^{C_k}, B^m) = \\ &\sum_{s^m \in \mathbf{S}} \Pr\{s^m | C_k, a^{C_k}, B^m\}(r^m(t) + \gamma V^m(C_k, B^m(\mathbf{T}^{C_k}))) \\ &\qquad = \sum_{\mathbf{T}^{C_k} \in \mathbb{T}^{C_k}} B^m(\mathbf{T}^{C_k}) Q_t^m(C_k, a^{C_k}, B^m | \mathbf{T}^{C_k}) \end{aligned} \tag{6.14}$$

and

$$\begin{aligned} &Q_t^m(C_k, a^{C_k}, B^m | \mathbf{T}^{C_k}) = \\ &\sum_{s^m \in \mathbf{S}} \Pr\{s^m | C_k, a^{C_k}, \mathbf{T}^{C_k}\}(r^m(t) + \gamma V^m(C_k, B^m_{s^m}(\mathbf{T}^{C_k}))), \end{aligned} \tag{6.15}$$

where $r^m(t) = r^m(\mathbf{s}', \mathbf{T}^{C_k})$ expresses the reward that MG m receives at time t in the coalition C_k with the members of type $\mathbf{T}^{C_k}$ in the current state vector $\mathbf{s}'$. $\mathbb{T}^{C_k}$ denotes set of all possible type vectors for the members of coalition C_k. The probability of transition from the current state $\mathbf{s}'$ to next state s^m by taking coalitional action a^{C_k} by members of type $\mathbf{T}^{C_k}$ is denoted by $\Pr\{s^m | C_k, a^{C_k}, \mathbf{T}^{C_k}\} = \Pr\{s^m | \mathbf{s}', C_k, a^{C_k}, \mathbf{T}^{C_k}\}$. $B^m_{s^m}(\mathbf{T}^{C_k})$ expresses the updated belief after transition to the next state s^m about the types of other coalition members, $\mathbf{T}^{C_k}$, which can be estimated using the Bayesian theorem as follows:

$$B^m_{s^m}(\mathbf{T}^{C_k}) = \psi \Pr\{s^m | C_k, a^{C_k}, \mathbf{T}^{C_k}\} B(\mathbf{T}^{C_k}), \tag{6.16}$$

where ψ expresses a normalizing constant. Consequently, we can find the optimal value-function V^m with a modified Bellman's equation as follows:

$$V^m(C_k, \mathbf{T}^{C_k}) = \sum_{C_k|m\in C_k, \boldsymbol{\theta}^{C_k}} \Pr\{C_k, a^{C_k}, \boldsymbol{\theta}^{C_k}|B^m\} \times Q^m_{t-1}(C_k, a^{C_k}, B^m). \tag{6.17}$$

Unlike the original form of the Bellman's equation, in our problem, MG m cannot find the optimal V^m by maximizing Q^m_t as the coalitional process does not have full control of the coalition formation process. Therefore, MG m should estimate the probability $\Pr\{C_k, a^{C_k}, \boldsymbol{\theta}^{C_k}|B^m\}$ instead to find a specific coalition agreement $(C_k, \boldsymbol{\theta}^{C_k})$ that all coalition members will accept. Therefore, by considering Eq. (6.14) , Eq. (6.15) and the belief update Eq. (6.16), each MG can learn the long-term expected value of any agreement $(C, \boldsymbol{\theta})$ to find the optimal decision based on its belief about the types of other MGs.

6.4.2 Computational Approximations

As it has been mentioned in the previous part, it is not straightforward to estimate Eq. (6.17), since, on the one hand, we need to approximate the transition probability $\Pr\{s^m|C_k, a^{C_k}, \mathbf{T}^{C_k}\}$ and the acceptance probability $\Pr\{C_k, a^{C_k}, \boldsymbol{\theta}^{C_k}|B^m\}$. On the other hand, by considering the size of types and states spaces, it is not possible to directly compute Eq. (6.14), Eq. (6.15), and Eq. (6.17). Therefore, a realistic simplification is needed to approximate Eq. (6.17). To this end, we employ the BEB method to estimate the transition probability $\Pr\{s^m|C_k, a^{C_k}, \mathbf{T}^{C_k}\}$ [131]. In this method, we deploy counter parameters to determine how many times each transition occurs at each iteration t. The exploration bonus is used in order to put more weight on the paths that are not visited enough. Let us define the total number of transitions as:

$$\mu^m_0(C_k, a^{C_k}, \mathbf{T}^{C_k}) = \sum_{s^m\in\mathbf{S}} \mu^m(s^m, C_k, a^{C_k}, \mathbf{T}^{C_k}), \tag{6.18}$$

here, $\mu^m(s^m, C_k, a^{C_k}, \mathbf{T}^{C_k})$ is a counter that shows how many time transition to state s^m is accrued. Then, we can estimate $\Pr\{s^m|C_k, a^{C_k}, \mathbf{T}^{C_k}\}$ as:

$$\tilde{\Pr}\{s^m|C_k, a^{C_k}, \mathbf{T}^{C_k}\} = \frac{\mu^m(s^m, C_k, a^{C_k}, \mathbf{T}^{C_k})}{\mu^m_0(C_k, a^{C_k}, \mathbf{T}^{C_k})}. \tag{6.19}$$

Therefore, we can estimate the action value function in Eq. (6.14) as follows:

$$Q^m_t(C_k, a^{C_k}, B^m) = \sum_{\mathbf{T}^{C_k}\in\mathbb{T}^{C_k}} B^m(\mathbf{T}^{C_k}) \sum_{s^m\in\mathbf{S}} \tilde{\Pr}\{s^m|C_k, a^{C_k}, \mathbf{T}^{C_k}\} \times (r^m(t) + BEB + \gamma V^m(C_k, \tilde{B}^m_{s^m}(\mathbf{T}^{C_k}))) \ , \tag{6.20}$$

where BEB is given by:

$$BEB = \frac{\beta}{1 + \mu_0^m(C_k, a^{C_k}, \mathbf{T}^{C_k})} \tag{6.21}$$

β is a tuning parameter to adjust the chance of exploring less-visited transitions in transition probability. $\beta = 0$ means we skip the effect of BEB in our calculations. To estimate the acceptance probability $\Pr\{C_k, a^{C_k}, \boldsymbol{\theta}^{C_k} | B^m\}$, we need $\lambda_0^m(C_k, \boldsymbol{\theta}^{C_k})$, which defines the times that agreement $(C_k, \boldsymbol{\theta}^{C_k})$ has been proposed, and $\lambda^m(C_k, \boldsymbol{\theta}^{C_k})$ shows how many times this agreement has been accepted. Therefore, we can estimate the $\Pr\{C_k, a^{C_k}, \boldsymbol{\theta}^{C_k} | B^m\}$ as follows:

$$\tilde{\Pr}\{C_k, a^{C_k}, \boldsymbol{\theta}^{C_k} | B^m\} = 0.5 + \xi \frac{\lambda^m(C_k, \boldsymbol{\theta}^{C_k})}{\lambda_0^m(C_k, \boldsymbol{\theta}^{C_k})} \tag{6.22}$$

The initial value that is set for the acceptance probability is 0.5. Additionally, we assume that $0 < \xi < 0.5$.

The BRLC algorithm as applied to our MG coalition formation problem is given in Algorithm 6.1. The algorithm is divided into an initialization step and the main loop. Each MG's initial power level, location, and coalition are assigned randomly in the initialization step. Then the initial demand of each MG is set proportional to their non-cooperative operation where they only trade energy with the macrogrid. After all initial stages, the (C_k, θ_m, T^m) tuple will be transferred to all MGs.

The main loop consists of two phases: the learning phase (lines 11–14) and the coalition formation phase. In the learning phase, the values of all coalitions, the current reward of each MG, transition probabilities, and the value function of each MG will be updated. Then, in the coalition formation phase (lines 16–32), we assume that each time the power level of one random MG changes and that specific MG is given a chance to propose. The proposer MG makes a proposal $\pi_k^m = (C_k, \boldsymbol{\theta}^{C_k})$ in which it decides about the coalition to join (or stay in the same coalition) and proposes a new demand in a way that maximize its own belief about Q_t^m. The proposal will then be transmitted to the member of the target coalition. Suppose all the members in the coalition find that their value function will be higher considering the new proposal. In this case, the proposal will be accepted, and the proposer MG will join/stay in the targeted coalition with the new demand. Otherwise, the proposal will be rejected, and the proposer MG will stay in its previous coalition with the previous demand. After forming the new coalition structure, each coalition uses a greedy algorithm, introduced in Section 3.3, to exchange energy among the members of the coalition.

6.5 Performance Evaluation

In this section, at first, we briefly introduce our benchmark models and then examine the performance of the proposed model compared to the benchmarks.

6.5.1 Benchmarks

Maximum a Posterior Estimation (MAPE)

In this model, the estimation of the action-value function is simplified to the most probable belief type, believed by agent m based on its current belief vector B^m as follows:

$$\tilde{\mathbf{T}}_m^{C_k} = \arg\max_{T^n \in \mathbf{T}^n} \{B^m(T^n)\} \ \ , \forall n \in C_k. \tag{6.23}$$

The main advantage of MAPE with respect to BCRL is its lower complexity due to ignoring the expected coalition value MGs. To this end, in MAPE, MGs reduce their action-value function as follows:

$$\tilde{Q}_t^m(C_k, a^{C_k} | \tilde{\mathbf{T}}_m^{C_k}) = \sum_{s^m \in \mathbf{S}} \tilde{\Pr}\{s^m | C_k, a^{C_k}, \tilde{\mathbf{T}}_m^{C_k}\}(r^m(t) + BEB). \tag{6.24}$$

Since this method is a relaxed estimation of the proposed BCRL, we call it BCRLMAPE in the rest of the chapter. We adopted BCRLMAPE from [51, 133].

Fully Myopic Estimation (FME)

Similar to BCRLMAPE, the FME model has a lower complexity since in this model, only the instantaneous action-value function is considered, and the experience history is discarded. The action-value function is given by:

$$\tilde{Q}_t^m(C_k, a^{C_k}, B^m) = \sum_{\mathbf{T}^{C_k} \in \mathbb{T}^{C_k}} B^m(\mathbf{T}^{C_k}) \sum_{s^m \in \mathbf{S}} \tilde{\Pr}\{s^m | C_k, a^{C_k}, \mathbf{T}^{C_k}\}(r^m(t) + BEB). \tag{6.25}$$

The FME model, same as MAPE, is a reduced version of the BCRL; therefore, in the rest of this chapter, this model is called BCRLFME. We adopted BCRLFME from [51, 133]. BCRLMAPE and BCRLMAPE have less computational complexity comparing to the proposed BCRL. The computational complexity of BCRL method is $O\left(|\mathbb{T}|^{|C_k|}\left(n^2m^2 + n^3\right)\right)$, where n is equal to the number of digits in $\mu_0^m(C_k, a^{C_k}, \mathbf{T}^{C_k})$ and m is equal to the number of digits in $\lambda_0^m(C_k, \boldsymbol{\theta}^{C_k})$ [133]. In BCRLMAPE and BCRLFME, the computational complexities are reduced to $O\left(|\mathbb{T}|\,|C_k| + n^3\right)$ and $O\left(|\mathbb{T}|^{|C_k|}n^3\right)$, respectively [133].

Algorithm 6.1 Coalition formation with BCRL for distributed energy trading among MGs

Initialization:
for all MG m, $m \in M$ **do**
    Randomly assigns the power level.
    Randomly assign the location.
    Randomly assign to the coalition C_k.
    Initializes demand θ^m using direct power loss to macrogrid.
    Broadcast (C_k, θ^m, T^m) to all MGs and set $Q_t^m = 0$
end for
Main Loop:
for time slot $t = 1 : \mathcal{T}$ **do** all MG m, $m \in M$
    Update coalition action $a^{C_k} \rightarrow f(C_k, \boldsymbol{\theta}^{C_k})$ according to the agreement $(C_k, \boldsymbol{\theta}^{C_k})$.
    Update current reward $r^m(t)$.
    Update transition probabilities and beliefs.
    Estimate Eq. (6.20)
    BR Coalition Formation:
    Randomly selects a proposer MG m with the probability $1/M$.
    Make a proposal $\pi_k^m = (C_k, \boldsymbol{\theta}^{C_k})$ which maximize MG m beliefs about Q_t^m.
    Send $\pi_k^m = (C_k, \boldsymbol{\theta}^{C_k})$ to all MG n, $n \in C_k$.
    for all MG n, $n \in C_k$ **do**

        if $Q_t^n(C_k, a^{C_k}, \boldsymbol{\theta}^{C_k} \circ \theta^m) \geq Q_t^n(C_k, a^{C_k}, \boldsymbol{\theta}^{C_k})$ **then**
                Set a response $\Upsilon_k^n = 1$ and send $(\Upsilon_k^n, \theta^m, T^m)$ to MG m
        else
                Set a response $\Omega_k^n = 0$
        end if
    end for
    if $\prod_{n \in C_k} \Upsilon_k^n = 1$ **then**
            Update agreement $(C_k, \boldsymbol{\theta}^{C_k}) \rightarrow (C_k, \boldsymbol{\theta}^{C_k} \circ \theta^m)$
            Set the state $s^m \rightarrow C_k$
            Set the type T^m
            Broadcast T^m to all MG n, $n \in C_k$
    end if
end for

Q-Learning-Based Method

We compare our work with the Q-learning-based algorithm developed in previous chapters (see Section 2.3.2). We assume that MGs are agents and an agent's action is to refuse or accept the proposition of another MG to join their coalition. The state is the vector of coalition memberships, and the reward function is the same as Eq. (6.5).

BCG-based Method

We implement a BCG based approach for coalition formation which is proposed in Chapter 3 and 4. In this scheme, each MG makes a belief system about the types of other agents; however, agents do not learn from past experiences.

CG-based Method

A game theory-based coalition formation approach has been proposed in [7]. In this scheme, by employing a random merge and split technique, the system reaches a stable coalition formation which is not necessarily optimal or sub-optimal.

Note that the proposed method, BCRLMAPE, BCRLFME, and Q-learning benchmarks use the epsilon-greedy policy to increase the chance of exploration. The epsilon-greedy policy helps with the trade-off between exploration and exploitation. Agents attempt to improve their long-term benefits through exploration, while exploitation can be achieved by performing greedy actions. Algorithm 6.1 is also used for BCRLMAPE, BCRLFME.

6.5.2 Simulation Parameters and Setup

In this section, for numerical evaluation, we consider a network of 4 to 10 MGs within an area of 20 km by 20 km where MGs and macrogrid are located randomly. We divided the entire day into 240 time slots, where the load and generation patterns are randomly generated and periodically repeated every day with slight variations as in [7]. The proposed BCRL is compared with BCRLMAPE, BCRLFME, Q-learning, BCG, and CG-based benchmarks. Initial values of BCRL, BCRLMAPE and BCRLFME parameters are selected based on [133]. The results are averaged over 15 runs. The simulation parameters are presented in Table 6.1.

Table 6.1: Summary of simulation parameters.

Parameters	**Value**
Line resistance ($\mathbb{R}_{mn}$)	0.2
Medium voltage (U_0)	50 kV
Low voltage (U_i)	22 kV
Transformer loss fraction (ρ)	0.02
Threshold distance (D_{tr})	5 km
Virtual cost parameter (ω_s)	0.02
Virtual cost parameter (ω_l)	0.04
Virtual cost parameter (ω_0)	0.08
Scaling parameter (δ)	0.95

6.5.3 Numerical Results and Discussions

In Fig. 6.2, the average cost per user versus the number of MGs ranging from 4 to 10 is presented. As expected, increasing the number of MGs will reduce the cost since MGs have more chance to make local coalitions in a dense network, resulting in less power transmission with the macrogrid, resulting in lower cost. Moreover, since BCRL is designed to overcome the uncertainty, it demonstrates less cost compared to the other algorithms. The proposed algorithm shows 16% to 8% improvement compared to BCG and the sub-optimal BCRLMAPE, respectively.

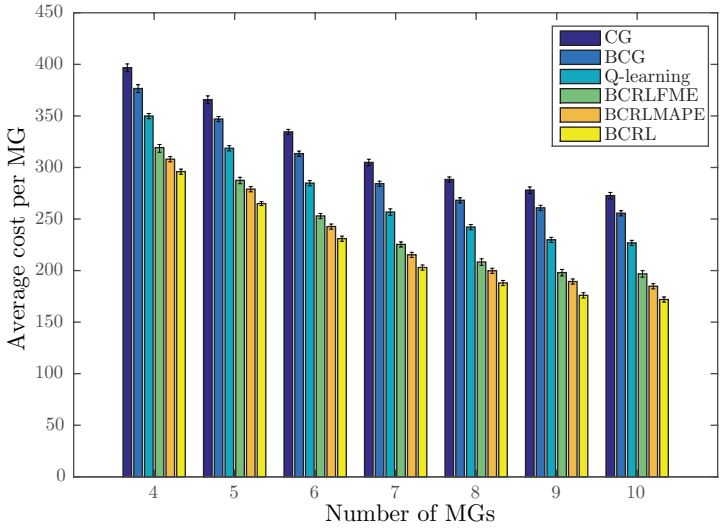

Figure 6.2: Average cost versus number of MGs.

In Fig. 6.3, to evaluate the effect of increasing power levels, the average cost per user versus the power levels is demonstrated. It should be noted that, in this figure, Q-learning and CG models cannot be compared with other methods since the power levels are not considered in these models. As is shown, when the power levels increase, the average cost decreases as expected. As we increase the number of power levels, the quantization error will reduce, and as a result, all the approaches perform better. As we can see in Fig. 6.3, at different power levels, BCRL reduces the average cost per MG to 7% and 15% compared to the BCRLMAPE and BCG methods, respectively.

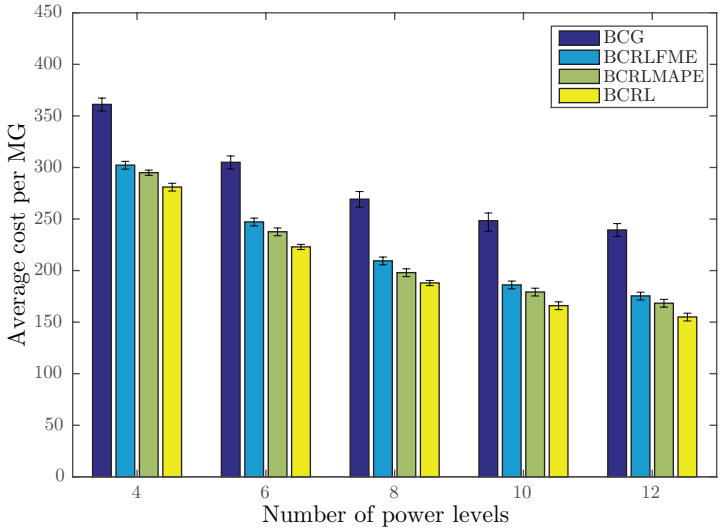

Figure 6.3: Average cost per MG versus the power levels.

In Fig. 6.4, we present the average power loss per user versus the number of MGs. As the number of MGs increases, the distance between MGs will be reduced, reducing the power loss in the system. Moreover, since BCRL is designed to overcome the uncertainty, it demonstrates better power loss results than benchmark approaches, with an up to 50% improvement compared to conventional CG. While Q-learning benefits from past experiences to make the best decision, BCG relies on beliefs about other players' types. The CG method only performs based on the random join and split iterations in coalitions to reach a stable coalition formation, which is not necessarily optimal.

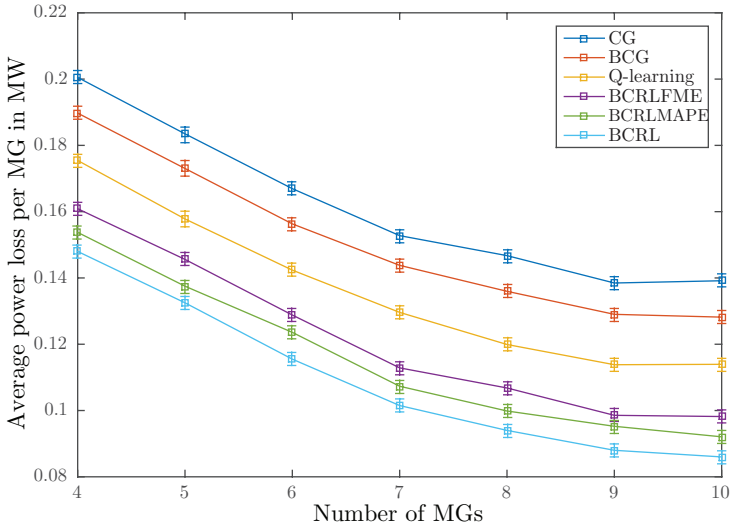

Figure 6.4: Average power loss versus number of MGs.

Fig. 6.5 shows the average power loss per user versus the number of power levels. As expected, by increasing the number of power levels, the power loss will be reduced due to the lower quantization error. As can be seen, the BCRL method is less prone to quantization errors due to its comprehensive estimation model for the expected action value. The BCRL method gained up to 20% on average compared to the BCG method.

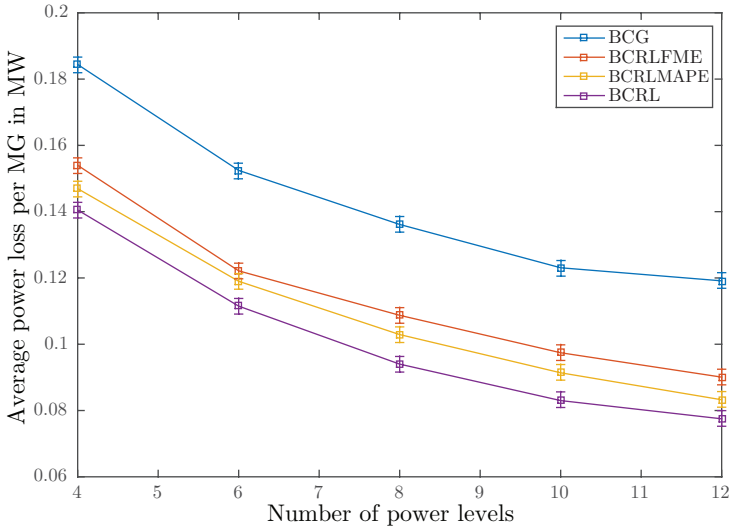

Figure 6.5: Average power loss per MG versus the number of power levels.

In Fig. 6.6, the average amount of energy transferred to the macrogrid versus the number of MGs is presented. As we can see, BCRL requires a lower amount of energy exportation to or importation from the macrogrid compared to the benchmark techniques. Additionally, due to the lower power loss between nearby MGs, the probability of nearby MGs joining the same coalition will increase by increasing the number of MGs, which can reduce the power exported to (imported from) the macrogrid as well.

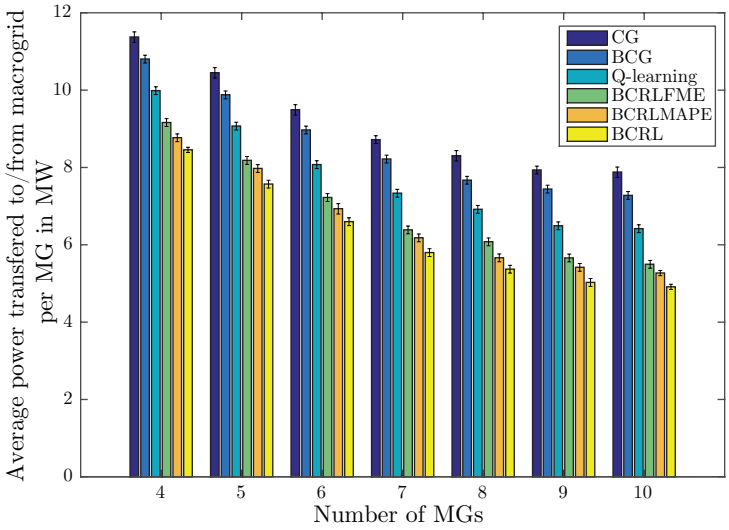

Figure 6.6: Average energy transfer with macrogrid versus number of MGs.

Fig. 6.7 shows the impact of increasing the cost of transferring energy with the macrogrid versus virtual cost weighting parameter ω_0. Here, the range of weighting parameter ω_0 varied between 0.02 to 0.22. As we can see, when ω_0 increases, the average energy transfer with macrogrid decreases, giving a chance to the coalition of MGs to operate in islanding mode. We can see that the proposed BCRL model always performs better in making independent coalitions that rely less on macrogrid. BCRL decreases the energy transactions with macrogrid up to 10% compared to the CG technique.

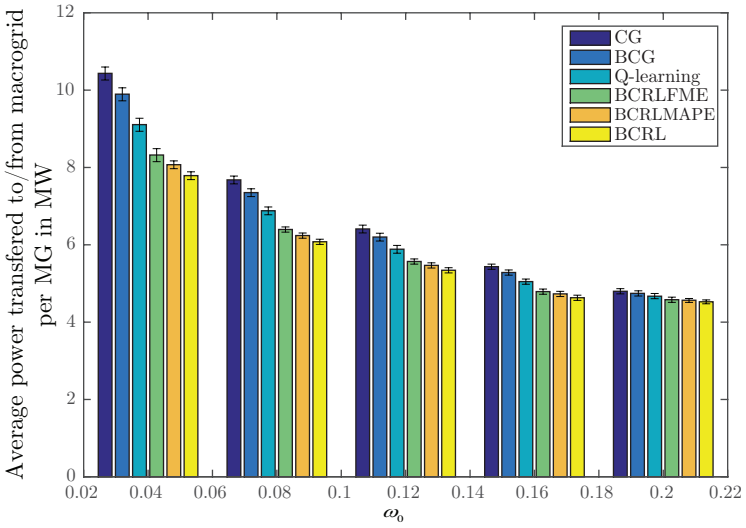

Figure 6.7: The average energy transfer with macrogrid versus virtual cost weighting parameter ω_0.

Figure 6.8 shows the convergence of the BCRL technique in terms of the average cost per user. As shown, the proposed model will be converged after 12,000 iterations.

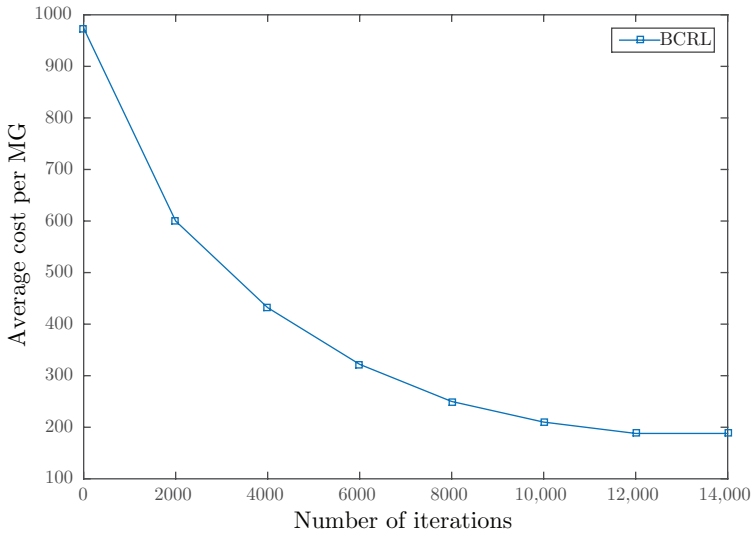

Figure 6.8: Convergence of the average cost per user versus the number of iteration for BCRL.

In Figure 6.9, the average number of iterations that are needed for the convergence of accumulative average cost as the number of power levels increases in the BCRL scheme is demonstrated.

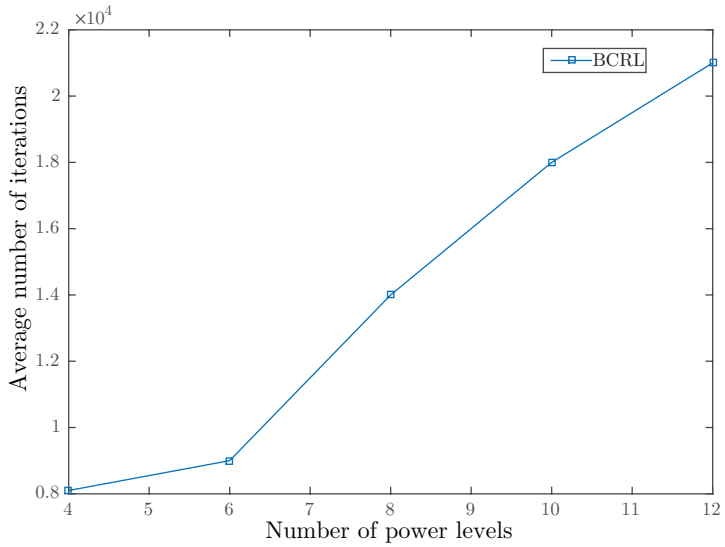

Figure 6.9: Number of iterations to convergence versus the power levels.

6.6 Conclusions

In this chapter, we proposed a BCRL-based approach for learning the optimal policy to minimize the cost of distributed energy trading among MGs. Each MG is modeled as an agent that can compete and cooperate with other agents. We model this problem as a POMDP, which aims to maximize the reward for each agent to overcome the uncertainties caused by MGs based on their generation and demand. The proposed algorithm helps each agent systematically propose joining a new coalition and gives the coalition members the chance to accept or reject the proposal according to their expected long-term rewards. With the proposed algorithm, MGs reach stable coalitions where the energy export to the macrogrid and distant MGs are reduced in the system. To evaluate the performance of the proposed model, we compared our results with five benchmark schemes, which showed that the proposed scheme reduced the cost and power loss more than the others, reaching to 23% reduction in cost and a 28% decrease in power loss. Large state-action spaces in the BRL-based methods can impact their scalability when the number of agents increases. In the next chapter, we propose a DRL-based method to address this issue.

Chapter 7

Deep Reinforcement Learning-Based Coalition Formation for Energy Trading among MGs

7.1 Introduction

In this chapter, we address the problem of minimizing cost in the coalitional MG communities considering the dynamic nature of the system. We propose a DRL approach that helps minimize the total cost by forming efficient coalitions [134]. The proposed method aims to address the issue of large state-action spaces associated with the reinforcement learning methods used in the earlier chapters. The results show 16% to 30% improvement in cost minimization compared to an existing Q-learning-based scheme and a conventional CG-based approach, respectively.

The rest of this chapter is organized as follows. In Section 7.2, the system model is described. In Section 7.3, we introduce our DRL-based coalition formation scheme. Numerical results are provided in Section 7.4 and finally, the conclusion is presented in Section 7.5.

7.2 System Model

In this chapter, we consider a system of M interconnected MGs, and all the MGs are connected to the main macrogrid. The MGs can participate in an energy transaction with

other nearby MGs. The energy trading opportunity gives the MGs the chance to export their surplus energy generation or import energy when they have unsatisfied loads. In each epoch of our considered system, some MGs are sellers (exporters) of energy, and others are buyers (importers) of energy. We represent the generated energy and the demand of m-th MG by e_m^g and e_m^d, respectively. Therefore, the equivalent surplus/shortage can be defined as $e_m^q = e_m^g - e_m^d$. e_m^q shows the amount of energy that m-th MG is required to import from or export to the grid, which makes MGs involve in an energy trading transaction. This value changes in each iteration and imposes uncertainty on the system. Energy trading transaction among MGs occurs with some cost. We define the total costs in the following as it is explained in Section 4.2:

$$S_{mn} = wd_{mn}E_{mn} + \delta PL(E_{mn}). \tag{7.1}$$

Consequently, the objective of this chapter is to minimize the total costs as follows:

$$\begin{aligned} &\min \sum_{m=0}^{M}\sum_{n=0}^{M} S_{mn} \\ &s.t: \sum_{m=1}^{M} e_m^q + \sum_{m=0}^{M}\sum_{n=0}^{M} P(E_{mn}) = \sum_{n=0}^{M} E_{0n} - \sum_{m=0}^{M} E_{m0}. \end{aligned} \tag{7.2}$$

Considering the above-mentioned objective function and criteria, we can conclude that MGs prefer to trade energy with their close neighbors rather than further ones or with macrogrid. Having energy trading in short distances can help reduce costs associated with distance, and also, there will be zero transformer loss in any energy trading between MGs. Therefore, forming a group of MGs that can trade energy among themselves can be a promising method to reduce cost. In this chapter, we use the coalition formation methodology to divide MGs into such groups. In the following section, we introduce the coalition formation.

7.2.1 Coalition Formation

We define coalitions as a group of members with a coalition leader l_C. We assume that we have C coalitions and consequently C leaders. The leader is responsible for approving new members for the coalitions. Also, all communications with members happen through leaders. We represent each coalition with a pair (C, v_C). All the members of coalition C cooperate to maximize the total profit of the coalition, known as the coalition value $v(C)$. We define the coalition value as the negative form of the cost. In this manner, we ensure the minimization of cost as the members of coalitions, which are MGs in our problem,

attempt to maximize their coalition values. The coalition value of each coalition consists of the cost of energy trading among the coalition members plus the cost of the energy transaction with the macrogrid. Therefore, the coalition value can be defined as:

$$v(C) = -\sum_{m=0}^{|C|}\sum_{n=0}^{|C|} S_{mn}, \tag{7.3}$$

here, the number of coalition members of coalition C is denoted by $|C|$. We use the index 0 to consider energy transactions (import or export) with the macrogrid. After forming the coalition, the coalition leader schedules energy trading in a manner that minimizes the total cost of energy trading among coalition members. Hence, the coalition payoff can be defined as the maximum achievable coalition value as follows:

$$v_{max}(C) = \max\{-\sum_{i=0}^{|C|}\sum_{j=0}^{|C|} S_{mn}\}. \tag{7.4}$$

7.3 DQN-Based Coalition Formation

In this section, we introduce our DQN-Based Coalition Formation (DQN-CF) scheme, where each coalition leader uses DQN to decide whether to accept an MG to the coalition or not based on minimizing the total cost of the coalition. In the following, we first introduce the conventional Q-learning method and then extend it to present the proposed DQN-CF algorithm. We will later use the Q-learning method as a benchmark to evaluate the DQN-CF.

7.3.1 Q-Learning-Based Coalition Formation

The tuples of Q-learning can be defined as agents l, states $s \in \mathbf{S}$, actions $a \in \mathbf{A}$ and reward function r. In the proposed scheme, we define each element as follows:

- *Agents*: In this scheme, we consider the leaders to be the agents of the system. Since more than one agent performs actions, we have a multi-agent scenario.
- *Actions*: In each time step, one MG has the chance to propose to join a new coalition randomly. The corresponding leader decides to accept or reject the joining proposal. Therefore, action space of leader l is a binary variable a^l.

- *States*: We define the state of the agent l in the system as $s^l = \{\hat{q_1}, \hat{q_2}, .., \hat{q_M}, P_{index}\}$. The $\hat{q_m}$ is the quantized total surplus energy or shortfall of energy of the m-th MG. The $\hat{q_m}$ is zero for non member MGs except for the proposer and P_{index} represents the index of the proposer.

- *Reward*: The aim of the reward design is to minimize the total cost. The reward function is considered to be in the negative form of the coalition cost as follows:

$$r^l = v_C^{max} = \max\{-\sum_{i=0}^{|C|}\sum_{j=0}^{|C|} S_{mn}\}. \tag{7.5}$$

Q-learning chooses actions that maximize the expected current and future rewards to reach a sub-optimal policy. We update Q-value in Q-learning considering Bellman's equation as below:

$$Q_\pi(s,a) = Q_\pi(s,a) + \alpha(r(s,a) + \gamma V_\pi(s') - Q_\pi(s,a)), \tag{7.6}$$

where α and γ denote the learning rate, and discount factor that shows the significance of future rewards respectively. The epsilon-greedy method is used to guarantee the action exploration as in [124].

7.3.2 DQN-CF

Q-table is used in regular Q-learning to keep the record of cumulative reward corresponding to each action and state pair a and s and then the agent decides to choose the next action according to the Q-values. Although this works well for small Q-tables with limited action-state space, the memory will be larger when the action-state space is large, and consequently, time complexity increases dramatically. To this end, DQN has been proposed to address these problems and employed in many studies where a neural network is used to estimate the Q-values [58, 60]. DQN method is explained in 2.3.2. We summarize the proposed DQN-CF scheme in Algorithm 7.1. In every epoch of the system, one MG is selected at random to be the proposer. The proposer can choose randomly to stay in the current coalition or join a new coalition. The coalition leader, as the agent of the DQN, will take action to accept or reject the joining proposal according to the estimated value function. Consecutive merges and splits happen until the system reaches a stable coalition formation in each state.

Algorithm 7.1 Coalition formation with DQN for energy trading among MGs

Initialization: Initialize parameters α and γ.
At time $t = 0$:
for MG $m = 1$ to M **do**
    Randomly select coalition C, set the power level e_m^q.
    Broadcast C to all MGs
end for
Main loop:
for Each time slot $t = 1$ to $\mathcal{T}$ **do**
    for leader $l = 1$ to L **do**
        Update current coalition reward $r^l(t)$
        Update the DQN estimation
    end for
    DQN Coalition formation With probability of $1/M$ the proposer MG_i is selected from the set M and randomly choose coalition C;
    For the leader of coalition C:
    Take an action a^l according to DQN estimation
    Send a^l to the proposer i
    If $a^l = Yes$ then set $MG_i \in C_k$ and update the r^l.
end for

7.3.3 Baseline Algorithms

Baseline I - Q-Learning based Coalition Formation: We introduced the Q-learning approach in Section 7.3.1, and we use this method to compare the proposed DQN-based method with a well-known reinforcement learning-based technique.

Baseline II - Conventional CG: The CG-based method is used as a benchmark, as explained in previous chapters.

7.4 Results

7.4.1 Simulation Parameters and Setup

In this work, MATLAB Toolbox is used as simulation software. A region of 25km by 25km is considered in which MGs are assumed to be distributed randomly. The number of MGs varies between 6 and 18. A 24-hour interval is assumed for the simulation. Load and generation patterns are generated randomly, considering the Gaussian random variable. The same pattern with slight variations is repeated for each day periodically to simulate a longer time horizon [7]. We compare the proposed DQN-CF technique with Q-learning and CG-based methods. The results are achieved for 15 runs with at least 1500 iterations, and the average results are reported. The simulation values are summarized in Table 7.1. We use Adam optimizer in our DQN-CF method.

7.4.2 Simulation Results

In Fig. 7.1, we present the average cost per MG versus the number of MGs ranging from 6 to 18. As expected, increasing the number of MGs will reduce the cost since MGs have more chance to make local coalitions in a dense network, resulting in less power transmission with a macrogrid. Moreover, DQN-CF demonstrates better results in terms of cost compared to the other algorithms. The proposed algorithm shows 16% and 4% improvement compared to CG and the suboptimal Q-learning-based approach, respectively.

In Fig. 7.2, to evaluate the effect of increasing power levels, we demonstrate the average cost per user versus the number of quantized power levels. As shown, when the number of power levels increases, the average cost decreases, as expected. As we increase the number of power levels, the quantization error will be reduced, and as a result, the performance of DQN-CF and Q-learning-based approaches will improve.

Table 7.1: Summary of simulation parameters.

parameters	**value**
Line resistance (R_{ij})	0.2
Medium voltage (U_0)	50 kv
Low voltage (U_i)	22 kv
Transformer loss fraction (ρ)	0.02
Threshold distance (D_{tr})	5 km
Virtual cost parameter (ω_s)	0.02
Virtual cost parameter (ω_l)	0.04
Virtual cost parameter (ω_0)	0.08
Scaling parameter (δ)	0.95
Learning rate of Q-learning (α)	0.5
Discount factor of Q-learning (γ)	0.8
Size of hidden layers	2
Number of hidden units	25
Training batch size	160
Size of replay memory	60
Training interval	60

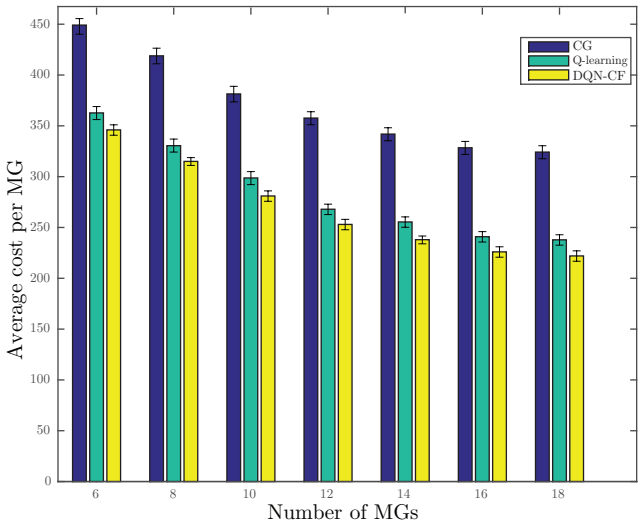

Figure 7.1: Average cost versus number of MGs.

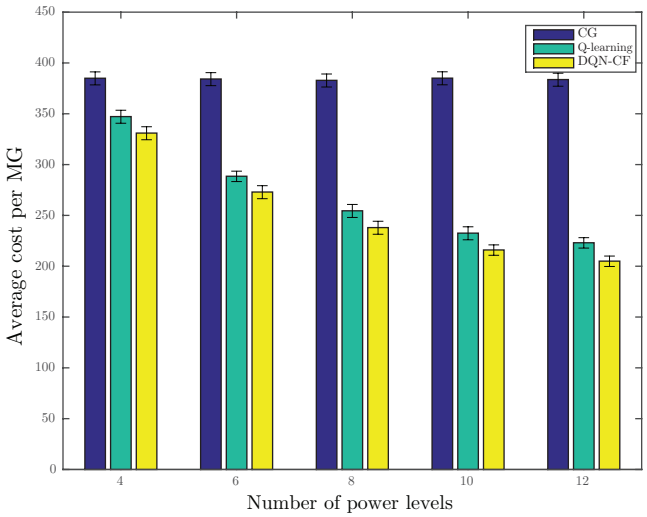

Figure 7.2: Average cost versus number of power levels.

In Fig. 7.3, we present the average power loss per user versus the number of MGs. As the number of MGs increases, the power loss decreases. Moreover, since DQN-CF is designed to overcome the uncertainty, it demonstrates better results in terms of power loss compared to benchmark approaches, up to 24% improvement compared to conventional CG.

In Fig. 7.4, the average cost per MG is plotted versus the number of iterations. Since we considered the cost as the reward, less value shows better performance. As expected, DQN-CF converges faster with better performance compared to the Q-learning-based approach.

7.5 Conclusion

In this chapter, we investigated the coalitional energy trading problem with the aim of cost minimization in a system with uncertainties. We proposed a DRL approach to overcome the uncertainties that arise from MGs' power level, which results in less energy transfer from macrogrid or distant MGs. We compared the proposed approach with Q-learning and CG schemes, and a significant reduction in terms of cost (almost 16% and 30%) has been achieved.

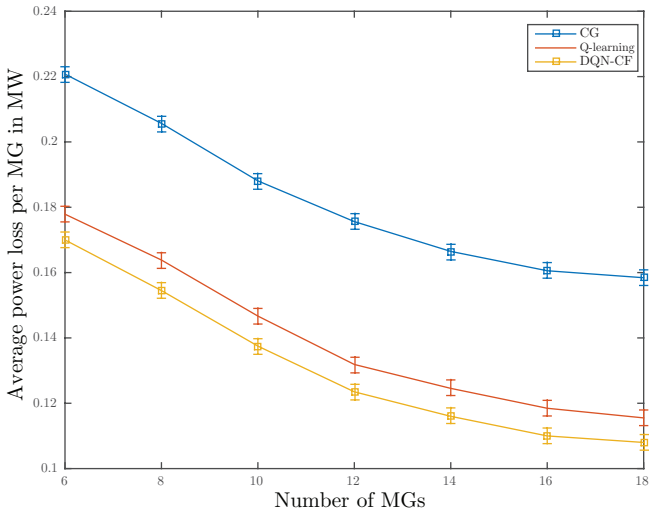

Figure 7.3: Average power loss versus number of MGs.

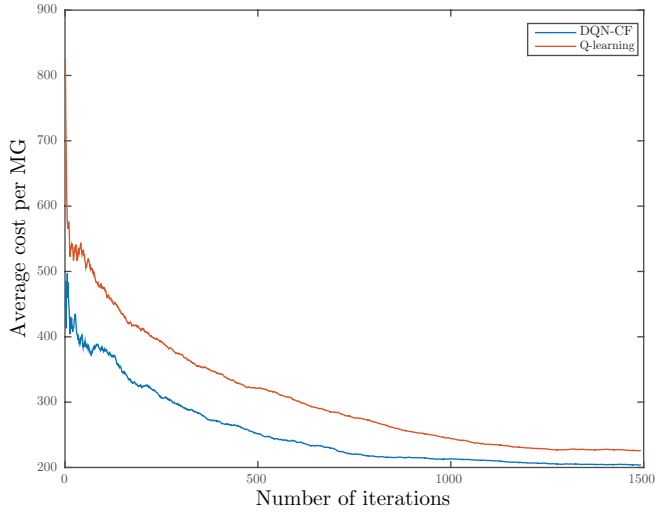

Figure 7.4: Average cost per MG is plotted versus the number of iterations.

Chapter 8

Conclusion and Future Directions

The power grid has been evolving in the last two decades. During this transformation, MGs emerged as promising electricity distribution systems that can revolutionize the smart grid, particularly when they can autonomously form coalitions or communities. MGs are equipped with generators that can provide part of their own energy demand or even the demand of other MGs through peer-to-peer energy trading.

The necessity of a more distributed MGs system that depends less on the utility grid to satisfy their demand makes energy trading among MGs a critical problem. On the other hand, variations in the generation of renewables, demand and energy price in the MGs market impose major uncertainty on these systems, which cannot be neglected in the design of MGs and energy trading methods. To this end, in this dissertation, we proposed BCG-based, BRL-based and DRL-based methods to tackle the energy trading problem among MGs under uncertainty. The contribution of each chapter is summarized in the following:

- In chapter 2, we provided comprehensive background information on the proposed systems and conducted a thorough literature review on the problem of MG energy trading.
- In chapter 3, we studied the problem of energy trading among MGs and EVs with the aim of power loss minimization where there is an uncertainty resulting from the type of agents (moving/fixed). We proposed a novel BCG-based algorithm that allows the MGs and EVs to reduce the overall power loss by forming coalitions intelligently. The proposed scheme was compared with a conventional CG-based approach and

a Q-learning-based approach, and a reduction of 16% in power loss was achieved compared to the Q-learning-based approach.

- In chapter 4, the problem of energy trading among MGs was revisited to minimize the cost, considering the uncertainties that are imposed by the lack of information about whether the MGs can fulfill their commitment or not. A BCG-based scheme was proposed, which helps the MGs to minimize the overall cost by forming stable coalitions. The results showed a 30% improvement in terms of cost minimization compared to the conventional CG method.

- In chapter 5, we aimed to minimize power loss while addressing and overcoming the uncertainties from the energy level of agents in the system. We proposed a BRL based algorithm that enables MGs to learn from past experiences toward maximizing their expected long-term reward, consequently reducing the overall power loss. We compared this scheme with a CG-based approach, Q-learning-based approach, random coalition formation approach, and a case with no coalitions. Our results showed that more than 50% reduction in power loss compared to no coalitions and less power loss than the other approaches are achieved.

- In chapter 6, a comprehensive BCRL approach was proposed that includes a belief update system and helps agents learn from past experiences to take the best action through interaction with the environment gradually. Our results showed that the proposed scheme in this study reduced the cost and power loss more than the others, reaching to 23% reduction in cost and a 28% reduction in power loss compared to the Q-learning method.

- In chapter 7, we proposed a DQN-based approach to address the scalability problem that impacts previously mentioned reinforcement learning techniques. We compared the proposed approach with Q-learning and CG schemes, and a significant reduction of almost 16% and 30% in terms of cost has been achieved.

8.1 Future Direction

As the future grids will be more complex, more advanced machine learning methods are required to address energy trading problems among MGs, which brings about unprecedented challenges to be dealt with in future studies. A summary of these challenges is briefly presented in the following.

The long convergence time places a significant burden on the execution of reinforcement learning methods in highly dynamic MG systems. Therefore, a thorough analysis of the convergence issue and the elements that affect the convergence are critical. Faster reinforcement learning approaches with shorter convergence time and online learning capabilities can significantly improve the performance of MG systems.

In addition, the uncertainty imposed by the stochastic nature of the MG systems requires continuous adaptation of the parameters of the reinforcement learning methods or even the technique itself over time. For example, a highly populated MG community that supports diverse users has an extremely dynamic function. Moreover, users enter or leave the system with distinct structures and different energy generation and demands. Thus, a careful investigation of whether a generalized approach is implementable in real-world scenarios is necessary.

Furthermore, the scalability of reinforcement learning algorithms needs to be studied. Reinforcement learning algorithms can become impractical for relatively large data, especially in multi-agents scenarios. Therefore, employing scalable learning methods capable of controlling the dense smart grid system is important.

On the other hand, the simulation design of such dense and diverse scenarios needs careful attention. This calls for considering more practical elements and constraints to simulate realistic scenarios, which will pave the way for achieving an optimal result in real-life implementations.

In summary, joint Bayesian coalition formation methods and reinforcement learning techniques are promising solutions to tackle energy trading problems in smart MG communities under uncertainty. While as these communities move toward denser and more heterogeneous structures, considering more advanced simulation design and the use of more intelligent learning techniques is critical.

References

[1] X. Fang, S. Misra, G. Xue, and D. Yang. Smart grid; the new and improved power grid: A survey. *IEEE Communications Surveys Tutorials*, 14(4):944–980, Fourth 2012.

[2] M. H. Rehmani et al. Ieee access special section editorial smart grids: a hub of interdisciplinary research. *IEEE Access*, 3:3114–3118, 2015.

[3] J. Gao et al. A survey of communication/networking in smart grids. *Future Gener. Comput. Syst.*, 28(2):391–404, feb 2012.

[4] R. H. Lasseter and P. Paigi. Microgrid: a conceptual solution. In *2004 IEEE 35th Annual Power Electronics Specialists Conference (IEEE Cat. No.04CH37551)*, volume 6, pages 4285–4290 Vol.6, 2004.

[5] M. Erol-Kantarci, B. Kantarci, and H. T. Mouftah. Reliable overlay topology design for the smart microgrid network. *IEEE Network*, 25(5):38–43, Sep. 2011.

[6] L. Lei, J. Li, M. Erol-Kantarci, and B. Kantarci. An integrated reconfigurable control and self-organizing communication framework for community resilience microgrids. *The Electricity Journal*, 30(4):27–34, 2017.

[7] W. Saad, Z. Han, and H. V. Poor. Coalitional game theory for cooperative microgrid distribution networks. In *IEEE International Conference on Communications Workshops*, pages 1–5, 2011.

[8] K. Su, J. Li, and H. Fu. Smart city and the applications. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, pages 1028–1031, Sep. 2011.

[9] T. Shelton et al. The 'actually existing smart city'. *Cambridge Journal of Regions, Economy and Society*, 8(1):13–25, 10 2014.

[10] H. Mouftah et al. *Transportation and Power Grid in Smart Cities: Communication Networks and Services*. Wiley, 2018.

[11] A. Ibrahim et al. The role of big data in smart city. *International Journal of Information Management*, 36(5):748 – 758, 2016.

[12] V. Coelho et al. A communitarian microgrid storage planning system inside the scope of a smart city. *Applied Energy*, 201:371–381, 2017.

[13] S. Khan et al. Artificial intelligence framework for smart city microgrids: State of the art, challenges, and opportunities. In *2018 Third International Conference on Fog and Mobile Edge Computing (FMEC)*, pages 283–288. IEEE, 2018.

[14] S. Obara and J. Morel. *Clean energy microgrids, vol. 1*. United Kingdom: IET-The institution of Engineering and Technology, 2017.

[15] D. Olivares et al. Trends in microgrid control. *IEEE Transactions on smart grid*, 5(4):1905–1919, 2014.

[16] F. Martin-Martínez et al. A literature review of microgrids: A functional layer based classification. *Renewable and Sustainable Energy Reviews*, 62:1133–1153, 2016.

[17] D. Ton and M. Smith. The us department of energy's microgrid initiative. *The Electricity Journal*, 25(8):84–94, 2012.

[18] L. Mariam et al. Microgrid: Architecture, policy and future trends. *Renewable and Sustainable Energy Reviews*, 64:477–489, 2016.

[19] E. Alegria et al. Certs microgrid demonstration with large-scale energy storage and renewable generation. *IEEE Transactions on Smart Grid*, 5(2):937–943, 2013.

[20] W. W. Anderson. Smart power infrastructure demonstration for energy reliability and security (spiders) final report. Technical report, Naval Facilities Engineering Command Joint Base Pearl Harbor-Hickam United . . . , 2015.

[21] R. Tonkoski, L.A.C. Lopes, and D. Turcotte. Active power curtailment of pv inverters in diesel hybrid mini-grids. In *2009 IEEE Electrical Power Energy Conference (EPEC)*, pages 1–6, 2009.

[22] European research project cluster. microgrids and more microgrids projects [online]. http://www.microgrids.eu/default.php.

[23] G. Kariniotakis, A. Dimeas, and F. Van Overbeeke. Pilot sites: success stories and learnt lessons. *Microgrids*, pages 206–274, 2013.

[24] Leonardo energy. the first microgrid in the netherlands [online]. http://www.olino.org/blog/nl/wp-content/uploads/2009/10/the-first-micro-grid-in-the-netherlands-bronsbergen.pdf.

[25] Nice grid. project architecture and diagram [online]. http://www.nicegrid.fr/en/diagram/.

[26] National renewable energy centre. atenea microgrid [online]. http://www.cener.com/en/areas/renewable-energy-grid-integration-department/infrastructures-and-technical-resources/microgrid/.

[27] Jofemar corporation. factory microgrid – description [online]. http://www.factorymicrogrid.com/es/el-proyecto/descripcion-del-proyecto.aspx.

[28] European commission. 2020 climate energy package [online]. https://ec.europa.eu/clima/policies/strategies/2020/.

[29] N. Scott. Microgrids a guide to their issues and value. *Highlands and Islands Enterprise*, 2016.

[30] Microgrid media. microgrid projects map [online]. http://microgridprojects.com/.

[31] P. Punjad et al. Case study of micro power grid applications in remote rural area of thailand. In *AORC Technical Meeting*, pages 1–6, 2014.

[32] M. Sadeghi, M. Erol-Kantarci, and H.T. Mouftah. *Connected and Autonomous Electric Vehicle Charging Infrastructure Integration to Microgrids in Future Smart Cities*, chapter 1. CRC Taylor & Francis, 2020.

[33] H. Kamankesh et al. Optimal scheduling of renewable micro-grids considering plug-in hybrid electric vehicle charging demand. *Energy*, 100:285–297, 2016.

[34] S. Bahramara and H. Golpîra. Robust optimization of micro-grids operation problem in the presence of electric vehicles. *Sustainable cities and society*, 37:388–395, 2018.

[35] A. Kavousi-Fard et al. Impact of plug-in hybrid electric vehicles charging demand on the optimal energy management of renewable micro-grids. *Energy*, 78:904–915, 2014.

[36] J. Trovão and C. Antunes. A comparative analysis of meta-heuristic methods for power management of a dual energy storage system for electric vehicles. *Energy conversion and management*, 95:281–296, 2015.

[37] J. Trovão et al. A multi-level energy management system for multi-source electric vehicles–an integrated rule-based meta-heuristic approach. *Applied Energy*, 105:304–318, 2013.

[38] S. Derakhshandeh et al. Coordination of generation scheduling with pevs charging in industrial microgrids. *IEEE Transactions on Power Systems*, 28(3):3451–3461, 2013.

[39] I. Zenginis et al. Cooperation in microgrids through power exchange: An optimal sizing and operation approach. *Applied energy*, 203:972–981, 2017.

[40] F. Ahmad et al. Developments in xevs charging infrastructure and energy management system for smart microgrids including xevs. *Sustainable Cities and Society*, 35:552–564, 2017.

[41] A. Karnama et al. Optimal management of battery charging of electric vehicles: A new microgrid feature. In *2011 IEEE Trondheim PowerTech*, pages 1–8, June 2011.

[42] JA. P. Lopes et al. Identification of control and management strategies for lv unbalanced microgrids with plugged-in electric vehicles. *Electric Power Systems Research*, 80(8):898–906, 2010.

[43] S. Rahman et al. A vehicle-to-microgrid (v2m) framework with optimization-incorporated distributed ev coordination for a commercial neighborhood. *IEEE Transactions on Industrial Informatics*, 2019.

[44] B. Li, M. C. Kisacikoglu, C. Liu, N. Singh, and M. Erol-Kantarci. Big data analytics for electric vehicle integration in green smart cities. *IEEE Communications Magazine*, 55(11):19–25, Nov 2017.

[45] S. Mousavian, M. Erol-Kantarci, L. Wu, and T. Ortmeyer. A risk-based optimization model for electric vehicle infrastructure response to cyber attacks. *IEEE Transactions on Smart Grid*, 9(6):6160–6169, Nov 2018.

[46] S. Mousavian, M. Erol-Kantarci, and T. Ortmeyer. Cyber attack protection for a resilient electric vehicle infrastructure. In *2015 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6, Dec 2015.

[47] T. Başar and G. J. Olsder. *Dynamic noncooperative game theory*. SIAM, 1998.

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

[49] O. Morgenstern and J. Von Neumann. *Theory of games and economic behavior*. Princeton university press, 1953.

[50] R. J. Aumann and B. Peleg. Von neumann-morgenstern solutions to cooperative games without side payments. *Bulletin of the American Mathematical Society*, 66(3):173–179, 1960.

[51] G. Chalkiadakis and C. Boutilier. Sequentially optimal repeated coalition formation under uncertainty. *Autonomous Agents and Multi-Agent Systems*, 24(3):441–484, 2012.

[52] G. Chalkiadakis and C. Boutilier. Bayesian reinforcement learning for coalition formation under uncertainty. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, pages 1090–1097, 2004.

[53] G. Chalkiadakis and C. Boutilier. Coalitional bargaining with agent type uncertainty. In *IJCAI*, pages 1227–1232, 2007.

[54] E. Alpaydin. *Introduction to machine learning*. MIT Press, 2010.

[55] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998.

[56] M. Ghavamzadeh et al. Bayesian reinforcement learning: A survey. *arXiv preprint arXiv:1609.04436*, 2016.

[57] M. O. Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002.

[58] V. Mnih et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[59] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[60] M. Elsayed and M. Erol-Kantarci. Ai-enabled future wireless networks: Challenges, opportunities, and open issues. *IEEE Vehicular Technology Magazine*, 14(3):70–77, 2019.

[61] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[62] A. Luth et al. Local electricity market designs for peer-to-peer trading: The role of battery flexibility. *Applied Energy*, 229:1233–1243, 2018.

[63] T. Morstyn and M. D. McCulloch. Multiclass energy management for peer-to-peer energy trading driven by prosumer preferences. *IEEE Transactions on Power Systems*, 34(5):4005–4014, 2019.

[64] S. Nguyen et al. Optimizing rooftop photovoltaic distributed generation with battery storage for peer-to-peer energy trading. *Applied Energy*, 228:2567–2580, 2018.

[65] F. Si et al. Cost-efficient multi-energy management with flexible complementarity strategy for energy internet. *Applied Energy*, 231:803–815, 2018.

[66] M. Khorasany, Y. Mishra, and G. Ledwich. A decentralized bilateral energy trading system for peer-to-peer electricity markets. *IEEE Transactions on Industrial Electronics*, 67(6):4646–4657, 2020.

[67] Y. Wang et al. Shadow price based co-ordination methods of microgrids and battery swapping stations. *Applied Energy*, 253:113510, 2019.

[68] J. Lee et al. Distributed energy trading in microgrids: A game-theoretic model and its equilibrium analysis. *IEEE Transactions on Industrial Electronics*, 62(6):3524–3533, June 2015.

[69] A. Jadhav and N. Patne. Priority based energy scheduling in a smart distributed network with multiple microgrids. *IEEE Transactions on Industrial Informatics*, PP(99):1–1, 2017.

[70] S. Park et al. Contribution-based energy-trading mechanism in microgrids for future smart grid: A game theoretic approach. *IEEE Transactions on Industrial Electronics*, 63(7):4255–4265, July 2016.

[71] W. Tushar et al. Prioritizing consumers in smart grid: A game theoretic approach. *IEEE Transactions on Smart Grid*, 5(3):1429–1438, May 2014.

[72] H. Wang et al. Reinforcement learning in energy trading game among smart microgrids. *IEEE Transactions on Industrial Electronics*, 63(8):5109–5119, Aug 2016.

[73] W. Tushar et al. Grid influenced peer-to-peer energy trading. *IEEE Transactions on Smart Grid*, 11(2):1407–1418, 2020.

[74] B. Ahmad Bhatti and R. Broadwater. Energy trading in the distribution system using a non-model based game theoretic approach. *Applied Energy*, 253:113532, 2019.

[75] A. Paudel et al. Peer-to-peer energy trading in a prosumer-based community microgrid: A game-theoretic model. *IEEE Transactions on Industrial Electronics*, 66(8):6087–6097, 2019.

[76] K. Chen, J. Lin, and Y. Song. Trading strategy optimization for a prosumer in continuous double auction-based peer-to-peer market: A prediction-integration model. *Applied Energy*, 242(C):1121–1133, 2019.

[77] W. Liu, D. Qi, and F. Wen. Intraday residential demand response scheme based on peer-to-peer energy trading. *IEEE Transactions on Industrial Informatics*, 16(3):1823–1835, 2020.

[78] J. Kang et al. Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains. *IEEE Transactions on Industrial Informatics*, 13(6):3154–3164, 2017.

[79] C. Feng et al. Coalitional game based transactive energy management in local energy communities. *IEEE Trans. on Power Systems*, 2019.

[80] R. Lahon, C. P. Gupta, and E. Fernandez. Coalition formation strategies for cooperative operation of multiple microgrids. *IET Generation, Transmission Distribution*, 13(16):3661–3672, 2019.

[81] J. Mei, C. Chen, J. Wang, and J. L. Kirtley. Coalitional game theory based local power exchange algorithm for networked microgrids. *Applied Energy*, 239:133 – 141, 2019.

[82] C. Essayeh, M. R. El Fenni, and H. Dahmouni. Optimization of energy exchange in microgrid networks: A coalition formation approach. *Protection and Control of Modern Power Systems*, 4(1):24, 2019.

[83] E. Winter. The shapley value. In *Handbook of Game Theory with Economic Applications*, volume 3, pages 2025–2054. Elsevier, 2002.

[84] S. Misra, P. V. Krishna, V. Saritha, and M. S. Obaidat. Learning automata as a utility for power management in smart grids. *IEEE Communications Magazine*, 51(1):98–104, January 2013.

[85] B. Jiang and Y. Fei. Dynamic residential demand response and distributed generation management in smart microgrid with hierarchical agents. *Energy Procedia*, 12:76 – 90, 2011.

[86] Y. Xu, W. Zhang, W. Liu, and F. Ferrese. Multiagent-based reinforcement learning for optimal reactive power dispatch. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1742–1751, Nov 2012.

[87] C. Guan, Y. Wang, X. Lin, S. Nazarian, and M. Pedram. Reinforcement learning-based control of residential energy storage systems for electric bill minimization. In *2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC)*, pages 637–642, Jan 2015.

[88] B. G. Kim, Y. Zhang, M. van der Schaar, and J. W. Lee. Dynamic pricing for smart grid with reinforcement learning. In *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 640–645, April 2014.

[89] B. G. Kim, Y. Zhang, M. van der Schaar, and J. W. Lee. Dynamic pricing and energy consumption scheduling with reinforcement learning. *IEEE Transactions on Smart Grid*, 7(5):2187–2198, Sept 2016.

[90] W. Liu et al. Cooperative neural fitted learning for distributed energy management in microgrids via wireless networks. In *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*, pages 1–5, Sept 2017.

[91] H. Zhou et al. Multi-agent bayesian deep reinforcement learning for microgrid energy management under communication failures. *IEEE Internet of Things Journal*, pages 1–1, 2021.

[92] X. Xiao et al. Energy trading game for microgrids using reinforcement learning. In *Game Theory for Networks*, pages 131–140. Springer International Publishing, 2017.

[93] L. Xiao et al. Reinforcement learning-based energy trading for microgrids [online]. https://arxiv.org/abs/1801.06285.

[94] T. Chen and S. Bu. Realistic peer-to-peer energy trading model for microgrids using deep reinforcement learning. In *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, pages 1–5, 2019.

[95] T. Chen and W. Su. Indirect customer-to-customer energy trading with reinforcement learning. *IEEE Transactions on Smart Grid*, 10(4):4338–4348, 2019.

[96] M. Sedighizadeh et al. Stochastic multi-objective economic-environmental energy and reserve scheduling of microgrids considering battery energy storage system. *International Journal of Electrical Power Energy Systems*, 106:1–16, 2019.

[97] M. Bornapour et al. Optimal stochastic scheduling of chp-pemfc, wt, pv units and hydrogen storage in reconfigurable micro grids considering reliability enhancement. *Energy Conversion and Management*, 150:725–741, 2017.

[98] P. Fazlalipour, M. Ehsan, and B. Mohammadi-Ivatloo. Risk-aware stochastic bidding strategy of renewable micro-grids in day-ahead and real-time markets. *Energy*, 171:689–700, 2019.

[99] A. Najafi et al. Medium-term energy hub management subject to electricity price and wind uncertainty. *Applied Energy*, 168:418–433, 2016.

[100] Y. Wang et al. A stochastic-cvar optimization model for cchp micro-grid operation with consideration of electricity market, wind power accommodation and multiple demand response programs. *Energies*, 12(20), 2019.

[101] M. Tavakoli et al. Cvar-based energy management scheme for optimal resilience and operational cost in commercial building microgrids. *International Journal of Electrical Power Energy Systems*, 100:1–9, 2018.

[102] M. R. Ebrahimi and N. Amjady. Adaptive robust optimization framework for day-ahead microgrid scheduling. *International Journal of Electrical Power Energy Systems*, 107:213–223, 2019.

[103] Y. Xiang, J. Liu, and Y. Liu. Robust energy management of microgrid with uncertain renewable generation and load. *IEEE Transactions on Smart Grid*, 7(2):1034–1043, 2016.

[104] C. Zhang et al. Robustly coordinated operation of a multi-energy microgrid with flexible electric and thermal loads. *IEEE Transactions on Smart Grid*, 10(3):2765–2775, 2019.

[105] Z. Shi et al. Distributionally robust chance-constrained energy management for islanded microgrids. *IEEE Transactions on Smart Grid*, 10(2):2234–2244, 2019.

[106] Y. Li et al. Optimal scheduling of an isolated microgrid with battery storage considering load and renewable generation uncertainties. *IEEE Transactions on Industrial Electronics*, 66(2):1565–1575, 2019.

[107] L. Bai et al. Interval optimization based operating strategy for gas-electricity integrated energy systems considering demand response and wind uncertainty. *Applied Energy*, 167:270–279, 2016.

[108] Y. Li et al. Multi-objective optimal dispatch of microgrid under uncertainties via interval optimization. *IEEE Transactions on Smart Grid*, 10(2):2046–2058, 2019.

[109] M. Sola and G. M. Vitetta. Demand-side management in a smart micro-grid: A distributed approach based on bayesian game theory. In *2014 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, pages 656–661, 2014.

[110] X. Han et al. Distributed energy-sharing strategy for peer-to-peer microgrid system. *Journal of Energy Engineering*, 146(4):04020033, 2020.

[111] A. Anvari-Moghaddam et al. A multi-agent based energy management solution for integrated buildings and microgrid system. *Applied Energy*, 203:41–56, 2017.

[112] P. R. Jeyaraj, S. P. Asokan, and A. C. Karthiresan. Optimum power flow in dc microgrid employing bayesian regularized deep neural network. *Electric Power Systems Research*, 205:107730, 2022.

[113] P. Wang et al. Uncertainty-aware energy management of extended range electric delivery vehicles with bayesian ensemble. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1556–1562, 2020.

[114] Y. Ji et al. Real-time energy management of a microgrid using deep reinforcement learning. *Energies*, 12(12), 2019.

[115] T. Yang et al. Dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning. *Energy*, 235:121377, 2021.

[116] M. Sadeghi, S. Mollahasani, and M. Erol-Kantarci. Power loss-aware transactive microgrid coalitions under uncertainty. *MDPI Energies*, page to appear, 2020.

[117] J. Machowski et al. *Power system dynamics: stability and control*. John Wiley & Sons, 2020.

[118] P. A. Lipka et al. Constructing transmission line current constraints for the ieee and polish systems. *Energy Systems*, 8(1):199–216, 2017.

[119] S. Ghahramani. *Fundamentals of Probability: With Stochastic Processes*. Chapman and Hall/CRC, 2018.

[120] W. Saad et al. Coalitional game theory for communication networks. *IEEE Signal Processing Magazine*, 26(5):77–97, 2009.

[121] K. Akkarajitsakul, E. Hossain, and D. Niyato. Coalition-based cooperative packet delivery under uncertainty: A dynamic bayesian coalitional game. *IEEE Transactions on Mobile Computing*, 12(2):371–385, 2013.

[122] M. Pourahmadi. *Foundations of time series analysis and prediction theory*, volume 379. John Wiley & Sons, 2001.

[123] K. Apt and A. Witzel. A generic approach to coalition formation. In *Proc. of the Int. Workshop on Computational Social Choice (COMSOC)*, pages 1–6, Dec 2006.

[124] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[125] Husheng Li and Weiyi Zhang. Qos routing in smart grid. In *2010 IEEE Global Telecommunications Conference GLOBECOM 2010*, pages 1–6, 2010.

[126] J. Machowski et al. *Power system dynamics: stability and control*. John Wiley & Sons, 2020.

[127] M. Sadeghi, S. Mollahasani, and M. Erol-Kantarci. Cost-aware dynamic bayesian coalitional game for energy trading among microgrids. In *IEEE International Conference on Communications Workshops (ICC)*, pages 1–6, June 2021.

[128] S. Ieong and Y. Shoham. Bayesian coalitional games. In *AAAI*, pages 95–100, 2008.

[129] D. S. Watkins. *Fundamentals of matrix computations*, volume 64. John Wiley & Sons, 2004.

[130] M. Sadeghi and M. Erol-Kantarci. Power loss minimization in microgrids using bayesian reinforcement learning with coalition formation. In *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pages 1–6, Sep. 2019.

[131] J. Zico Kolter and Andrew Y. Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 513–520, 2009.

[132] M. Sadeghi, S. Mollahasani, and M. Erol-Kantarci. Cost-optimized microgrid coalitions using bayesian reinforcement learning. *Energies*, 14(22), 2021.

[133] A. Asheralieva. Bayesian reinforcement learning-based coalition formation for distributed resource sharing by device-to-device users in heterogeneous cellular networks. *IEEE Transactions on Wireless Communications*, 16(8):5016–5032, 2017.

[134] M. Sadeghi and M. Erol-Kantarci. Deep reinforcement learning based coalition formation for energy trading in smart grid. In *2021 IEEE 4th 5G World Forum (5GWF)*, pages 200–205, 2021.

[135] S. Kotz et al. *Continuous multivariate distributions, Volume 1: Models and applications*. John Wiley & Sons, 2004.

[136] K. W. Ng, G. Tian, and M. Tang. *Dirichlet and related distributions: Theory, methods and applications*. John Wiley & Sons, 2011.

[137] Y. Xiao et al. Bayesian hierarchical mechanism design for cognitive radio networks. *IEEE Journal on Selected Areas in Communications*, 33(5):986–1001, 2015.

[138] Y. Xiao et al. Bayesian reinforcement learning for energy harvesting communication systems with uncertainty. In *2015 IEEE International Conference on Communications (ICC)*, pages 5398–5403, 2015.

[139] Y. W. Teh. *Dirichlet Process*, pages 280–287. Springer US, Boston, MA, 2010.

APPENDICES

Appendix A

Expected Utility Estimation and Belief Update

A.1 Expected Utility Estimation

According to Eq. (3.5), expected utility realized by i-th agent is given by:

$$\overline{u}^i\left(C, \mathbb{T}^C\right) = E\left[u^i\left(C, \mathbb{T}^C\right)\right] = \sum_{k=1}^{2^{|C|-1}} B^i\left(\mathbf{T}_k^C\right) u^i\left(C|\mathbf{T}_k^C\right), \tag{A.1}$$

where $B^i\left(\mathbf{T}_k^C\right)$ can be expressed as:

$$B^i\left(\mathbf{T}_k^C\right) = \prod_{j \in C \setminus \{i\}} p_{ij}\left(T_x\right). \tag{A.2}$$

To clarify the presented equations, we consider an example in which agent 1, agent 2 and agent 3 make a coalition for energy trading. The expected utility realized by agent 1 can be summation over the utility of 4 possible scenarios that can be considered by agent 1 multiplied by their chance of happening. Considering that agent 1 is an agent of type fixed, then we have:

$$u^1\left(C, \mathbb{T}^C\right) = \left\{ \begin{array}{l} u^1(\{T_{fixed}, T_{fixed}, T_{fixed}\}), u^1(\{T_{fixed}, T_{fixed}, T_{moving}\}), \\ u^1(\{T_{fixed}, T_{moving}, T_{fixed}\}), u^1(\{T_{fixed}, T_{moving}, T_{moving}\}) \end{array} \right\}. \tag{A.3}$$

Consequently, $B^1\left(\mathbf{T}_1^C\right)$ can be obtained as:

$$B^1\left(\mathbf{T}_1^C\right) = p_{12}(T_{fixed})p_{13}(T_{fixed}). \tag{A.4}$$

Therefore, we can calculate the expected utility realized by agent 1 as following:

$$\begin{aligned}\overline{u}^1\left(C, \mathbf{T}_C^1\right) &= E\left(u^1\left(C, \mathbf{T}_C^1\right)\right) \\ &= \begin{array}{l} p_{12}\left(T_{fixed}\right) p_{13}\left(T_{fixed}\right) u^1(\{T_{fixed}, T_{fixed}, T_{fixed}\})+ \\ p_{12}\left(T_{fixed}\right) p_{13}\left(T_{moving}\right) u^1(\{T_{fixed}, T_{fixed}, T_{moving}\})+ \\ p_{12}\left(T_{moving}\right) p_{13}\left(T_{fixed}\right) u^1(\{T_{fixed}, T_{moving}, T_{fixed}\})+ \\ p_{12}\left(T_{moving}\right) p_{13}\left(T_{moving}\right) u^1(\{T_{fixed}, T_{moving}, T_{moving}\}) \end{array}.\end{aligned} \tag{A.5}$$

In each iteration, belief probabilities are updated. In the following, the process of belief update is explained in more detail.

A.2 Belief Update

Considering an agent j sold (or purchased) energy, then the location of agent j for the traded energy can be either of the two observations, which are denoted as O_{fixed} and O_{moving}. Consequently, agent i is interested in calculation of posterior belief $p_{ij}(T_{fixed}|O_{fixed})$ or $p_{ij}(T_{fixed}|O_{moving})$ after having this new observation. According to Bayes rule, agent i can update its belief about agent j if a fixed behavior is observed as follows:

$$p_{ij}(T_{fixed}|O_{fixed}) = \frac{p_{ij}(O_{fixed}|T_{fixed})p_{ij}(T_{fixed})}{p_{ij}(O_{fixed}|T_{fixed})p_{ij}(T_{fixed}) + p_{ij}(O_{fixed}|T_{moving})p_{ij}(T_{moving})}, \tag{A.6}$$

where $p_{ij}(T_{fixed}) = b_{ij}^{\tau_{ij}}$ is a prior belief at about the probability of type fixed. $p_{ij}(O_{fixed}|T_{fixed})$ means the chance that we have the correct observation which is equal to $1 - p_e$. Since $p_{ij}(T_{fixed}) + p_{ij}(T_{moving}) = 1$, then $p_{ij}(T_{moving}) = 1 - b_{ij}$. $p_{ij}(O_{fixed}|T_{moving})$ is the probability that a fixed behavior from an agent of type moving be observed. This can happen in two cases. The first case is when a moving agent transfers energy from charging station 1, which will be observed by other agents as a fixed behavior when the observation is error free. This case can happen with the chance $\epsilon_{ij}\left(1 - p_c\right)$. The second case is when a moving agent transfers energy from charging station 2, which is supposed to be observed by other agents as a moving behavior, but an observation error occurs and fixed behavior is observed. This case can happen with the chance $\left(1 - \epsilon_{ij}\right) p_e$. Therefore $p_{ij}(O_{fixed}|T_{moving})$ is given by:

$$p_{ij}(O_{fixed}|T_{moving}) = \epsilon_{ij}\,(1-p_c) + (1-\epsilon_{ij})\,p_e. \tag{A.7}$$

Consequently, posterior $p_{ij}(T_{fixed}|O_{fixed})$ can be obtained as

$$p_{ij}(T_{fixed}|O_{fixed}) = \frac{b_{ij}(1-p_e)}{b_{ij}(1-p_e) + (1-b_{ij})((1-\epsilon_{ij})(1-p_e) + \epsilon_{ij}p_c)} \tag{A.8}$$

A.3 Generalizing Belief Update Mechanism

In this section, we consider N charging stations. We assume that agents observe charging or discharging at any station other than station 1 as moving behavior. If the agent j is a moving agent, then the belief probability of agent i that agent j will transfer energy from charge station n is denoted by ϵ_{ij_n}, where $\sum_{n=1}^{N} \epsilon_{ij_n} = 1$. Also, we assume two sets of observation errors as follows:

> The first set of observation errors happens when charging/discharging at station 1 is observed as charging/discharging at other stations. We show these probabilities of observation error with p_{e_n} where $2 \leq n \leq N$. p_{e_n} shows the probability of observation error when charging/discharging at station 1 is observed as charging/discharging at station n. The second set of observation errors happens when charging/discharging at any station other than station 1 is observed as charging/discharging at station 1. We show these probabilities of observation error with p_{c_n}, where $2 \leq n \leq n$. p_{c_n} shows the probability of observation error when charging/discharging at station N is observed as charging/discharging at station 1.

Considering the assumptions above, the chance that we have the correct fixed observation is $p_{ij}(O_{fixed}|T_{fixed}) = 1 - \sum_{n=2}^{N} p_{e_n}$. The chance that a moving agent transfers energy from charging station 1, given it will be observed by other agents as a fixed behavior and the observation is error-free, is equal to $\epsilon_{ij_1}(1 - \sum_{n=2}^{N} p_{c_n})$. Also, the chance that a moving agent transfers energy from any station other than station 1, which is supposed to be observed by other agents as a moving behavior, but an observation error occurs and fixed behavior is observed, can be found as $\sum_{n=2}^{N} \epsilon_{ij_n} p_{e_n}$. Therefore $p_{ij}(O_{fixed}|T_{moving})$ is given by:

$$p_{ij}(O_{fixed}|T_{moving}) = \epsilon_{ij_1}(1 - \sum_{n=2}^{N} p_{c_n}) + \sum_{n=2}^{N} \epsilon_{ij_n} p_{e_n}. \tag{A.9}$$

Consequently, posterior $p_{ij}(T_{fixed}|O_{fixed})$ can be obtained as:

$$p_{ij}(T_{fixed}|O_{fixed}) = \frac{b_{ij}(1 - \sum_{n=2}^{N} p_{e_n})}{b_{ij}(1 - \sum_{n=2}^{N} p_{e_n}) + (1 - b_{ij}) \left(\epsilon_{ij_1}(1 - \sum_{n=2}^{N} p_{c_n}) + \sum_{n=2}^{N} \epsilon_{ij_n} p_{e_n} \right)}. \quad \text{(A.10)}$$

Posterior probability $p_{ij}(T_{fixed}|O_{moving})$ can be found in the same way.

Appendix B

Dirichlet Distribution for Modeling States Transition Probabilities

The agents begin the learning process by assigning a prior density function to the state transition probabilities $H(s', a, s)$. The initial belief of the agent about H is reflected by this prior function. The agent then interacts with the environment, observes different realizations of the state, and consequently updates its belief over H using Bayes' rule. As the number of observations and interactions with the environment grows, the updated belief of the agent will be more accurate, and the effect of the prior density function will be negligible. Therefore, the selection of the prior density function does not degrade the learning process in the long term. In chapter 5, Dirichlet distribution is chosen as the prior density function for the state transition function H. In the following, we first explain the Dirichlet distribution, and then we elaborate on the idea of selecting Dirichlet distribution as a prior distribution over H.

Dirichlet distribution has been widely selected as the prior distribution in statistics inference processes [135–138]. Dirichlet distribution is a continuous probability distribution with the following PDF [135]:

$$f(x_1, ..., x_K; \alpha_1, ..., \alpha_K) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \tag{B.1}$$

where $0 < x_i < 1$, $\sum_{i=1}^{K} x_i = 1$ and $\Gamma(.)$ is the gamma function. Dirichlet distribution is defined

by K parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ and formally can be denoted as $f(x_1, ..., x_K; \alpha_1, ..., \alpha_K) \sim \text{Dir}(\boldsymbol{\alpha})$.

Dirichlet distribution can be considered as the probability distribution for the parameters of multinomial distributions [139]. Since we have a limited set of discrete state transition (s', a, s) and H is the probability distribution over this limited set of outcomes, the probability distribution of discrete state transition (s', a, s) during the first t iterations can be modeled with a multinomial distribution as following:

$$(s', a, s) \,|H \sim \text{mul}(\boldsymbol{\varphi}, t), \tag{B.2}$$

where $\boldsymbol{\varphi} = \{\varphi_{(s',a,s)}\}_{s\in\mathbf{S},s'\in\mathbf{S},a\in\mathbf{A}}$ and $\varphi_{(s',a,s)}$ represent the probability of transition from state s to s' by taking action a. The Dirichlet distribution can therefore be used to model the probability distribution φ for the multinomial distributions of H as:

$$H \sim \text{Dir}(\boldsymbol{\alpha}), \tag{B.3}$$

where $\boldsymbol{\alpha} = \{\alpha_{(s',a,s)}\}_{s\in\mathbf{S},s'\in\mathbf{S},a\in\mathbf{A}}$ and

$$f(\boldsymbol{\varphi}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{|\mathbf{S}|^2|\mathbf{A}|} \alpha_i\right)}{\prod_{i=1}^{|\mathbf{S}|^2|\mathbf{A}|} \Gamma(\alpha_i)} \prod_{i=1}^{|\mathbf{S}|^2|\mathbf{A}|} \varphi_i^{\alpha_i - 1}. \tag{B.4}$$

Also, Dirichlet distribution known as the conjugate prior distribution of the multinomial distributions. This means that if a random variable follows a multinomial distribution, and the prior distribution of the multinomial parameter is distributed as a Dirichlet, then the posterior distribution is Dirichlet as well [139]. Therefore, the posterior distribution over φ is $\text{Dir}(\boldsymbol{\alpha} + \boldsymbol{\lambda}_t)$ where $\boldsymbol{\lambda}_t = \{\lambda(s', a, s)\}_{s\in\mathbf{S},s'\in\mathbf{S},a\in\mathbf{A}}$ is the number of times that transition from state s to s' occurred during past t iterations.

In our problem in chapter 5, we are interested in computing $\Pr(s_{t+1}, \boldsymbol{\alpha}_t | s_t, a_t, \boldsymbol{\alpha}_{t-1})$. Using the chain rule we can obtain:

$$\Pr(s_{t+1}, \boldsymbol{\alpha}_t | s_t, a_t, \boldsymbol{\alpha}_{t-1}) = \Pr(s_{t+1} | s_t, a_t, \boldsymbol{\alpha}_{t-1}) \Pr(\boldsymbol{\alpha}_t | s_{t+1}, s_t, a_t, \boldsymbol{\alpha}_{t-1}). \tag{B.5}$$

Considering $H \sim \text{Dir}(\boldsymbol{\alpha})$, and the fact that expected value of the Dirichlet distribution is $E(X_i) = \frac{\alpha_i}{\sum_j \alpha_j}$, we can calculate $\Pr(s_{t+1}, \boldsymbol{\alpha}_t | s_t, a_t, \boldsymbol{\alpha}_{t-1})$ as following:

$$\begin{aligned}
&\Pr\left(s_{t+1}, \boldsymbol{\alpha}_t | s_t, a_t, \boldsymbol{\alpha}_{t-1}\right) = \Pr\left(s_{t+1} | s_t, a_t, \boldsymbol{\alpha}_{t-1}\right) \Pr\left(\boldsymbol{\alpha}_t | s_{t+1}, s_t, a_t, \boldsymbol{\alpha}_{t-1}\right) \\
&= \int \Pr\left(s_{t+1} | s_t, a_t, \boldsymbol{\alpha}_{t-1}, H\right) \Pr\left(H | s_t, a_t, \boldsymbol{\alpha}_t\right) dH . \mathbf{1}\left(\boldsymbol{\alpha}_t = \boldsymbol{\lambda}_t\right) \\
&= E(T) = \frac{\lambda_t(s', a, s)}{\sum\limits_{s'' \in S} \lambda_t(s'', a, s)}
\end{aligned} \quad , \qquad \text{(B.6)}$$

where $\mathbf{1}(.)$ is the indicator function.