REINFORCEMENT LEARNING FOR MULTI-SCALE DEMAND SIDE ENERGY MANAGEMENT

HU ZE

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Electrical and Electronic Engineering

Reinforcement Learning for Multi-Scale Demand Side Energy Management

HU ZE

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

August 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

HU Ze (Name of student)

Abstract

The global energy system is undergoing a significant transition driven by climate change and global warming, largely resulting from substantial carbon emissions associated with expanding industrial production. This transition is characterized by a shift from fossil-fuel-based thermal generation toward renewable energy sources on the supply side and from centralized large-scale generation toward decentralized and distributed generation on the demand side. Consequently, there is a growing emphasis on unlocking demand-side flexibility to provide dispatchable resources for multiple uses of the grid. In this context, this thesis investigates optimal decision-making problems, particularly focusing on energy management faced by entities on the demand side within the distribution network.

In operational energy management problems, demand-side entities typically aim to minimize their energy costs by strategically adjusting load profiles and managing energy devices subject to operational constraints. The diversity and distributed nature of demand-side entities—including individual buildings, energy communities, and retail electricity markets with responsive consumers—present unique challenges in energy management that require tailored solutions rather than a universal approach. In other words, optimization of energy management at different scales emphasizes different issues: individual consumers face uncertainty in energy prices and distributed generation; community systems grapple with complexity arising from diverse energy consumption profiles and non-convex network constraints involving multiple energy types; collective participation in the retail electricity market (REM) involves strategic interactions under dynamic pricing schemes. Therefore, energy management strategies adapted to scenarios with different scales need to be developed individually. For this reason, this thesis specifically addresses multi-scale energy management problems on the demand side with multi scales to provide adaptive, scenario-specific solutions, ultimately contributing to the broader goals of energy transition and carbon emission mitigation.

Meanwhile, machine learning (ML) has become a useful and reliable technique for multiple uses, e.g., forecasting, anomaly detection, and decision-making. As one of the most popular categories of ML techniques, reinforcement learning (RL) has been gaining much attention as a decision-making tool for multiple scenarios in power systems. RL can enable the algorithm as a smart agent to learn from interactions with the environment by "trial and error" in a Markovian environment. Given the inherent uncertainties in electricity prices, energy demands, and distributed generation, these operational decision-making problems can naturally be formulated as stochastic processes and modeled as MDPs, making RL particularly suitable for automating energy management decisions on the demand side. For multi-scale demand side operation problems, RL can be implemented as a smart energy management system to optimize energy consumption decisions automatically, reducing the need for sophisticated manual calculation to lower energy costs. To make the most of RL techniques in demand-side energy management problems, this thesis thus develops novel RL algorithms specifically tailored to address multi-scale, scenario-specific objectives within demand-side decision-making contexts.

Specifically, this thesis advances the state-of-the-art by developing three novel RL algorithms tailored specifically to different scales and scenarios of demand-side energy management. At the individual building level, a forecast-enhanced RL approach is proposed to optimally dispatch integrated energy devices based on predictive models of loads, renewable generation, and prices, achieving cost reduction while satisfying multi-energy demands. At the community level, a safe RL method is introduced, enabling the Lagrangian method in the RL algorithm to reduce network constraint violations within integrated community energy systems (ICES), significantly improving operational safety. In the retail electricity market scenario, interactions between consumers and the utility are modeled as a dynamic Stackelberg game, where a novel multi-agent RL (MARL) algorithm is developed to estimate the multiple equilibria of this game, providing possible market outcomes in the REM. Finally, the three novel RL algorithms are validated by using real-world datasets and provide

superior performance to baseline approaches. The numerical results of this thesis underscore the transformative potential of the RL technique to empower energy consumers as active and efficient participants within modern energy distribution systems.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude and deep appreciation to my supervisors, Prof. Siqi Bu and Dr. Kevin K.W. Chan, for their consistent support, invaluable guidance, and endless patience throughout my PhD journey. It is my real pleasure to work with Dr. Kevin Chan and Prof. Siqi Bu. The kindness, vast knowledge, and creative insights of Dr. Kevin Chan have lightened both my research and personal growth. The constructive advice and thoughtful comments from Prof. Siqi Bu on my research papers have enriched my future career.

My PhD journey has been really memorable and fruitful thanks to the wonderful people I had the privilege to work alongside. I am particularly grateful to Dr. Ziqing Zhu, whose careful guidance and generous support have greatly influenced my research. Sincere thanks also go to Dr. Shiwei Xia, Dr. Zilin Li, Dr. Zhaoyuan Wang, Mr. Xiang Wei, and Mr. Yafeng Zhao, whose support has been invaluable both academically and personally.

I express my deep gratitude to Prof. Lang Tong and Prof. Timothy Mount for their generous mentorship and valuable insights during my unforgettable visit to Cornell University. My sincere appreciation also extends to my labmates at Cornell—Dr. Cong Chen, Dr. Xinyi Wang, Dr. Ahmed Alahmed, Mr. Minjae Jeon, Ms. Siying Li, and Ms. Valentina Norambuena—for their stimulating discussions and wonderful companionship.

My deepest appreciation and love are reserved for my beloved parents and my girlfriend, who have unconditionally supported me in pursuing my dreams.

Last but not least, I gratefully acknowledge the financial support provided by the PhD studentship and attachment fellowship awarded by The Hong Kong Polytechnic University, which made this academic endeavor possible.

Table of Contents

Lists of Figures, Tables and Acronyms

List of Figures

List of Tables

List of Acronyms

ADMM	Alternating Direction Method Of Multipliers
BaS-MAQL	Bayesian Stackelberg Game Context
BIES	Building Integrated Energy Systems
BLSTM	Bidirectional LSTM
BNE	Bayesian Nash Equilibrium
CHP	Combined Heat & Power
C-MDP	Constrained Markov Decision Process
CNN	Convolutional Neural Network
CPP	Critical Peak Pricing
DDPG	Deep Deterministic Policy Gradient
DER	Distributed Energy Resources
DG	Dispatchable Generation
DR	Demand Response
EBS	Electric Battery Systems
EB	Electric Boiler
ERB	Experience Replay Buffer
ERB	Experience Replay Buffer
ESS	Energy Storage Systems
EU	End User
EV	Electric Vehicles
FOR	Feasible Operation Region
GB	Gas Boiler
GLU	Gated Linear Units
GRN	Gated Residual Networks
GSS	Gas Storage Systems
HVAC	Heating Ventilation Air-Conditioning
ICES	Integrated Community Energy Systems

IDR	Integrated Demand Response
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MARL	Multi-Agent RL
MDP	Markov Decision Process
MEU	Multiple Energy Users
MILP	Mixed-Integer Linear Programming
ML	Machine Learning
MNC-ICES	Multi-Network Constrained Ices
MP	Markov Process
MPPT	Maximum Power Point Tracking
MPC	Model Predictive Control
MSE	Mean Square Error
MSNE	Mixed-Strategy Nash Equilibrium
NE	Nash Equilibrium
NSI	Net Solar Irradiation
OAT	Outdoor Air Temperature
PD-TD3	Primal-Dual Twin Delayed Deep Deterministic Policy Gradient
PDF	Probabilistic Distribution Function
PV	Photovoltaic
REM	Retail Electricity Market
RS	Renewable Energy Sources
RF	Rainfall
RH	Relative Humidity
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RO	Robust Optimization
RTP	Real-Time Pricing
SAC	Soft Actor-Critic

SAT	Surface Air Temperature
S-DDPG	Safe Deep Deterministic Policy Gradient
SI	Solar Irradiation
SO	Stochastic Optimization
SoC	State Of Charge
SOTA	State-Of-The-Art
SP	Stochastic Programming
SPE	Subgame Perfect Equilibrium
TES	Thermal Energy Storage
ToU	Time-of-Use
TFT	Temporal Fusion Transformer
TD3	Twin Delayed Deep Deterministic Policy Gradient
THW	Temperature-Humidity-Wind
UV	Ultraviolet
VFTC	Variable Flow Temperature Constant
VRE	Variable Renewable Energy
VSN	Variable Selection Networks
WA	Wet Appliance
WEM	Wholesale Electricity Market
WT	Wind Turbine

Chapter I
Introduction

1.1 Research Background

Since the first intermittent power system was developed to supply thousands of lamps in New York City in 1882 by Thomas Edison, the energy system has continued to evolve from the steam-powered machine era to an electricity-powered era due to electricity's fast transmission speed at the speed of light and direct usage in machines [1]. To this day, the modern power system, mainly formed by power generators, transformers, transmission lines and consumption devices, keeps growing in scale and has covered most places where human lives, even in the Arctic and on satellites [2, 3]. The power production and consumption have been dramatically growing due to population growth and the popularization of electricity-based technologies around the world, exhibiting the brightest technological developments and big-bang like prosperity that have ever been made on Earth.

However, the usage of a major source for power generation---fossil fuel---produces about 34 billion tonnes (Gt) of carbon dioxide per year, which is over 40% of energy-related carbon dioxide (CO_2) emissions, resulting in a severe greenhouse effect and thus global warming [4]. The global warming caused by the cumulative burning of fossil fuel and other carbon-emission activities has heated the atmosphere to a temperature increase of over 1°C since 1900, which has endangered the lifelong existence of human beings [5]. For this reason, people who have realized the seriousness of this problem have tried to reduce carbon emissions. One of the well-known actions is the Paris Agreement made by around 196 countries in 2015, aiming to limit global warming to 1.5°C above pre-industrial levels by 2030 and reach net zero by 2050 [6]. Even so, the temperature is still increasing, which makes it doubtful if the target of the Paris agreement will be reached and bring huge uncertainty and danger to the future of all

life on Earth. It still requires more great work in not only policy regulation but also technology development to limit the global warming effect.

From the perspective of the power system, one of the most intuitive and efficient approaches is to substitute the thermal generator consuming fossil fuel with renewable generators with almost zero carbon emission. Solar, wind, and hydro power have been greatly promoted by policy stimulation in recent years. For instance, renewable power now account for over 20% of generation in the United States [7]. Meanwhile, technological development leads to the wide and increasing installation of energy devices, including distributed energy resources (DERs), combined heat and power (CHP) units, etc., which also leads to high renewable penetration on the demand side and provides conditions for energy conversion on the demand side to meet various applications. All these implementations endow high operational flexibility for energy end users, who thus become active and crucial participants in the energy system operation. Since then, multiple operating paradigms have arisen from the demand side [8]. For instance, with the help of DERs and integrated energy devices, energy consumers can schedule their consumption plan flexibly to minimize their electricity bill or even arbitrage using energy storage systems (ESS). Consumers with high demand in various energy forms and multiple energy devices can have more space for action to further optimize their devices' operation schedules to reach higher profits. In this context, energy management on the demand side becomes a significant and meaningful topic [9].

1.2 Incentives and Literature Review

To optimize energy management on the demand side, this work aims to develop the optimal scheduling/operation methods for multi-scale entities (e.g., grid-connected building systems and community systems) and analyze the outcome of the optimal decisions in multi-agent environments of the RTP-DR problem. However, there are specific problems in different level systems at the size of single building customers, community operators, and utility companies, which are illustrated as follows.

One severe challenge for energy management on the demand side is the requirement for scenarios tailored solution rather than a uniform solution because the demand side entities in different scales face different energy management problems. In other words, optimization problems in multi-scale energy management may emphasize different issues: individual consumers face uncertainty in energy prices and distributed generation; community systems grapple with complexity arising from diverse energy consumption profiles and non-convex network constraints involving multiple energy types, while the collective participates in the retail electricity market (REM) and involve strategic interactions under dynamic pricing schemes. Therefore, energy management strategy adapted to multi-scale entities needs to be developed individually.

The most common and also the most minor scale case is a single consumer with integrated energy demand and multiple devices, which can be characterized as building-integrated energy systems (BIES) and account for about 40% of global energy use [10]. By coordinating multiple energies, including power, gas, and heat, the BIES consumer dispatches devices like CHP and ESS efficiently according to the profile demand, price, and renewable generation to obtain more abundant flexibility and achieve sufficient renewable usage. Energy management in such systems emphasizes robust operation under uncertainties from energy prices, DER production, and multi-energy demand, which is not only for self-profit-maximization but also vital to improving operational flexibility and maximizing renewable energy use in the whole energy system.

On a larger scale, a bunch of consumers who are geographically located nearby to each other can form an energy community and operate in a cooperative way for a lower energy cost. Such local energy systems can potentially contribute to the overall energy and climate objectives, helping reverse energy consumption and emissions trends worldwide [11]. Furthermore, the proliferation of distributed energy devices and energy integration of multi-energy lay a solid foundation for better cooperation in a community system for satisfying consumers' demand for both power and heat. It is necessary and pressing to increase energy efficiency and utilization with the means of energy integration in the whole system. The challenges in such systems mainly lie in the safe

operation that satisfies the operation constraints of multi-energy networks, which could be multiple and non-convex.

The energy management problem with the largest scale in this thesis is the real-time pricing (RTP) - demand response (DR) problem between an electricity retailer (utility company) and multiple consumers (end users, EUs). When the electricity retailer implements real-time pricing in REM, consumers can learn the pricing behavior of utility companies to optimize their plans, leading to a non-cooperative game [12]. This problem may not be an energy management problem in the common sense, but it involves the demand-side management in the distribution network and also the energy scheduling of consumers, indicating the great potential of demand-side flexibility and also the risk of uncertainty in consumption patterns.

Furthermore, machine learning (ML) has become a reliable and useful technique for multiple uses, e.g., forecasting, anomaly detection, and decision-making [13]. As one of the most popular categories of ML techniques, reinforcement learning (RL) has been gaining much attention as a decision-making tool for multiple scenarios in power systems. RL can enable the algorithm as a smart agent to learn from the interaction with the environment by "trial and error" with limited information [14]. For multi-scale demand side operation problems, RL can be installed as an energy management system to optimize the energy consumption decision automatically, reducing the sophisticated calculation by hand to lower energy costs. It will be especially useful in scenarios with dynamic environments with uncertain information like electricity price, renewable generation, energy demand, etc. The further implementation of these techniques can realize better use of energy on the demand side, which is a crucial part of promoting the carbon neutral career.

Overall, entities with multi-scale in the demand side may have different objectives and are subject to different sets of constraints. This makes power consumption behavior in different scales more complex, thus harder to capture. Furthermore, the collective behavior of consumers may adversely change the whole picture of the energy system. On the other hand, using state-of-the-art (SOTA) ML and RL techniques, which are

game-changers for the traditional power system problem, provide essential tools to manage the system operation on the demand side. Advancing RL application in multi-scale demand-side energy management can assist in a better understanding of the demand side behavior and is beneficial to promoting energy transition and decarbonization [15].

1.3 Primary Contributions

The work presented in this thesis contributes to several key issues of decision-making in the demand side of the power system, specifically surrounding the application of RL in multi-scale systems: (i) scheduling the multi-energy devices under uncertainties of the renewable profile and energy demand for an integrated energy consumer like BIES, (ii) achieving a safe operation subject to multi-energy network constraints for collective community consumers in Integrated Community Energy Systems (ICES), (iii) estimating the mixed-strategy Nash equilibrium for the RTP-DR problem between an electricity retailer and multiple energy consumers in REM.

The main contributions of this thesis can be summarized as follows.

1) A hybrid data-driven approach integrating TFT and SAC algorithm, TFT-SAC, is developed to schedule the day-ahead operation strategy for BIES accounting for uncertainty in renewable output and energy demands. The TFT is used to forecast the renewable generation and energy demand based on historical data, and the forecasts that are obtained are then utilized by the SAC algorithm to solve the scheduling problems. Unlike conventional black-box forecasting methods, the TFT provides interpretability through the attention mechanism, enhancing the trustworthiness of forecasting results for decision-making. Furthermore, the SAC algorithm, trained to maximize the policy entropy, can learn an operational strategy with superior robustness and generalization capabilities. The proposed TFT-SAC approach is trained and tested on a real-world dataset to validate its superior performance in reducing energy costs and computational time compared with the benchmark approaches. The generalization

performance for the learned scheduling policy and the sensitivity analysis are examined in various scenarios.

2) A novel MNC-ICES model is proposed to interpret the concept of ICES. The proposed model accounts for the constraints of multi-network, which captures the physical characteristics of energy flow and imposes security operational constraints for the distribution level energy transmissions. Energy devices are modeled in high fidelity to describe the realistic physical operating attributes in practice. Additionally, the renewable uncertainty and integrated demand elasticity are considered to describe the novel characteristics of modern distribute-level energy systems. A constrained optimization problem is formulated to denote the operation problem in the proposed MNC-ICES model and then transformed into a Constrained Markov decision process (C-MDP) for the application of RL approaches. Specifically, the C-MDP is formulated from the constrained operational optimization problem in MNC-ICES with multi-energy integration. Constraints on voltage in the power network, gas flow, gas pressure and gas injection in the gas network, pipeline flow, and nodal flow in the district heat network are considered security constraints and imposed safety requirements, being modelled as the cost term in a tuple of C-MDP.

3) A safe RL algorithm, Primal-Dual Twin Delayed Deep Deterministic Policy Gradient (PD-TD3), based on a C-MDP) is proposed to optimize the decisions of ICES operators for profits-maximization subject to multi-energy network constraints. The PD-TD3 algorithm using double networks reduces the over-estimation problem of the action value for both the reward and cost, and the delayed update stabilizes the training process of policy and its dual variable. With such an accurate estimation of Q values, the proposed algorithm converges to the optimal solution that balances the maximal profits and the lowest constraint violation. In addition, the training processes of the policy and its dual variable are stabilized by delayed updates, which contributes to the training efficiency and helps to converge to the global optimal.

4) A 1-leader, N-follower dynamic Bayesian Stackelberg game is developed to represent the sequential decision-making RTP-DR problem. This game is assumed to

be an incomplete information environment in a non-cooperative game between an electricity retailer and multiple EUs. All players learn others' strategies dynamically to maximize their own profits in certain sequential RTP-DR problems. The proposed game is then re-formulated into a MDP for reinforcement learning solutions.

5) A multi-agent RL (MARL) algorithm is developed to estimate the mixed-strategy Nash equilibrium(MSNE) of the RTP-DR problem. By solving the MDP for each player, the subgame perfect equilibrium (SPE) of the dynamic Stackelberg game is reached, and the convergence conditions are almost identical to the equilibrium conditions (No player can benefit from deviating from current decisions). Compared to typical MAQL, the proposed approach utilizes probability distributions to represent Q-values, enhancing the algorithm's learning speed and strategic depth, leading to a more accurate equilibrium point. The results show that the optimal decision trajectories of both the retailer and end users are multiple, indicating the equilibrium for the proposed game is indeed MSNE.

1.4 Thesis Layout

This thesis comprises six chapters in total, including this introductory chapter. The remaining chapters are organized as follows.

Chapter II carefully reviews the past research and critical challenges in multi-scale demand side energy management problem in terms of different scales. A review of fundamentals and advances of RL is also provided. In depth discussion in challenges in both demand side energy management problems across different scales and current RL algorithms are discussed, which are to be addressed in Chapter III, IV, and V.

Chapter III presents the application of a novel transformer-based RL model, namely TFT-SAC, in energy forecasting and afterward energy management in a BIES. The models of a modern BIES system constitute the energy devices of micro-CHP, ESS, photovoltaic (PV), gas boiler (GB), and uncertain demand of power and heat. The novel method, TFT-SAC, adopts TFT for interpretable energy forecasting and SAC for follow-up operational optimization. The proposed hybrid data-driven approach is

trained and tested on a real-world dataset to validate its superior performance in reducing energy costs and better generalization performance compared with the benchmark approaches.

Chapter IV provides an overview of the state-of-the-art concepts for techno–economic modeling of ICES by establishing a Multi-Network Constrained ICES (MNC-ICES) model. The proposed model underscores the diverse energy devices at community and consumer levels and multiple networks for power, gas, and heat in a privacy-protection manner. The corresponding operational optimization/energy management problem in the proposed model is formulated into a C-MDP and solved by a Safe RL approach. A novel Safe RL algorithm, PD-TD3, is developed to solve the C-MDP. By optimizing operations and maintaining network safety simultaneously, which is tested against benchmark approaches.

Chapter V employs MSNE to analyze the multiple equilibria in the non-convex game of the RTP-DR problem, which is considered as a combination of demand-side management problem of the retailer and energy management problem of EUs, providing a comprehensive view of the potential transaction results in REM. A novel multi-agent Q-learning algorithm is developed to estimate SPE in the proposed game. The proposed algorithm has a bi-level structure and adopts probability distributions to denote Q-values, representing the belief in environmental response. Through validation on a Northern Illinois utility dataset, the proposed approach demonstrates notable advantages over benchmark algorithms.

Finally, the concluding remarks of the thesis are summarized in Chapter VI, and some prospective extensions and possible directions for future research work are also presented.

1.5 List of Publications

Journal paper published:

1. **Z. Hu**, K. W. Chan, Z. Zhu, X. Wei, W. Zheng, and S. Bu, “Techno–Economic Modeling and Safe Operational Optimization of Multi-Network Constrained Integrated Community Energy Systems,” in *Advances in Applied Energy*, vol. 15, pp. 100183-, 2024, doi: 10.1016/j.adapen.2024.100183
2. **Z. Hu**, P. Zheng, K.W. Chan, S. Bu, Z. Zhu, X. Wei et al., "A Hybrid Data-Driven Approach Integrating Temporal Fusion Transformer and Soft Actor-Critic Algorithm for Optimal Scheduling of Building Integrated Energy Systems," in *Journal of Modern Power Systems and Clean Energy*, doi: 10.35833/MPCE. 2024.000909 (Early Access)
3. **Z. Hu**, Z. Zhu, X. Wei, K. W. Chan, S. Bu, “Mixed Strategy Nash Equilibrium Analysis in Real-Time Pricing and Demand Response for Future Smart Retail Market,” in *Applied Energy* (Accepted)
4. Z. Zhu, **Z. Hu**, K. W. Chan, S. Bu, B. Zhou, and S. Xia, “Reinforcement learning in deregulated energy market: A comprehensive review,” *Applied Energy*, vol. 329, pp. 120212-, 2023, doi: 10.1016/j.apenergy.2022.120212
5. Z. Zhu, K. W. Chan, S. Bu, **Z. Hu**, and S. Xia, “An Imitation Learning Based Algorithm Enabling Priori Knowledge Transfer in Modern Electricity Markets for Bayesian Nash Equilibrium Estimation,” *IEEE Transactions on Power Systems*, vol. 39, no. 4, pp. 5465–5478, 2024, doi: 10.1109/TPWRS.2023.3341456
6. X. Wei, K. W. Chan, G. Wang, **Z. Hu**, Z. Zhu, and X. Zhang, “Robust preventive and corrective security-constrained OPF for worst contingencies with the adoption of VPP: A safe reinforcement learning approach,” in *Applied Energy*, vol. 380, pp. 124970-, 2025, doi: 10.1016/j.apenergy.2024.124970
7. X. Wei, X. Zhang, G. Wang, **Z. Hu**, Z. Zhu, and K. W. Chan, “Online Voltage Control Strategy: Multi-Mode Based Data-Driven Approach for Active Distribution

Networks," *IEEE Transactions on Industry Applications*, vol. 61, no. 1, pp. 1569–1580, 2025, doi: 10.1109/TIA.2024.3462891

Journal paper under review or preparation:

8. **Z. Hu**, Z. Zhu, L. Zhu, X. Wei, S. Bu, K. W. Chan, "Advancing Hybrid Quantum Neural Network for Alternative Current Optimal Power Flow," submitted to *Journal of Modern Power Systems and Clean Energy* (Under Review)

Conference paper presented:

9. **Z. Hu**, K. W. Chan, S. Bu, Z. Zhu and X. Wei, "A novel peer-to-peer electricity market mechanism with the participation of electricity retailers," 12th IET International Conference on Advances in Power System Control, Operation and Management (APSCOM 2022), Hybrid Conference, Hong Kong, China, 2022, pp. 347-352, doi: 10.1049/icp.2023.0125
10. **Z. Hu**, Z. Wang, Z. Zhu, Z. Li, S. Bu and K. W. Chan, "Dispatch of Virtual Inertia and Damping via Deep Reinforcement Learning for Improving System Frequency Stability," 2024 IEEE Power & Energy Society General Meeting (PESGM), Seattle, WA, USA, 2024, pp. 1-5

Chapter II
Literature Review

2.1 Overview

In modern power systems, the demand side entities are varied in scales and also operation methods. This chapter aims to introduce the state-of-the-art multi-scale demand side energy management models. Furthermore, fundamentals and current challenges of RL techniques are also provided and discussed to cover the necessary concepts that would be used in the thesis. Firstly, energy forecasting and day-ahead optimal operation for BIES are reviewed as a single self-schedule proactive consumer in the distribution network. Secondly, safe operation and constrained optimization for an ICES containing a group of integrated energy consumers and multiple energy devices are reviewed at the community level of demand side energy management problems. In addition, the interaction and game between an electricity retailer and corresponding end users of energy are reviewed at the level of demand side management in REM. Lastly, the fundamentals of RL, which mainly include MDP and the Bellman function, are presented. The categories and challenges of RL algorithms are also briefly discussed.

2.2 State-of-the-Art Multi-Scale Demand Side Energy Management

Multi-scale demand side energy management involves dynamic decision-making problems from a single energy building to integrated energy community and the whole REM. In the following, a literature review on the models and solutions of corresponding scenario and problems are presented in detail.

2.2.1 Energy Forecasting and Day-ahead Optimal Decision Making for Building Integrated Energy System

The BIES operates to meet multiple energy demands using both internal energy devices and external energy resources. Specifically, the electric system, which comprises PV panels, micro-CHP units, and BESSs, is grid-connected to satisfy the power demands of the building. Typically, BIESs purchasess electricity from the external power market when the demand exceeds renewable generation and may sell electricity when renewable generation is surplus. The BESS enhances the operational flexibility and adds complexity to the decision-making process. PV and BESS, as components of DC systems, are connected to the building and power grid through electronic interfaces. The maximum power point tracking (MPPT) is used to control the inverter between the DC and AC systems, maximizing energy extraction from PV panels despite fluctuating solar conditions. For simplicity, the dynamics inside the power converters are neglected, as the focus is on optimizing the hourly operational strategy. Additionally, independent heating systems, consisting of micro-CHP units and GBs, are commonly deployed in building complexes, campuses, and industrial parks, particularly in regions with high heat demands [16]. These localized heating systems reduce the significant transmission losses associated with centralized heating. The BIES model also assumes a connection to an external natural gas market as the fuel source for the micro-CHP units. Detailed models of these devices are provided as follows. The energy management of BIES is hindered by two key challenges: 1) high operational risk due to the intermittent and uncertain nature of PV power generation and energy demand [17], and 2) intractable optimization caused by the non-convexity of the CHP unit [18]. For the former, PV generation and demand uncertainty has been shown to bring significant profit loss and endanger the system stability by leading to energy shortage or renewable curtailment [19]. The problem even gets more severe in large buildings with high peak demands or high solar capacity. Accurate forecasting of PV output and demand is thus crucial for smart scheduling in energy devices (e.g., energy storage) to avoid profit loss and system blackout. Much of the existing research has focused on developing model-based frameworks for optimal operation in multi-carrier energy systems. These optimization problems generally rely on precise models and

estimated exogenous factors such as weather-dependent renewable generation and energy loads. To address uncertainties, techniques like robust optimization (RO) and stochastic optimization (SO) have been used, where RO models uncertainties as bounded sets, and SO uses a set of scenarios to represent uncertainty. While these conventional methods are effective for managing the scheduling of multi-carrier systems, they face challenges in handling highly nonlinear units, particularly in competitive markets. Stochastic programming (SP) becomes inefficient as the number of scenarios increases, and RO often yields overly conservative results by focusing on worst-case scenarios. Both SP and RO also suffer from the "curse of dimensionality," where increased actions, decision variables, and constraints lead to exponentially growing computational requirements, limiting their scalability for real-world energy management applications involving multiple devices and uncertainties [20].

As for the latter, CHP is well known for providing flexibility in power and heat in a feasible operation region (FOR), which is non-convex in practice and makes the optimization non-tractable. FOR convexification is a widely adopted solution but sacrifices considerable operational flexibility [21]. The optimal scheduling of CHP remains an open question in BIES optimal operation. Moreover, the variable renewable/demand forecast and non-convex operation optimization are not independent of each other, e.g., the flexible dispatch of CHP can provide compensation for the renewable uncertainty. This indicates a deep correlation between the forecast and downstream non-convex scheduling in BIES.

2.2.2 Constrained Optimization in Grid-Connected Integrated Community Energy Systems

ICES have emerged as a promising approach for efficient multi-energy coordination and utilization, particularly in managing demand flexibility and increasing renewable energy penetration [22]. The energy management of ICES is, therefore, essential for integrating diverse energy transactions and enhancing overall energy efficiency. However, the concept of ICES, which represents an integrated energy system at the

community level, is still under discussion and lacks a clear definition. Some researchers have described ICES as a modern development that reorganizes local energy systems to integrate distributed energy resources and engage local communities [11]. Others have focused on its role in managing local energy generation, delivery, and exchange to meet local demand, with or without grid connection [23]. However, these descriptions do not fully capture the operational logic and model structure of ICES. Inspired by the concept of energy communities [24, 25], *ICES* is defined as follows: ICES is a socio-economic unit rooted in a physical community, characterized by cooperative multi-energy production and consumption through either shared or unshared integrated energy devices, and functioning as a non-commercial market actor that amalgamates economic, environmental, and social community objectives. While sharing the goal of maximizing social welfare through energy device scheduling and demand response stimulation, ICES extends beyond electric energy to include the integration of power, gas, and heat, emphasizing coordination among both energy devices and demands. Thus, ICES represents an effective strategy for maximizing social welfare and facilitating decarbonization.

Based on a review of previous studies, the modeling of ICES can be divided into three main components: community-level devices, consumer-level devices, and network constraints. Devices in ICES can be divided into community-level and consumer-level. Community-level devices typically include dispatchable generation (DG) units, ESS, and renewable energy sources (RES). DG units comprise CHP systems [26-30], power-only units [26, 28, 31], and heat-only units [26, 28, 30, 31]. CHP systems, which serve as critical energy converters across power, gas, and heat, are modeled simplistically with fixed energy conversion rates in most works of ICES [26, 28, 30, 31]. However, the realistic and physical characteristics of CHP are always overlooked, which describes the multi-energy conversion as a FOR but presents computational challenges with non-convexity[32]. Power-only and heat-only units are rarely used, which are typically modeled using a linear [26-30] or quadratic generation cost function [31, 33]. For ESS, [34-36]electric battery systems (EBS) [25-28, 34-36],

thermal energy storage (TES)[25, 34-36], and gas storage systems (GSS) are considered. Compared to prevalent EBS and TES, which have variable costs, GSS is less common due to static gas prices. Typical simplified ESS models are usually employed with static value for charging and discharging efficiency. This is because ESS does not directly participate in multi-energy conversion and is the core part of the ICES, although ESS is deemed necessary. RES, such as PV systems and wind turbines (WT), introduce renewable power output with uncertainty, constituted of an energy conversion model and forecast errors. The energy conversion model provides the output given the solar irradiation or wind speed and other external conditions (e.g., temperature [36]), while the forecast errors are sampled from specific probabilistic distribution functions (PDFs). For example, Weibull and Beta PDFs can be used to represent the forecast error distribution for the WT and PV, respectively [34]. At the consumer level, the modeling focuses on the energy demand for electricity and heat, typically based on a quadratic energy utility function to represent demand response characteristics. Households may possess energy conversion or flexible devices like micro-CHP, ESS, and boilers, sometimes overlapping with community-level devices. The boilers employed on the demand side enable energy conversion to realize a more flexible integrated demand response (IDR) [30]. Moreover, some studies have extended ICES modeling to include more detailed flexible devices, for example, electric vehicles (EVs) [27], to explore the unique characteristics of ICES in various scenarios. Additionally, as research accounting for network constraints in ICES is very limited, previous works in network modeling are reviewed in the following operational research part rather than separately.

The existing literature on the energy management of ICES primarily addresses the coordination of two energy systems, a focus partly due to the complexity and computational intensity involved. For power and heat systems within ICES, prior research has primarily concentrated on leveraging thermal demand characteristics, given their direct impact on human comfort. IDR strategies for power and heat have been employed to manage uncertainty and enhance profitability without compromising comfort levels. For instance, in [27], the coordination between flexible IDR for power

and heat and electric vehicle charging stations is explored in ICES under the uncertainty of renewable generation. A bi-level model predictive control (MPC) based approach is utilized in [29] to optimally integrate thermal demand and flow dynamics into the ICES scheduling problem. [30] optimizes the distributed scheduling problem of multiple energy hubs in ICES. Furthermore, [37] considers the impact of the thermal inertia of detailed space heating loads to model the thermal demand response character in the IDR problem of multiple energy users (MEUs) in ICES. On the other hand, literature on power and gas systems in ICES is limited due to the non-convex nature of gas flow, focusing on the coordination of energy flow in distribution networks. While research on two-network coupling systems is extensive, the coordination and interaction among multiple networks are still underexplored. Notably, comprehensive modeling and mathematical optimization of multi-networks are proposed in a multi-energy district [38], which shares a similar scale with ICES. However, multi-energy districts primarily concentrate on network operations without addressing energy device scheduling and are considered centrally controlled entities, in contrast to the community-oriented nature of ICES. As a result, the multi-network constrained scheduling of ICES operators and the interaction (e.g., IDR) between ICES operators and MEUs remain critical yet underexplored aspects—the modeling and operational optimization of multi-network constrained ICES warrants further investigation.

The constrained optimization/energy management problem in ICES is challenging to solve in terms of non-convexity, privacy protection, and computational burden, which are caused by non-convex constraints of devices and network, the distributed operation manner of MEUs, and the increasing scale of the modern community, respectively. It can be solved by multiple approaches, including heuristic algorithms [27] and mathematical programming [28, 29, 37]. Heuristic algorithms are a class of optimization algorithms that are designed to explore solution spaces to find near-optimal solutions efficiently. Therefore, this approach is particularly useful for problems with non-convexity. A metaheuristic algorithm, chaotic differential evolution, is adopted to schedule and price for multiple ICESs in [27]. However, it fails to

guarantee optimality theoretically and is easy to fall into suboptimal, especially in large-scale and complex problems like ICES operation. In contrast, mathematical programming guarantees the solution optimality with rigorous proof but falls short in dealing with non-convexity, which requires complicated convexification. Works in [28] formulate a convex problem by employing the Big-M method and inequality second-order cone constraint, after which active and reactive dispatching for ICES is solved to the global optimal. Similarly, network constraint nonlinearity was tackled using the big-M method and piece-wise linearization [31]. The two-stage optimization of the power and heat system in ICES is then solved by a robust method subject to energy price uncertainty. However, these approaches above require global information for solutions, violating the privacy protection of MEUs in an ICES. To overcome this drawback, the alternating direction method of multipliers (ADMM), was adopted to schedule the sub-systems of ICES in a decentralized manner [37]. Even though these approaches can partially deal with non-convexity and realize privacy protection, they still face increasing computational burdens with the growing scalability of the consumers and devices, which is known as the "curse of dimensionality [20]."

2.2.3 RTP-DR problem between utility companies and energy consumers

The practical utilization of RTP- DR by retailers and end users (EUs) offers a viable approach for enhancing grid efficiency and providing flexibility resources [39]. RTP, which continuously adjusts the price of electricity based on current supply and demand conditions [40], is a popular dynamic pricing approach employed by electricity retailers. Compared to other dynamic pricing programs such as time-of-use (ToU) and critical peak pricing (CPP), RTP demonstrates superior performance in reflecting the intrinsic value of electricity across different transaction periods, thereby improving market intelligence and efficiency [41]. EUs, in response to the fluctuating price signals in REM, intelligently adjust their power consumption among various transaction time slots based on appliance utility and electricity prices. This behavior, known as "price-based demand response", effectively utilizes the inherent flexibility of power

appliances in EU households [42]. Despite the intuitive conflict of interest between retailers and EUs, the RTP-DR mechanism exploits the flexibility of EUs' power consumption and facilitates its supply to the grid, thus ensuring incentive compatibility within the market [43]. Furthermore, RTP-DR results in direct energy savings for consumers and empowers them to manage their energy usage more effectively, leading to more informed and active participation in the energy market.

Research on RTP-DR can be divided into two primary categories: optimal strategy development for retailers and aggregators, and market equilibrium estimation.

The first category involves developing mathematical models and algorithms to optimize pricing strategies, including ToU [44], CPP [45], and learning-based methods. In ToU, the retailer segments the day into peak, off-peak, and mid-peak hours with predetermined rates, enabling EUs to plan their consumption. Under CPP, prices are significantly increased during pre-announced critical periods to reduce peak demand, while lower rates apply during non-critical periods. Learning-based methods optimize pricing in dynamic environments with uncertain demand and wholesale prices, leveraging preference modeling or model-free deep reinforcement learning. For the EUs side, Load-shedding and load-shifting are the main methods of EUs to change their load profile, while they may have certain strategies under different RTP schemes. Under ToU, EUs optimize consumption by shifting usage to lower-cost periods in day-ahead scheduling. For CPP, EUs avoid consumption during critical periods to minimize costs. When learning-based pricing is used, EUs solve real-time multi-device operation problems considering demand characteristics and price uncertainties [46].

For the latter category of market equilibrium estimation, the Stackelberg game is a popular means to construct a bi-level decision-making model, and features sequential interactions between a single leader and multiple followers [44]. In recent research, it has been widely adopted to model RTP-DR problems in various scenarios [47], [48], [49], [50], [51]. A key challenge in the Stackelberg game modeling fitted in the RTP-DR problem is to incorporate network constraints. Previous research, such as [47] and [48], usually overlooked physical constraints due to computational complexity and

temporal correlation of EUs' power consumption characteristics for simplicity. However, this results in over-ideal solutions that did not comply with network constraints and thus did not ensure the safe operation of the distribution network [49]. What's worse, neglecting the temporal correlation of EUs' power consumption eliminated one of the most important functions of DR, i.e., peak shaving and valley filling among time slots [52]. The cross-impact of network constraints and temporal-related non-linear power consumption characteristics may render the RTP-DR problem non-convex [50]. As a result, the market equilibrium may deviate from the unique Nash Equilibrium (NE), and result in multiple equilibria in the game [51]. Reference [49] claimed the problem formulated from the aforementioned game was non-convex and NP-hard, so commercial solvers could not find a good solution. To date, the non-convex RTP-DR problem with multiple equilibria remains unsolved.

Nonetheless, the non-convex Stackelberg game has multiple equilibria in nature and is suitable to be analyzed from the view of MSNE. Compared to pure-strategy NE, MSNE is a set of probability distributions on several possible local equilibria. The adoption of MSNE allows players to randomize their strategy in a probability distribution. There are several underlying reasons to implement the mixed strategy and MSNE. 1) The non-convex game may have more than one NE especially when there are non-linear temporal-correlated power consumption constraints and network constraints [53]. 2) The nature of mixed strategy complies with the uncertain action of both the human and learning algorithm [54]. 3) The strategic behavior of players in a game with multiple equilibria should be various and stochastic. 4) A probability distribution over several equilibria should exist accounting for optimal stochastic (mixed) strategies and multiple equilibria [55]. Additionally, the analysis of MSNE is able to provide a more comprehensive understanding of potential strategic behaviors in non-cooperative games. Especially, when solving the RTP-DR problem, social welfare over several equilibria may vary significantly, thus the presence of MSNE can guide further regulation development. Although the works in [56] exploit multiple-equilibria to indicate the market equilibrium accounting for the demand uncertainty, there is no

current research focusing on the multiple equilibria in the non-convex Stackelberg game that results from the physical constraints and non-linear power demand.

In most works above, mathematical programming methods are employed to estimate NE in the Stackelberg game. For instance, the bi-level Stackelberg game has been transformed into mathematical programming with equilibrium constraints and solved using traditional mathematical methods after being reformulated to a linear problem [57]. In [50], a network-constrained Stackelberg game is solved centrally for the optimal prices and demand in the RTP-DR problem. However, with the increasing scale of the problem, the huge burden on convexification and the possible exponentially growing demand for the computation resources, which is called the "curse of dimensionality," put significant barriers to implementing the mathematical methods [58].

2.3 Reinforcement Learning Algorithms

In this subsection, a brief review of RL fundamentals is provided, covering necessary concepts and algorithms that will be further employed while elaborating RL applications on marketized power systems in subsequent sections. The MDP as the most simplified formulation of RL is introduced, and RL algorithms are reviewed and discussed.

2.3.1 The Concept of Markov Decision Process

To start with, the concept of Markov Property is introduced as a foundation of Markov Process (MP) and MDP. The Markov Property refers to the conditional probability of s_{t+1} occurring given that s_t has already occurred, being independent to the previous states from s_{t-1} and beforehand. Intuitively, the MP is defined as the state sequence (a random process) with such a property. The MP is always formulated as a two-tuple (S, P), where S is a set of states with Markov Property, and P denotes transition probabilities among states. Consider a smart agent that can take different

actions given different states. Similarly, the MDP can be formulated as a tuple (S, A, R, P), considered as a MP with the incorporation of actions and rewards, in which:

- S is a set of states that contains environment information related to the decision-making at this time.
- A is a set of actions that can be taken by the agent at the corresponding state.
- P represents the transition probability, denoted as: $P: S \times A \rightarrow (S)$, which is a probability distribution over the set S.
- $R(S, A)$ indicates the reward of selecting action a in the previous state s.

Once the MDP is observed (assume the MDP is fully-observable, while the condition of partially-observable MDP also exists), the aim of agents is to obtain the optimal policy π^*, which refers to the sequential decisions [33], for maximizing accumulative rewards in the MDP. However, the "performance" of a given policy cannot be simply evaluated by the immediate reward after the action for the sake of long-term benefits. Instead, the state-value function as formulated in (2.1), which is the so-called "expected accumulative reward", is adopted to evaluate the performance of the policy π at the state s until the termination of the entire episode. When evaluating a policy, the mapping from actions to states (state transition) is still uncertain (see the definition of transition probability). Intuitively, the action-value function, which implies the expected accumulated reward by executing the policy π after taking action a at the state s, can be defined in (2.1). Note that the discount factor γ is intended to discount the future reward indicating the uncertainty.

$$V_\pi(\mathrm{s}) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{k+t+1} | S_t = s\right] \tag{2.1}$$

$$q_\pi(s, a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{k+t+1} | S_t = s, A_t = a\right] \tag{2.2}$$

Furthermore, (2.2) indicates that the action-value constitutes two parts: the immediate reward, and the sum of possible state-values at s_{t+1} weighted by their probabilities. Inspired from (2.2), the "optimal" policy can be induced by taking the action with maximum expected reward in an iterative manner once the action-value

function is available. By substituting $G_{t+1} = V_\pi(s_{t+1})$, the iterative expression of $V_\pi(s)$, which is the so-called Bellman Equation, can be written as:

$$V_\pi(\mathrm{s}_t) = E_\pi[R_{t+1} + \gamma V_\pi(\mathrm{s}_{t+1})] \tag{2.3}$$

$$q_\pi(s_t, a_t) = E_\pi\left[R_{t+1} + \gamma q_\pi\left(\mathrm{s}_{t+1}, \mathrm{a}_{t+1}\right)\right] \tag{2.4}$$

By definition, the optimal state-value function and action-value function are the maximum values of $V_\pi(s)$ and $q_\pi(s,a)$, i.e., $V(s) = max\, V_\pi(s)$ and $q^*(s,a) = \max q_\pi(s,a)$. Then, the Bellman Equation of the optimal state-value function and action-value function can be formulated as follows.

$$V^*(s_t) = \max_a R_{s_t}^{a_t} + \gamma \sum P_{s_t s_{t+1}}^{a_t} \max\nolimits_{a_t} V^*(s_{t+1}) \tag{2.5}$$

$$q^*(s_t, a_t) = R_{s_t}^{a_t} + \gamma \sum P_{s_t s_{t+1}}^{a_t} \max\nolimits_a q^*(s_{t+1}, a_{t+1}) \tag{2.6}$$

With (2.6) in hand, the optimal policy can be intuitively derived by maximizing the action-value $q^*(s,a)$ as formulated in (2.7), while such procedure is implemented with assistance of most useful techniques including DP and RL algorithms.

$$\pi^*(a|s) = \begin{cases} 1, if & a = \mathrm{argmax} q^*(s,a) \\ 0\ , & o.w. \end{cases} \tag{2.7}$$

2.3.2 Reinforcement Learning Algorithms

In recent years, RL algorithms have gained great attention for addressing optimization problems [59]. By interacting with the external environment, RL algorithms enable intelligent agents to iteratively learn optimal strategies with only partial environmental information. Compared to traditional mathematical programming, RL offers advantages in scalability with high computational efficiency and generalization to various scenarios [60].

As illustrated in the previous subsection, an RL algorithm learns a better strategy by making decision based on the current observation and update the decision-making strategy with the received reward. Based on whether an environment model is learnt, RL algorithms can be roughly divided into two categories, model-based RL and model-free RL. Specifically, model-based RL always requires modeling the transition probability and the reward function, while model-free RL does not learn an explicit environment model, but learns all the environment implicitly within the strategy. Since

the tasks in power systems always need to receive highly uncertainty or high-order signals, works in this thesis mainly focus on the model-free environment to better adapt to the complex scenarios in power systems.

By leveraging DNN to estimate value functions, model-free RL algorithms can handle complicated optimization problems by estimating non-convex Q-functions or policies. DRL has been successfully applied in diverse domains [61-69]. For example, a model-free DRL algorithm, DDPG, optimizes the energy management of an integrated energy hub in [32]. Similarly, SAC algorithms optimize the scheduling of islanded energy systems, accounting for multi-uncertainties and hydrothermal simultaneous transmission [62]. RL algorithms can also be extended to multi-agent environment in peer-to-peer multi-energy trading [70, 71], showing their strong adaptability to different settings.

However, conventional RL algorithms suffer from several challenges.

1) In most real-world decision-making problems, the RL agent makes sequential decisions based on observed state information. However, whole state information is always not fully observed and may require forecasting, for example, power system dispatching orders may require forecast on renewable and load. Although RL algorithms can learn from current state to make decisions, there is no explicit forecasting procedure in the design of RL algorithms, resulting in a poor ability to deal with future uncertainties.

2) RL algorithms are designed for unconstrained optimization problems. Even though they are applicable to some optimization problems with soft constraints, their efficacy may diminish when applied to most constrained optimization problems. The lack of consideration for network constraint violations restricts the application of RL algorithms in industrial practice, as it can lead to economic losses and even system blackouts [63].

To solve the first challenge, some literature has tended to integrate decision-making with upstream forecasting for a holistic data-driven tool for scheduling in integrated energy systems. For instance, [72] adopted a long short-term memory (LSTM) method

to extract temporal features and assist the decision-making of the DRL algorithm in integrated energy management. [32] combined a convolutional neural network (CNN) and bidirectional LSTM (BLSTM) to forecast solar output in an energy hub by analyzing sky images. The predicted value is then imported into the DDPG algorithm for further scheduling decision-making. Although these methods have shown good performance, the LSTM struggles with capturing complex temporal patterns and dependencies that span multiple time steps effectively [73], and related research is still limited.

To address the latter challenges mentioned above, Safe RL algorithms have been developed to solve constrained optimization problems, which are designed to maximize reward while complying with hard constraints. Specifically, Safe RL approaches can be classified into three categories: 1) Penalizing constraint violations in the reward function by adding a penalty term [64]. However, this requires choosing a suitable penalty value, which is a difficult and sensitive task that depends on the reward scale, the number and scale of constraints, and the degree of safety [65]. 2) Projecting unsafe actions to safe ones by solving a projection problem, for example, approximated Lyapunov constraints [66]. This method relies on a projection model, which is based on predefined DNNs or matrices with potentially large approximation errors [67]. Therefore, the resulting actions could be overly conservative. 3) Penalizing the constraint violations in the action-value function dynamically by introducing the Lagrangian multiplier, instead of using a fixed penalty value in the reward [68, 69]. The multiplier is stochastically updated as a dual variable of the policy during the agent training based on the cost value function. However, the Lagrangian method-based Safe RL algorithm may not converge to the optimal solution because of the cost value function overestimation.

Chapter III
A Forecasted-Enhanced Reinforcement Learning Method for Optimal Scheduling of Building Integrated Energy Systems

3.1 Overview

This chapter focuses on the scheduling/energy management problem of BIES that suffers from uncertainty of DER and energy demands, and also complex operation characteristics of integrated energy devices. In the context of energy management problems in BIES, the RL algorithm is one of the promising candidates, which learns from historical data and receives available environment information to make operational decisions. Such scheduling is based on day-ahead/hour-ahead prediction for required variables, including renewable output, energy demand, etc. Although RLs can learn from the current state to make decisions, there is no explicit forecasting procedure in the design of RL algorithms, resulting in a poor ability to deal with future uncertainties. Integrating decision-making with upstream forecasting for a holistic operational tool is a natural idea to improve operational efficiency. Recently, some literature has tended to integrate decision-making with upstream forecasting for a holistic data-driven tool for scheduling in integrated energy systems. For instance, [72] adopted a long short-term memory (LSTM) method to extract temporal features and assist the decision-making of the DRL algorithm in integrated energy management. [32] combined a convolutional neural network (CNN) and bidirectional LSTM (BLSTM) to forecast solar output in an energy hub by analyzing sky images. The predicted value is then imported into the deep deterministic policy gradient (DDPG) algorithm for further scheduling decision-making. Although these methods have shown good performance, the LSTM struggles with capturing complex temporal patterns and dependencies that span multiple time steps effectively[32], and related research is still limited. Many

studies employ DRL techniques in conjunction with black-box forecasting tools, raising concerns about model transparency and reliability. The opacity of these models can lead to significant profit losses [74], thereby limiting the real-world applicability of data-driven strategies.

In order to deal with the decision-making of BIES under uncertainty, a hybrid data-driven method for forecast-enhanced reinforcement learning is developed, in which a temporal fusion transformer (TFT) model performs time-series forecasting of uncertain DER output and energy demands while a soft actor-critic (SAC) learns the optimal strategy at the downstream. The optimal scheduling problem of BIES is formulated as a MDP for the solution of the SAC algorithm. Finally, the forecasting accuracy, generalization performance, robustness to exogenous uncertainty, and sensitivity to external signals are analyzed, validating the applicability and advancement of the proposed approach.

Differing from the previous literature in model and methodology, the main contributions of this chapter are highlighted as follows:

1) *System Modeling and Markov Decision Process Formulation*: This chapter presents a detailed mathematical model for BIES, including micro-CHP unit, BESSs, PV panels, and gas boilers (GBs). The non-convex scheduling/energy management problem in BIES is formulated into an optimization problem and then reformulated into an MDP for the application of RL algorithms.

2) *A Hybrid Data-Driven Method for Forecasted-enhanced Reinforcement Learning*: A hybrid data-driven approach integrating TFT and SAC algorithm, namely TFT-SAC approach, is proposed to tackle the non-convex operational optimization problem in BIES. The TFT is used to forecast the renewable generation and energy demand based on historical data, and the obtained forecasts are then utilized by the SAC algorithm to solve the scheduling problems. Unlike conventional black-box forecasting methods, the TFT provides interpretability through the attention mechanism, enhancing the trustworthiness of forecasting results for decision-making. Furthermore, the SAC

algorithm, trained to maximize the policy entropy, can learn an operational strategy with superior robustness and generalization capabilities.

3) *Algorithm Validation and Optimal Scheduling Analysis*: The proposed TFT-SAC approach is trained and tested on a real-world dataset to validate its superior performance in reducing the energy cost and computational time compared with the benchmark approaches. The generalization performance for the learned scheduling policy and the sensitivity analysis are examined in various scenarios.

The remainder of this chapter is organized as follows. Section 3.2 covers system description, the optimization problem, and MDP formulation. Section 3.3 introduces the proposed hybrid data-driven approach integrating TFT and SAC algorithm. Section 3.4 validates the proposed TFT-SAC approach with simulations, and Section 3.5 concludes this chapter

3.2 Problem Formulation

3.2.1 System Description

Work in this chapter focuses on a modern BIES that encompasses grid-connected electric systems and independent heating systems, as illustrated in Fig. 3.1. In practice, such systems can be found in university campuses, residential complexes, and industrial parks. The BIES operates to meet multiple energy demands using both internal energy devices and external energy resources. Specifically, the electric system, which comprises PV panels, micro-CHP unit, and BESSs, is grid-connected to satisfy the power demands of the building. Typically, BIESs purchases electricity from the external power market when the demand exceeds renewable generation and may sell electricity when renewable generation is surplus. The BESS enhances the operational flexibility and adds complexity to the decision-making process. PV and BESS, as components of DC systems, are connected to the building and power grid through electronic interfaces. For the purposes of this study, the dynamics inside the power converters are neglected, as the focus is on optimizing the hourly operational strategy.

Additionally, independent heating systems, consisting of micro-CHP units and GBs, are commonly deployed in building complexes, campuses, and industrial parks, particularly in regions with high heat demands (e.g., most of North America and northern China). These localized heating systems reduce the significant transmission losses associated with centralized heating. The BIES model also assumes a connection to an external natural gas market as the fuel source for the micro-CHP units.

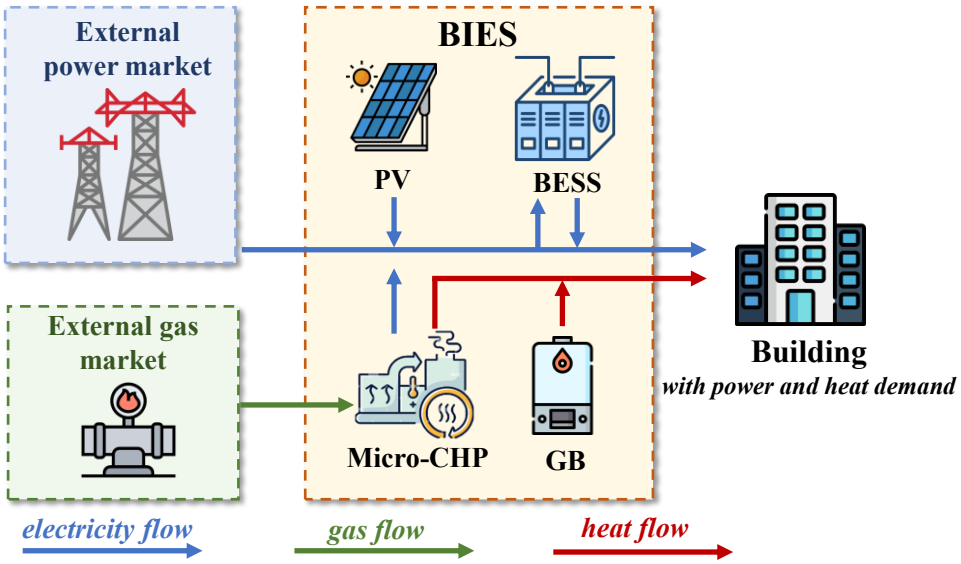


Fig. 3.1 Illustration of BIES systems

3.2.2 Device modeling

1) Micro-CHP Unit Modeling

The micro-CHP unit is a crucial component of BIESs, functioning as a single-input multi-output energy converter. It is highly efficient in converting natural gas to power and heat, and a key element in enhancing the energy efficiency of the system. Typically, the micro-CHP unit is modeled with constant energy conversion efficiencies for both power and heat. However, the generation of heat and power by micro-CHP units is interdependent, resulting in a feasible operating region (FOR). In this section, a non-convex operational model is employed for the micro-CHP unit. The non-convex FOR of this model is depicted in Fig. 3.2, bounded by the curve ABCDEFG. This FOR is considered to comprise two convex subregions, labeled as I and II.

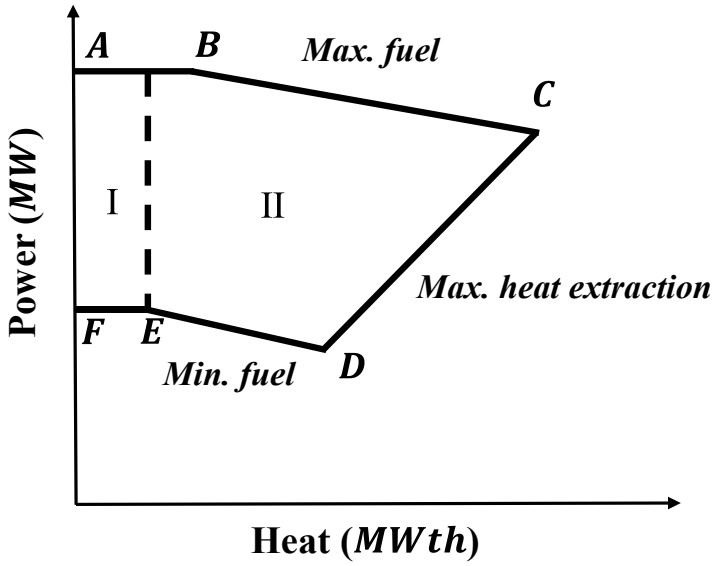


Fig. 3.2 FOR of micro-CHP unit.

The mathematical representation of the FORs for the micro-CHP unit is given by (3.1)-(3.8), as detailed in [32].

$$\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{B}} - \frac{\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{B}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{C}}}{\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{B}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{C}}} \times \left(\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{B}}\right) \le 0, \forall t \in T \tag{3.1}$$

$$\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{C}} - \frac{\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{C}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{D}}}{\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{C}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{D}}} \times \left(\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{C}}\right) \le 0, \forall t \in T \tag{3.2}$$

$$-\left(1 - \overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma \le$$

$$\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{E}} - \frac{\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{E}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{F}}}{\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{E}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{F}}} \times \left(\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{E}}\right), \forall t \in T \tag{3.3}$$

$$-\left(1 - \underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma \le$$

$$\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{D}} - \frac{\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{D}} - \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{E}}}{\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{D}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{E}}} \times \left(\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{D}}\right), \forall t \in T \tag{3.4}$$

$$\overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}} + \underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}} = \mathrm{I}_{\mathrm{CHP}}^{\mathrm{t}}, \forall t \in T \tag{3.5}$$

$$-\left(1 - \underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma \le \mathrm{H}_{\mathrm{CHP.h}}^{\mathrm{t}} - \mathrm{H}_{\mathrm{CHP.h}}^{\mathrm{E}} \le \left(1 - \overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma, \forall t \in T \tag{3.6}$$

$$0 \le \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{t}} \le \mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{A}} \times \mathrm{I}_{\mathrm{CHP}}^{\mathrm{t}}, \forall t \in T \tag{3.7}$$

$$0 \le \mathrm{H}_{\mathrm{CHP.h}}^{\mathrm{t}} \le \mathrm{H}_{\mathrm{CHP.h}}^{\mathrm{A}} \times \mathrm{I}_{\mathrm{CHP}}^{\mathrm{t}}, \forall t \in T \tag{3.8}$$

where $\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{t}}$ presents the output power of micro-CHP unit at time t, and $\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{t}}$ represent the output heat. $\mathrm{P}_{\mathrm{CHP.e}}^{\mathrm{A}}$ and $\mathrm{P}_{\mathrm{CHP.h}}^{\mathrm{A}}$ are the generated power and heat of the micro-CHP at point *A*, those at other points *B, C, D, E*, and *F* similarly defined; $\overline{\mathrm{X}}$ and $\underline{X}$ are the operating statuses in the convex subregions I and II, respectively: If the micro-CHP unit operates in the convex subregion *I*, $\overline{\mathrm{X}} = 1$ and $\underline{X} = 0$; otherwise, $\underline{X} = 1$ and

$\overline{X} = 0$; Γ is a sufficiently large number used to assist in the model description; and I^t_{CHP} is the commitment status of the micro-CHP unit. $T = \{1, \dots, 24\}$ is the set of operational hours.

The total operation cost of the micro-CHP unit at time t is expressed as:

$$C^t_{CHP}(P^t_{CHP.e}, P^t_{CHP.h}) = \overline{\alpha}_{CHP}{P^t_{CHP.e}}^2 + \overline{\beta}_{CHP}P^t_{CHP.e} + \overline{\gamma}_{CHP} + \underline{\alpha}_{CHP}{P^t_{CHP.h}}^2 + \underline{\beta}_{CHP}P^t_{CHP.h} + \underline{\gamma}_{CHP}P^t_{CHP.e}P^t_{CHP.h} \quad (3.9)$$

where $\overline{\alpha}_{CHP}$, $\underline{\alpha}_{CHP}$, $\overline{\beta}_{CHP}$, $\underline{\beta}_{CHP}$, $\overline{\gamma}_{CHP}$, and $\underline{\gamma}_{CHP}$ are the cost coefficients.

2) BESS Modeling

The BESS is conceptualized as a battery capable of charging and discharging with distinct efficiencies. The operational strategy of the BESS is designed with a granularity of one hour, corresponding to one time slot. This means that all charging and discharging activities of the BESS within a time period are aggregated into a single operation. Consequently, the BESS can either charge or discharge in any given time slot, but not both simultaneously [75].

$$E^t_{BESS} = (1-\beta)E^{t-1}_{BESS} + P^t_{BESS.c}\eta_{BESS.c} - P^t_{BESS.d} \quad (3.10)$$

$$0 \le P^t_{BESS.c} \le S^t_{BESS.c}P_{BESS.c.max} \quad (3.11)$$

$$0 \le P^t_{BESS.d} \le S^t_{BESS.d}P_{BESS.d.max} \quad (3.12)$$

$$S^t_{BESS.c} + S^t_{BESS.d} \le 1 \quad (3.13)$$

$$E_{BESS.min} \le E^t_{BESS} \le E_{BESS.max} \quad (3.14)$$

where E^t_{BESS} is the state of charge (SoC) of BESS at time t; β and $\eta_{BESS.c}$ are the predetermined loss factor and charging efficiency, respectively; $P^t_{BESS.c}$ and $P^t_{BESS.d}$ are the charging power and discharging power of BESS at time t, respectively; $S^t_{BESS.c}$ and $S^t_{BESS.d}$ are the charging state and discharging state of BESS at time t, respectively; and the subscripts max and min represent the maximum and minimum value of corresponding variables, respectively.

The SoC is calculated in (3.10). The charging power and discharge power of BESS are constrained by (3.11) and (3.12), respectively. Constraint (3.13) is employed to determine the charging or discharging state of BESS. The total capacity of BESS is constrained by (3.14).

3) GB Modeling

The GB is modelled as an energy device transforming natural gas to heat with a fixed rate. The model of GB can be described as:

$$P^t_{\mathrm{GB}.h} = \eta_{\mathrm{GB}} P^t_{\mathrm{GB}.g} \tag{2.15}$$

$$P_{\mathrm{GB}.g.min} \le P^t_{\mathrm{GB}.g} \le P_{\mathrm{GB}.g.max} \tag{2.16}$$

$$P_{\mathrm{GB}.h.min} \le P^t_{\mathrm{GB}.h} \le P_{\mathrm{GB}.h.max} \tag{2.17}$$

where η_{GB} is the natural gas conversion efficiency; $P^t_{\mathrm{GB}.g}$ is the consumed natural gas of GB at time *t*; and $P^t_{\mathrm{GB}.h}$ is the generated heat of GB at time *t*.

3.2.3 Optimization Problem

Considering all the models of *devices in* BIES presented above, the primary objective of BIES is to minimize the total cost of system operation. Specifically, the operational cost encompasses several components, including the cost of purchasing electricity and gas from the external markets (EM), the degradation of BESSs, and the penalty incurred for unfulfilled energy demand. Consequently, the optimization problem for BIES operator can be formulated as:

$$\min_{\delta^t} C_b = \sum\nolimits_{\mathrm{t}=1}^{\mathrm{T}} \left\{ x^t_{w.e} \begin{pmatrix} P^t_{\mathrm{CHP}.e} + P^t_{\mathrm{BESS}.c} \\ -P^t_{\mathrm{BESS}.d} - P^t_{\mathrm{PV}.e} \end{pmatrix} + x^t_{w.g} P^t_{w.g} \right\} \tag{3.18}$$

$$s.t.\ \forall t \in T$$

$$(2.1) - (2.17)$$

$$P^t_{w.\mathrm{e}} + P^t_{\mathrm{DER}} + P^t_{\mathrm{CHP}.e} + P^t_{\mathrm{CHP}.d} - P^t_{\mathrm{BESS}.c} = P^t_e \tag{3.19}$$

$$P^t_{\mathrm{CHP}.h} = P^t_h \tag{3.20}$$

$$P^t_{w.g} = P^t_{\mathrm{GB.g}} \tag{3.21}$$

where $\{P^t_{\mathrm{CHP}.e}, P^t_{\mathrm{BESS}.d}, P^t_{\mathrm{BESS}.c}, P^t_{w.e}, P^t_{\mathrm{CHP}.h}, P^t_{\mathrm{GB}.h}\}$ is the set of decision variables. $P^t_{BESS,d}$ and $P^t_{BESS,c}$ are the discharge and charge power; $x^t_{w.e}$ and $x^t_{w.g}$ are the wholesale electricity and natural gas market price; $P^t_{w.\mathrm{e}}$ is power purchased from the wholesale electricity market; $P^t_{\mathrm{PV.e}}$ is the power output of PV penal. P^t_e and P^t_h are the power and heat demands within the BIES. The objective function aims to minimize the costs for purchasing electricity and operation of devices. Also, the objective is constrained by (3.1)-(3.17), and (3.19)-(3.21), where (3.1)-(3.17) are operating

constraints for micro-CHP unit, BESS, and GB, and (3.19)-(3.21) indicate the multi-energy balance.

3.2.4 Markov Decision Process

To optimize the decision-making process of BIES operator, an MDP is leveraged to describe the optimization problem. The BIES operator is an intelligent agent whose objective is to improve the operation decisions by minimizing the total cost in (3.18). The MDP can be denoted by a tuple $< S^t, A^t, R^t(s,a), P^t(s,a), \mu, \gamma^t >$, where$S^t = \{x^t_{w.e}, x^t_{w.g}, E^t_{\text{BESS}}, P^t_{e.fore}, P^t_{h.fore}, P^t_{\text{PV}.fore}\}$ is the state, which encompasses electricity market price $x^t_{w.e}$, natural gas market price $x^t_{w.g}$, SoC of BESS E^t_{BESS}, forecast of power demand $P^t_{e.fore}$, forecast of heat demand $P^t_{h.fore}$, and forecast of PV generation $P^t_{\text{PV}.fore}$; $A^t = \{P^t_{\text{CHP}.e}, P^t_{\text{BESS}.d}, P^t_{\text{BESS}.c}, P^t_{w.e}, P^t_{\text{CHP}.h}, P^t_{\text{GB}.h}\}$ is the action, including the available actions as the decision variables in (3.18); $R^t(s,a)$ is the reward quantifying the agent performance, which is presented by the opposite of objective function in (3.18); μ is the policy of the MDP, which contains a series of action for each state; and γ^t is the discount factor that discounts all rewards in the future state.

As the main objective of the agent is to identify the optimal policy that maximizes the accumulated return, the value of each state using the state value function $V^\mu(s)$ is evaluated as given in (3.22). Moreover, the state-action value function $Q^\mu(s,a)$ that captures the joint value of a particular action a at a state s is demonstrated in (3.23), where $\mathbb{E}(\cdot)$ is the expectation function, s_0 and a_0 are the initial state and action, respectively.

$$V^\mu(s) = \mathbb{E}\left[\sum_{t\in T} \gamma^t R^t \,|s_0 = s\right] \tag{3.22}$$

$$Q^\mu(s,a) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R^t \,|s_0 = s, a_0 = a\right] \tag{3.23}$$

3.3 Proposed TFT-SAC algorithm

In this section, a novel TFT-SAC approach to solve the optimal scheduling problem of BIES is introduced. The structure of the proposed TFT-SAC approach is depicted in Fig. 3.3. Specifically, the TFT uses historical PV power generation and energy consumption data alongside meteorological and static covariates (e.g., geographical coordinates and energy types) to forecast future trends. Variable selection networks (VSNs) identifies relevant features, while an LSTM network captures long-term dependencies. A multi-head self-attention layer focuses on crucial time steps, enhancing the forecasting accuracy. These forecasts inform subsequent optimization tasks. The SAC algorithm uses forecasting data to generate the optimal operation strategies for the BIES. These strategies are implemented, and the resulting state transitions (state, action, reward, next state) are stored in the experience replay buffer (ERB). The experiences are sampled to train the critic and actor networks until the SAC algorithm converges, producing an optimal operation strategy for BIES. The details of the TFT and SAC algorithm are presented in the following subsections.

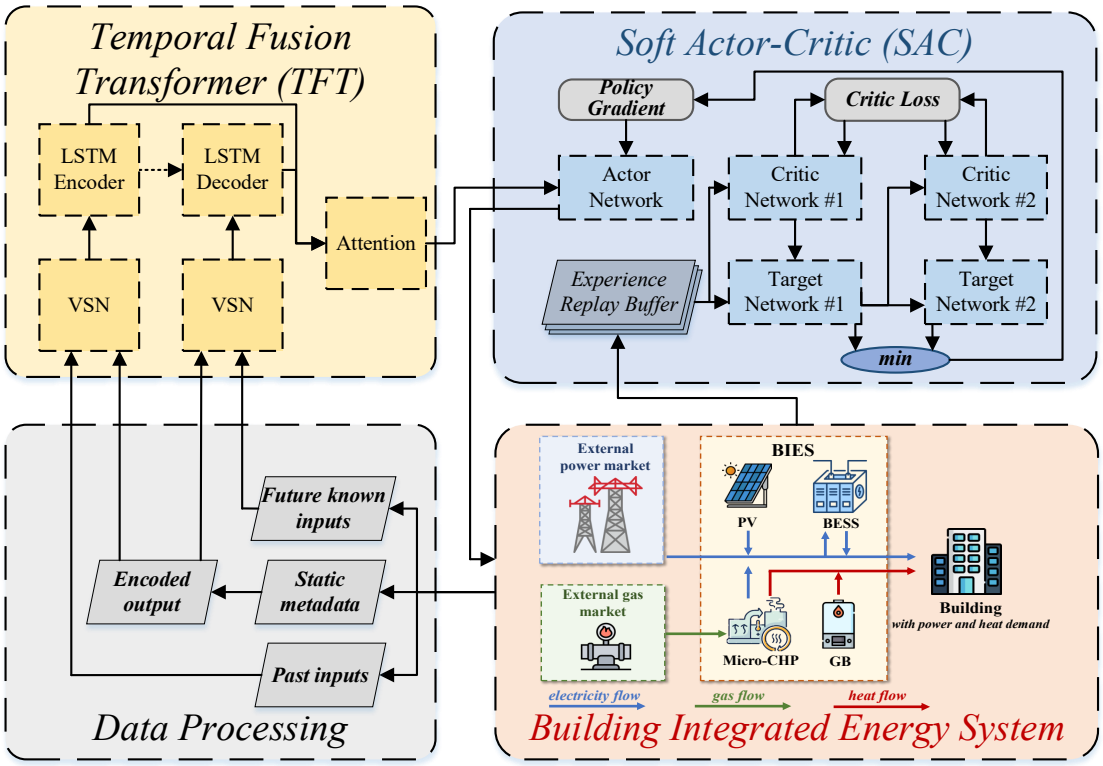


Fig. 3.3 Structure of proposed TFT-SAC approach

3.3.1 TFT Model

This subsection introduces the TFT model, an interpretable deep learning model designed for time-series forecast. The TFT model effectively captures complex temporal relationships and delivers reliable forecasts, which are essential for managing BIES. Specifically, the interpretability of the multi-head self-attention mechanism and VSN stems from its ability to assign VSN weight v_{χ_t} and attention weight $\tilde{A}(\boldsymbol{Q}, \boldsymbol{K})$ to input data points, thereby visualizing the most influential time steps and features in the prediction process. Detailed algorithm design is covered in the following subsections.

1) Quantile Outputs

The TFT model generates quantile forecasts, which are particularly useful for estimating the uncertainty of future *forecast*s. The quantile forecasts are obtained through a linear transformation of the outputs from the temporal fusion decoder. The mathematical representation of this process is given as:

$$\hat{\mathrm{y}}_{\mathrm{i}}(\mathrm{q}, \mathrm{t}, \tau) = \mathrm{f}_{\mathrm{q}}\left(\tau, \mathbf{y}_{\mathrm{i,t-k:t}}, \mathbf{z}_{\mathrm{i,t-k:t}}, \mathbf{x}_{\mathrm{i,t-k:t+\tau}}, \mathrm{s}_{\mathrm{i}}\right) \tag{3.24}$$

where $\hat{\mathrm{y}}_{\mathrm{i}}(\mathrm{q}, \mathrm{t}, \tau)$ is the q^{th} quantile value for predicting the future τ steps at time point *t*; $\mathrm{f}_{\mathrm{q}}(.)$ is the forecasting model; $\mathbf{y}_{\mathrm{i,t-k:t}}$ is the vector of historical target variables from time points $\mathrm{t} - \mathrm{k}$ *to* t; $\mathbf{z}_{\mathrm{i,t-k:t}}$ is the vector of past-observed inputs from time points *t-k* to *t*; $\mathbf{x}_{\mathrm{i,t-k:t+\tau}}$ is the vector of priori-known future inputs; and s_{i} is the static metadata, which is the covariate in energy forecast.

The training of TFT model involves minimizing the quantile loss [76], which is designed to penalize the overestimations and underestimations differently based on the quantile level. The quantile loss function is formulated as:

$$\mathcal{L}(\Omega, W) = \sum_{y_t \in \Omega} \sum_{q \in \Omega} \sum_{\tau=1}^{\tau_{\max}} \frac{QL(y_t, \hat{y}(q, t-\tau, \tau), q)}{M_{\tau_{\max}}} \tag{3.25}$$

where $\mathcal{L}(\Omega, W)$ is the quantile loss of single time series at the average prediction point; y_t is the actual data; $\hat{y}$ is predictions; Ω is the domain of training data containing $M_{\tau_{\max}}$ samples; W is the weight of TFT model; $\tau_{\max}$ is the maximum step; and the function $QL(.)$ can be expressed as:

$$QL(y, \hat{y}, q) = q(y - \hat{y})_+ + (1 - q)(\hat{y} - y)_+ \tag{3.26}$$

where QL is the output quantiles ($q = \{0.1, 0.5, 0.9\}$ in the experiments); and $(.)_+ = \max(0, .)$. To ensure consistency in prediction dimensions across different prediction points, the regularization is applied as:

$$q_{risk} = \frac{2 \sum_{y_t \in \widetilde{\Omega}} \sum_{\tau=1}^{\tau_{\max}} QL(y_t, \hat{y}(q, t-\tau, \tau), q)}{\sum_{y_t \in \widetilde{\Omega}} \sum_{\tau=1}^{\tau_{\max}} |y_t|} \tag{3.27}$$

where $\tilde{\Omega}$ is the domain of test samples; q_{risk} is the normalized quantile losses across the entire forecasting horizon.

2) Gating Mechanism

In the time-series forecast, especially with multiple regression, identifying relevant variables and the extent of non-linear processing is challenging. The TFT model uses gated residual networks (GRNs) for adaptive non-linear processing as needed in (3.28), and the gated linear units (GLUs) are shown in (3.31).

$$\mathrm{GRN}_{\omega} = \mathrm{LayerNorm}\big(\mathbf{a} + \mathrm{GLU}_{\omega}(\boldsymbol{\eta}_1)\big) \tag{3.28}$$

$$\boldsymbol{\eta}_1 = \boldsymbol{W}_{1,\omega}\boldsymbol{\eta}_2 + \boldsymbol{b}_{1,\omega} \tag{3.29}$$

$$\boldsymbol{\eta}_2 = ELU\big(\boldsymbol{W}_{2,\omega}\boldsymbol{a} + \boldsymbol{W}_{3,\omega}\mathbf{c} + \boldsymbol{b}_{2,\omega}\big) \tag{3.30}$$

$$GLU_{\omega}(\boldsymbol{\gamma}) = \sigma\big(\boldsymbol{W}_{4,\omega}\boldsymbol{\gamma} + \boldsymbol{b}_{4,\omega}\big) \odot \big(\boldsymbol{W}_{5,\omega}\boldsymbol{\gamma} + \boldsymbol{b}_{5,\omega}\big) \tag{3.31}$$

where $LayerNorm(.)$ is the layer normalization function; $\mathbf{a}$ is the vector of primary inputs to GRN; and $\mathbf{c}$ is an optional context vector; $ELU(.)$ is the Exponential Linear Unit activation function; $\sigma(.)$ is the sigmoid activation function, $\boldsymbol{W}_{1,\omega}$, $\boldsymbol{W}_{2,\omega}$, $\boldsymbol{W}_{3,\omega}$, $\boldsymbol{W}_{4,\omega}$, and $\boldsymbol{W}_{5,\omega}$ are index to denote weight sharing respectively; $\boldsymbol{b}_{1,\omega}$, $\boldsymbol{b}_{2,\omega}$, $\boldsymbol{b}_{4,\omega}$, and $\boldsymbol{b}_{5,\omega}$ are index to denote bias sharing respectively. The GRN layer is controlled by the GLU layer, which may skip the layer entirely if GLU outputs are close to 0. $\boldsymbol{a} + GLU_{\omega}(\boldsymbol{\eta}_1)$ represents linear and nonlinear contributions, with GLU controlling the degree of nonlinearity.

3) VSN

The VSN is a key component of the TFT that improves the performance by selecting important features and filtering out noises. It assigns weights to features, which are used to combine the processed inputs:

$$\upsilon_{\chi_t} = \text{Softmax}\left(\text{GRN}_{\upsilon_\chi}(\boldsymbol{\Xi}_t, \mathbf{c}_s)\right) \tag{3.32}$$

where v_{χ_t} is the set of weights corresponding to the features; $\boldsymbol{\Xi}_\mathbf{t}$ is the flattened vector; and $\boldsymbol{c}_s$ is obtained from the static covariate encoder. The processed features are weighted by their corresponding variable selection weights and combined.

5) Temporal Self-attention Layer

The TFT model employs a temporal self-attention layer that plays a key role in capturing long-term dependencies in time-series data. This layer not only improves the model's ability to understand complex temporal relationships but also enhances the interpretability of forecasts. The self-attention layer used here is a masked and interpretable multi-head attention layer combined with a gating mechanism to selectively control information flow.

The core concept behind the temporal self-attention layer is to calculate the relevance, or "attention", of different time steps to each other, enabling the TFT model to focus on important events or sequences within the data. This is done using the following equation for attention:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = A(\boldsymbol{Q}, \boldsymbol{K})\boldsymbol{V} \tag{3.33}$$

where $\boldsymbol{V}$ is the value of input based on the similarity between the query vector $\boldsymbol{Q}$ and key vector $\boldsymbol{K}$; and $A(\cdot)$ is a normalization function that determines the attention weights of value $\boldsymbol{V}$. The scaled dot-product mechanism for calculating attention is defined as:

$$A(\boldsymbol{Q}, \boldsymbol{K}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_{attn}}}\right) \tag{3.34}$$

Multi-head self-attention mechanism enhances the power of the self-attention mechanism by allowing the model to jointly focus on information from different representation subspaces at different positions. Instead of using a single set of queries, keys, and values, the multi-head self-attention mechanism splits them into multiple sets, each of which is processed independently. Each head computes attention separately, and the results are then concatenated and linearly transformed to produce the final output. By having multiple heads, the TFT model can capture a richer set of

relationships and nuances in the data compared with a single self-attention mechanism, which are presented as:

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \left[\boldsymbol{H}_1,\boldsymbol{H}_2 \dots,\boldsymbol{H}_{m_H}\right]\boldsymbol{W}_H \tag{3.35}$$

$$\boldsymbol{H}_h = Attention\left(\boldsymbol{Q}\boldsymbol{W}_{\boldsymbol{Q}}^{(\boldsymbol{h})},\boldsymbol{K}\boldsymbol{W}_{\boldsymbol{K}}^{(\boldsymbol{h})},\boldsymbol{V}\boldsymbol{W}_{\boldsymbol{V}}^{(\boldsymbol{h})}\right) \tag{3.36}$$

where $\boldsymbol{W}_Q^{(h)} \in \mathbb{R}^{d_{\text{model}}\times d_{\text{attn}}}$, $\boldsymbol{W}_K^{(h)} \in \mathbb{R}^{d_{\text{model}}\times d_{\text{attn}}}$, and $\boldsymbol{W}_V^{(h)} \in \mathbb{R}^{d_{\text{model}}\times d_V}$ are the head-specific weights for queries, keys, and values, respectively; and $\boldsymbol{W}_H \in \mathbb{R}^{(m_H d_v)\times d_{\text{model}}}$ linearly combines outputs concatenated from all heads $\boldsymbol{H}_h$ ($H = 1, 2, \dots, m_H$). m_H is the number of heads, d_{model}, d_{attn} and d_V are the dimension of model, attention layer and weight $\boldsymbol{V}$.

One of the main issues with traditional multi-head attention *mechanism* is that each head uses different value vectors, making it difficult to directly determine the feature importance from the attention weights. By modifying the mechanism to share the same value vector across all heads, the TFT model can produce a unified set of attention weights, thereby improving interpretability:

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \widetilde{\boldsymbol{H}}\boldsymbol{W}_H \tag{3.37}$$

$$\widetilde{H} = \tilde{A}(\boldsymbol{Q},\boldsymbol{K})\,\boldsymbol{V}\,\boldsymbol{W}_V = \left\{\frac{1}{m_H}\sum_{h=1}^{m_H} A\left(\boldsymbol{Q}\,\boldsymbol{W}_Q^{(h)},\boldsymbol{K}\,\boldsymbol{W}_K^{(h)}\right)\right\}\boldsymbol{V}\,\boldsymbol{W}_V$$

$$= \frac{1}{m_H}\sum_{h=1}^{m_H} Attention\left(\boldsymbol{Q}\,\boldsymbol{W}_Q^{(h)},\boldsymbol{K}\,\boldsymbol{W}_K^{(h)},\boldsymbol{V}\,\boldsymbol{W}_V\right) \tag{3.38}$$

where $MultiHead(.)$ is interpretable multi-head, $\widetilde{\boldsymbol{W}_{\boldsymbol{H}}} \in \mathbb{R}^{d_{model}\times d_{attn}}$ denotes the final linear mapping used across $\boldsymbol{W}_H$, and $\boldsymbol{W}_V \in \mathbb{R}^{d_{\text{model}}\times d_V}$ is the value weights shared across all heads. Compared to $A(\boldsymbol{Q},\boldsymbol{K})$ in (2.34), this modification allows each attention head to share the same set of values $\tilde{A}(\boldsymbol{Q},\boldsymbol{K})$, resulting in a single and interpretable set of attention scores that can be analyzed to determine feature importance [77].

3.3.2 SAC Algorithm

In this subsection, the SAC algorithm as a state-of-the-art maximum-entropy-based off-policy DRL algorithm is described to solve the optimization problem of BIES.

Typical DRL algorithms generally suffer from limited robustness in real-world applications due to ineffective exploration. In contrast, the SAC algorithm uses entropy as a regularization term in the objective function to enhance adaptability and generalization performance.

1) Algorithm Description

As a DRL algorithm with an actor-critic structure, the SAC algorithm outperforms most algorithms, e.g., DDPG, in convergence performance. The SAC algorithm maximizes both accumulative rewards and policy entropy. The entropy function $H(\cdot)$ is defined in (3.39), where $\pi(\cdot\,|s_t)$ is the strategy conditioned on the state s_t. The state value function $V_r^{\mu}(s)$ and state-action value function are $Q_r^{\mu}(s,a)$ presented in (3.40) and (3.41), respectively, where the temperature parameter α determines the relative importance of the entropy term against the reward, and thus controls the stochasticity of the optimal policy.

$$H\big(\pi(\cdot\,|s_t)\big) = -\sum_{a} \pi(a|s_t) \ln \pi(a|s_t) \tag{3.39}$$

$$V_r^{\mu}(s) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \Big(R_t + \alpha H\big(\pi(\cdot\,|s_t)\big)\Big)\,|s_0 = s\right] \tag{3.40}$$

$$Q_r^{\mu}(s,a) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(R_t + \alpha \sum_{t\in T} H\big(\boldsymbol{\pi}(\cdot\,|s_t)\big)\right) |s_0 = s, a_0 = a\right] \tag{3.41}$$

At the same time, the value functions can be expressed as (3.42) according to the relationship between (3.39) and (3.40). Equation (3.40) allows us to derive the solution for the policy as (3.43).

$$V_r^{\mu}(s_t) = \mathbb{E}\big[Q_r^{\mu}(s,a)\big] + \alpha H\big(\boldsymbol{\pi}(\cdot\,|s_t)\big) \tag{3.42}$$

$$\boldsymbol{\pi}^*(\cdot\,|s_t) = arg\max_{\boldsymbol{\pi}\in\Delta} V_r^{\mu}(s) = \frac{e^{Q_h^{\pi}(s,\cdot)/\alpha}}{\sum_a e^{Q_h^{\pi}(s,a)/\alpha}} \tag{3.43}$$

where $\Delta = \{\boldsymbol{\pi}|\boldsymbol{\pi} \geq 0,\ \mathbf{1}\cdot\boldsymbol{\pi} = 1\}$. When the Q value converges to the optima, the optimal policy achieves the optimal state value function. Therefore, the updating of Q-value function can be realized by using the closed form solution in an off-policy scheme.

2) Algorithm Implementation

The SAC algorithm adopts an actor-critic structure with DNNs to estimate the policy (actor) and Q-value functions (critic). The actor network is represented by the policy function $\mu(s|\theta^{\mu})$ parameterized by θ^{μ}. The critic employs clipped double Q network Q_1 and Q_2 parameterized by θ^{Q_1} and θ^{Q_2}, and also their target networks, parameterized by $\theta^{Q_1'}$ and $\theta^{Q_2'}$ Therefore, the target y_t for the Q value is expressed as (3.44), where $\tilde{a}_{t+1}$ is the action under the current policy in the next state s_{t+1} and π_θ is the executed policy. Then, the L2 loss is used to update the Q-network in (3.45) for $j = \{1,2\}$.

$$y_t = r_t + \gamma\left(\min_{j\in\{1,2\}} Q(s_{t+1}, \tilde{a}_{t+1}|\theta^{Q_j}) - \alpha \log \pi_\theta(\tilde{a}_{t+1}|s_{t+1})\right) \tag{3.44}$$

$$\nabla_{\theta^Q} L = \frac{1}{N}\sum_{n\in N}[y_t - Q(s, a|\theta^{Q_j})] \tag{3.45}$$

To train these networks, the agent randomly samples tuples (s_j, a_j, r_j, s_{j+1}) from the ERB to form nth mini batch for experience replay learning, where $n \in N$, and N is the set of all batches. The online critic networks are updated by one step of gradient descent to the mean square error (MSE) θ^{Q_j} in (3.45), while the actor network is updated by one step of gradient ascent using (3.46). To stabilize the training, the target network parameters are soft updated with (3.47).

$$\nabla_{\theta^\mu} L = \nabla_{\theta^\mu}\frac{1}{N}\sum_{n\in N}\left[\min_{j\in\{1,2\}} Q\big(s_t, \tilde{a}_t(s)\big) - \alpha \log \pi_\theta(\tilde{a}_t|s_t)\right] \tag{3.46}$$

$$\theta^{Q'} \leftarrow \rho\theta^{Q} + (1-\rho)\theta^{Q'} \tag{3.47}$$

where $\tilde{a}_t(s)$ is a sample from $\pi_\theta(\cdot\,|s_t)$; ρ is the soft update parameter.

3.3.3 Discussions

The use of the proposed TFT-SAC approach is unique and effective for the dynamic operation and control of BIES. This combination offers several advantages and potential shortcomings compared to other traditional approaches.

1) Integrated forecasting and operation: the TFT provides accurate and data-driven forecasts of PV generation and energy demand, which allows the SAC algorithm to make informed decisions. This integration reduces uncertainty in the decision-making process, leading to more reliable system operations. Moreover, the most important part in bridging TFT and SAC is not the model itself, but to consider how the forecast or

what kind of forecast can be helpful to the decision-making of the RL algorithm. The TFT can be helpful by providing the interpretability of forecasting results, which is more valuable than forecast accuracy in this case.

2) Offline training and efficient online operation: The proposed TFT-SAC approach allows for offline training using historical data, enabling the development of a robust policy before deployment. Once trained, the algorithm operates in real time with minimal computational overhead, which is a significant advantage over approaches like SO or RO that require repeated recalculation.

3) Handling non-convexity: The operation of BIES involves non-convex constraints such as the FOR. The SAC algorithm, leveraging DNNs, can effectively learn non-convex optimal operating policies due to the powerful representation capabilities of DNNs. In comparison, traditional mathematical programming approaches, such as mixed-integer linear programming (MILP), address non-convexity by linearizing nonlinear relationships and explicitly formulating integer constraints, facing scalability and computational challenges particularly in large, dynamic systems like BIES. Heuristic algorithms can explore complex optimization landscapes and are often more flexible than mathematical programming. However, they may suffer from high computational demands, especially in large-scale systems, and may converge to local optima rather than finding the global solution.

4) Training complexity: The proposed TFT-SAC approach requires extensive offline training, which can be computationally expensive and time-consuming, particularly for large datasets. The performance highly relies to a high-quality training dataset, which is typically hard to acquire in the real world.

5) Dependence on forecasting accuracy: The effectiveness of SAC algorithm in making optimal decisions depends heavily on the forecasting accuracy provided by TFT. If the forecasts are inaccurate due to unexpected external factors, the quality of the operational decisions may be compromised.

Overall, the proposed TFT-SAC approach provides an effective solution for BIES operation. The integrated forecast and optimize structure, capability to handle non-

convexity, and efficient implementation make this approach a compelling alternative to traditional approaches, despite some challenges related to training complexity and dependence on forecasting accuracy.

3.4 Case Study

3.4.1 Simulation Setup

To validate the effectiveness of the proposed TFT-SAC approach, case studies are conducted using data from a real building located in Zhenjiang, China. The BIES under study comprises a micro-CHP unit, PV panels, BESSs, and a GB device to meet both heat and power demands.

The micro-CHP unit, with a rated output of 25.3 kWh, is designed to satisfy the heat demand of the building while partially covering its power demand. The PV system includes 610 PV panels, each with a capacity of 280 W, resulting in a theoretical maximum output of 170.8 kWh. However, due to practical limitations, the actual capacity is 153 kWh. The BESS consists of 24 LiFePO4 batteries, each with a storage capacity of 5.12 kWh, providing a maximum output of 72 kWh. This setup enables the BESS to support peak power demand for up to 4 hours. Detailed information on micro-CHP and BESS is shown in Appendix A.

The proposed TFT-SAC approach is implemented in Python, and the neural networks are developed using PyTorch. To achieve the optimal performance, the neural network parameters and hyperparameters are carefully chosen based on empirical values and adjusted throughout the training process. The complete configuration details for SAC algorithm are presented in Tables 3.1 and 3.2, while hyperparameter settings of TFT for forecasts of energy demand and PV generation are shown in Table 3.3. The Adam optimizer is used as the training algorithm to update the network weights.

Table 3.1 Neural network architectures setting of sac algorithm

Neutral Networks	Number of hidden layers	Number of neurons	Learning rate	Soft update parameter	Optimizer
Actor	3	[512,32]	1×10^{-4}	11×10^{-2}	Adam
Critic	2	[512,32]	1×10^{-3}	11×10^{-2}	Adam

Table 3.2 Hyperparameter setting of sac algorithm

Training parameter	Number
Replay buffer size	1×10^{6}
Replay start size	128
Batch size	128
Discount factor	0.99

Table 3.3 Hyperparameter setting of TFT for forecasts of energy demand and PV generation

Parameter	Forecast of energy demand	Forecast of PV generation
Learning rate	1×10^{-4}	3.5×10^{-3}
Grad clip value	0.1	0.9
Patience	10	2
Batch size	16	16
Drop out	0.2	0.1
Time step	168	24
Hidden size	128	32
Number of LSTM layers	6	4
Number of attention heads	6	3
Loss function	Quantile Loss	Quantile Loss

3.4.2 Computational Performance of Different Algorithms

This subsection compares the SAC algorithm with baseline algorithms such as TD3 and DDPG. Each algorithm is trained for 10000 episodes on sampled days from the training set. Figure 3.4 shows the episodic reward evolution of different algorithms during the offline training process. Considering the fluctuations in state features, the data have been smoothed using a 100-episode moving average method. This is because the oscillations caused by the exogenous state features cannot be addressed by the operational strategies even if the policy is optimal.

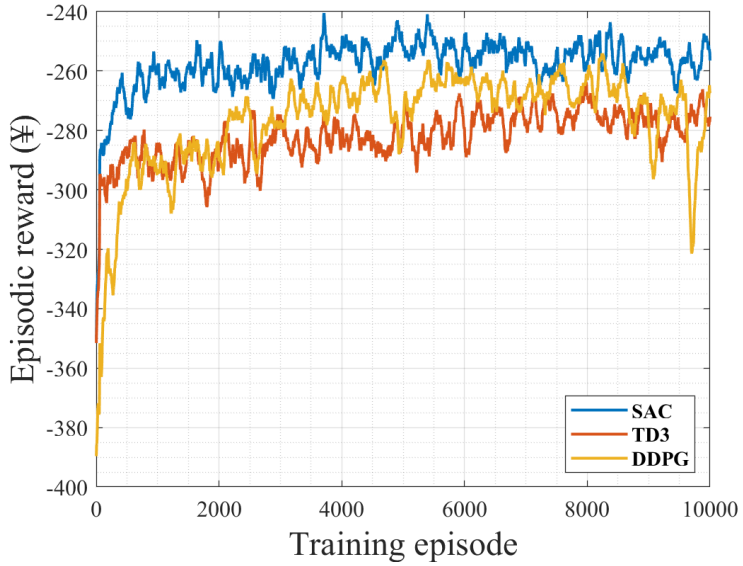


Fig. 3.4 Episodic reward evolution of different algorithms during offline training process.

Fig 3.4 shows that initially, the learning curves of different algorithms are similar due to randomly selected energy schedules and Gaussian noise. Early on, rewards are low for all algorithms. As training progresses, rewards increase as agents learn and refine their policies. The reward of SAC algorithm grows the fastest initially, followed by TD3 and then DDPG. Around 2000 iterations, the reward of DDPG increases sharply, surpassing TD3 but still remaining lower than the SAC algorithm, which is close to converging. DDPG and TD3 converge around 5000 iterations. The SAC algorithm achieves a significantly higher final reward compared with DDPG and TD3,

with the reward of DDPG slightly higher than that of TD3. This indicates the superior offline training performance of the SAC algorithm.

To evaluate the performance of the proposed TFT-SAC approach, the trained actor network parameters are used to generate operational strategies for the BIES over 50 test days. The proposed forecast-enhanced RL approach is compared with benchmark approaches: typical RL approaches (TD3, DDPG, and SAC) and another forecast-enhanced RL approach (LSTM-SAC). Fig 3.5 compares the cumulative costs for energy consumption with different approaches over 50 test days. The results indicate that the cumulative costs with typical RL approaches are significantly higher than those with forecast-enhanced RL approaches. The cost gap increases with more training episodes, highlighting the differences between different approaches. For forecast-enhanced approaches, the cumulative costs are similar, showing that combining forecasting with RL is effective. Notably, the proposed TFT-SAC approach achieves lower costs than LSTM-SAC, demonstrating its superior performance. However, the difference between the proposed TFT-SAC approach and LSTM-SAC is small compared with their differences from typical RL approaches, suggesting limited room for improvement in current forecast-enhanced RL approaches.

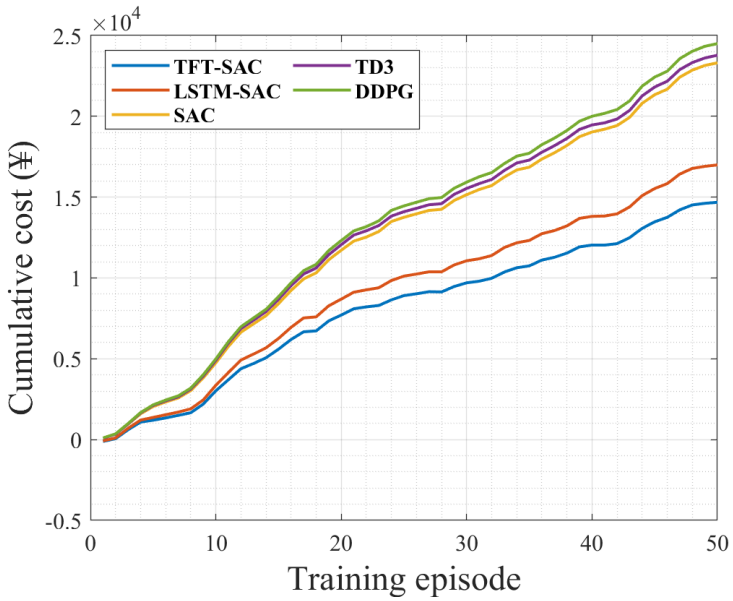


Fig. 3.5 Cumulative cost for energy consumption with different approaches over 50 test days.

Notably, the proposed TFT-SAC approach achieves lower costs than LSTM-SAC, demonstrating its superior performance. However, the difference between the proposed TFT-SAC approach and LSTM-SAC is small compared with their differences from typical RL approaches, suggesting limited room for improvement in current forecast-enhanced RL approaches. The performance differences between TFT-SAC and benchmark algorithms may vary in scenarios with different settings, and be affected by the uncertainty resources significantly.

3.4.3 Forecasting Performance Analysis

As shown in Table 3.4, the TFT model outperforms the LSTM model across three performance metrics, i.e., mean absolute error (MAE), root mean squared error (RMSE), and R^2 in forecasts of both PV generation and building energy demand.

Fig. 3.6 and 3.7 show that the forecasting curves of TFT model closely fit the target curves, demonstrating its effectiveness in capturing time-series patterns. The TFT model particularly excels in forecasting PV generation, accurately capturing peaks and valleys, which is crucial for energy forecasting. In summary, the TFT model shows superior forecasting accuracy and pattern recognition compared with the LSTM model, which is crucial for energy management in BIES, guiding energy allocation, optimizing resource utilization, and improving overall energy efficiency.

Table 3.4 Performance metrics of TFT and LSTM models

Forecast object	Model	MAE	RMSE	R^2
PV generation	LSTM	3.66	12.23	0.8402
	TFT	5.22	11.24	0.8721
Energy demand	LSTM	3.37	4.6	0.9407
	TFT	2.20	3.26	0.9670

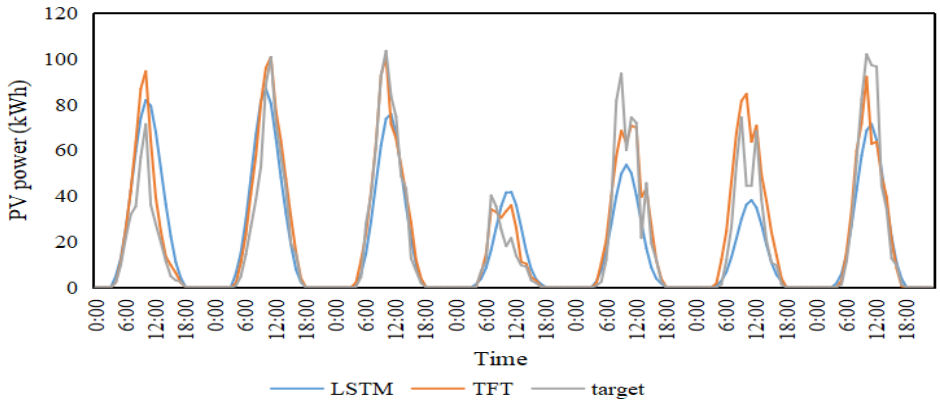


Fig. 3.6 Performance of LSTM and TFT models in forecasting PV generation

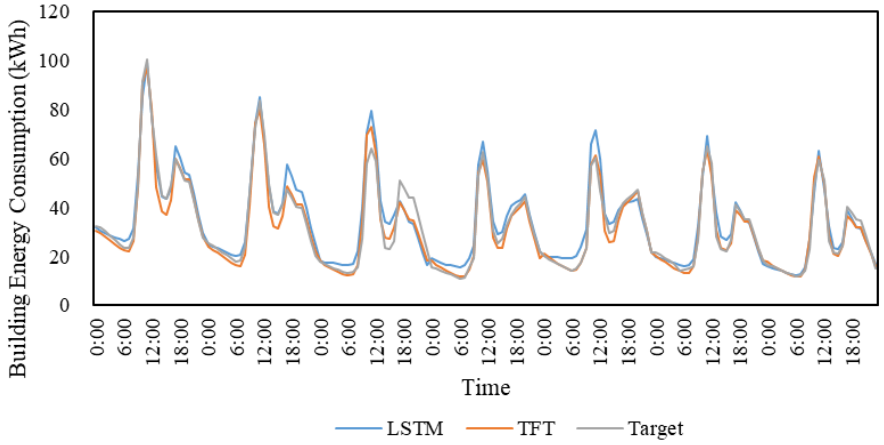


Fig. 3.7 Performance of LSTM and TFT models in forecasting building energy demand

The meteorological data include net solar irradiation (NSI), solar irradiation (SI), ultraviolet (UV), outdoor air temperature (OAT), rainfall (RF), relative humidity (RH), temperature-humidity-wind (THW), and surface air temperature (SAT). Fig 3.8 and illustrates the relative importance of different features in the TFT model for forecasting PV generation. In the encoder, SI appears as the most significant factor, indicating that direct sunlight intensity plays a crucial role in forecasting PV generation. Meanwhile, in the decoder, longitude emerges as the most important feature, highlighting the importance of geographical positioning in the forecasting process. This is intuitive because the position affects the angle of sunlight and daylight duration, which ultimately impacts PV generation.

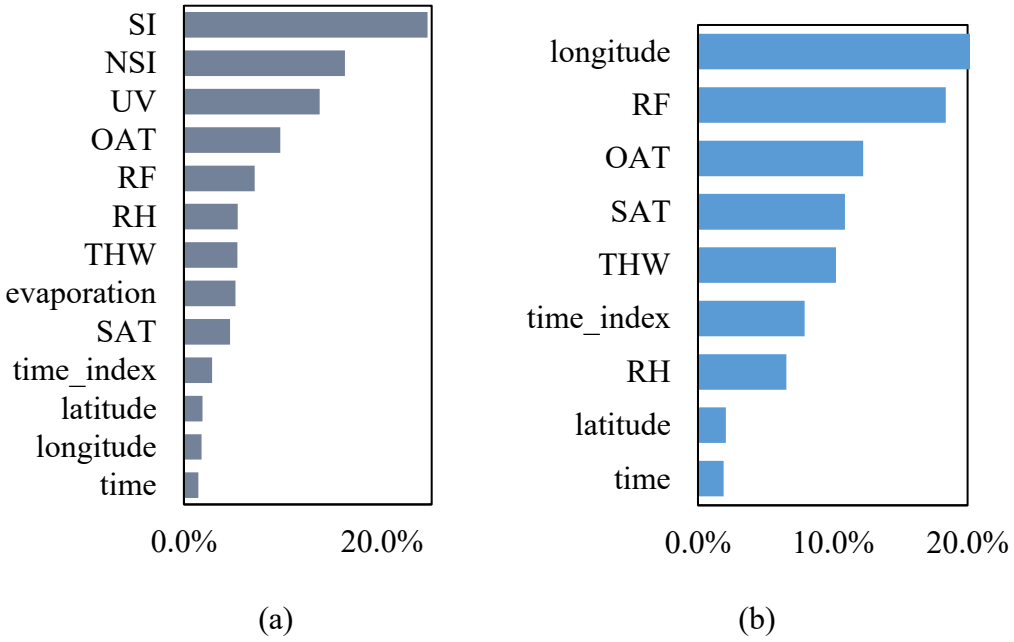


Fig. 3.8. Relative importance of different features in TFT model for forecasting PV generation. (a) Encoder. (b) Decoder.

Fig 3.9 depicts relative importance of different features in TFT model for forecasting building energy demand. Unlike PV generation, which predominantly relies on weather-related factors, building energy demand is highly influenced by calendar-based information. Features such as hour of the day, workday status, and specific time-based attributes are ranked highly, reflecting the relationship between user behavior and energy usage. These calendar-related features indicate the impact of typical human activities and routines—such as work schedules and holidays—on building energy demand.

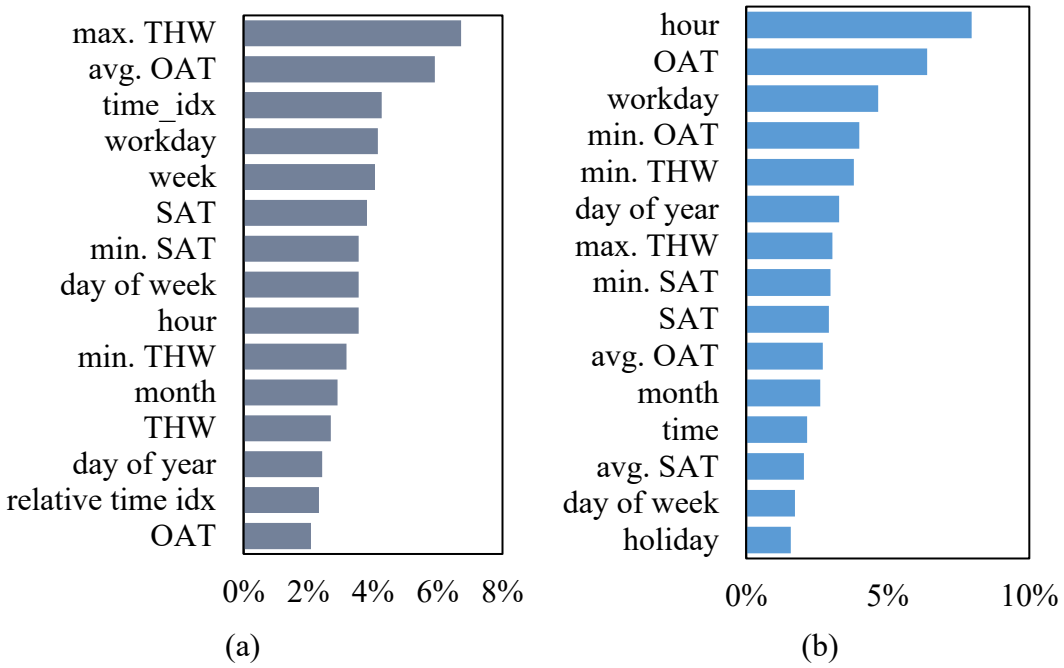


Fig. 3.9 Relative importance of different features in TFT model for forecasting building energy demand. (a) Encoder. (b) Decoder.

The importance ranking reveals that the TFT model considers both weather conditions and temporal attributes to accurately predict energy demands. This is crucial because user activities are often influenced by the time of day or specific events on the calendar, and these behavioral patterns significantly affect energy usage in buildings. The model's attention to these aspects shows its ability to learn from diverse data sources and focus on the most impactful features during the training process, resulting in a more reliable forecast.

Fig 3.10 and 3.11 illustrate the attention distribution of TFT model over the past 7 days (indexed by -7 to -1) during the forecasting process. Fig 3.10 shows that the attention of TFT model is concentrated on the recent past, especially the previous day, reflecting the strong daily cyclic patterns of PV generation. Minor peaks indicate consideration of earlier time steps, but these have lower weights due to the influence of short-term environmental factors like SI.

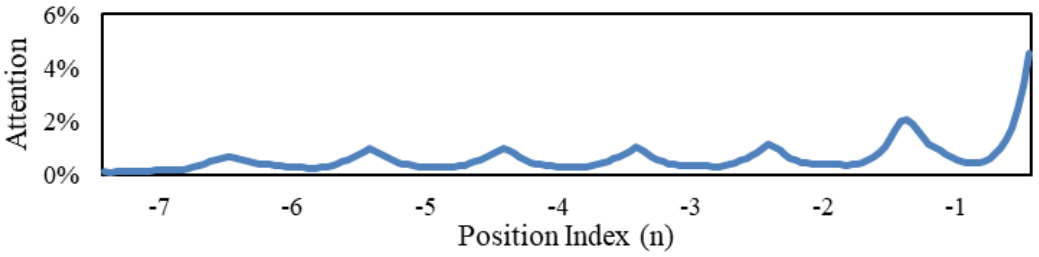


Fig. 2.10 Attention of TFT model over past 7 days for forecasting PV generation.

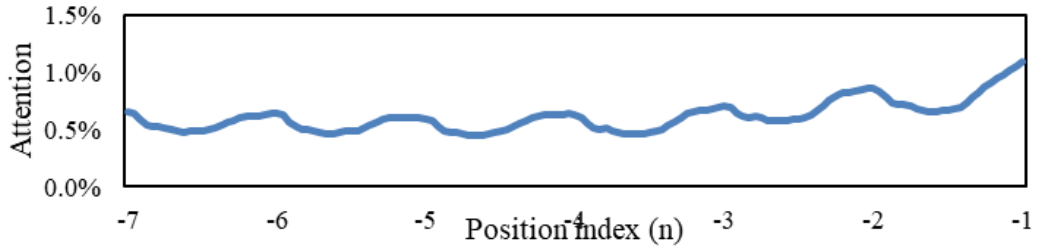


Fig. 3.11 Attention of TFT model over past 7 days for forecasting building energy demand.

Fig 3.11 shows a smooth distribution across various historical time steps with a gradual increase. This suggests the TFT model considers a range of past data, reflecting that the high complexity and irregularity of building energy demands are influenced by factors like user behavior, daily activities, and weather conditions.

In comparison, the TFT model for forecasting PV generation focuses on recent time steps due to daily cyclic patterns, while that for forecasting building energy demands has a broad attention span over the entire historical cycle, balancing long-term trends and short-term impacts. The gradual increase in attention weights indicates the emphasis on recent information for imminent forecasts.

The uniform attention distribution for building energy demand suggests its cyclical patterns are less pronounced or more complex than those of PV generation. This highlights the importance of extracting information from multiple time scales for accurate forecasts and underscores the need for effective energy management strategies to optimize BIES operational efficiency.

In summary, the TFT model provides accurate and interpretable forecasts for both PV generation and building energy demand, supporting the RL algorithm in formulating efficient scheduling strategies.

3.4.4 Generalization Performance

To validate the generalization performance, different approaches are tested over a test set that shows different statistical characteristics compared with the training set. The test set is represented by several typical weeks labeled W-1 to W-4 for comparative analysis. These typical weeks include scenarios with extreme PV generation or energy demand. Table 3.5 presents the daily operational costs of BIES across different weeks. The results clearly demonstrate that forecast-enhanced RL approaches achieve significantly lower operational costs compared with typical RL approaches, underscoring the effectiveness of combining forecasting and decision-making. Furthermore, the average operational cost of the proposed TFT-SAC approach is lower than that of LSTM-SAC, indicating that the proposed TFT-SAC approach outperforms all the comparable approaches across a range of scenarios, thereby demonstrating its strong generalization capabilities. Although the daily cost improvements may appear marginal, the cumulative benefits of the proposed TFT-SAC approach over extended operation could result in substantial additional profits.

Table 3.5 Comparison of daily average operational cost of BIES across different weeks

Week	Daily average operational cost (¥)				
	DDPG	TD3	SAC	LSTM-SAC	TFT-SAC
W-1	500.14	499.3	490.19	328.02	**325.79**
W-2	361.75	361.2	347.92	232.76	**231.6**
W-3	450.34	449.66	431.4	318.91	**311.03**
W-4	733.25	732.44	715.75	521.3	**520.99**

3.4.5 Robust Operation

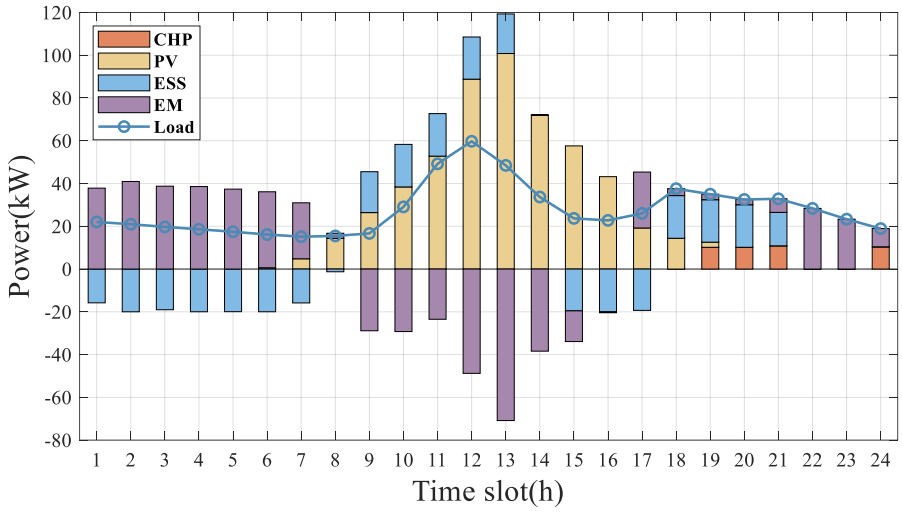
To compare the robustness of the proposed TFT-SAC approach with other RL approaches, independent Gaussian noise is introduced to real PV generation and energy demand to represent uncertain scenarios. The average daily operational costs of BIES under different noise levels are presented in Table 3.6. Across all noise levels, the typical RL approaches incur significantly higher operational costs than forecast-enhanced RL approaches, with cost differences ranging from ¥60 to ¥100. Among all the tested approaches, the proposed TFT-SAC approach demonstrates the lowest average operational costs, indicating superior robustness. However, the cost variations between the proposed TFT-SAC approach and LSTM-SAC remained small, in the range of ¥10 and ¥20. In contrast, the cost difference of the proposed TFT-SAC approach with N=0.01 and N=0.05 is approximately ¥5, and that of TD3, SAC, and LSTM-SAC is ¥3. This larger cost variation suggests that the proposed TFT-SAC approach is more sensitive to forecasting accuracy than other approaches, even though it consistently achieves the lowest average operational costs among all approaches.

Table 3.6 Comparison of daily average operational cost of BIES across different noise levels

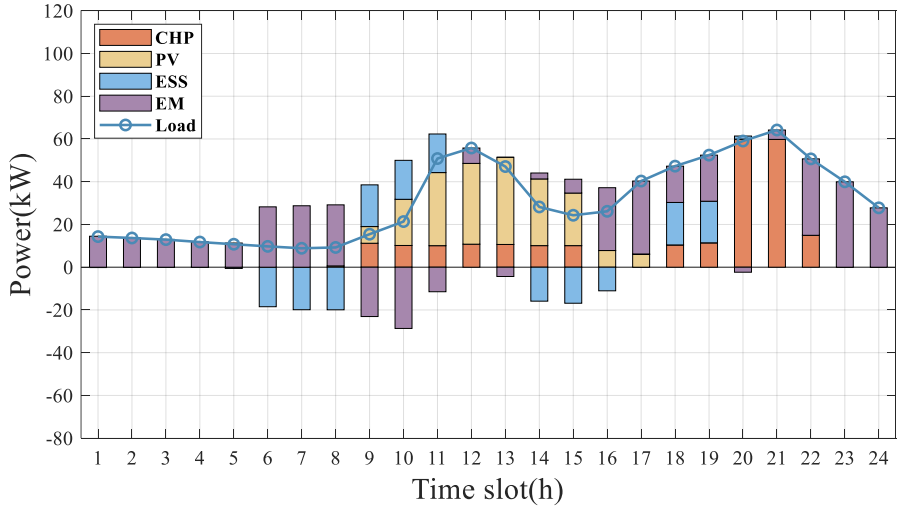
Noise level N	Daily average operational cost (¥)				
	DDPG	TD3	SAC	LSTM-SAC	TFT-SAC
0.01	596.07	557.56	557.49	505.12	**490.04**
0.02	596.38	558.24	558.18	505.82	**491.88**
0.03	597.37	559.02	558.96	506.62	**494.91**
0.04	599.8	559.85	559.78	507.47	**495.13**

3.4.6 Operational Analysis

To evaluate the generalization of the optimal energy management policy learned by the proposed TFT-SAC approach, two typical scenarios are applied: a summer day (August 27) and a winter day (December 25). Figures 3.12 and 3.13 show the power and heat profiles on the two typical days, respectively, where bars above the horizontal axis represent power generation/purchase and bars below indicate storage discharge/power sold.

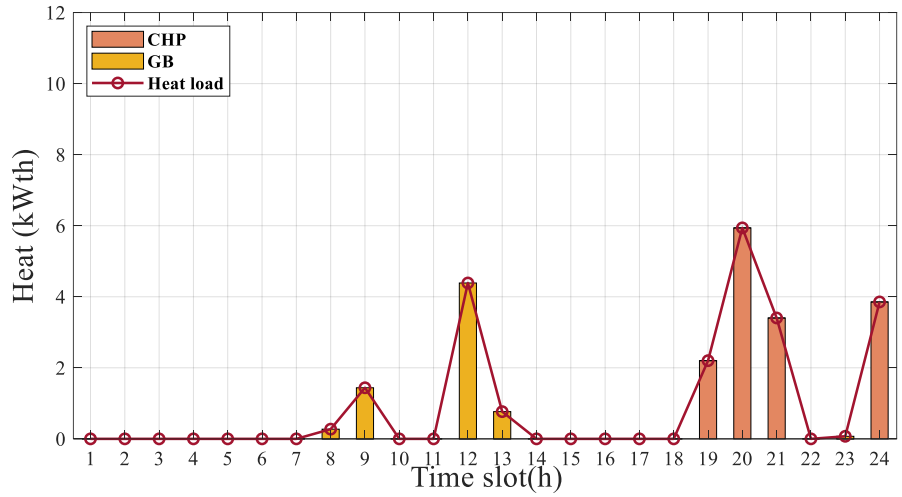


(a)

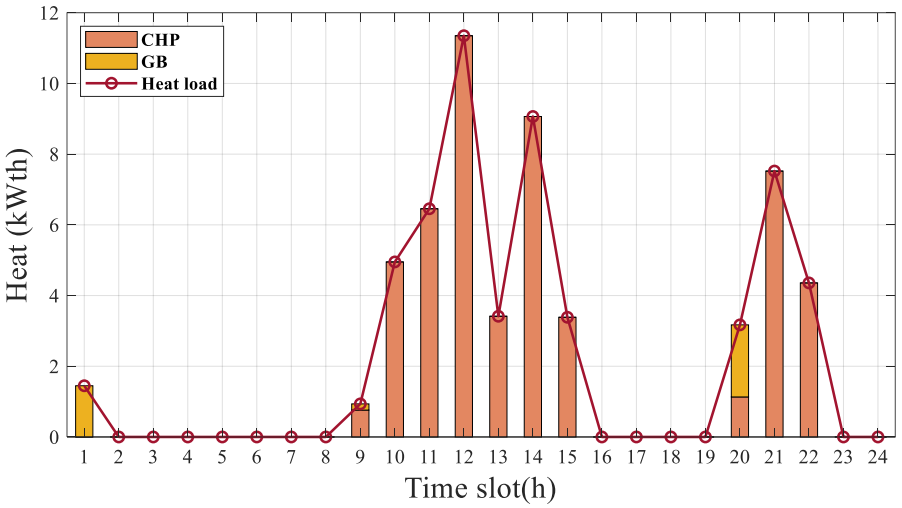


(b)

Fig. 3.12 Power generation and consumption of BIES. (a) A typical summer day. (b) A typical winter day.



(a)



(b)

Fig. 3.13 Heat generation and consumption of BIES. (a) A typical summer day. (b) A typical winter day.

Both scenarios share common trends. Initially, from 00:00 to 8:00, the BIES purchases electricity due to zero PV generation and low SoC of ESS. ESS charges at low prices for future demands. From 09:00 to 15:00, PV generation and ESS discharge could meet most power demands, with excess power sold at high electricity prices. From 18:00 to 24:00, the BIES does not sell electricity, and the micro-CHP unit becomes the primary power source due to high demand.

Nevertheless, there are some evident differences between the two typical days. On the winter day, the micro-CHP unit operates from 09:00 to 15:00 to meet high heat

demands and support the power demands due to low PV generation. On the summer day, the micro-CHP unit is inactive as PV and BESS can meet the demands and the excess power is sold. The policy effectively uses the micro-CHP unit in winter and BESS in summer, charging at low prices and discharging at peak prices to maximize economic benefits.

Finally, it can be concluded that the proposed TFT-SAC approach can learn an effective policy and can generalize to variable state information on different test days. Also, the flexibility of BIES is investigated on two typical winter and summer days. Specifically, the summer day has higher PV generation and lower heat demand, so it has higher energy export and makes use of more flexibility of BESS. Due to lower PV generation and higher heat demand, the winter day has higher power import and higher utilization of the micro-CHP unit, which also provides significant flexibility to BIES.

3.4.7 Sensitivity Analysis

In this subsection, a detailed sensitivity analysis is conducted to evaluate the impact of changes in key factors on the operation and performance of BIES. Specifically, the sensitivities of the episodic reward to variations in electricity price, PV generation, power demand, and heat demand are analyzed, as shown in Fig. 3.14.

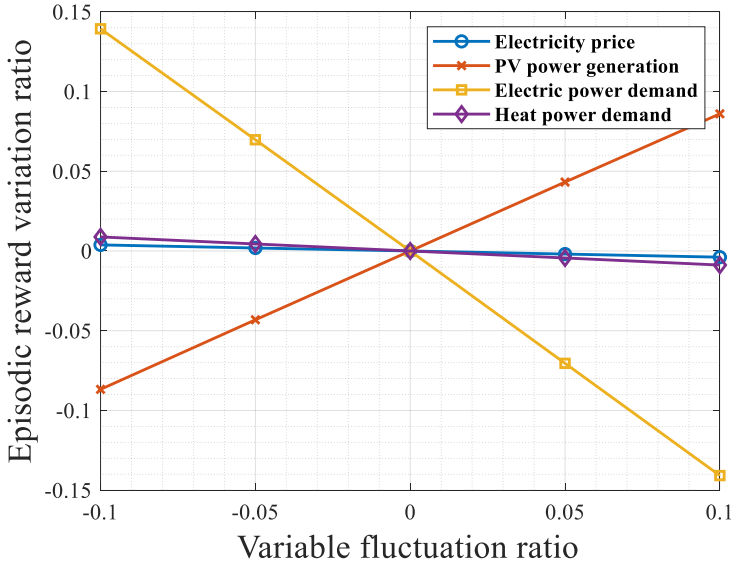


Fig. 3.14 Sensitivity analysis of the proposed model on key factors

The sensitivity analysis is performed by varying each parameter independently from 90% to 110% of the initial configured value, with a granularity of 5%. This range is selected to represent potential fluctuations in market and operational conditions, and the granularity is chosen to provide a balanced level of detail without excessive computational overhead.

The results in Fig. 3.14 indicate the following. The episodic rewards of BIES are negatively correlated with electricity price, which is expected given that higher electricity prices increase the cost of purchasing electricity. There is a positive correlation between PV generation and episodic reward, as increased PV generation reduces the need for power from EM and allows for more excess power to be sold back to EM. Both power and heat demands negatively impact the rewards, with power demand having a particularly significant effect. This can be attributed to the fact that meeting higher demands requires more energy procurement, which incurs additional costs.

Interestingly, the power demand has a greater effect on the episodic reward compared with PV generation. This is because the total daily PV generation is lower than the total power demand. As a result, any reduction in power demand has a larger marginal impact on profitability, either through reduced procurement or allowing more energy to be sold during peak periods.

In terms of scheduling policies, the changes in power demand and PV generation lead to noticeable shifts in action prioritization. For instance, increased PV generation results in more frequent utilization of battery storage for energy arbitrage, while fluctuations in electricity price affect decisions regarding energy procurement timing. These findings emphasize the importance of accurate forecasts for PV generation and energy demand to optimize the operational strategies of BIES effectively.

3.5 Summary

In conclusion, a novel hybrid data-driven approach, namely TFT-SAC, is developed in this chapter for the energy management problem in BIES. Specifically, the TFT

model enhances the forecasting accuracy and transparency through attention mechanisms and the VSN, enhancing interpretability and trustworthiness of forecasting results. The integration of the SAC algorithm for optimization further strengthens the proposed framework by ensuring more effective exploration during training, leading to strategy that exhibits robustness and generalization capabilities. Simulation results demonstrate the superior performance of the proposed TFT-SAC approaches compared with existing approaches. The interpretability of the TFT model and the generalization performance of SAC algorithm are analyzed. The sensitivity analysis of reward on several key factors in BIES is also conducted.

Chapter IV
A Safe Reinforcement Learning algorithm for Operational Optimization of Multi-Network Constrained Integrated Community Energy Systems

4.1 Overview

This chapter focuses on comprehensive techno–economic modeling and energy management in ICES, which considers multi-network constraints and the complex behavior of integrated energy consumers. To this end, the work in this chapter present a novel MNC-ICES model that considers network constraints of integrated energy, including electricity, natural gas, and heat. In the proposed model, the ICESO secures the safe operation of the MNC-ICES by accounting for non-convex energy devices, renewable uncertainties, and IDR of MEUs. A constrained optimization problem is formulated to represent the operation problem in the proposed MNC-ICES model and then transformed into a C-MDP for the application of RL approaches. Compared to existing software programs for regional IES operation, this research highlights operational safety regarding detailed multi-network constraints and detailed energy device models. Moreover, a SOTA Safe RL algorithm, namely PD-TD3, is developed based on the Lagrangian-based Safe RL method to optimize the scheduling and pricing strategies in MNC-ICES. The proposed algorithm shows great potential for Safe RL to become a useful energy management tool in modern ICES regarding operational safety with multi-network constraints.

The contributions of this chapter are as follows:

1) *Comprehensive Modeling of Community Energy System*: A novel MNC-ICES model is proposed to interpret the concept of ICES. The proposed model accounts for the constraints of multi-network, which captures the physical characteristics of energy flow and imposes security operational constraints for the distribution level energy

transmissions. Energy devices are modeled in high fidelity to describe the realistic physical operating attributes in practice. Additionally, the renewable uncertainty and integrated demand elasticity are considered to describe the novel characteristics of modern distribution-level energy systems. Overall, the proposed model can be implemented as a basis for practical network-constrained community operation tools.

2) *Constrained-Markov Decision Process Modeling*: A C-MDP is formulated from the constrained operational optimization problem in MNC-ICES with multi-energy integration. Constraints on voltage in the power network, gas flow, gas pressure and gas injection in the gas network, pipeline flow, and nodal flow in the district heat network are considered security constraints and imposed safety requirements, being modelled as the cost term in a tuple of C-MDP.

3) *Novel Safe Reinforcement Learning Algorithm and Validation*: A novel Safe RL algorithm, namely PD-TD3, is proposed to solve the C-MDP and the constrained operational optimization problem in MNC-ICES. The PD-TD3 algorithm using double networks reduces the over-estimation problem of the action value for both the reward and cost, and the delayed update stabilizes the training process of policy and its dual variable. With such an accurate estimation of Q values, the proposed algorithm converges to the optimal solution that balances the maximal profits and the lowest constraint violation. In addition, the training processes of the policy and its dual variable are stabilized by delayed updates, which contributes to the training efficiency and helps to converge to the global optimal.

The remaining chapter is organized as follows. The mathematical models of MNC-ICES, including integrated networks, energy devices, and MEUs, are presented in Section 4.2. The constrained operational optimization problem and the corresponding C-MDP are formulated in Section 4.3. The novel Safe RL algorithm is proposed in Section 4.4 to solve the C-MDP. Finally, several scenarios are simulated to verify the algorithm performance and analyze the simulation result in Section 4.5. The whole chapter is concluded in Section 4.6.

4.2 System Modeling

This section proposes an MNC-ICES model, including various types of energy sectors and corresponding network models. Specifically, the MNC-ICES model, as depicted in Fig. 4.1, operates as a localized integrated energy system catering to MEUs on the demand side. The proposed model consists of 1) two types of DERs, WT, and PV; 2) two types of energy storage systems, EBS and TES; 3) CHP as a power generation unit, as well as 4) MEU consisting of electric boiler (EB), GB, and energy demand for power and heat. More importantly, the modeling of physical integrated energy networks for electricity, natural gas, and heat within the MNC-ICES model is presented. These networks are foundational components and are vital for the efficient transmission and distribution of energy resources. To this end, multi-network constraints are proposed to govern the behavior of each network, complying with physical constraints in real-world operation. The cooling system (including CHP cooling generation, cooling network, and cooling load) is omitted for simplicity, since its similar operational characteristics to the heating system. The loads for MEUs are consequently modelled in terms of EB and GB, which is a simplified model but sufficient to reflect the basic consumption behavior in the regime of ICES operation.

The MNC-ICES model is assumed to encompass a singular operator, i.e., ICESO, scheduling energy devices and conducting energy transactions. The ICESO should manage the energy schedules of energy devices and determine the energy prices for MEUs to maximize the total profits without violating the network constraints. Therefore, the ICESO needs to schedule the energy devices dynamically for local energy conversion and price-integrated energy to mobilize the IDR resources of MEUs. In contrast, MEUs adjust energy consumptions due to IDR oriented from energy flexibilities. The whole period of operation and transaction can be divided into 24 intervals ($t = \{1,2,\dots,T\}$), and N MEUs are represented by $i = \{1,2,\dots,N\}$. In each step, the ICESO should read the wholesale prices information, observe the local information on energy devices, and evaluate the state of charge of energy storage

systems of TES and EBS before scheduling. Then, the energy prices for MEUs need to be set, the operation status of energy devices needs to be scheduled, and the TES and EBS need to be charged or discharged at each time interval. The detailed models of MNC-ICES are presented as follows.

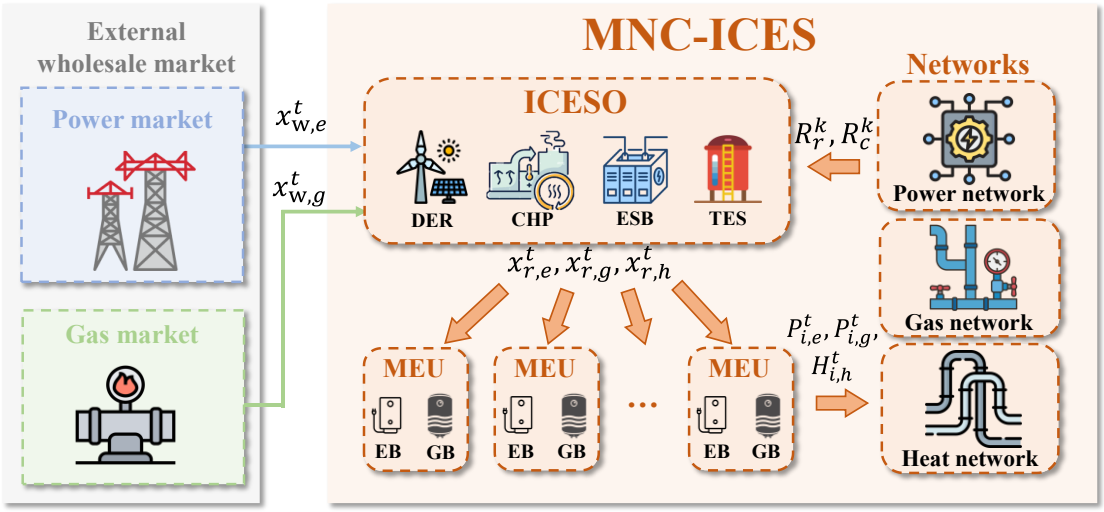


Fig. 4.1 Illustration of the proposed multi-network constrained integrated community energy system model

4.2.1 Electricity Distribution Network

In the distribution of electricity networks, the prevailing topology is often radial, which lends itself well to representation as a tree graph. In this representation, the root point corresponds to the connection with the transmission network. The distribution network can thus be visualized as an interconnected web of nodes and transmission lines, embodying the essential structure of a tree graph.

Let $n \in N_e$ denote the set of nodes within the distribution network, and $(n, m) \in P_e$ represent the set of transmission lines governing the interconnection of these nodes. Fowllowing the paradigm of radial distribution electricity networks, this network configuration captures the hierarchical nature of power flow from the root point, linked to the transmission network, branching out to various nodes within the distribution system. To govern and constrain the dynamics of real power, reactive power, and voltage within the radial distribution network, the linearized DistFlow approach is

adopted [78]. The ensuing sections delve into the specifics of how linearized DistFlow constraints shape and guide the real power, reactive power, and voltage considerations in the context of distribution system operation.

$$P_1^t = \sum_{\forall n} x_n^t, t \in T \tag{4.1}$$

$$P_{n+1}^t = P_n^t - p_{n+1}^t, \forall n \in N_e, t \in T \tag{4.2}$$

$$Q_{n+1}^t = Q_n^t - q_{n+1}^t, \forall n \in N_e, t \in T \tag{4.3}$$

$$V_{n+1}^t = V_n^t - (b_n^1 P_n^t + b_n^2 Q_n^t), \forall n \in N_e, t \in T \tag{4.4}$$

$$\underline{V_n} < V_n^t < \overline{V}_n, \forall n \in N_e, t \in T \tag{4.5}$$

$$0 \le p_i^t \le \overline{p}_i^t, \forall i \in I, t \in T \tag{4.6}$$

$$0 \le q_i^t \le \overline{q}_i^t, \forall i \in I, t \in T \tag{4.7}$$

$$0 \le P_{nm}^t \le \overline{P}_{nm}, \forall n, m \in N_e, \forall (n,m) \in P_e, t \in T \tag{4.8}$$

$$0 \le Q_{nm}^t \le \overline{Q}_{nm}, \forall n, m \in N_e, \forall (n,m) \in P_e, t \in T \tag{4.9}$$

In (4.1)-(4.9) P_{nm}^t and Q_{nm}^t indicate the real power and reactive power flow from bus n to node m at time t. V_n^t is the voltage magnitude at the bus n at time t. p_n^t and q_n^t are the real and reactive power exchange at bus n. b_n^1 and b_n^2 are the resistance and reactance between the bus n and $n+1$. $\overline{V}_n/\underline{V_n}$ are upper/lower bound for voltages of each bus. $\overline{P}_{nm}/\underline{P_{nm}}$ and $\overline{Q}_{nm}/\underline{Q_{nm}}$ denote the upper/lower limits for active and reactive power of the transmission line between bus n and bus m.

4.2.2 Natural Gas Distribution Network

The natural gas network, renowned for its intricate network of pipelines enabling bidirectional gas flow, constitutes a critical infrastructure for the dissemination of energy resources. Traditionally, the directionality of gas flow is contingent upon the interplay of gas pressure differentials and injections at discrete nodes. However, in this work, the scope is limited to the dynamics of unidirectional gas flow within this network. This assumption is made based on operational constraints whereby consumers exclusively draw upon gas resources, with the absence of gas production and storage facilities.

Within this defined framework, let $n \in N_g$ denote the set of nodes, and $(n,m) \in P_g$ represent the set of gas pipelines intricately threading through the natural gas network. To model the dynamics of unidirectional gas flow, this study employs the Weymouth equation [79, 80]. The network-wide constraints for natural gas networks are given as (4.10)-(4.14).

$$gf_{mn}^t = \text{sgn}(Pr_m^t, Pr_n^t)\, C_{mn}\sqrt{|(Pr_m^t)^2 - (Pr_n^t)^2|}, \forall (n,m) \in P_g, \forall t \in T \quad (4.10)$$

$$-\overline{gf}_{mn} \le gf_{mn}^t \le \overline{gf}_{mn}, \forall (n,m) \in P_g \quad (4.11)$$

$$G_n^t = -\sum_{m \in N_g} gf_{mn}^t, \forall (n,m) \in P_g, \forall t \in T \quad (4.12)$$

$$\underline{Pr}_n \le Pr_n^t \le \overline{Pr}_n, \forall n \in N_g, \forall t \in T \quad (4.13)$$

$$0 \le G_n^t \le \overline{G}_n, \forall n \in N_g, \forall t \in T \quad (4.14)$$

In equations above, gf_{mn}^t is the gas flow in the pipeline from node m to node n. Pr_n^t is the gas pressure of the node n. G_n^t is the gas consumption in the node n. C_{mn} is the line pack constant of gas pipeline mn. $\text{Sgn}(\cdot)$ is the signal function to determine the direction of the gas flow. Equations (4.10)–(4.12) show the constraints for nodal natural gas flow balance with the setting of $P_{REFg,t} = P_{n,max}$. In (4.11), $\overline{gf}_{mn}$ is the limitations for the gas flow in the network. Equations (4.13)–(4.14) limit the nodal pressure and gas sources within its threshold, where $\overline{Pr}_n$ and $\underline{Pr}_n$ are the upper and lower bounds of gas pressure at node n, $\overline{G}_n$ is the limitation for gas consumption in node n. It is worth noting that (4.10) is a non-convex equation constraint in an optimization problem, being hard to tackle by using a mathematical programming approach.

4.2.3 District Heating Network

Heat networks are vital for transmitting thermal energy through hot water via water pipelines, which are conventionally comprised of supply and return pipelines. The generation of heat energy, typically by CHP systems within the MNC-ICES model, initiates the flow of water in the supply pipelines to consumers at each node. After the consumer utilizes the heat energy, the water, now cooled, is directed back to the CHP through return pipelines. This unidirectional water flow mirrors the direction of heat

flow. Notably, the temperature and pressure of water decrease along the heat transmission direction, indicating both heat loss during transmission and the propulsive force for water flow. The heat flow is roughly described by Fig. 4.2.

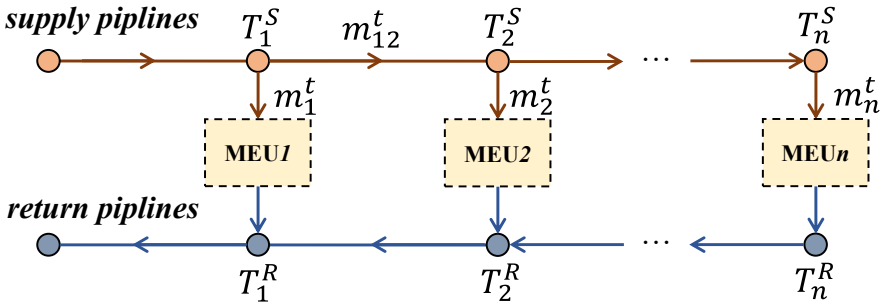


Fig. 4.2 Representation of district heating network

Variable Flow Temperature Constant (VFTC) method is employed to model the heating network [81]. The temperature at the supply and return sides of each node is considered constant over time. During heat transmission, a fixed proportion of heat injected into a pipeline is lost as the water progresses to the next node. Denoting nodes as $n \in N_h$ and direct supply and return pipelines as $(n,m) \in S_n^+$ and $(n,m) \in S_n^-$, respectively, the heat network model is formulated as follows.

$$\sum_{(n,m)\in S_n^+} M_{nm}^t - \sum_{(n',m')\in S_n^-} M_{n'm'}^t = M_n^t, \forall (n,m) \in S_n^-, \forall (n',m') \in S_n^+, n \in N_h, \forall t \in T \quad (4.15)$$

$$\sum_{i\in I_n} H_i^t = -c_f M_n^t (T_n^S - T_n^R), n \in N_h, \forall t \in T \quad (4.16)$$

$$\underline{M}_n^N \leqslant M_n^t \leqslant \overline{M}_n^N, n \in N_h, \forall t \in T \quad (4.17)$$

$$0 \leqslant M_{nm}^t \leqslant \overline{M}_{nm}^S, \forall (n,m) \in (S_n^- \cup S_n^+), \forall t \in T \quad (4.18)$$

In (4.15)-(4.18), M_{nm}^t represents the pipeline heating flow, M_n^t denotes nodal heating flow, and H_i^t signifies the nodal power injection of a consumer. c_f denotes the heat capacity of water, while T_n^S and T_n^R indicate temperatures of node n in the supply and return networks, respectively. $\overline{M}_n^N$ and $\underline{M}_n^N$ represent the upper and lower bounds of nodal flow. The heat flow $\overline{M}_{nm}^S$ in the pipeline (n,m) is positive if the direction aligns with water flow and negative otherwise. In the proposed model, (4.15) is the equality constraints for nodal flow, while (4.16) describes the nodal power injection given the nodal flow. Equations (4.17) and (4.18) impose inequality constraints on

nodal flow and pipeline flow. Importantly, the constraints reveal bidirectional nodal flow and unidirectional water/heat flow within the pipeline.

4.2.4 Energy Devices Modeling

1) Combined heat and power (CHP)

CHP, a single-input multi-output energy converter, assumes a crucial part of the MNC-ICES model due to its high energy conversion efficiency from natural gas to electricity and heat [82, 83]. CHP is characterized by two constant energy conversion efficiencies for electricity and heat. The detailed operation model of CHP, depicted by a non-convex FOR enclosed by the boundary curve ABCDEFG, is adopted and shown in Fig. 4.3. $\mathrm{P}_{\mathrm{CHP}}^{\mathrm{t}}$, $\mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}}$ are generated power and heat for the CHP in time slot t. The FOR of the CHP is divided into two convex sections and is represented as follows [32].

$$\mathrm{P}_{\mathrm{CHP.n}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP.n}}^{\mathrm{B}} - \frac{\mathrm{P}_{\mathrm{CHP.n}}^{\mathrm{B}} - \mathrm{P}_{\mathrm{CHP.n}}^{\mathrm{C}}}{\mathrm{H}_{\mathrm{CHP.n}}^{\mathrm{B}} - \mathrm{H}_{\mathrm{CHP.n}}^{\mathrm{C}}} \times \left(\mathrm{H}_{\mathrm{CHP.n}}^{\mathrm{t}} - \mathrm{H}_{\mathrm{CHP.n}}^{\mathrm{B}}\right) \le 0, \forall t \in T \tag{4.19}$$

$$\mathrm{P}_{\mathrm{CHP}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP}}^{\mathrm{C}} - \frac{\mathrm{P}_{\mathrm{CHP}}^{\mathrm{C}} - \mathrm{P}_{\mathrm{CHP}}^{\mathrm{D}}}{\mathrm{H}_{\mathrm{CHP}}^{\mathrm{C}} - \mathrm{H}_{\mathrm{CHP}}^{\mathrm{D}}} \times \left(\mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}} - \mathrm{H}_{\mathrm{CHP}}^{\mathrm{C}}\right) \le 0, \forall t \in T \tag{4.20}$$

$$-\left(1 - \overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma \le \mathrm{P}_{\mathrm{CHP}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP}}^{\mathrm{E}} - \frac{\mathrm{P}_{\mathrm{CHP}}^{\mathrm{E}} - \mathrm{P}_{\mathrm{CHP}}^{\mathrm{F}}}{\mathrm{H}_{\mathrm{CHP}}^{\mathrm{E}} - \mathrm{H}_{\mathrm{CHP}}^{\mathrm{F}}} \times \left(\mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}} - \mathrm{H}_{\mathrm{CHP}}^{\mathrm{E}}\right), \forall t \in T \tag{4.21}$$

$$-\left(1 - \underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma \le \mathrm{P}_{\mathrm{CHP}}^{\mathrm{t}} - \mathrm{P}_{\mathrm{CHP}}^{\mathrm{D}} - \frac{\mathrm{P}_{\mathrm{CHP}}^{\mathrm{D}} - \mathrm{P}_{\mathrm{CHP}}^{\mathrm{E}}}{\mathrm{H}_{\mathrm{CHP}}^{\mathrm{D}} - \mathrm{H}_{\mathrm{CHP}}^{\mathrm{E}}} \times \left(\mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}} - \mathrm{H}_{\mathrm{CHP}}^{\mathrm{D}}\right), \forall t \in T \tag{4.22}$$

$$\overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}} + \underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}} = \mathrm{I}_{\mathrm{CHP}}^{\mathrm{t}}, \forall t \in T \tag{4.23}$$

$$-\left(1 - \underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma \le \mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}} - \mathrm{H}_{\mathrm{CHP}}^{\mathrm{E}} \le \left(1 - \overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) \times \Gamma, \forall t \in T \tag{4.24}$$

$$0 \le \mathrm{P}_{\mathrm{CHP}}^{\mathrm{t}} \le \mathrm{P}_{\mathrm{CHP}}^{\mathrm{A}} \times \mathrm{I}_{\mathrm{CHP}}^{\mathrm{t}}, \forall t \in T \tag{4.25}$$

$$0 \le \mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}} \le \mathrm{H}_{\mathrm{CHP}}^{\mathrm{A}} \times \mathrm{I}_{\mathrm{CHP}}^{\mathrm{t}}, \forall t \in T \tag{4.26}$$

In equations above, $\mathrm{P}_{\mathrm{CHP}}^{\mathrm{t}}$, $\mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}}$ are generated power and heat for the CHP in time slot t. As the region is described by a non-convex polygon, $\mathrm{P}_{\mathrm{CHP}}^{\mathrm{A}}$ and $\mathrm{H}_{\mathrm{CHP}}^{\mathrm{A}}$ indicate the power and heat output of the CHP at point A in the feasible region, and the same applied to the other points BCDEF. $\overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\left(\underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right)$ states the operating status in the first (second) convex section, when the CHP operate in the first (second) section, $\overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\left(\underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) = 1$, and $\underline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\left(\overline{\mathrm{X}}_{\mathrm{CHP}}^{\mathrm{t}}\right) = 0$. Γ denotes a sufficiently large number to assist model description, while $\mathrm{I}_{\mathrm{CHP}}^{\mathrm{t}}$ is the commitment status of the CHP. The total operation cost

of the CHP unit at time t can be expressed by equation (4.27), where a_{CHP}, b_{CHP}, c_{CHP}, d_{CHP}, e_{CHP} and f_{CHP} represent the cost coefficients.

$$C^t_{CHP}(P^t_{CHP}, H^t_{CHP}) = a_{CHP}{P^t_{CHP}}^2 + b_{CHP}P^t_{CHP} + c_{CHP} + d_{CHP}{H^t_{CHP}}^2 + e_{CHP}H^t_{CHP} + f_{CHP}P^t_{CHP}H^t_{CHP} \quad (4.27)$$

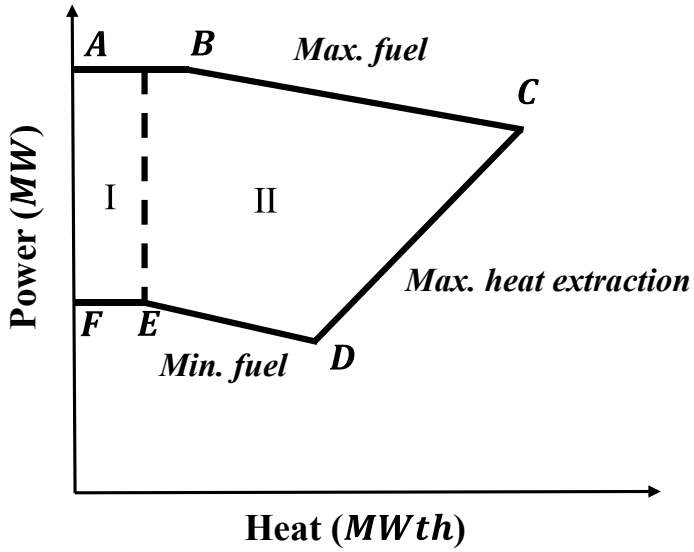


Fig. 4.3 Feasible operation region (FOR) of CHP units

2) Distributed energy resources (DER)

The power output of DER, denoted as P^t_{DER}, is defined in equation (4.28) by incorporating the power generation of PV and WT. The power generation function for DER accounts for power output uncertainty, modeled by probabilistic distribution functions for PV and WT, respectively.

$$P^t_{DER} = P^t_{PV} + P^t_{WT} \quad (4.28)$$

As variable renewable energy (VRE), wind power inherently carries high uncertainty. The wind speed (ω), directly influencing power output, is predicted with an unavoidable error $\Delta\omega$, which is modelled by a Weibull PDF [84]. The power output P^t_{WT} of WT is positive if and only if the wind speed exceeds the starting speed (ω^c_{in}); otherwise, P^t_{WT} is always zero. The upper limit for WT power is $P^t_{WT.rated}$ when $\omega^c_{rated} \leq \omega \leq \omega^c_{out}$. If the wind speed surpasses the cutout speed ω^c_{out}, WT will be cut out, resulting in $P^t_{WT} = 0$. Additionally, the Weibull PDF is employed to estimate the uncertainty parameter due to wind speed prediction errors. The wind speed (ω), directly influencing power output, is predicted with an unavoidable error $\Delta\omega$ in (4.29), which

is modelled by a Weibull PDF [84]. The power output P_{WT}^{t} of WT is modeled in equation (3.30), where it is positive if and only if the wind speed exceeds the starting speed (ω_{in}^{c}); otherwise, P_{WT}^{t} is always zero. The upper limit for WT power is $P_{WT.rated}^{t}$ when $\omega_{rated}^{c} \le \omega \le \omega_{out}^{c}$. If the wind speed surpasses the cutout speed ω_{out}^{c}, WT will be cut out, resulting in $P_{WT}^{t} = 0$. Additionally, the Weibull PDF is employed to estimate the uncertainty parameter due to wind speed prediction errors in (4.31)

$$\omega = \omega_{fs} + \Delta\omega \tag{4.29}$$

$$P_{WT}^{t}(\omega) = \begin{cases} 0, & \omega \le \omega_{in}^{c} \quad or \quad \omega \ge \omega_{out}^{c} \\ \dfrac{\omega + \omega_{in}^{c}}{\omega_{rated} + \omega_{in}^{c}} P_{WT.rated}^{t}, & \omega_{in}^{c} \le \omega \le \omega_{rated}^{c} \\ P_{WT.rated}^{t}, & \omega_{rated}^{c} \le \omega \le \omega_{out}^{c} \end{cases} \tag{4.30}$$

$$F_{\omega}(\Delta\omega;\lambda,k) = \begin{cases} \dfrac{k}{\lambda}\left(\dfrac{\Delta\omega + 0.5}{\lambda}\right)^{k-1} e^{-\left(\frac{\Delta\omega+0.5}{\lambda}\right)^{k}} & \Delta\omega \ge -0.5 \\ 0, & \Delta\omega < -0.5 \end{cases} \tag{4.31}$$

For photovoltaic power generation, the prediction error ΔI of PV is introduced in (4.32). PV generates electricity by converting solar radiation energy, and power generation is directly related to solar irradiance in (4.33). The Beta PDF is employed to estimate uncertain parameters with minimal error in (4.34).

$$I = I_{fs} + \Delta I \tag{4.32}$$

$$P_{PV}^{t} = \sum_{n \in N_{pv}} \eta_{pv_n} S_{pv_n} I^{t} \tag{4.33}$$

$$F_{S}(\Delta I;\alpha,\beta) = \frac{(\Delta I + 0.5)^{\alpha-1}\left(1 - (\Delta I + 0.5)\right)^{\beta-1}}{\int_{0}^{1} u^{\alpha-1}(1-u)^{\beta-1} du} \tag{4.34}$$

3) Energy Storage Systems (ESS)

ESS contains the EBS and TES for the energy storage of power and thermal energy, respectively. The EBS functions as a charge-dischargeable battery with varying efficiency [85]. The operational strategy of EBS is modeled at a granularity of one hour, i.e., one time interval. Charging and discharging operations are consolidated into a single activity within one time slot [86].

The detailed model of EBS is shown as follows.

$$E_{EBS}^{t} = (1 - \beta_{EBS}) E_{EBS}^{t-1} + P_{EBS.c}^{t} \eta_{EBS.c} - P_{EBS.d}^{t} \tag{4.35}$$

$$0 \le P_{EBS.c}^{t} \le S_{EBS.c}^{t} P_{EBS.max} \tag{4.36}$$

$$0 \leq P_{EBS.d}^{t} \leq S_{EBS.d}^{t} P_{EBS.max} \tag{4.37}$$

$$S_{EBS.c}^{t} + S_{EBS.d}^{t} \leq 1 \tag{4.38}$$

$$0 \leq E_{EBS}^{t} \leq E_{EBS.max} \tag{4.39}$$

In equations above, E_{EBS}^{t} is the battery capacity at time interval t. β and $\eta_{EBS.c}$ are predetermined parameters representing the loss factor and charging efficiency, respectively. $P_{EBS.c}^{t}$ and $P_{EBS.d}^{t}$ represents the charging power and discharging power at time step t, respectively. $S_{EBS.c}^{t}$ and $S_{EBS.d}^{t}$ represent the charging state and discharging state at time step t, respectively. $P_{EBS.c.max}$ and $P_{EBS.d.max}$ are the maximum charging and discharging power, respectively. $E_{EBS.min}$ and $E_{EBS.max}$ represent the upper and lower limits of battery capacity, respectively.

In the model above, the representation of the SoC is shown in (4.35). The maximal charge and discharge power are constrained by (4.36) and (4.37), respectively. (4.38) is employed to determine the charge of discharge state of EBS. (4.39) constraints the range of total capacity of the energy in EBS.

In the model of EBS, E_{EBS}^{t} is the battery capacity at time interval t. β and $\eta_{EBS.c}$ are predetermined parameters representing the loss factor and charging efficiency, respectively. $P_{EBS.c}^{t}$ and $P_{EBS.d}^{t}$ represents the charging power and discharging power at time step t, respectively. $S_{EBS.c}^{t}$ and $S_{EBS.d}^{t}$ represent the charging state and discharging state at time step t, respectively. $P_{EBS.c.max}$ and $P_{EBS.d.max}$ are the maximum charging and discharging power, respectively. $E_{EBS.min}$ and $E_{EBS.max}$ represent the upper and lower limits of battery capacity, respectively. TES has a similar model to EBS. Please refer to Appendix for the detailed description.

A generalized energy storage system model is applied to address TES. This model aligns with that of EBS and is not detailed here for the sake of brevity.

$$E_{TES}^{t} = (1-\beta)E_{TES}^{t-1} + H_{TES.c}^{t}\eta_{TES.c} - H_{TES.d}^{t} \tag{4.40}$$

$$0 \leq H_{TES.c}^{t} \leq S_{TES.c}^{t} H_{TES.max} \tag{4.41}$$

$$0 \leq H_{TES.d}^{t} \leq S_{TES.d}^{t} H_{TES.max} \tag{4.42}$$

$$S_{TES.c}^{t} + S_{TES.d}^{t} \leq 1 \tag{4.43}$$

$$0 \leq E_{TES}^{t} \leq E_{TES.max} \tag{4.44}$$

4.2.5 Multi-Energy User (MEU) Modeling

As rational integrated energy consumers, MEUs engage in the procurement of energy to fulfill their energy demands for both power and heat [87]. MEUs are conceptualized to possess elastic electricity consumption appliances and energy conversion devices, including EB and GB, which enable them to adjust energy consumption dynamically across various time periods and multiple energies.

$$E_{MEU}\left(P_{i,e}^{t}\right)=\begin{cases}\omega_i^t-\dfrac{\lambda_i^t}{2}\left(P_{i,e}^{t}\right)^2, 0\le P_{i,e}^{t}\le\dfrac{\omega_i^t}{\lambda_i^t}\\ \dfrac{(\omega_n^t)^2}{2\lambda_i}, P_{i,e}^{t}>\dfrac{\omega_i^t}{\lambda_i}\end{cases} \tag{4.45}$$

$$H_{MEU}\left(H_{i,eb}^{t},H_{i,gb}^{t},H_{i,h}^{t}\right)=-\sigma_i^t\left(H_{i,eb}^{t}+H_{i,gb}^{t}+H_{i,h}^{t}\right)^2+\varsigma_i^t\left(H_{i,eb}^{t}+H_{i,gb}^{t}+H_{i,h}^{t}\right) \tag{4.46}$$

In (4.45), $P_{i,e}^{t}$ is the power consumption of MEU i during time interval t. ω_i and λ_i are preset parameters reflecting the preference of MEU in energy consumption. In (4.46), $H_{i,eb}^{t}, H_{i,gb}^{t}$ and $H_{i,h}^{t}$ are heat power from EB, GB, and ICESO, respectively. Similarly, σ_i^t and ς_i^t are preset parameters. By considering the utility above, the objective function of MEUs is modelled by (4.47) and is constrained by (4.48)-(4.52).

$$\max_{P_{i,e}^{t},P_{i,eb}^{t},P_{i,gb}^{t},H_{i,h}^{t}} U_{MEU}=\sum_{t=1}^{T}\left\{\begin{array}{l}\underbrace{\left(E_{MEU}\left(P_{i,e}^{t}\right)+H_{MEU}\left(H_{i,eb}^{t},H_{i,gb}^{t},H_{i,h}^{t}\right)\right)}_{Utility\ for\ energy\ consumption}\\ -\underbrace{\left(x_{r,e}^{t}\left(P_{i,e}^{t}+P_{i,eb}^{t}\right)+x_{r,g}^{t}P_{i,gb}^{t}+x_{r,h}^{t}P_{i,h}^{t}\right)}_{Cost\ for\ energy\ purchase}\end{array}\right\} \tag{4.47}$$

$s.t.$

$$H_{i,eb}^{t}=\eta_{EB,i}P_{i,eb}^{t} \tag{4.48}$$

$$H_{i,gb}^{t}=\eta_{GB,i}P_{i,gb}^{t} \tag{4.49}$$

$$0\le P_{i,e}^{t}\le P_{i,e.max}^{t} \tag{4.50}$$

$$0\le P_{i,eb}^{t}\le P_{i,eb.max}^{t} \tag{4.51}$$

$$0\le P_{i,gb}^{t}\le P_{i,gb.max}^{t} \tag{4.52}$$

In these equations, $P_{i,eb}^{t}$ and $P_{i,gb}^{t}$ are power and gas consumed by EB and GB of MEU i during time interval t, respectively. $\eta_{EB,i}$ and $\eta_{GB,i}$ are the energy conversion rates of EB and GB for MEU i. $P_{i,e.max}^{t}$, $P_{i,eb.max}^{t}$ and $P_{i,gb.max}^{t}$ are the upper bounds for the corresponding power consumption of electric appliance, power consumption of EB, and gas consumption of GB.

Equation (4.47) is the objective function of the MEU with the decision variables of $P_{i,e}^t, P_{i,eb}^t, P_{i,gb}^t, H_{i,h}^t$. In (4.47), the first term is the utility for integrated energy consumption, and the second term is the cost for integrated energy purchase from ICESO. Equations (4.48) and (4.49) are the equality constraints for the power conversion of EB and GB. Inequalities (4.50)-(4.52) are constraints for electricity consumption, and power input for EB and GB, respectively.

4.3 Problem Formulation

This section presents the multi-networks constrained operational optimization problem for the ICESO and reformulates it into a corresponding C-MDP for the implementation of Safe RL algorithm. Specifically, the cost term in the C-MDP is denoted by the network constraint violations. By solving this C-MDP, the ICESO can maximize its reward with the tolerated constraint violation.

4.3.1 Objective Function and Constraints

The profit of ICESO is mainly the difference between the revenue for selling energy to MEUs and the cost of energy purchasing, as well as the imbalance penalty. The corresponding objective function is presented in (4.53).

$$\max_{\varphi} U_{IESP} = \sum_{\mathrm{t}=1}^{\mathrm{T}} \left\{ \begin{array}{l} \underbrace{\left(x_{r,e}^t \sum_{i=1}^{N} P_{i,e}^t + x_{r,g}^t \sum_{i=1}^{N} P_{i,g}^t + x_{r,h}^t \sum_{i=1}^{N} P_{i,h}^t\right)}_{Revenue\ for\ selling\ energy} - \\ \underbrace{\left(x_{w,e}^t P_{w,e}^t + x_{w,g}^t P_{w,g}^t\right)}_{Cost\ for\ energy\ pruchase} - \underbrace{\left(\delta_e^t P_{imb.e}^t + \delta_g^t P_{imb.g}^t + \delta_h^t H_{imb}^t\right)}_{Cost\ for\ energy\ balance} \end{array} \right\} \quad (4.53)$$

$$s.t.\ \forall t \in T$$

$$(4.1) - (4.52)$$

$$P_{w.e}^t + P_{DER}^t + \sum_{n \in N_e} P_{CHP.n.e}^t + P_{EBS.d}^t - P_{EBS.c}^t + P_{imb.e}^t = \sum_{i \in I_n} P_{i.e}^t \quad (4.54)$$

$$\sum_{n \in N_h} H_{CHP,n}^t + H_{TES.d}^t - H_{EBS.c}^t + H_{imb}^t = \sum_{i \in I_n} H_i^t \quad (4.55)$$

$$P_{w.g}^t + P_{imb.g}^t = \sum_{i \in I_n} P_{i.g}^t + \sum_{n \in N_g} P_{CHP.g}^t \quad (4.56)$$

$$x_{min.e}^t \le x_{r.e}^t \le x_{max.e}^t \quad (4.57)$$

$$x_{min.g}^{t} \leq x_{r.g}^{t} \leq x_{max.g}^{t} \tag{4.58}$$

$$x_{min.h}^{t} \leq x_{r.h}^{t} \leq x_{max.h}^{t} \tag{4.59}$$

In (4.53), $\varphi = \left\{x_{r,e}^{t}, x_{r,g}^{t}, x_{r,h}^{t}, P_{CHP.e}^{t}, P_{EBS.d}^{t}, P_{EBS.c}^{t}, H_{CHP}^{t}, H_{TES.d}^{t}, H_{TES.c}^{t}\right\}$ is the set of decision variables, and several decision variables are omitted due to the energy balance among several variables. The objective function in (3.53-a) constitutes three parts, revenue for selling energy, cost for energy purchase, and cost for energy balance, where $P_{imb.e}^{t}, H_{imb}^{t}, P_{imb.g}^{t}$ are the imbalanced electricity, heat and natural gas for ICESO. Penalty indexes δ_{e}^{t}, δ_{g}^{t}, δ_{h}^{t} are preset parameters to penalize the energy imbalance and determined based on energy prices. Also, the objective is constrained by (4.1)-(4.52) and (4.54)-(4.59). Equality constraints (4.54)-(4.56) indicate the integrated energy balance. (4.57)-(4.59) are inequality constraints for the retail energy prices, where $x_{max.e}^{t}$, $x_{min.e}^{t}$, $x_{max.g}^{t}$, $x_{min.g}^{t}$, $x_{max.h}^{t}$, $x_{min.h}^{t}$ are preset parameters indicating the upper bounds and lower bounds for power, natural gas and heat, respectively.

The energy balance constraints are actually relaxed by introducing the penalty terms δ. However, network constraints are not directly relaxed to the objective function, as penalties for network constraint violations are hard to determine. Specifically, compared with energy imbalance that only decreases the profits from the economic perspective, the violation of network constraints is more serious and may affect the safe operation of the ICES. Moreover, determining penalties for network constraint violation to realize a fair tradeoff between improving profits and reducing violations is not straightforward. Therefore, it is assumed that the ICESO aims to guarantee safe operation rather than uplift the economic revenue. Consequently, the network-constrained operational optimization problem is formulated to C-MDP in the next subsection.

Moreover, it is worth noting that violating safety constraints in an ICES has immediate physical, operational, and regulatory consequences in practice. In the power network, exceeding line/transformer ampacity or voltage limits causes overheating, accelerated insulation aging, protection miscoordination, inverter trips, and potentially feeder outages. In the gas network, breaching pressure/flow bounds risks compressor

surge, line-pack depletion, and, in extreme cases, pipeline damage or supply interruption. In the district heating network, violating temperature/pressure/flow constraints leads to comfort violations, pump cavitation, thermal stress and leaks, or safety-valve discharge. Because these carriers are coupled (e.g., CHP, boilers), a violation in one layer can cascade to others.

4.3.2 Markov Decision Process (MDP)

To optimize the decision-making process of ICESO, a MDP is leveraged to describe the integrated energy transactions and then a DRL algorithm is used to solve it. This approach treats the ICESO as an intelligent agent that makes decisions based on the environmental observation of wholesale market prices (both electricity and gas), and power generation of DER. The objective is to improve the pricing decisions by maximizing the accumulated return, using a well-defined reward function in (4.53). The MDP can be denoted by $< S, A, R, P, \mu, \gamma >$. S is the set of states. $S = \{x_{w,e}^t, x_{w,g}^t, P_{WT,predict}^t, P_{PV,predict}^t, E_{EBS}^t, E_{TES}^t\}$, encompassing electricity market price, natural gas market price, forecast power generation of WT and PV, state of charge of EBS and TES. A is the set of actions. $A = \{x_{r,e}^t, x_{r,g}^t, x_{r,h}^t, \mathrm{P}_{\mathrm{CHP}}^{\mathrm{t}}, \mathrm{H}_{\mathrm{CHP}}^{\mathrm{t}}, \mathrm{P}_{\mathrm{EBS.c}}^{\mathrm{t}}, \mathrm{P}_{\mathrm{EBS.d}}^{\mathrm{t}}, \mathrm{H}_{\mathrm{TES.c}}^{\mathrm{t}}, \mathrm{H}_{\mathrm{TES.d}}^{\mathrm{t}}\}$ represents the available actions as the decision variables in (4.53). $R: S \times A \times S \mapsto \mathbb{R}$ is the reward function, which quantifies the action's performance and is presented by the objective function. $P: S \times A \times S \mapsto [0, 1]$ is the transition probability function. The state transition function is not considered due to the assumption of uncoupled state across time periods. $\mu: S \mapsto P(A)$ represents the policy of the agent, mapping from states to a probability distribution over actions [88]. $\gamma \in [0, 1]$ is the discount factor to discount the future reward.

The discounted accumulative reward under policy μ is denoted as (4.60). In (4.60), $\tau = (s_0, a_0, s_1, a_1 \cdots)$ is a trajectory of the agent with a series of actions, and $\tau \sim \pi$ indicate trajectories distribution under policy. To conclude, the aim of MDP is to find

the optimal policy μ^* that can maximize the discounted accumulative reward $R(\mu)$, as (4.61).

$$R(\mu) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t, a_t, s_{t+1})\right] \tag{4.60}$$

$$\mu^* = \arg\max_{\mu} R(\mu) \tag{4.61}$$

4.3.3 Constrained-Markov decision process (C-MDP)

To maintain the energy flow complying with the network constraints, a cost function is proposed for the C-MDP, indicating the violation of network constraints. The C-MDP can be denoted as $< S, A, R, C, P, \mu, \gamma >$, which is an ordinary MDP augmented by cost function $C(s, a)$. The cost function is denoted by $C: S \times A \times S \mapsto \mathbb{R}$, mapping from transition tuples to cost. The explicit expression of the cost function is given in (4.62). It comprises the standardized constraint violation of cost in three kinds of network constraints. As the transmission capacity is always designed to be large enough to carry the real and reactive power in the electric distribution network, only voltage constraints are considered in the following cost function.

$$C = \left\{ \begin{array}{l} \underbrace{\sum_{n\in N_e,\forall(n,m)\in P_e}\left[\left[\frac{V_{n,t}-\overline{V}_n}{\overline{V}_n}\right]^+ + \left[\frac{\underline{V}_n - V_{n,t}}{\underline{V}_n}\right]^+\right]}_{Cost\ for\ constraints\ violation\ in\ electricity\ network} + \\ \underbrace{\sum_{n\in N_g,\forall(n,m)\in P_g}\left[\left[\frac{|gf_{k,mn}|-\overline{gf}_{mn}}{\overline{gf}_{mn}}\right]^+ + \left[\frac{Pr_{k,n}-\overline{Pr}_n}{\overline{Pr}_n}\right]^+ + \left[\frac{\underline{Pr}_n - Pr_{k,n}}{\underline{Pr}_n}\right]^+ + \left[\frac{G_{k,n}-\overline{G}_n}{\overline{G}_n}\right]^+\right]}_{Cost\ for\ constraints\ violation\ in\ gas\ network} + \\ \underbrace{\sum_{n\in N_h,\forall(n,m)\in P_h}\left[\left[\frac{M_n^t-\overline{M}_n^N}{\overline{M}_n^N}\right]^+ + \left[\frac{\underline{M}_n^N - M_n^t}{\underline{M}_n^N}\right]^+ + \left[\frac{M_{nm}^t - M_{nm}^S}{\overline{M}_{nm}^S}\right]^+\right]}_{Cost\ for\ constraints\ violation\ in\ heat\ network} \end{array} \right\} \tag{4.62}$$

In (4.62), $[x]^+ = \max\{0, x\}$ is the projection function. The cost function is constituting costs for constraint violation in the electricity network, gas network, and heat network. To limit the constraint violation, the constraint for the cost function is proposed as (4.63), where d is the upper bound of the cost function.

$$C(\mu) \le d \tag{4.63}$$

The long-term discounted cost under policy μ is similarly defined as $C(\mu) = E_{\tau\sim\mu}[\sum_{t\in T}\gamma_t\, C(s_t, a_t, s_{t+1})]$, and the corresponding limit is d. In the C-MDP, the

goal is to select a policy μ that maximizes the long-term reward R(π) while satisfying the constraints on the long-term costs.

$$\mu^* = \arg\max_{\mu} R(\mu) \quad s.t.(9) \tag{4.64}$$

4.4 Proposed TD3 algorithm

In this section, a PD-TD3 algorithm is developed to solve the proposed C-MDP and learn the optimal operational strategy for ICESO. Specifically, the proposed C-MDP is formulated into a Lagrangian function, which is then converted to an unconstrained min-max problem and thus applicable to the solution of the iterative primal-dual TD3 algorithm. The PD-TD3 algorithm then solves the primal-dual problem by using the gradient descent to iteratively update the policy and Lagrangian multiplier.

4.4.1 Primal-Dual TD3 algorithm

The challenges of optimal operation of MNC-ICES model mainly come from the non-linear integrated network constraints. Conventional deep reinforcement learning algorithms do not directly consider these constraints in the learning process [88]. Moreover, traditional DRL algorithms still face the problem of overestimation of Q-value and cost value in C-MDP and instability during the training process. To overcome these drawbacks, the proposed PD-TD3 algorithm is able to address the challenges of constrained optimal operation problem in MNC-ICES model by solving the C-MDP, and mitigating the issue of value overestimation and training instability.

As a RL method, the key of the primal-dual algorithm is to augment constraints on the expected rewards, such that the training of the RL agent converges to the optimal constraints-satisfying policies. Therefore, the objective of primal-dual TD3 for cost minimization can be generally written as (3.65), where $\mathcal{L}(\mu, \lambda)$ as (4.66) is the augmented objective action-value function, λ denotes the multipliers of constraints.

$$(\mu^*, \lambda^*) = \arg\min_{\lambda>0} \max_{\mu} \mathcal{L}(\mu, \lambda) \tag{4.65}$$

$$\mathcal{L}(\mu, \lambda) = R(\mu) - \sum_{i} \lambda(C(\mu) - d) \tag{4.66}$$

In (4.66), $R(\mu)$ and $C(\mu)$ represent the reward and the cost for constraint violation of a DRL agent. For constrained optimal operation of MNC-ICES model, the reward and constraint violation can be the total profits of the ICESO and violation of physical constraints of integrated distribution networks, respectively. To solve the unconstrained minimax problem (4.65), the iterative primal-dual method is used as a canonical approach where in each iteration. In each iteration, the primal policy π and the dual variable λ are updated in turn. The primal-dual update procedures at iteration k are as follows:

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta \big(\mathcal{L}(\mu(\theta), \lambda_k)\big)|_{\theta=\theta_k} \tag{4.67}$$

$$\lambda_{k+1} = f_k\big(\lambda_k, \mu(\theta)\big) \tag{4.68}$$

In the proposed PD-TD3 algorithm, the primal variable, i.e., policy parameters, is updated by policy gradient, which is specified later. The dual variable, i.e., the Lagrangian multiplier, is updated by (4.69). In (4.69), β_k is the step size of the multiplier update. $[x]^+ = \max\{0, x\}$ is the projection onto the dual space $\lambda^k > 0$. Note that the dual variable is updated in a stable manner via (4.69), which applies a tunable step size for gradual adjustments

$$\lambda_{k+1} = [\lambda_k + \beta_k(C(\mu_k) - d)]^+ \tag{4.69}$$

4.4.2 Algorithm Design for Primal-Dual TD3

The proposed PD-TD3 algorithm is an off-policy DRL algorithm, enabling offline training of strategies in optimization problems and using DNN approximate action-value functions. The overall framework of the PD-TD3 is summarized in Algorithm 4.1. As a DRL algorithm based on the actor-critic framework, the PD-TD3 adopts DNNs to approximate the value functions and policy functions of the C-MDP, which denotes the critic and actor, respectively. To estimate both the reward and cost in the C-MDP, PD-TD3 employs two kinds of critic networks, namely the reward critic network and the cost critic network. Additionally, as PD-TD3 uses the trick of double networks, each type of critic consists of two online Q networks and their target networks,

mitigating the issue of Q-value overestimation observed in other value-based RL algorithms. Also, the target networks of the critic are delayed copies of the online network, which is supposed to mitigate the instability of the training process. Therefore, three sets of neural networks are employed: (1) two reward critic Q-networks $Q_{R1}(s,a|\theta_{R1}^{Q})$, $Q_{R2}(s,a|\theta_{R2}^{Q})$ and their target network $Q'_{R1}\left(s,a|\theta_{R1}^{Q'}\right)$, $Q'_{R2}\left(s,a|\theta_{R2}^{Q'}\right)$, (2) two cost critic Q-networks $Q_{C1}(s,a|\theta_{C1}^{Q})$, $Q_{C2}(s,a|\theta_{C2}^{Q})$ and their target networks $Q'_{C1}\left(s,a|\theta_{C1}^{Q'}\right)$, $Q'_{C2}\left(s,a|\theta_{C2}^{Q'}\right)$, and (3) the actor policy network $\mu(s|\theta^{\mu})$ and its target network $\mu'(s|\theta^{\mu'})$.

During the training process, the agent randomly samples transitions $(s_i, a_i, r_i, c_i, s_{i+1})$ from the ERB to form a mini-batch N for experience replay learning. Then, the target of the reward and cost critic networks are presented as (4.70) and (4.71), which are employed to update Q-functions.

$$y_i = r_i + \gamma \min_{j\in\{1,2\}} Q\left(s_{i+1}, \tilde{a}_{i+1}|\theta_{R_j}^{Q'}\right) \tag{4.70}$$

$$z_i = c_i + \gamma \min_{j\in\{1,2\}} Q\left(s_{i+1}, \tilde{a}_{i+1}|\theta_{C_j}^{Q'}\right) \tag{4.71}$$

In (4.70) and (4.71), $\tilde{a}_{t+1}$ is the clipped target action shown in (4.72). Here, target policy smoothing is employed by incorporating clipped Gaussian noise into the target action during the evaluation process. This technique promotes smoother and more stable policy updates, facilitating convergence and enhancing the quality of the learned policy.

$$\tilde{a}_{i+1} = \mu'(s|\theta^{\mu'}) + \tilde{\varepsilon}, \tilde{\varepsilon}\sim clip(\mathcal{N}(0,\tilde{\sigma}), -c, c) \tag{4.72}$$

Based on the target, the reward and cost critic networks are updated by minimizing the loss function, i.e., MSE between the value functions and their targets, proposed in (4.73) and (4.74), respectively.

$$L_R = \frac{1}{N}\sum_{i\in N}\left[y_i - Q_R(s_i, a_i|\theta_R^{Q})\right]^2 \tag{4.73}$$

$$L_C = \frac{1}{N}\sum_{i\in N}\left[z_i - Q_C(s_i, a_i|\theta_C^{Q})\right]^2 \tag{4.74}$$

To mitigate the training error caused by correlated samples, the primal variable, i.e., policy network, and dual variable, i.e., Lagrangian multiplier after a fixed number *e* of iterations by using (4.75) and (4.76), which is the so-called “delayed” update. This

delay in primal and dual variable updates reduces the correlation between successive updates and prevents rapid forgetting of previously learned policies. The policy is updated by one step of sampled gradient descent using (4.77). Also, it should be noted that the dual variable updated in (4.78) uses the minimized estimated Q-value for cost, alleviating the overestimation of the Q value to ensure a proper update descent.

$$\nabla_{\theta^\mu}\mathcal{L}(\theta^\mu,\lambda) \approx \frac{1}{N}\sum_{i\in N}\nabla_{\theta^\mu}\left[Q_{R1}\left(s_i,\mu(s_i|\theta^\mu)|\theta_{R1}^Q\right)-\lambda Q_{C1}\left(s_i,\mu(s_i|\theta^\mu)|\theta_{C1}^Q\right)\right] \quad (4.77)$$

$$\nabla_\lambda\mathcal{L}(\theta^\mu,\lambda) = \frac{1}{N}\sum_{i\in N}\left[\min_{j\in\{1,2\}} Q\left(s_{i+1},\tilde{a}_{i+1}|\theta_{C_j}^{Q'}\right)-d\right] \quad (4.78)$$

Additionally, all target networks of the actor and critic are updated by using the soft update presented in (4.79) and (4.80). It allows a small pace update in each iteration and ensures a gradual and stable convergence of the networks, where ρ represents the soft update parameter.

$$\theta^{Q'} \leftarrow \rho\theta_Q + (1-\rho)\theta^{Q'} \quad (4.79)$$

$$\theta^{\mu'} \leftarrow \rho\theta_\mu + (1-\rho)\theta^{\mu'} \quad (4.80)$$

4.4.3 Discussion of Potential Limitations

Previous subsections address the Q-value overestimation problem in the typical RL algorithms. Considering the PD-TD3 is developed based on the conventional TD3 algorithm, it also inherits the following drawbacks: 1) The PD-TD3 algorithm is more complex than the TD3 algorithm and requires more computing resources by increasing numbers of hyperparameters. 2) The TD3-based algorithm is relatively sensitive to the selection of hyperparameters. However, as the ultimate goal is to deploy this well-trained algorithm to online dispatch, this could not be a serious problem.

If this algorithm is deployed to real-world ICES for online dispatch, an important assumption is that, the environment (state transition) of the simulation (test system) should be similar to the real-world ICES. Otherwise, the algorithm will generate unsafe decisions because it cannot be adaptive to the unknown environment. Potential solutions are twofold. First, a comprehensive modelling of system state transition using

advanced deep learning methods is necessary. Second, the output of this RL algorithm should be corrected by real-time control algorithms, such as model predictive control.

Algorithm 4.1 PD-TD3 algorithm

1: Initialize policy parameters θ^{μ}, Q-function parameters θ_{R1}^{Q}, θ_{R2}^{Q}, θ_{C1}^{Q}, θ_{C2}^{Q}, and empty buffer R

2: Initialize target networks $\theta^{\mu'} \leftarrow \theta^{\mu}$, $\theta_{R1}^{Q'} \leftarrow \theta_{R1}^{Q}$, $\theta_{R2}^{Q'} \leftarrow \theta_{R2}^{Q}$, $\theta_{C1}^{Q'} \leftarrow \theta_{C1}^{Q}$, $\theta_{C2}^{Q'} \leftarrow \theta_{C2}^{Q}$

3: Initialize Lagrangian multiplier λ

4: **repeat**

5: Initialize a random process N for action exploration

6: Receive initial state s_0

7: **for** each transaction time slot $t = 1, \dots, T$ **do**

8: Select action

9: Execute action and observe

10: Store transition in the reply buffer

11: if s is terminal, reset environment state

12: Randomly sample a bath of transitions from N

13: Compute target actions using

$$\tilde{a}_{i+1} = \mu'(s|\theta^{\mu'}) + \tilde{\varepsilon}, \tilde{\varepsilon} \sim clip(\mathcal{N}(0, \tilde{\sigma}), -c, c)$$

14: Compute target using

$$y_i = r_i + \gamma \min_{j \in \{1,2\}} Q\left(s_{i+1}, \tilde{a}_{i+1} | \theta_{R_j}^{Q'}\right)$$

$$z_i = c_i + \gamma \min_{j \in \{1,2\}} Q\left(s_{i+1}, \tilde{a}_{i+1} | \theta_{C_j}^{Q'}\right)$$

15: Update Q-function by one step of gradient descent by minimizing

$$L_r = \frac{1}{N} \sum_{i \in N} \left[y_i - Q_r\left(s_i, a_i | \theta_R^Q\right)\right]^2$$

$$L_c = \frac{1}{N} \sum_{i \in N} \left[z_i - Q_c\left(s_i, a_i | \theta_C^Q\right)\right]^2$$

16: **if** k mod e **then**

17: Update policy by one step of gradient ascent using

$$\nabla_{\theta^{\mu}} \mathcal{L}(\theta^{\mu}, \lambda) \approx \frac{1}{N} \sum_{i \in N} \nabla_{\theta^{\mu}} \left[Q_{R1}\left(s_i, \mu(s_i|\theta^{\mu}) | \theta_{R1}^{Q}\right) - \lambda Q_{C1}\left(s_i, \mu(s_i|\theta^{\mu}) | \theta_{C1}^{Q}\right)\right]$$

18: Update Lagrangian multiplier by one step of gradient ascent using

$$\nabla_\lambda \mathcal{L}(\theta^\mu, \lambda) = \frac{1}{N}\sum_{i \in N}\left[\min_{j\in\{1,2\}} Q\left(s_{i+1}, \tilde{a}_{i+1} | \theta_{C_j}^{Q'}\right) - d\right]$$

19: Update target networks with

$$\theta^{Q'} \leftarrow \rho\theta_Q + (1-\rho)\theta^{Q'}$$

$$\theta^{\mu'} \leftarrow \rho\theta_\mu + (1-\rho)\theta^{\mu'}$$

20: **end if**

21: **end for**

22: **until** convergence

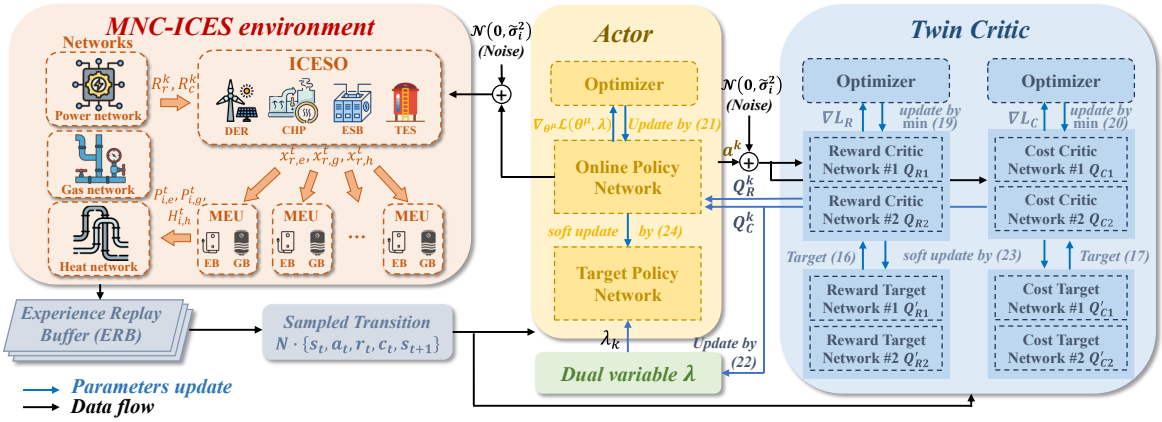


Fig. 4.4 Illustration of the proposed PD-TD3 algorithm

4.5 Case Study

To validate the performance of the proposed PD-TD3 algorithm, a test system consisting of 5 MEUs is adopted and is shown in Fig. 4.5. As shown in the test system, a standard IEEE-33 bus electricity network, a heat network, and a natural gas network are considered to model the whole integrated energy network structure in MNC-ICES. It should be mentioned that simulation based on the test system is a generalized scenario but not a representation of a specific real-world application. The numerical results obtained from the proposed test system only serve as a demonstration of the proposed models and methods and a foundation for further application. To bridge the gap between generalized scenarios and real-world applications of the ICES models and Machine Learning methods, future work will involve applying the developed methods to actual energy systems models or more detailed case studies.

It should be mentioned that simulation based on the test system is a generalized scenario but not a representation of a specific real-world application. The proposed algorithm is only tested on a small system to validate its functionality and compared performance to benchmark algorithms. The numerical results obtained from the proposed test system only serve as a demonstration of the proposed models and methods and a foundation for further application. To bridge the gap between generalized scenarios and real-world applications of the ICES models and Machine Learning methods, future work will involve applying the developed methods to actual energy systems models or more detailed case studies.

The key parameter settings for the test system are given as follows. The voltage constraints of the electricity network are set with an upper bound of 0.9p. u. and lower bound of 1.1p. u., and the other configuration data for the power network is taken from [89]. The natural gas network transmits the gas in the pipeline with an inside diameter of 0.3m and an efficiency of 90%, operating at a temperature of $288.15°\text{K}$. The gas compressibility factor is set as $0.9Pa^{-1}$. The allowable pressure for gas transmission is limited from $110kPa\ to\ 100kPa,$ and the maximal gas flow rate is $400m^3/h$. In the 10-node district heat system, the supply temperature at the most upstream node is $70℃$, and the return side temperature of the most downstream node is set to $30℃$. The temperature loss is assumed to be 0.1K/m on the supply side, and 0.05K/m in return side pipelines [86].

The hourly wholesale prices for electricity and natural gas are obtained using the real data of New England ISO. The constant natural gas price is set at $4.75£/\text{MMBtu}$, resulting in a natural gas price of approximately $16.2\ £/\text{MW}$ or $0.165\ £/\text{m}^3$. Moreover, the power output of WT and PV are adapted from real-world data in [90]. The pricing ranges for electricity, natural gas, and heat in the MNC-ICES are set as $0-50£/\text{MW}$, $0-50£/\text{MW}$, and $0-40£/\text{MW}$, respectively.

The proposed DRL algorithm is implemented on the Pytorch [91]. The neural networks are configured with the settings shown in Table 4.1. The hyperparameters of the algorithm shown in Table 4.2 are selected based on empirical values and adjusted

during the training process until the algorithm converges to the maximum profit. The quadratic programming problem for the comprehensive energy demand response of the MEUs is solved using commercial solver.

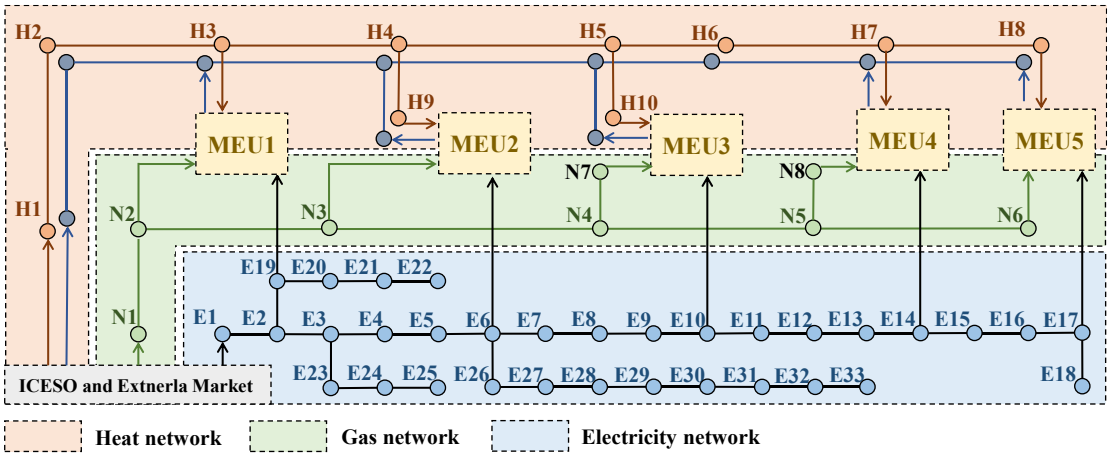


Fig. 4.5. Test system of integrated community energy system

Table 4.1 Neural network architectures settings

Neutral Networks	Actor	Critic
No. of Hidden Layers	3	2
No. of Neurons	[128,32]	[128,32]
Activation Function	tanh	ReLU
Learning Rate	4e-4	7e-4
Soft update parameter	1e-3	1e-3
Delayed update frequency	2	2
Optimizer	Adam optimizer	Adam optimizer

Table 4.2 Hyperparameter settings of the PD-TD3 algorithm

Training parameters	Parameters
Replay Buffer Size	1e+6
Replay Start Size	128
Batch Size	128
Discount Factor	0.99

4.5.1 Training Performance

This subsection aims to validate the convergence performance of the PD-TD3. Safe RL algorithms using both the Lagrangian method and the direct penalty method are employed as benchmark algorithms. The L-SAC [65] and S-DDPG [68] belong to the former, while typical TD3 with direct penalizing cost stands for the latter. The penalty index λ is set as 1, 10, 100, and 1000 for a comprehensive comparison. In this context, each algorithm is trained 1000 episodes to learn the optimal strategy in pricing and scheduling in MNC-ICES, while each episode contains 24 steps, indicating 24 hours per day. Figs. 4.6-4.7 present the evolution of cumulative reward and cost for each episode, respectively. The corresponding values are also listed in Table 4.3 for a clearer demonstration. The allowable operating range of the cumulative cost is 0~10, and the upper bound of 10 is marked in black in Fig. 4.7.

As illustrated in Fig. 4.6, the cumulative reward of the four algorithms has a similar trend, which fluctuates a lot in the initial stage of training, since the algorithm randomly chooses actions to explore the environment. Similarly, the initial cost for constraint violation in Fig. 4.7 is relatively higher and fluctuates in the initial stage. With the learning process going on, the policy is continuously trained and improved, resulting in an increasing trend in reward and a decreasing trend in cost. In the comparison between the algorithms using the Lagrangian-based method and direct penalty method, it can be observed that typical TD3 algorithms with fixed penalty index usually have lower reward and higher cost. The reward and cost of TD3 decrease as λ increases. Specifically, TD3 with $\lambda = 1000$ has the lowest reward among all algorithms even L-SAC, and the lowest cost among TD3 with all λ settings. Moreover, its cost reaches around the allowable requirement but is still unqualified. This demonstrates a worse performance of the direct penalty-based Safe RL algorithms compared to Lagrangian-based Safe RL algorithms.

In the comparison within algorithms using the Lagrangian Safe RL method, it can be observed that the L-SAC converges with the lowest cumulative reward, which is

nearly zero. This is thought as a local optimum and caused by the improper tradeoff between the reward and cost, indicating an over-conservative policy of L-SAC. However, the PD-TD3 shows a fast convergence to the highest reward among the three algorithms, and it can be observed in both Fig. 4.6 and Table 4.3 that the cumulative reward is about to converge around 200 episodes with a reward of almost 10000. This is driven by the delayed policy update that can update the policy without the training noise, training the policy networks effectively. In addition, the PD-TD3 deals with the physical constraints by directly adjusting the policy of the actor-network. On this account, the cost of PD-TD3 for an episode containing 24 operating hours is in the allowable range of 0~10 after 500 episodes. In comparison, the cost of S-DDPG is out of the allowable range during almost the whole training process, while the cost of L-SAC is about 0 and is thought of as over-conservative. This is owing to the double Q cost networks that assist in estimating the cost more precisely by eliminating the overestimation of cost and thus achieving a fair balancing of the tradeoff between the reward and cost, which has the highest reward and an allowable cost. Furthermore, it can be observed that the reward of PD-TD3 converges around 200 episodes, and the cost is operating in the allowable range after about 500 episodes. The convergence speed of PD-TD3 is similar to L-SAC but is much faster than S-DDPG twice. Also, the convergence process of the reward and cost in PD-TD3 demonstrates that the dual variable converges to optimal after the convergence of reward, since the dual variable is updated with delay in a small step to allow a high exploration in reward, avoiding getting stuck in a local optimum like L-SAC.

Overall, compared with direct-penalty TD3 baselines, PD-TD3 attains the highest average cumulative reward (~10,000 by around 200 episodes) while reducing the empirical constraint-violation probability, which is proxied by episode safety cost, into the allowable band (cost 0–10 per 24-h episode) by around 500 episodes; in contrast, penalty-TD3 yields lower reward and persistent over-limit violations even at $\lambda = 1000$. Among Lagrangian safe-RL methods, L-SAC achieves near-zero violations but collapses to near-zero reward (over-conservative), whereas S-DDPG keeps violations

out of range for most of training; thus PD-TD3 offers the best reward–safety trade-off, with convergence speed comparable to L-SAC and roughly twice as fast as S-DDPG.

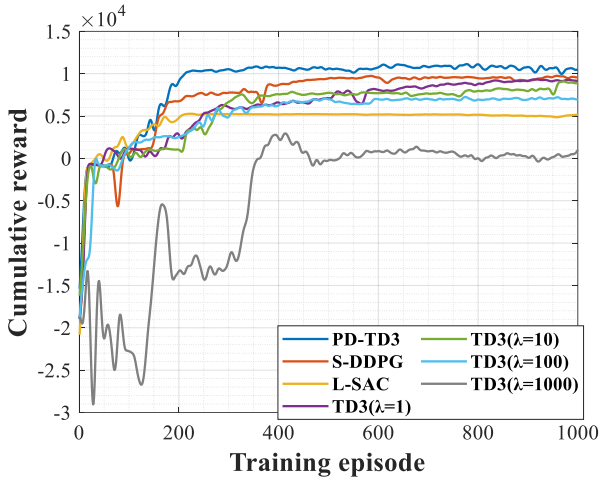


Fig. 4.6 The evolution of cumulative reward for different Safe RL algorithms.

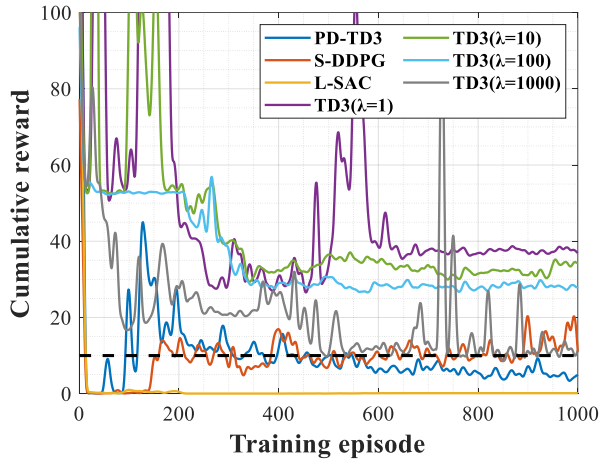


Fig. 4.7 The evolution of cumulative cost of constraint violation for different Safe RL algorithms.

Table 4.3 The cumulative values of reward and cost for constraint violation for different Safe RL algorithms.

Algorithms	Evaluation	Episode						
		1	100	200	400	600	800	1000
PD-TD3	Reward	-15402	-87	9113	10703	11105	10513	10473
	Cost	196	27	21	12	6	5	5
S-DDPG	Reward	-18964	556	6744	8665	9462	9596	9555
	Cost	77	1	14	17	11	10	11

L-SAC	Reward	-20826	1282	4778	5179	5151	5145	5141
	Cost	101	1	0.5	0	0.1	0.1	0.1
TD3	Reward	-18965	1151	2786	6593	8794	9162	7412
(λ =1)	Cost	170	60	46	30	38	37	37
TD3	Reward	-16229	335	1022	7562	7964	8922	6925
(λ =10)	Cost	126	105	54	32	35	32	34
TD3	Reward	-18967	1266	2530	6254	6932	6987	6614
(λ =100)	Cost	96	53	53	29	29	28	28
TD3	Reward	-18964	-22773	-13324	2363	7	1015	6614
(λ =1000)	Cost	8990	1830	2102	1993	2072	4210	4471

4.5.2 Generalization Performance

To demonstrate the generalization performance of the proposed approach, two scenarios are simulated from the data set and analyzed in Figs. 4.11-4.12. Two scenarios are characterized as typical days for summer and winter with different demands and renewable generation. Despite the two scenarios having different energy production and consumption characteristics, there are some similarities during the operation. Firstly, during periods without lower demands for electric and heat power, the CHP unit is turned off due to its high operational cost; the demands are satisfied mainly by imported power and gas. Secondly, electricity prices show a similar trend to the homogenous demand and wholesale prices for electric power across scenarios. These show the generic strategy of the learned policy when facing similar conditions in ICES operation.

Nevertheless, the learned policies and energy resources show more differences across scenarios. As shown in Fig. 4.11, the CHP unit in winter day is turned on for about 15 hours on winter days (0:00-11:00 and 18:00-22:00) since profits caused by high demand for electricity and heat power across most hours can cover the operational cost of CHP. However, CHP only works one hour on summer day in Fig. 4.10 due to the lower demands and potential uneconomic operation. This results in heavy reliance

on the external market on the summer day. Moreover, Fig. 4.9 shows lower prices for electric and heat power on winter day. This is a consequence of the power generation of the CHP unit and wind turbine, which has a lower cost than the external prices or zero generation cost. On the other hand, the flexibility of ESS is more efficiently realized on the summer day in Fig. 4.8. Although the export of power is not allowed in the proposed model, the larger dependence on the external market on the summer day provides more arbitrary chances for ESS.

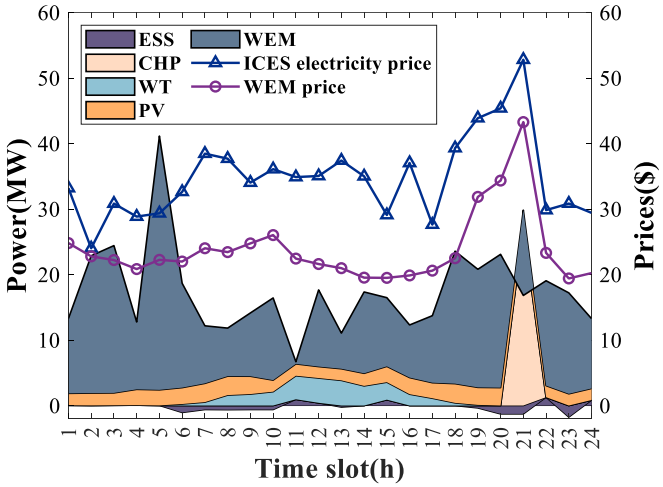


Fig. 4.8 Energy sources and prices for electric power with PD-TD3 method in the summer day

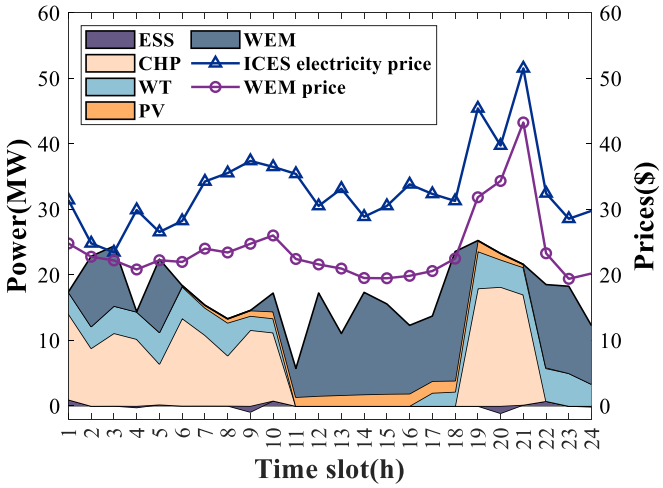


Fig. 4.9 Energy sources and prices for electric power with PD-TD3 method in the winter day

Finally, it can be concluded that the proposed PD-TD3 algorithm is able to learn an effective policy for profit-maximization and safe operation in an MNC-ICES that can generalize to different scenarios. Furthermore, the proposed method investigates the flexibility potential of energy sources for two typical summer and winter days. More specifically, compared to the summer day, the ICES reports on CHP electric and heat power generation on winter day due to its less renewable power generation and higher demands. In contrast, summer day imports more electric power and natural gas from the external market, leading to higher energy prices for MEUs.

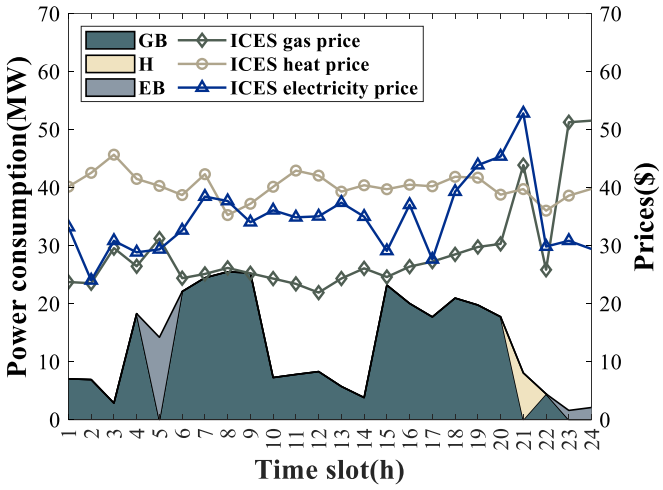


Fig. 4.10 Energy sources and prices for heat power with PD-TD3 method in the summer day

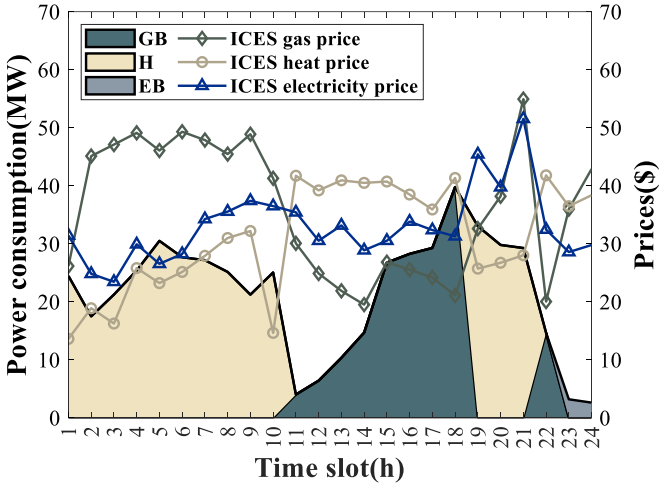


Fig. 4.11 Energy sources and prices for heat power with PD-TD3 method in the winter day

4.5.3 Analysis of Pricing and Operation Decisions

For a more in-depth analysis of the learned pricing and device scheduling policies of the three algorithms above, the energy resources and prices for satisfying power demand and heat demand in the typical test day (the winter scenario) are presented in Figs. 4.12-4.13 for comparison of PD-TD3 and S-DDPG. As illustrated in Fig. 4.8 and Fig. 4.12, two transaction results show a similar trend in electricity prices in the whole transaction period because of the inherent impact of the same wholesale electricity prices, but the prices in Fig. 4.8 are higher than those in Fig. 4.12, therefore resulting in a smoother power consumption curve. The higher prices in Fig. 4.8 also leads to a low electric power purchase in the WEM in most periods except 11:00-18:00 due to the low power demand and low prices in WEM. However, the agent of S-DDPG poses a much lower ICES power price, which makes the power consumption curve much steeper. The ICESO of both algorithms determines lower heat prices in periods with a turning on CHP due to its low marginal cost for heat production, while having a lower natural gas price compared to heat in the rest periods.

The devices scheduling is also illustrated in Fig. 4.12. It can be observed that the PD-TD3 operates CHP for a longer period and purchases less electricity from the WEM compared to S-DDPG. Specifically, the CHP is turned on during the periods of 0:00-11:00 and 18:00-22:00 to sell power and heat to MEUs, since WEM prices, power, and heat demands are relatively higher. In the rest of the hours, the ICESO tends to sell electricity and natural gas from the external market due to the low demand and high cost of the CHP operation. During some periods, the ICESO can provide all the power demand through the power generation from the CHP, DER, and EBS, and the power demand is also cut down or shifted, which is thought of as operating in high energy efficiency. However, the S-DDPG relies on the external market much more by purchasing power and gas to satisfy demand in most periods except for 18:00-22:00, when the CHP is turned on. This is because the S-DDPG algorithm cannot learn the tiny difference in WEM prices and demand between the periods of 0:00-11:00 and

11:00-18:00, even though the market environment during the former periods offers a positive profits uplift for operating the CHP. This not only demonstrates the superior policy of the PD-TD3 compared to S-DDPG but also indicates the high energy efficiency of the MNC-ICES operated by the PD-TD3 algorithm.

Nevertheless, the policies generated by the PD-TD3 and S-DDPG also show differences in physical constraint violations. As the power consumption under PD-TD3 is much smoother and lower compared to the S-DDPG shown in Fig. 4.12, it is intuitive that the latter may have more constraint violations in the electrical distribution network due to higher power consumption in peak hours (8:00-9:00 and 18:00-21:00). On the other hand, there is a lower value in a single kind of power consumption for satisfying heat demand in a single time slot, which is operated by S-DDPG and shown in Fig. 4.13. The network safety of gas and heat is easier to guarantee in the former algorithm, while the policy generated by the latter algorithm may transfer the burden of power transmission from the electrical distribution network to the gas and the thermal networks during peak energy consumption periods, which verify the inherent logic of maintaining the operational safety accounting for multi-energy networks.

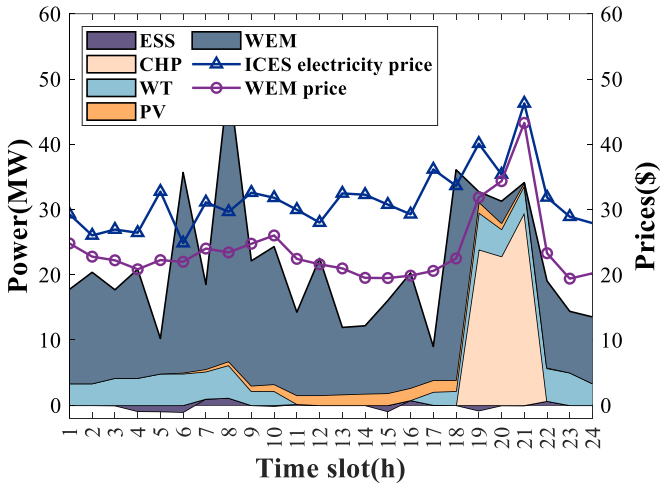


Fig. 4.12 Energy sources and prices for electric power with S-DDPG

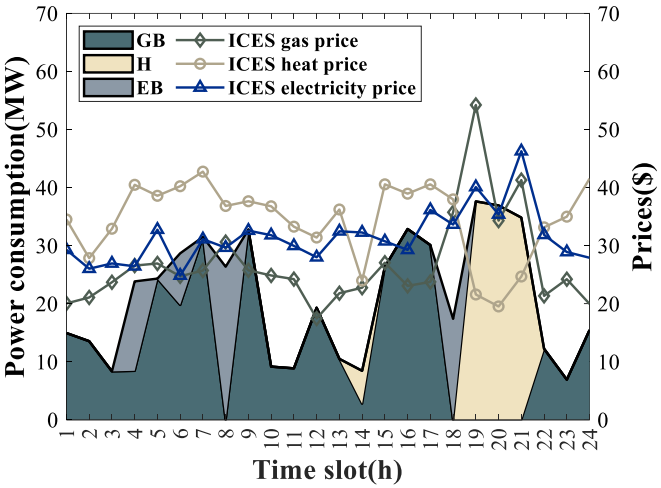


Fig. 4.13 Energy sources and prices for heat power with S-DDPG

4.5.4 Impact of CHP Models

As discussed in Section 4.2, the CHP with a non-convex operating region is effective in generating both electricity and heat, providing system flexibility in reducing the network constraint violation. This subsection examines the influence of the non-convex CHP model on the operation region. For this purpose, a comparison between the scenario with the simplified and detailed CHP model is made, along with the subsequent analysis of the impact on energy transactions and ICES operation. The energy resources and prices in the typical winter day by using simplified CHP are presented in Figs. 4.14-4.15. The total reward and network constraint violations with different CHP models are presented in Table 4.4.

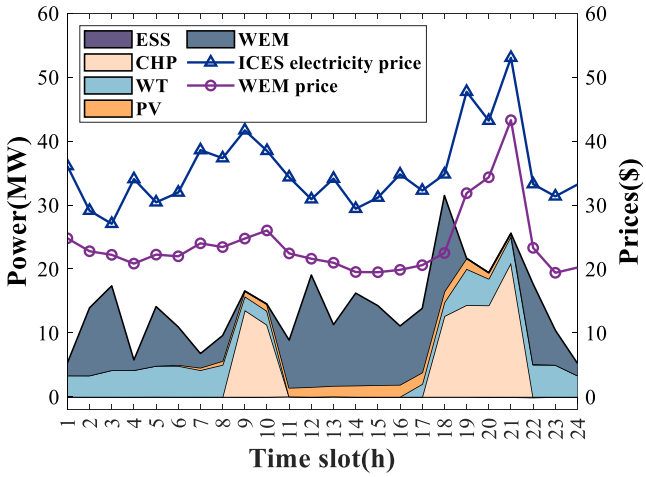


Fig. 4.14 Energy sources and prices for electric power by using simplified CHP

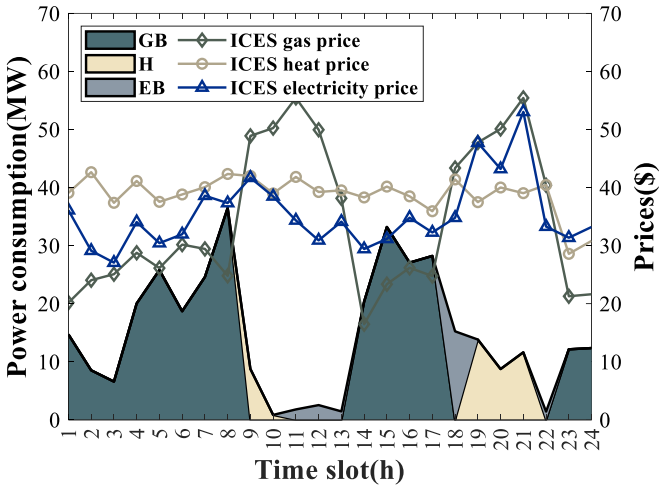


Fig. 4.15 Energy sources and prices for heat power by using simplified CHP

In Fig. 4.14, CHP's operation periods are cut down from 0:00-7:00 since CHP's power and heat output are constrained, leading to its non-profitability during that period. However, this change also results in higher electricity prices and a different power consumption portfolio with a lower average value in MNC-ICES. It should be noted that the less operational profitability of CHP increases the prices for heat but decreases the gas prices, especially during the 0:00-7:00 in Fig. 4.15, when non-convex CHP is in operation but simplified one is not, since the ICESO aims to stimulate the MEUs to consume natural gas instead of heat with turned down CHP. Moreover, the heat consumption is affected even in peak hours of heat demand, which is 18:00-22:00. As the heat generation of CHP is constrained to be linearly related to the power generation, and the heat demand is higher than the power demand during 18:00-22:00, the heat generation is limited to a low level, leading to a significant cut-down in heat consumption in Fig. 4.15 comparing those in Fig. 4.9. Nevertheless, the implementation of simplified CHP model decreases the total generation of both power and heat, which is shown in Table 4.4. The generations of power and heat decrease around two times and seven times, respective, while the generation cost only decrease no more than three times since the detailed CHP model has a small marginal generation cost for heat. Finally, the implementation of the simplified model results in a lower cumulative reward of 7698.97, compared to 11593.98 in the detailed CHP model.

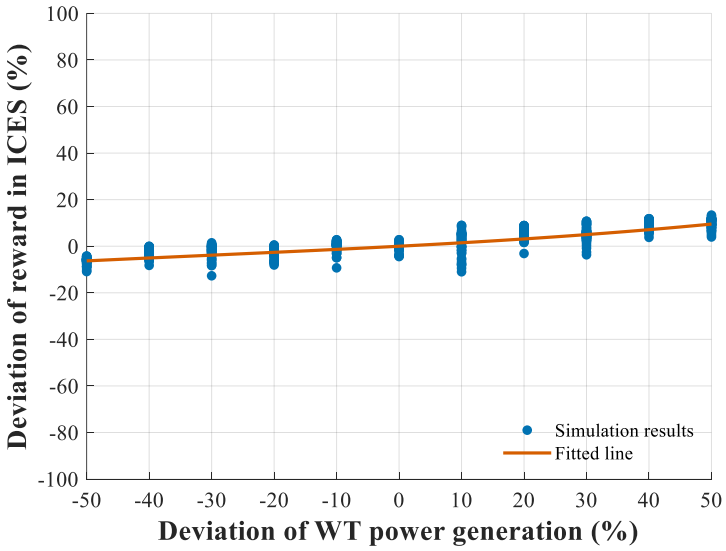
Even though the simplified model decreases the profits of ICESO significantly, it results in a smaller physical constraint violation of electricity and heat networks and has a similar violation in terms of gas networks in comparison to the detailed one, and is shown in Table 4.4. Especially, the cost for heat network violation decreases from 4.91 to 0, because of the significant decrease in heat generation of CHP and the consequent reduced heat consumption of MEUs. On the other hand, due to the substitute effect, MEUs tend to consume more natural gas, putting a burden on the gas network operation. This leads to a slightly higher cost for gas network constraint violations. In summary, the implementation of a simplified model cut down the power and heat generation, and cumulative reward significantly by narrowing the operation region, indicating that the simplified model deviates from the detailed model to a great extent and reflects an unfaithful simulation of the ICES operation in reality.

Table 4.4 Results comparison of implementing detailed and simplified CHP models

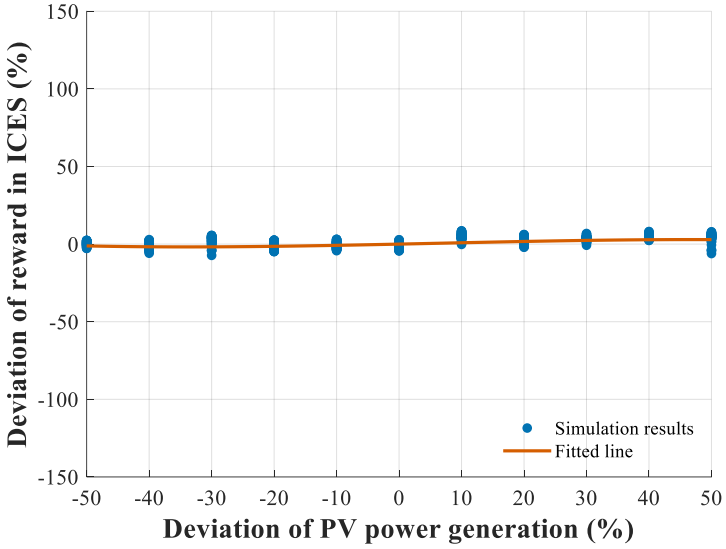
	Total reward	Network violation			CHP output		
		E	G	H	E	H	Cost
Realistic CHP model with non-convex feasible region	11593.98	0.65	0.10	4.91	163.61	501.25	5728.65
Simplified CHP model with fixed conversion rate	7698.97	0.56	0.20	0.00	87.04	69.63	2558.4

4.5.5 Sensitivity Analysis

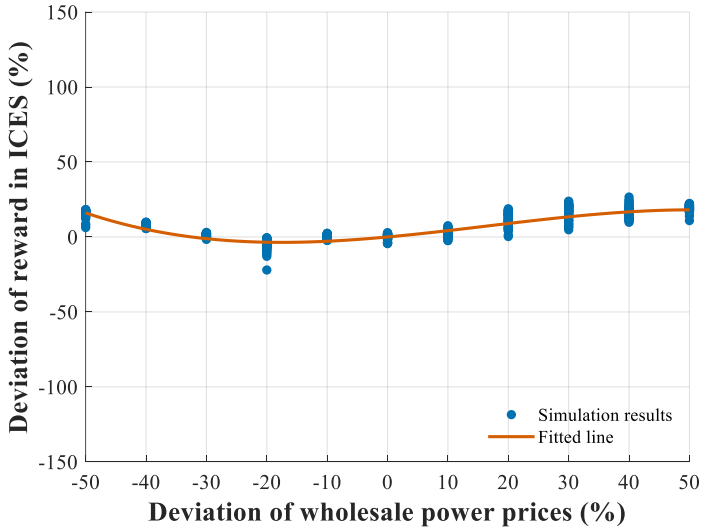
In this subsection, a sensitivity analysis of operational profits (reward) and network constraint violations (cost) is conducted to evaluate the impact of various factors on the operational performance of the MNC-ICES model and the proposed Safe RL approach. The tested factors include renewable power generation (wind turbines and photovoltaic systems), wholesale energy prices (electricity and gas), and integrated energy demand levels (electricity and heat), which are considered to introduce the most uncertainties into the MNC-ICES model. Additionally, as the algorithm's performance is significantly influenced by the random seed, which determines the sequence of random numbers generated, the system is simulated 50 times for each scenario using different random seeds. The sensitivities of reward and cost to these factors are evaluated and illustrated in Fig. 4.16-4.17, respectively. The horizontal axis represents the variable fluctuation ratio of the factors, ranging from 50% to 150% of the initial configured value in increments of 10%. The vertical axis represents the rate of change in the episodic reward/cost of the ICES. Each data point corresponds to a simulation result for a specific scenario under one random seed with a specific factor adjustment, and the trend line is plotted by fitting to the given data points.



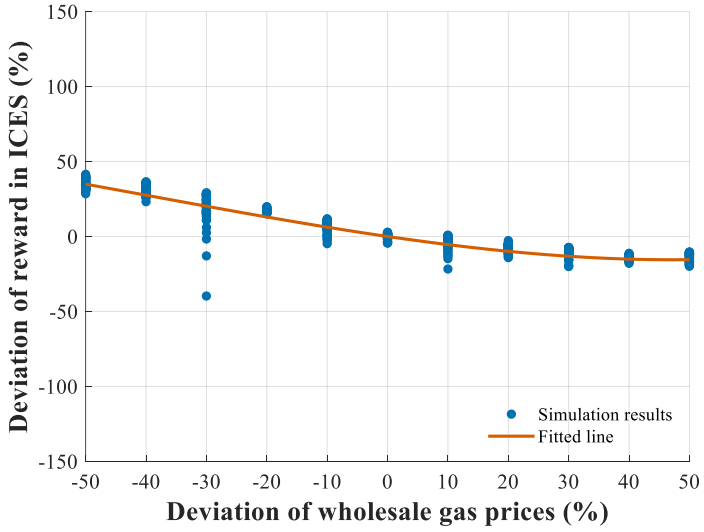
(a)


Deviation of reward in ICES (%)
150
100
50
0
-50
-100
-150
-50 -40 -30 -20 -10 0 10 20 30 40 50
Deviation of PV power generation (%)
Simulation results
Fitted line

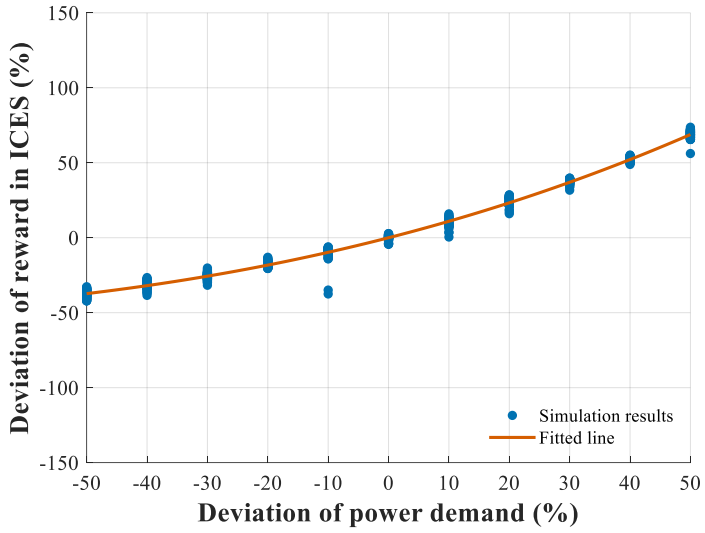
(b)


Deviation of reward in ICES (%)
150
100
50
0
-50
-100
-150
-50 -40 -30 -20 -10 0 10 20 30 40 50
Deviation of wholesale power prices (%)
Simulation results
Fitted line

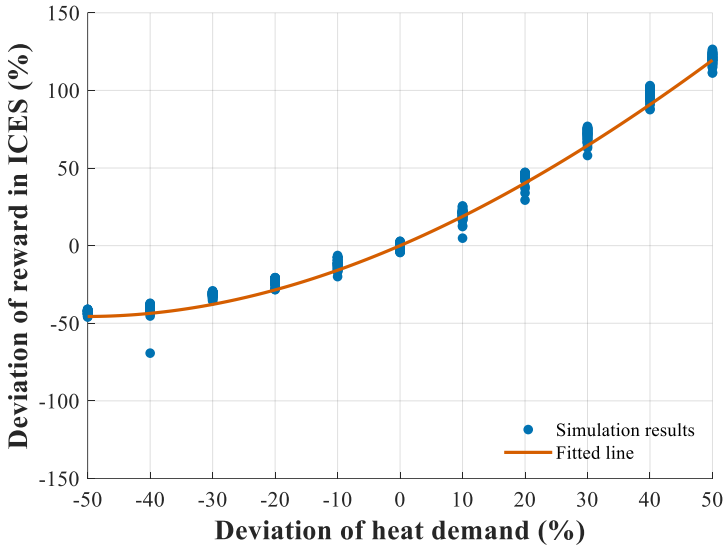
(c)


Deviation of reward in ICES (%)
150
100
50
0
-50
-100
-150
-50 -40 -30 -20 -10 0 10 20 30 40 50
Deviation of wholesale gas prices (%)
Simulation results
Fitted line

(d)



(e)



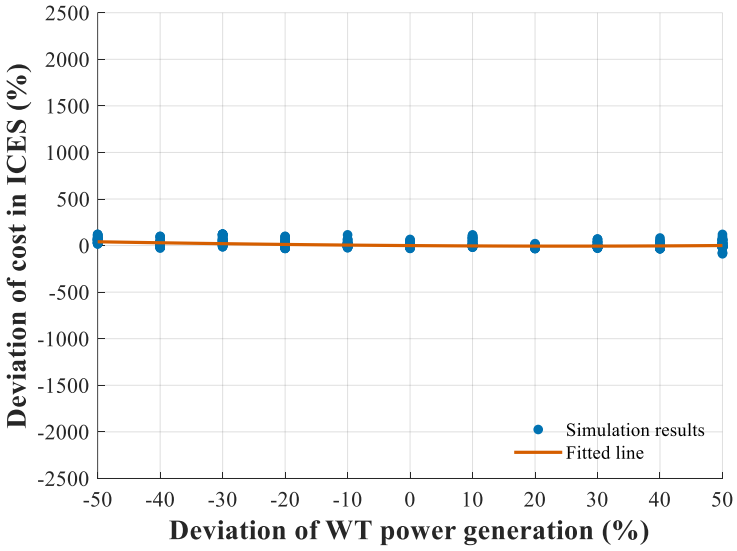
(f)

Fig. 4.16 Sensitivity analysis of ICES operation reward on different factors

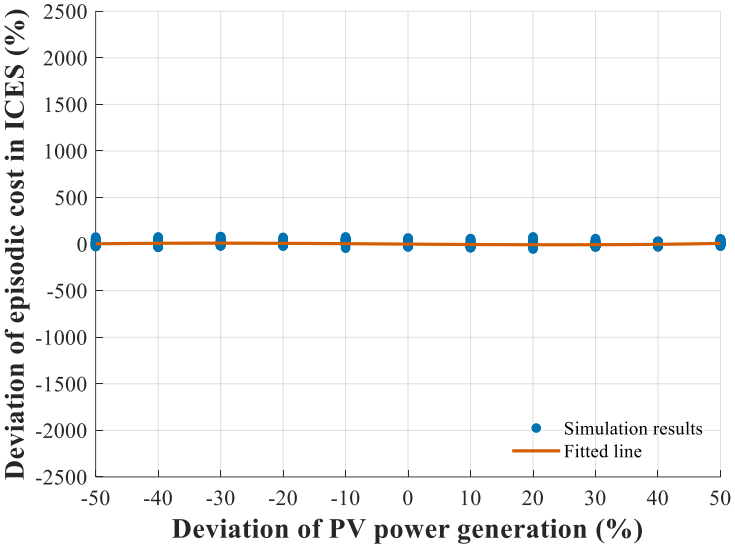
Fig. 4.16 depicts the sensitivity of reward to changes in various factors. Renewable generation and energy demand positively correlate with reward, exhibiting an approximately linear relationship. Specifically, deviations in energy demand most significantly affect the reward, whereas the reward is least sensitive to renewable generation due to its small proportion in the ICES energy mix. Wholesale gas prices exhibit an approximately linear negative correlation with reward, as higher gas prices

increase operational costs. Notably, the wholesale power price shows a nonlinear, likely hyperbolic, relationship with reward. The initial decline in reward corresponds to the natural negative correlation between reward and external energy prices. Conversely, the latter part of the hyperbolic curve is likely due to the increased energy storage arbitrage opportunities created by larger wholesale price gaps.

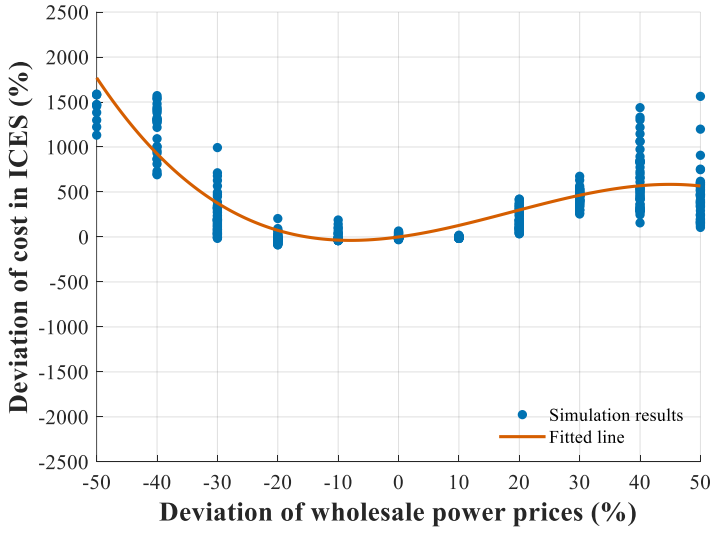
Fig. 4.17 illustrates the sensitivity of cost to changes in various factors. Among these, renewable generation has the weakest negative correlation with cost, with its impact being almost negligible. Energy prices demonstrate a certain hyperbolic relationship: initially, an increase in the price of a single energy source decreases costs for a network with lower energy consumption, while an increase in energy price improves the consumption of alternative energies, thereby raising network constraint violations for other energies. Fluctuations of energy demands show a piecewise linear relationship to the cost; energy demand below a certain level result in almost zero network violations, whereas demand above this threshold leads to linear growth in network constraints. Notably, thermal energy demand and wholesale power prices impact network constraint costs most. In contrast, wholesale gas prices and power demand have a weaker impact. The influence of PV and WT is relatively minor and can almost be disregarded.



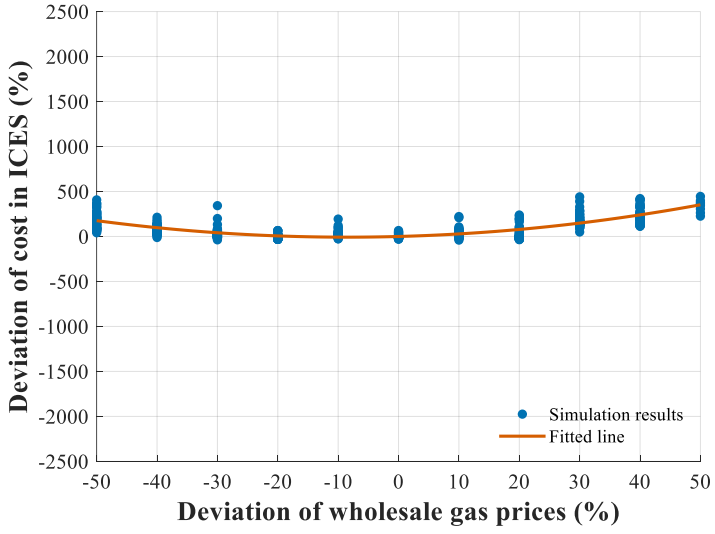
(a)


Deviation of episodic cost in ICES (%)
2500
2000
1500
1000
500
0
-500
-1000
-1500
-2000
-2500
-50 -40 -30 -20 -10 0 10 20 30 40 50
Deviation of PV power generation (%)
Simulation results
Fitted line

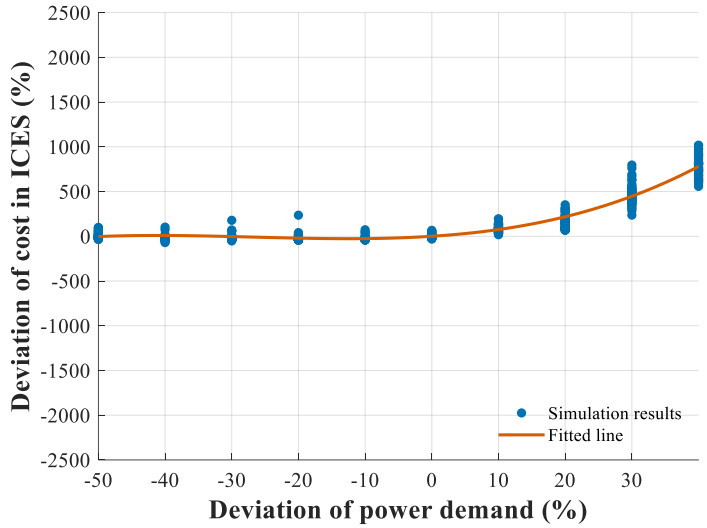
(b)


Deviation of cost in ICES (%)
2500
2000
1500
1000
500
0
-500
-1000
-1500
-2000
-2500
-50 -40 -30 -20 -10 0 10 20 30 40 50
Deviation of wholesale power prices (%)
Simulation results
Fitted line

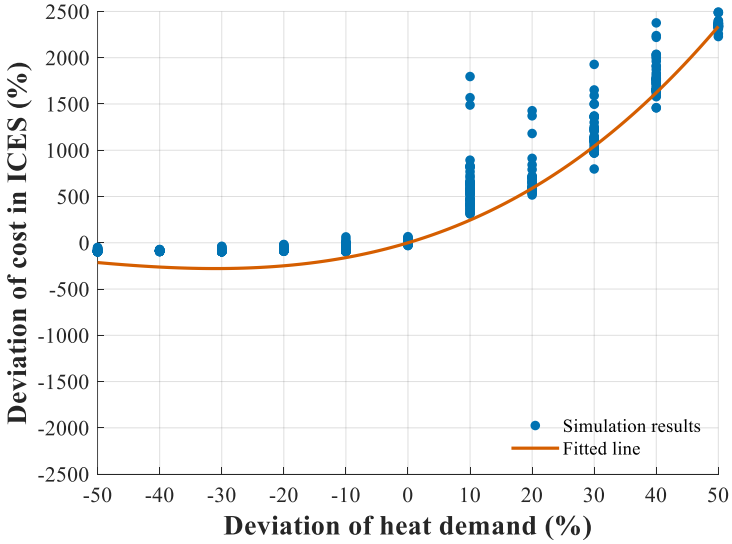
(c)


Deviation of cost in ICES (%)
2500
2000
1500
1000
500
0
-500
-1000
-1500
-2000
-2500
-50 -40 -30 -20 -10 0 10 20 30 40 50
Deviation of wholesale gas prices (%)
Simulation results
Fitted line

(d)



(e)



(f)

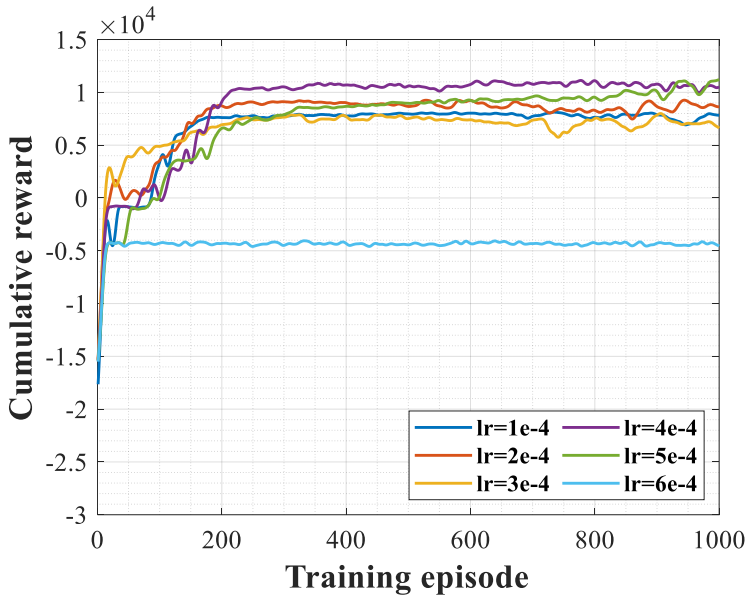
Fig. 4.17 Sensitivity analysis of ICES operation network constraints violation (cost) on different factors

4.5.6 Impact of Hyperparameters

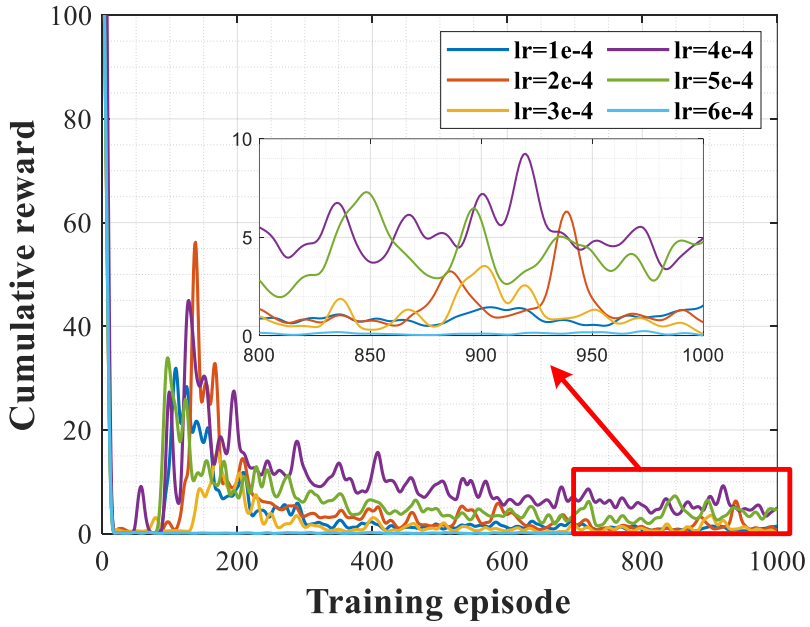
As hyperparameters have a great impact on algorithm performance, sensitivity analysis is conducted on selected hyperparameters, mainly including the Q-network learning rate and actor-network learning rate. Fig. 4.18 shows the evolution of cumulative reward and cost under different actor-network (policy) learning rates. It can

be observed in Fig. 4.18 a) that the episode reward can converge to a high value fast with a learning rate lying from 1e-4 to 5e-4 but may converge to a low value, which is a local optimal, in several episodes with a policy learning rate from 6e-4. Also, the curve of the cumulative reward increases faster in the initial stage with a lower learning rate in policy, which means the exploration in the initial stage contains a higher proportion of useless stochastic noise compared to the later stage. Among these parameter settings, the policy learning rate of 4e-4 (purple) can assist in achieving the highest accumulative reward. When comparing Fig. 4.18 a)-b), the policy learning rate with a higher cumulative reward always results in a higher cost for constraint violation. This demonstrates that the converged cumulative reward has a positive correlation with the cumulative cost, while the policy learning rate plays a key role in the tradeoff of the reward and cost. As all of these parameter settings satisfy the safe operating range of 0~10, the policy learning rate with the highest cumulative reward, which is 4e-4, is selected to be the final setting of the algorithm.

As for the critic network learning rate shown in Fig. 4.19, the converged reward, as well as the cost, firstly increases and then decreases with the growing actor-network

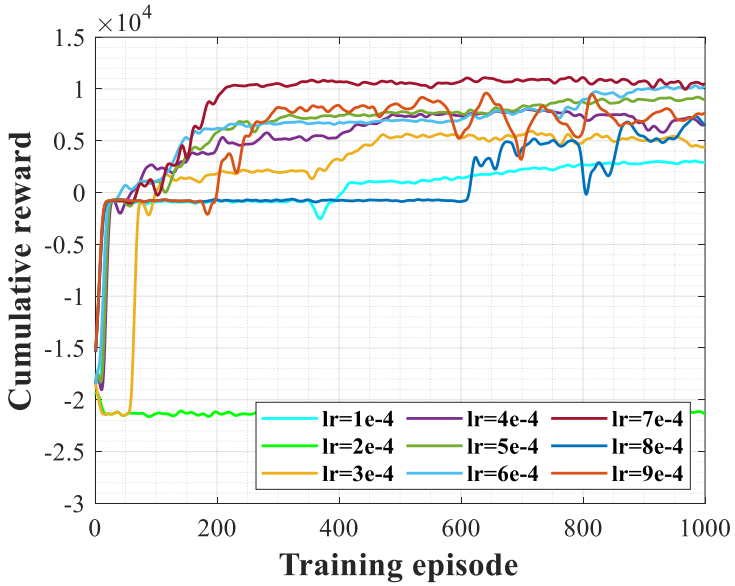


a) reward

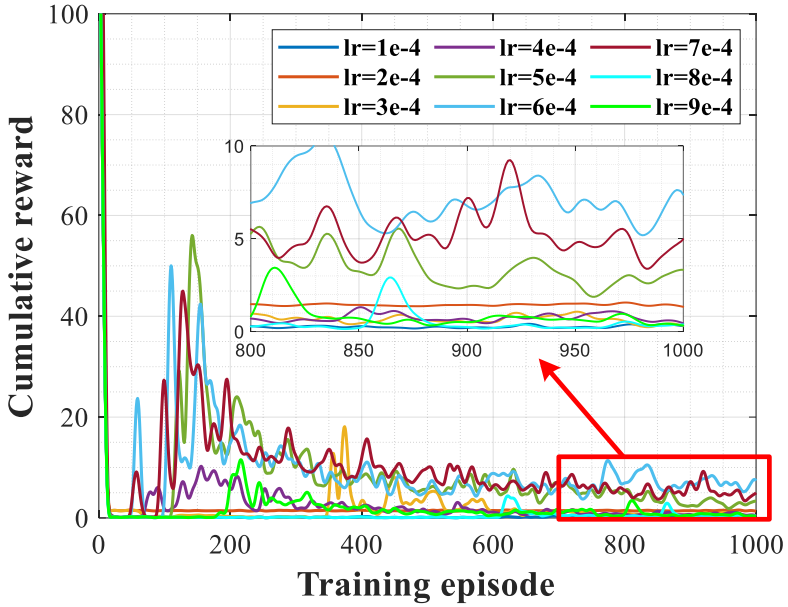


b) cost

Fig. 4.18 Evolution of cumulative reward and cost under different actor network (policy) learning rate



a) reward



b) cost

Fig. 4.19 Evolution of cumulative reward and cost under different critic network (Q-value) learning rate

learning rate, and the learning rate of 7e-4 shows the highest cumulative reward. In general, the evolution curve with a lower learning rate tends to increase gently, while it shows a steep increase or decrease in reward and cost with a higher learning rate. Moreover, a positive correlation between the reward and cost can also be observed for different settings in Fig. 4.19. Interestingly, the cumulative cost with the learning rate of 5e-4 and 6e-4 may exceed the tolerated cost during the training process, while the cost curve with 8e-4 is nearly zero and is considered too conservative. Among these parameter settings of the critic learning rate, the settings of 7e-4 can balance the reward and cost and increase the algorithm performance to the greatest extent.

4.6 Summary

In conclusion, this chapter proposes an MNC-ICES model to describe community-level energy systems. The proposed model comprehensively models multi-networks for integrated energy, realistic energy devices, renewable uncertainty, and IDR of MEUs. Within the community, the ICESO schedules energy devices and prices integrated

energy to maximize operational profits while securing the system operation within the safety requirements imposed by integrated network constraints. This model provides a basis for practical network-constrained community operation tools and can be used as a reference for software development in energy system operation. Numerical results reveal that the realistic model significantly differs from and can attain a higher economic value than simplified models. A novel Safe RL algorithm, PD-TD3, is developed to solve the constrained optimization problem in MNC-ICES and learn the optimal scheduling strategies to maximize profits without violating safety constraints dramatically. The proposed algorithm is based on the Lagrangian method, utilizing a Lagrangian multiplier to penalize the constraint violation during the policy updates. Double networks are employed to mitigate the Q value over-estimation issue of both reward and cost, enabling accurate updates of the Lagrangian multiplier and achieving a balanced tradeoff between the reward and cost. The simulation results demonstrate the superior computational performance and the optimality of the proposed algorithm compared with several benchmarks. Finally, the sensitivity of the MNC-ICES models and the proposed algorithm to model factors and hyperparameters is also analyzed. This work is impactful with potential beneficiaries, including ICES operators and residents, as well as reinforcement learning researchers and practitioners.

Chapter V
Multi-agent Reinforcement Learning for Mixed Strategy Nash Equilibrium Estimation in Real-Time Pricing and Demand Response

5.1 Overview

This chapter focuses on the RTP-DR problem as a combination of demand-side management for the electricity retailer and energy management for multiple EUs in the REM. A dynamic Bayesian Stackelberg game is first applied to the RTP-DR problem, describing the sequential transactions between the retailer (leader) and EUs (followers) under conditions of incomplete information. To solve this game, a novel Multi-Agent Q-Learning algorithm adapted for the dynamic Bayesian Stackelberg game context (BaS-MAQL) is proposed. In the proposed algorithm, both retailer and EUs are able to learn their strategies from dynamic interactions across a day (24 hours transactions). The estimated equilibrium of the game is not unique, indicating the MSNE of the proposed game.

Simulation results of BaS-MAQL algorithm illustrate its computational efficacy by analyzing several SPE in the Bayesian Stackelberg game. Findings reveal that while the retailer, as the game's leader, can predetermine the final equilibrium, the equilibria often yield comparable profits for the retailer but diverse outcomes for EUs, in terms of both profits and power consumption. This discrepancy underscores the necessity of developing new market policies to steer the market toward an equilibrium that maximizes overall social welfare, thereby contributing to the field of smart electricity markets. In addition, the implementation of the proposed algorithm can assist in making transaction decisions with maximized individual profits, which also stimulates consumer active participation and thus improves the market efficiency.

1) *Bayesian Stackelberg Game Model for RTP-DR Problem*: A 1-leader, N-follower dynamic Bayesian Stackelberg game is developed to represent the sequential decision-making RTP-DR problem. This game is assumed to be an incomplete information environment in a non-cooperative game between an electricity retailer and multiple EUs. All players learn the strategies of others dynamically to maximize their own profits in the sequential of RTP-DR problem. The proposed game is then re-formulated into a MDP for reinforcement learning's solutions.

2) *Novel Multi-agent Reinforcement Learning Algorithm*: A BaS-MAQL algorithm is proposed to solve the MDP. By solving the MDP for each player, the SPE of the dynamic Stackelberg game is reached, and the convergence conditions are almost identical to the equilibrium conditions (No player can benefit from deviating from current decisions). Compared to typical MAQL, this approach utilizes probability distributions to represent Q-values, enhancing the algorithm's learning speed and strategic depth, leading to more accurate equilibrium point. The results show that the optimal decisions trajectories of both the retailer and end users are multiple, indicating the equilibrium for the proposed game is indeed MSNE.

The rest of this paper is organized as follows. A hierarchical RTP-DR framework and a novel mixed strategy Bayesian Stackelberg game are established to capture the non-cooperative game between a single retailer and multiple EUs in Section 5.2. A corresponding MDP is formulated based on the game model, and the BaS-MAQL algorithm is proposed to solve the MDP in Section 5.3. The analysis of the MSNE is conducted in Section 5.4, using a case study comprising one retailer and multiple EUs. Finally, this chapter is concluded in Section 5.5.

5.2 Problem Formulation

In this section, the hierarchical RTP-DR framework containing the model of retailer and EU is established. Both retailer and EUs are strategic player that aims to maximize their profits given uncertain behavior of each other. The transactions under this framework are then formulated into a mixed-strategy Bayesian Stackelberg game. The

Bayesian property is introduced to the game by modeling the belief of the retailer in EUs' uncertain behaviors in a limited information environment. The advantage of mixed strategy adoption is to ensure the optimality of the strategy and demonstrate the inherent randomness and uncertainty of the power consumption behavior.

5.2.1 Hierarchical RTP-DR Framework

Before developing the utility functions, the following assumptions are required for the trading mechanism of WEM and REM: (1) Both the trading and pricing in WEM and REM are on an hourly basis. The "time-slot" $k \in \{1,2, \ldots\ldots, K\}$ mentioned hereinafter would correspond to discrete hours in a single day. (2) Each EU cluster constituted with n EUs ($n \in \{1, \ldots\ldots, N\}$) is managed by one single RE, who purchases the electricity from WEM and sells it to EUs with a dynamic price to be determined. (3) EUs need to determine their own electricity consumption (i.e., the amount of electricity to be purchased) of each time slot after the price announced by the RE. (4) All REs and EUs are rational players who make decisions to maximize their own profits. The operation of the proposed electricity market is roughly described in Fig. 5.1.

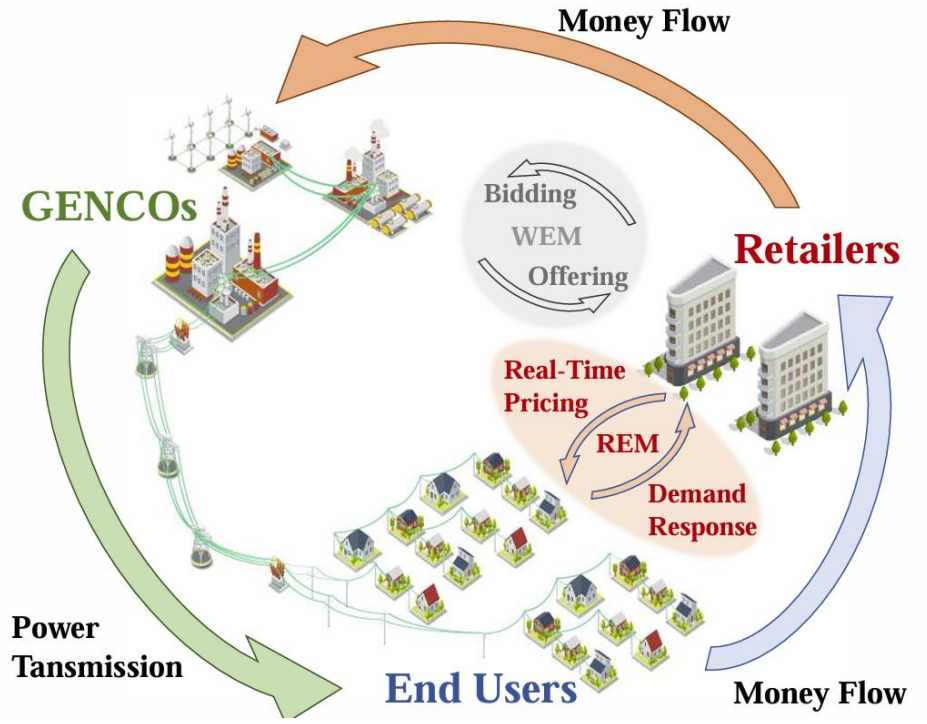


Fig. 5.1 Hierarchical RTP-DR framework within the electricity market

In the proposed market, the several assumptions are made for the trading mechanism of WEM and REM: (1) Transactions in WEM and REM are on an hourly basis. The "time-slot" $k \in \{1,2, \ldots\ldots, K\}$ mentioned hereinafter would correspond to discrete hours in a single day. (2) Each EU cluster constituted with ith EUs ($i \in \{1, \ldots\ldots, I\}$) is managed by one retailer. (3) EUs need to decide their own electricity consumption (i.e., the amount of electricity to be purchased) of each time slot after receive the electricity price from the retailer. (4) All retailers and EUs are rational players who make decisions to maximize their own profits. (5) The set of buses in the network is denoted by Ω_N, where $n \in \{1, \ldots\ldots, \Omega_N\}$, while the set of lines is denoted by Ω_L. The proposed market structure is roughly described in Fig.5.1.

1) Model of Retailers

Retailers are participants in both WEM and REM, acting as a "broker" between GENCOs and EUs. The clearing prices in WEM, λ_w^k, are assumed to be deterministic for the whole transaction period, as the strategic behavior of a single retailer has hardly impact on the wholesale clearing result. Based on λ_w^k, the retailer determines a uniform retail electricity price λ_r^k, which will be responded by EUs with total power consumption of ith EU during times slot k, denoted by p_i^k. The objective of the retailer is to maximize profits expressed by (5.1), where the $\lambda_r^k - \lambda_w^k$ indicates the difference between the retail and wholesale electricity price, so-called "price gap" in time slot k.

$$\max_{\lambda_r^k} \sum_{\forall k} \sum_{\forall i} (\lambda_r^k - \lambda_w^k) p_i^k \tag{5.1}$$

2) Model of EUs

EUs are demand-side (REM) participants who determine their power consumption based on the retail electricity price. Here, there are smart meters in households to response and determine the power consumption instead of the real "End Users". According to the load being fixed in a specific time slot or not, loads of EUs are divided into baseline load and elastic load. The baseline appliances of ith EU, including the must-run appliances like lights and refrigerators, are considered to consume fixed power $p_n^{bs.k}$ in time slot \$k\$. In contrast, the power consumption of the elastic load

$p_n^{el.k}$ mainly comes from elastic appliances, including heating ventilation air-conditioning (HVAC) and wet appliance (WA), which can be dynamically modified to satisfy both utility demand and economic considerations. Therefore, objective function of each EU is given by (5.2).

$$\max_{p_i^k} \sum_{\forall k} \left(U_i\left(p_i^k\right) - p_i^k \lambda_r^k\right) \tag{5.2}$$

$$s.t. \forall k \ \forall i$$

$$p_i^k = p_i^{bs.k} + p_i^{el.k} \tag{5.3}$$

$$0 \le p_i^{el.k} \le x_i^{el.\max} \tag{5.4}$$

$$e_{i.min} \le \sum_{\forall k} p_i^{el.k} \le e_{i.max} \tag{5.5}$$

In equations above, $U_i\left(p_i^k\right)$is the utility function for each EU, while $p_i^k \lambda_r^k$ stands for the cost for EUs with p_i^k power consumption. $U_i\left(p_i^k\right)$ modeled by a widely employed quadratic function as (5.6) [92]. It is concave and highlights the marginal decreasing utility with the increase of power consumption.

Moreover, the power consumption of the ith EU p_i^k in time slot k can be computed as (5.3). $p_i^{el.k}$ is constrained by (5.4), denoting the non-negative baseline power consumption. Total elastic power consumption $\sum_{\forall k} p_i^{el.k}$ among k periods is constrained by (5.5), where $e_{i.min}$ and $e_{i.max}$ are the lower bound and the upper bound, respectively.

$$U_i\left(p_i^k\right) = \begin{cases} \omega_i^k - \frac{\lambda_i}{2}\left(p_i^k\right)^2, 0 \le p_i^k \le \frac{\omega_i^k}{\lambda_i} \\ \frac{\left(\omega_i^k\right)^2}{2\lambda_i} - p_i^k \lambda_r^k, p_i^k > \frac{\omega_i^k}{\lambda_i} \end{cases} \tag{5.6}$$

Specifically, parameters ω_i^k, λ_i varying in different time slots represents the energy-consuming preferences of each EU. $\omega_i^k - \frac{\lambda_i}{2}\left(p_i^k\right)^2$ and $\frac{\left(\omega_i^k\right)^2}{2\lambda_i} - p_i^k \lambda_r^k$ indicate the valuation of power consumption p_i^k in different intervals.

5.2.2 The Mixed-Strategy Bayesian Stackelberg Game

To mathematically present the game between the retailer and EUs, a mixed strategy Bayesian Stackelberg game model formulated based on the aforementioned models of

retailers and EUs. The game model in compact-form is presented in (5.7), where the retailer is the leader acting first, and EUs are followers that act after the action of the leader.

$$T_R =< N, \phi_r, < \underbrace{S_r^k, \Theta_r^k, A_r^k, R_r^k, \Omega_r^k, \mu_r}_{Upper-Level} >, \\ < \underbrace{S_i^k, \Theta_i^k, A_i^k, R_i^k, \mu_i}_{Lower-Level} >_{i \in I} > \tag{5.7}$$

where I denotes the total number of EUs served by the RE, and Φ_r indicates the environment, i.e., transaction rules of REM. T_R is considered as a finite game, javing limited number of players and pure strategies available to each player. For the retailer with tuple $< S_r^k, \Theta_r^k, A_r^k, R_r^k, \Omega_r^k, \mu_r >$ in time slot k, S_r^k presents the set of observed states information, which specifically refer to the wholesale electricity price, Θ_R^k denotes the set of mixed strategies ϑ_r^k (probabilistic distribution over pure strategies) of the retailer, A_r^k is the set of pure actions (retail electricity prices) taken by the retailer, R_r^k represents the payoff functions (objective), Ω_r^k indicate the set of the retailer's belief P_r^k on the strategies of all EUs, indicating the estimated total power consumption, and μ_r denotes a set of executed strategies constituting strategies among all time slots, termed as the policy of the retailer. In a finite game, strategy set A_r^k in each time slot k for the retailer can be presented as $a_r^k \in A_r^k = \left\{a_1^k, a_2^k, \dots, a_{m_r^k}^k\right\}$, where m_r^k is the number of pure strategies during time slot k. The game allows players to play mixed strategy $\vartheta_r^k = \left\{\rho_1^k, \rho_2^k, \dots, \rho_{m_r^k}^k\right\}$, which is a probability distribution vector over the pure strategies. $\rho_{m_r^k}^k$ is the probability that action $a_{m_r^k}^k$ is chosen in the mixed strategy profile. The sum of the probability of each pure strategy in a mixed strategy ϑ_r^k should be equal to one.

Similarly, the tuple $< S_i^k, \Theta_i^k, A_i^k, R_i^k, \mu_i >_{i \in I}$ for each EU denotes the set of states (retail electricity prices), the mixed strategies, actions, payoff functions (utility function) and the strategies among all time slots. Also, finite pure strategy set A_i^k for EUs in each time slot k can be presented as $a_i^k \in A_i^k = \left\{a_1^k, a_2^k, \dots, a_{m_i^k}^k\right\}$, where m_r^k is the number of possible power consumption decisions during time slot k for EU. Also, the mixed

strategy of EUs is $\vartheta_i^k = \left\{\rho_1^k, \rho_2^k, \dots, \rho_{m_i^k}^k\right\}$. Θ_i^k represents the set of mixed strategies ϑ_i^k for $i \in I + r$ and let $\Theta^k \equiv \times_{i \in I+r} \Theta_i^k$.

5.2.3 Optimal Conditions and Equilibrium Analysis

For the RE, a mixed strategy ϑ_r and policy μ_r are optimal, if $\forall i \in I, k \in K$,

$$\vartheta_r^{k^*} \in \text{argmax}_{\vartheta_r^k \in \Theta_r^k} \sum_{a_i^k \in A_i^k} P_r^k\left(a_i^k | s_r^k\right) \vartheta_r^k(a_r^k) R_r^k\left(\vartheta_r^k, a_i^k\right) \tag{5.8}$$

$$\mu_r{}^* = \sum_{k \in K} \vartheta_r^{k^*} \tag{5.9}$$

In (5.8), $P_r^k\left(a_i^k | s_r^k\right)$ indicates the RE's belief on the type of EUs at state s_r^k, and is expressed as a reduced form because the types of EUs are set to be the action of EUs. While $\vartheta_r^k(a_r^k)$ means the mixed strategy of the retailer distributing probability on pure strategies a_r^k, $R_r^k\left(\vartheta_r^k, a_i^k\right)$ is the reward under the mixed strategy σ_r^k and EUs' action a_i^k in times slot *k*.

For all EUs, the retail electricity prices in each time slot can be seen as a state. By maximizing the profits of all EUs, a mixed strategy profile ϑ^* constituting I mixed strategies is defined optimal for EUs in time slot k, if $\forall n \in N, k \in K$,

$$\vartheta_n^{k^*} \in \text{argmax}_{\vartheta_i^k \in \Theta_i^k} \sum_{a_i^k \in A_i^k} \left(\prod_{i \in I} \sigma_i^k\left(a_i^k\right) \right) \vartheta_r^k(a_r^k) R_i^k\left(a_i^k\right) \tag{5.10}$$

$$\mu_i{}^* = \sum_{k \in K} \vartheta_i^{k^*} \tag{5.11}$$

In (5.10) and (5.11), $\vartheta_i^k\left(a_i^k\right)$ and $\vartheta_r^k(a_r^k)$ indicate mixed strategy of ith EU and the retailer at *k*. While $R_i^k\left(a_i^k\right)$ means the profit of the ith EU, it is acknowledged that the expected payoff for each participant obtained by the optimal strategy or policy must be larger than other strategies or policies. Also, the transaction result is both a MSNE and a Bayesian Nash Equilibrium (BNE) when the RE holds the belief on all EUs, being consistent with the EUs' strategy. Transaction result of the Bayesian Stackelberg game is MSNE when (5.8) and (5.10) are satisfied in time slot k, being identical to the equilibrium condition "the payoff (profits) of all participants cannot be improved by his own action deviation under the mixed strategy profile Θ constituting $\prod_{i=1}^{I} \vartheta_i^k$ and ϑ_r^k". Also, the MSNE is a BNE as the strategy of retailer is optimal if and only if the retailer

holds the correct belief $P_r\left(a_i^k|s_r^k\right)$ on each EU. While the $\mu_n{}^*$ is optimal if and only if each mixed strategy ϑ_i^k in $\mu_i{}^*$ is the optimal $\vartheta_i^{k^*}$ during its time slot. Given the game models above, the payoff function can be reformulated as (5.12) and (5.13) the retailer and EUs. With such continuous payoff function, MSNE always exists [93].

$$R_r^k = \sum_{k\in K}\sum_{n\in N} \vartheta_n^k(x_n^k)[P_r(x_n^k|p_r^k)\vartheta_r^k(p_r^k) - p_w^k] \tag{5.12}$$

$$R_n^k = \sum_{k\in K} [\vartheta_n^k(x_n^k)U_n(x_n^k,\omega_n) - \vartheta_n^k(x_n^k)\vartheta_r^k(p_r^k)] \tag{5.13}$$

5.3 Proposed MAQL algorithm

In this subsection, a bi-level MDP is formulated from the Bayesian Stackelberg game above, where the retailer and EUs act sequentially in the upper level and the lower level, respectively. A novel BaS-MAQL algorithm is developed to solve the proposed MDP and estimate SPE in the Bayesian Stackelberg game. This algorithm is bi-level and employs probabilistic distribution to denote the Q-value for the retailer, which can be updated with posterior experiences. Hence, the algorithm merits privacy protection aligning the reality because of the bi-level structure with incomplete information. Moreover, it is applicable in dealing with the massive non-convex multi-agent systems as the adoption of probabilistic Q-value distribution accelerate the training process in highly non-convex problem significantly.

5.3.1 Markov Decision Process

1) Upper-level problem

The upper-level problem is the RE level problem to maximize the profits of the RE in REM, where the RTP problem is formulated as a tuple $< S_r^k, A_r^k, R_r^k(s,a), \mu_r, \gamma_r >$. $S_r^k = \{p_w^k\}$, and $A_r^k = \{p_r^k\}$, are set to denote the state and action of the retailer in time slot k. The constraints of actions are set as (5.2), representing the retail electricity price is always higher than the wholesale electricity price to pursue RE's profits. For the selected action $a_r^k \in A_r^k$ and $s_r^k \in S_r^k$, reward $R_r^k(s,a)$ can be calculated as (5.14).

$$R_r^k = \sum_i^I \left[(\lambda_r^k - \lambda_w^k)(p_i^k) \right] \tag{5.14}$$

The policy μ_r refers to the optimal action to be taken at the given state. The aim of retailer is to find the optimal policy $\mu_r{}^*$ of the retail electricity prices. The state value function and state-action function are employed to value the state and state-action pair and further explore the optimal policy by calculating the accumulative rewards in (5.15) and (5.16). In these two equations, $\mathbb{E}$ is the expectation operator, discount factor $\gamma_r \in [0,1]$ is utilized to discount the future rewards in MDP for the uncertainty of rewards in future.

$$V_{\mu_r}(S_r^k) = \mathbb{E}\left[\sum_{t=0}^{K-k} \gamma_r R_r^{k+t} | S_r^k \right] \tag{5.15}$$

$$Q_{\mu_r}(S_r^k, A_r^k) = \mathbb{E}\left[\sum_{t=0}^{K-k} \gamma_r R_r^{k+t} | S_r^k, A_r^k \right] \tag{5.16}$$

The Bellman functions, including state value function and state-action value function, aim to find the optimal policy $\mu_r{}^*$ to maximize the Q-value of the retailer in each step. Thus, the MDP problem of the retailer level model is formulated as follows.

$$P1: \max_{\mu_r} \mathbb{E}\left[\sum_{t=0}^{K-k} \gamma_r R_r^{k+t} | S_r^k \right], \quad s.t. (5.3) - (5.5) \tag{5.17}$$

2) Lower-level problem

In the EU level (lower-level), the DR problem of each EU is formulated as a similar tuple $< S_i^k, A_i^k, R_i^k(s,a), \mu_i, \gamma_i >$, where $S_n^k = \{p_i^{bl.k}, \lambda_r^k\}$ denotes the set of states, where $p_i^{bl.k}$ is the baseline power consumption of the ith EU, and λ_r^k refers to the retail electricity price. Action denoted by $A_i^k = \{p_i^{el.k}\}$ refers to the elastic power consumption of the EU. While the reward function $R_i^k(s,a)$ has been proposed as (5.2) and (5.13). Policy μ_i is the set of actions which has been taken in each state for ith EU. Discount factor $\gamma_n \in [0,1]$ is utilized to discount the future rewards for the future uncertainty. Therefore, the lower-level model of the MDP can be formulated. To find the optimal policy $\mu_n{}^*$ to maximize Q in each time slot k, the RTP-DR problem is formulated as a bi-level model in the framework of MDP, which can capture the

interactive characters of RTP-DR problems, and be solved by MARL algorithm. In summarize, the states, actions, and rewards for the two proposed MDP are illustrated in Table 5.1.

Table 5.1 Variable Interpretations in Markov Decision Process

		Variable	Notation
Retailer	State S_r^k	λ_w^k	Wholesale electricity price in time slot k
	Action A_r^k	λ_r^k	Retail electricity price in time slot k
	Reward R_r^k	U_r^k	Utility of the RE gained from selling electricity to EUs in time slot k
EUs	State S_i^k	$p_i^{bl.k}$	Power consumption of the baseline appliances for EU i in time slot k
		p_r^k	Retail electricity price in time slot k
	Action A_i^k	$p_i^{el.k}$	Power consumption of the elastic appliances for EU n in time slot k
	Reward R_i^k	U_i^k	Utility of EU i gained from consuming and purchasing power in time slot k

5.3.2 Bayesian Stackelberg Multi-Agent Reinforcement Learning (BaS-MARL)

As one of the most popular model-free RL algorithm, Q-learning is a tabular algorithm that enables agents to learn to select the optimal action at each state, i.e., to generate the optimal policy [94]. The Q-learning merits a simple and precise structure, which makes it more reliable and explainable than that of state-of-the-art algorithms like TD3 and SAC. Moreover, it is believed to be stable and practical as the algorithm is not hyper-parameter-sensitive and is more flexible to be tuned aligning with different scenarios [95]. However, it is thought to be difficult to deal with large-scale problems and stochastic problems with uncertainty, which is basically caused by the non-proper way of Q-value update.

For these reasons, Q-learning is first improved to a bi-level algorithm in a multi-agent setting, which makes it applicable to the proposed Stackelberg game. Then, the upper-level algorithm is revised to be a Bayesian Q-learning for assisting the Q-value estimation and improving the computational performance in the proposed large-scale RTP-DR problem with demand uncertainties. The workflow of the BaS-MAQL algorithm is depicted in Fig 5.2. Compared to regular MAQL, there are two major modifications in the proposed BaS-MAQL, including bi-level structure and Q-value update, which are illustrated as follows.

1) Bi-level structure

The bi-level algorithm can be divided into upper-level and lower-level algorithms corresponding with the proposed MDPs, where the agents are the retailer and EUs, respectively. At the RE level (upper level), the retailer first selects the action, i.e., the retail electricity prices, based on ϵ-greedy strategy. These prices are then broadcast to EUs. After being informed of the retail electricity prices, at the EU level (lower level), each EU takes actions (determines power consumption) accordingly using ϵ-greedy strategy. The rewards for these selected actions are immediately received by each EU. With the state-action pair and the rewards, the Q-values of EUs are updated by a Q-value update strategy. EUs will follow these processes iteratively to generate actions, get rewards, and update Q-values until the termination criterion at the lower level is satisfied.

Once the iterations of EUs terminate, the total power consumption is determined and the retailer can receive his rewards, based on which updates his Q-value at the upper-level algorithm. Given the process above, the retailer repeats it until the termination criterion at the upper-level algorithm is satisfied. In this bi-level structure, the lower-level algorithm converges for one time during the training of one iteration of the upper-level algorithm. This process follows the nature of the Stackelberg game and ensures the optimality of convergence results.

2) Q-value update

Model-free Bayesian RL algorithm, like Bayesian-Q-learning, assumes that there is a prior probability distribution over each Q-value [96]. The Q-values are updated using the posterior probability distribution. Here, the Bayesian Q-learning is adopted in the upper-level algorithm of the bi-level MAQL [97]. In the proposed algorithm, a parametric Gaussian distribution $p(\mu_{s,a}, \delta_{s,a})$ is employed to denote the Q-value distribution of the action a at the state s. $\mu_{s,a}$ and $\delta_{s,a}$ are the mathematical expectation and the variance of the Q-value, respectively. Initially, the prior probability distribution can be normal distribution by default. It is assumed that r is the immediate reward of the chosen action a in the current state s, R_s is the discounted sum of rewards from the state s following the apparently optimal policy, and $R_{s,a} = r + \gamma R_s$ indicates the discounted reward of executing the actions a at the state s following the optimal action in the future states. γ is the discounted factor.

At the RE-level (upper-level), the Bayesian Q-learning is adopted. When receiving the immediate reward r and estimating future reward $R_s = x$, the updated posterior probability distribution $p(\mu_{s,a}, \delta_{s,a}|r + \gamma_r x)$ in the upper-level algorithm can be calculated by using (5.18). The uncertainty of reward is captured by weighting the probabilistic distribution that of $R_s = x$. This Q-value update method is called mixture updating, which is cautious and helpful in avoiding the over-estimation of Q-value.

$$p_{s,r}(\mu_{s,a}, \delta_{s,a}) = \int_{-\infty}^{+\infty} p(\mu_{s,a}, \delta_{s,a}|r + \gamma_r x) p(R_s = x) dx \tag{5.18}$$

At the EU level (lower-level), the Q-value is updated as the regular Q-learning by using (5.19), where $Q(S_i^k, A_i^k)$ denotes the Q-value function of EU, β_i and γ_i denotes the learning rate and discount factor for EUs, respectively. The Q-table, which stores the most recent updated Q-value, can help each EU optimize its policy by selecting the action with the greatest Q-value with a high probability in the following iteration. Finally, the iteration will come to an end when the termination criterion (5.20) is satisfied, i.e., the difference of updated and original Q-value is less than the specified threshold value τ_i.

$$Q(S_i^k, A_i^k) = Q(S_i^k, A_i^k) + \beta_i [R_i^k + \gamma \max Q(S_i^{k+1}, A_i^{k+1}) - Q(S_i^k, A_i^k)] \tag{5.19}$$

$$\left|Q\left(S_i^{k+1}, A_i^{k+1}\right) - Q\left(S_i^k, A_i^k\right)\right| \leq \tau_i \tag{5.20}$$

5.3.3 Discussions

This paper employ RL algorithm to solve the game and estimate its equilibrium because of its non-convexity, and strong exploration ability. The proposed game is actually non-linear and non-convex, since the end users have quadratic utility and the aggregated utility of all end-users with individual time correlated constraints are non-convex. Such game can only be solved by using methods that are capable of overcoming non-linearity and non-convexity.

Moreover, the proposed game is highly dynamic due to the strategic interaction between the retailer and end users, which probably has multiple equilibria. By owning randomness in exploration stage, the algorithm can better explore the action spaces to estimate all possible equilibriums, so that the multiple equilibria can be reached and analyzed from the perspective of mixed strategy Nash equilibrium. Compared to traditional mathematical methods, such randomness in RL methods can be thought of as a different starting point for deterministic optimization. As for the NE convergence, the termination condition for learning is the almost no change in accumulative reward or Q-values for the algorithm, which is almost identical to the NE definition "No players can get a higher payoff by deviating current actions." Therefore, the output results of the proposed algorithm are thought of NE.

Apart from merits of the typical RL algorithm, the proposed BaS-MAQL algorithm exhibits significant advantages in terms of solution robustness and adaptability compared with other RL algorithms, making it well-suited for smart electricity market simulations and policy-making. The pros and cons of the proposed algorithms are explained below.

1) The MAQL and single-agent Bayesian Q-learning have been developed for a long history. BaS-MAQL combines them by letting Bayesian Q-learning agents act as leaders, while typical agents in typical MAQL act as followers in the dynamic Stackelberg game to adapt to the proposed RTP-DR problem. Such

revision is to enhance the leader (retailer)’s learning ability to overcome the highly uncertain power consumption of end users by making Q-tables distributions rather than scalers. Therefore, computational convergence conditions and the learning philosophy of original algorithms are not changed after revision, which means the proposed algorithm can be adapted to any environment.

2) The BaS-MAQL algorithm mitigates the Q-value overestimation bias commonly associated with conventional value-based RL algorithms by representing the Q-value through probability distributions. This approach enables a more accurate estimation of Q-values, thereby facilitating the agents' learning of optimal strategies. Using probability distributions for Q-values enhances the algorithm's robustness by effectively quantifying and incorporating environmental uncertainties into the decision-making process.
3) Compared to other RL algorithms like MADDPG, BaS-MAQL requires fewer hyperparameters, simplifying its implementation and adaptation across different market contexts. This simplicity, combined with the algorithm's other merits, underscores its applicability and effectiveness in facilitating sophisticated simulations and analyses for electricity market design and strategy optimization.
4) Due to its tailored structure to the RTP-DR problem, the proposed BaS-MAQL is only applied in multi-agent problems that can be modeled as a one-leader multi-follower Stackelberg game. However, it is also possible to extend the algorithm into the multi-leader multi-follower structure, which may require sophisticated work to provide a convergence guarantee.
5) Most value-based RL algorithms, e.g., Q-learning, only enable discrete action space rather than continuous actions. This may require huge labor in action discretization to ensure algorithm performance in some environments. Also, with the increase in the number of actions, such an RL algorithm may require more iterations to converge, leading to a longer time in algorithm training.

In summary, the BaS-MAQL algorithm is an effective tool for simulating electricity market dynamics since it can mitigate the Q-value over-estimation and is easy to implement despite discrete action space and potential long convergence time. However, it may also be restricted to problems with specific structures and may require more computation labor with an increasing number of action spaces.

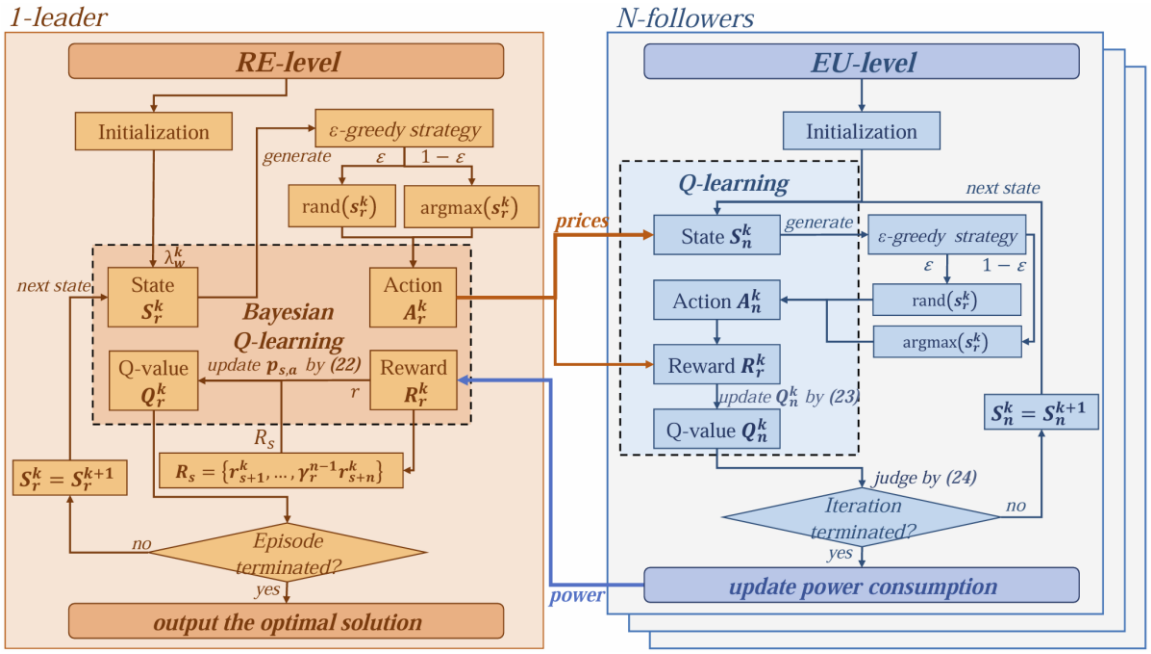


Fig. 5.2 Flowchart of the proposed BaS-MAQL algorithm

5.4 Case Study

5.4.1 Simulation Setup

The main objective of this case study is to investigate the MSNE of the MDP formulated in Section 5.3 by using the proposed BaS-MAQL algorithm, and verify the computational performance of the proposed algorithm. Therefore, a test system based on the IEEE 33-bus system consisting of 50 EUs is adopted to simulate the system operation [98].

The entire transaction period of one day is divided into 24 time-slots. Real transaction data of Commonwealth Edison (ComEd) are extracted herein as the 24 hours wholesale electricity price shown in Fig.4.3 to simulate the REM operation to the greatest extent and calculate the retailer's profits. Otherwise, the power demand of EUs

is set using real-world data [99], and the uncertain noise is added to the demand by following a normal Gaussian distribution. For implementing the proposed MAQL using the tabular-RL algorithm, the action spaces of both retailer and EUs are discretized within a granularity of \$0.5/kWh and 5kW, respectively. The retail electricity price limitation is set as \$5-7.8/kWh, which is equally discretized with a granularity of \$0.4/kWh.

Table 5.2 Parameter settings for the simulation

Parameter	Value	Parameter	Value
λ_i	1	ω_r^k	[7.6, 13.5]
β_i	0.3	τ_i	0.1
γ_i	0.7	γ_r	0.7

For EUs, the power consumption of baseline appliances for each hour is randomly chosen from (10, 15, 20, 25, 30) \$/kWh. The electricity consumptions of elastic appliances, as well as the action space of EUs, are set from [0,20] kWh within a granularity of 5 kWh.

The setting of hyper-parameters is summarized in Table.5.2. These parameters are fine-tuned when training algorithms, and are carefully selected until the algorithm can converge stably. The parameter ϵ of the greedy strategy is initially set to 0, and then increase 0.0005 per iteration until reaching 0.9950.

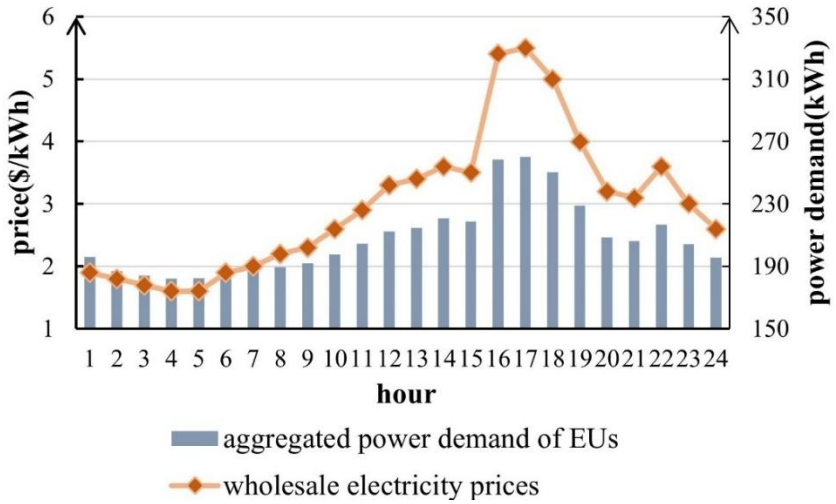


Fig. 5.3 Clearing prices set in WEM and power demand in each time slot

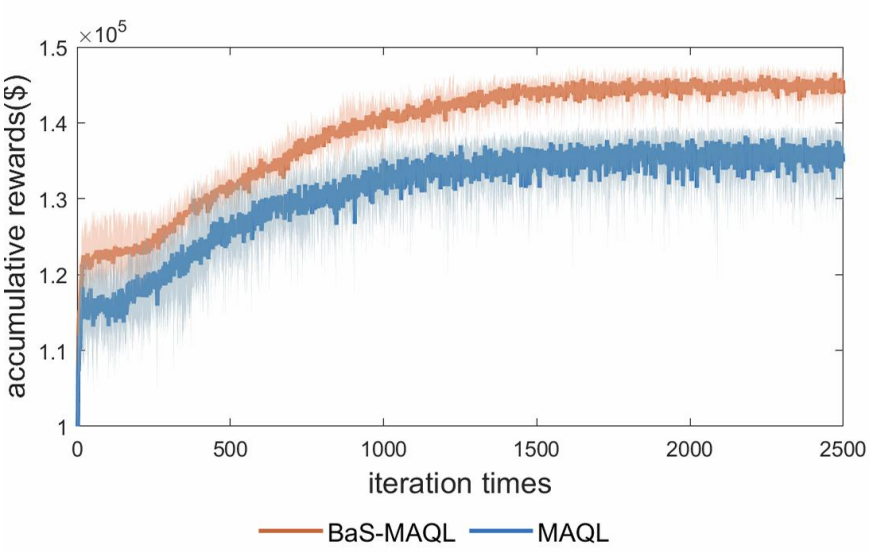
5.4.2 Convergence Analysis

In this section, the computational performance of the proposed BaS-MAQL algorithm with typical MAQL algorithm are compared. The comparison is based on the convergence speed and accumulative rewards.

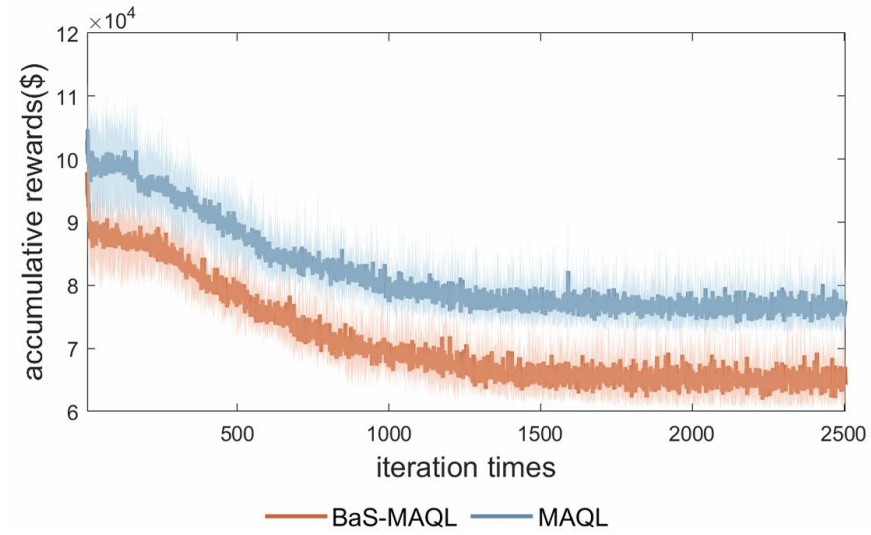
Fig.5.4 illustrates the convergence process of both the retailer and all EUs. It can be observed that the proposed BaS-MAQL algorithm achieves convergence with significantly higher profits for the retailer within 1500 episodes, whereas both the typical MAQL algorithms converge to a sub-optimal equilibrium. The retailer's profits in BaS-MAQL are 7\% higher than those in MAQL. However, the profits of the EUs converge to lower levels due to the near-zero-sum non-cooperative nature of the game.

The sub-optimality of the equilibrium in MAQL arises from its decision-making process, which is based on an environment with substantial uncertainty stemming from the mixed strategies of other agents. This uncertainty makes it easier to reach a sub-optimal equilibrium in a multi-agent system. In the context of the Stackelberg game, this sub-optimality manifests as the inability to achieve underestimated accumulative rewards for the leader. This is primarily due to the significant uncertainty brought about by multiple followers and the resulting large action spaces.

In contrast, the BaS-MAQL algorithm takes the probabilistic distribution of the Q-value into account. This modification can effectively help dealing with the demand uncertainty and estimating the best solution in a Stackelberg game with large-scale followers. As a result, the solution of the BaS-MAQL show a higher total reward for the retailer.



(a)

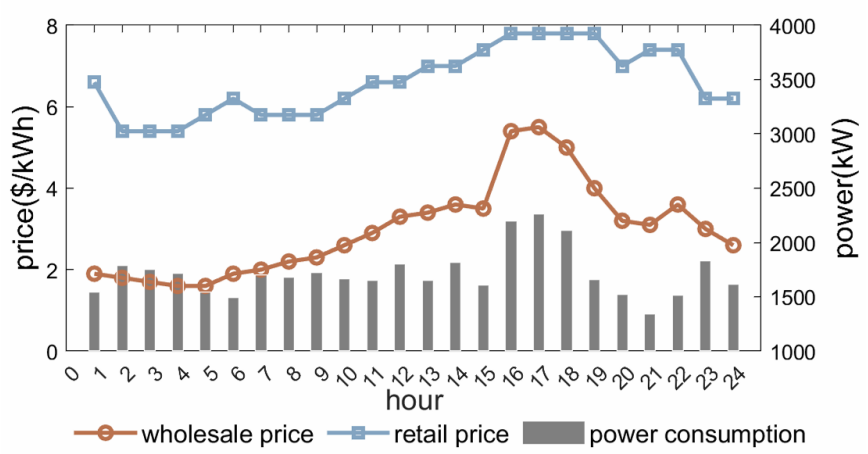


(b)

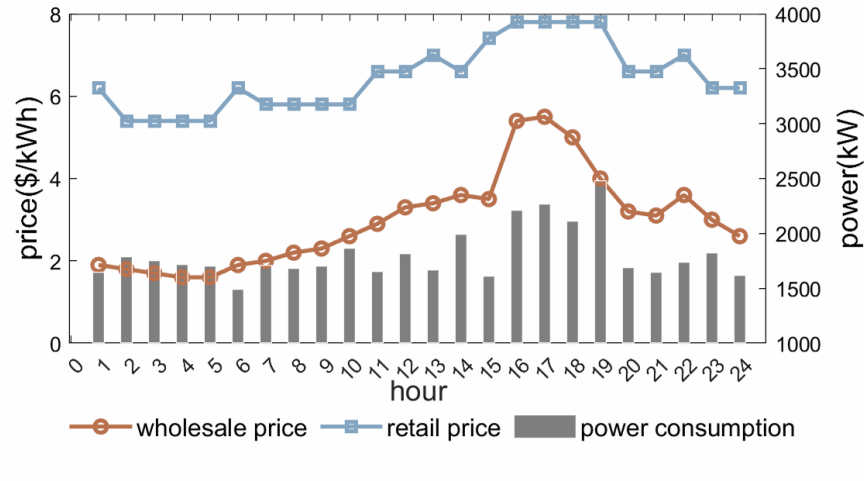
Fig.5.4 Accumulative rewards converge procedure of (a) the retailer and (b) EUs among different algorithms.

5.4.3 MSNE analysis

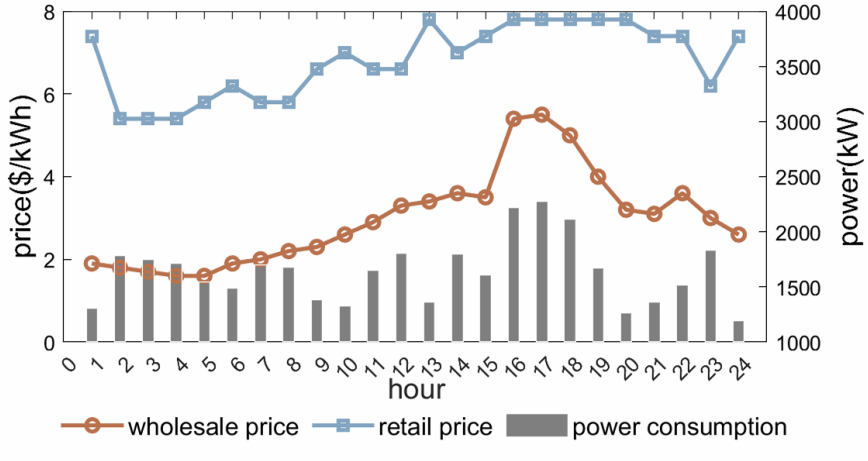
In this subsection, the transaction results to investigate the existence and impact of MSNE are analyzed. The converged transaction results mainly include retail electricity price and power consumption for the entire day (24 hours), and are depicted in Fig.5.5 (a)-(c) as three pure strategies denoted as SPE 1, 2, and 3. It should be noted that the aforementioned pure SPE may be similar to be thought of as the "local optima" in the



(a)



(b)



(c)

Fig.5.5 Convergence results containing (a) SPE 1, (b) SPE 2 and (c) SPE 3

optimization problem. However, as these SPE are all solutions for the Bayesian Stackelberg game consisting of multiple players, SPE cannot be seen as local optima, but an equilibrium where each player cannot improve his profits by changing his strategy.

Three SPE are achieved in the simulation, as shown in Fig.5.5 (a)-(c). The retail electricity prices exhibit similar fluctuations to the wholesale electricity prices, reflecting the retailers' costs in the wholesale market. At the beginning of the day (0:00-2:00), due to low power consumption, prices in all three SPE decrease and then stabilize at a low level ($5.4 or $5.8/kWh) from 2:00 to 9:00. Between 9:00 and 14:00, retail electricity prices experience a sharp increase, remaining relatively high ($6.2-$7.8/kWh) from 14:00 to 19:00. During the on-peak period of 14:00-19:00, the price gap, which represents the difference between retail and wholesale electricity prices and indicates retailers' profits, remains stable at around $3.50/kWh. However, during the peak hours between 16:00 and 19:00, the price gap narrows sharply due to the high wholesale price and the upward constraint on the retail price ($7.8/kWh), resulting in relatively low profits for retailers (around $5200) despite the highest peak power consumption (over 2000 kWh). The reduced price gap in this period is around $2.4/kWh.

To further illustrate the MSNE among the three SPE, the profits of the retailer and EUs in each SPE in Fig.5.5 (a) and (b) are compared. All time slots are classified into three scenarios based on the profit differences, providing insights into the appearance of MSNE. Scenario 1 represents a special case of MSNE where both the retailer and EUs play pure strategies. In Scenario 2, one player (the retailer or EUs) plays mixed strategies while the other plays pure strategies. Scenario 3 occurs when both the retailer and EUs employ mixed strategies. The scenario classification for each time slot throughout the day is summarized in Table 5.3.

Table 5.3 Scenario classification of each time slot

Scenarios	Time slots
1	1:00-2:00, 2:00-3:00, 3:00-4:00, 5:00-6:00, 6:00-7:00, 7:00-8:00, 10:00-11:00, 14:00-15:00, 15:00-16:00, 16:00-17:00, 17:00-18:00, 22:00-23:00
2	0:00-1:00, 8:00-9:00, 9:00-10:00, 11:00-12:00, 12:00-13:00, 13:00-14:00, 19:00-20:00, 21:00-22:00, 23:00-24:00
3	4:00-5:00, 18:00-19:00, 20:00-21:00

In Scenario 1 of MSNE, a pure strategy NE is considered as a special case where there is only one optimal strategy for both the retailer and EUs in each time slot. For example, all three SPE converge to the same NE at 2:00-3:00 with a retail electricity price of $5.4/kWh and a total power consumption of approximately 1980 kWh. If EUs consume more or less than 1980 kWh with a constant retail electricity price, they would incur profit losses and have no motivation to deviate from their actions. The same applies to the retailer when the power consumption of EUs is constant.

Scenario 2 of MSNE occurs when the optimal strategy of the retailer (or EUs) is a mixed strategy with a probability distribution over multiple pure strategies, while the optimal strategy of the opponent (EUs or the retailer) remains a pure strategy. In this case, the participant with MSNE will randomly choose different actions as these pure strategies result in the same profit. However, the opponent will suffer profit losses. For example, during 4:00-5:00, the retailer's profits are the same ($2700) despite setting different retail electricity prices of $5.8/kWh and $5.4/kWh in SPE 1 and 2, respectively. This is because EUs consume 1545 kWh and 1705 kWh during 5:00-6:00 under different retail electricity prices in SPE 1 and 2. However, for EUs, the total power consumptions of 1545 kWh and 1705 kWh are optimal strategies during 5:00-6:00 under the retail electricity prices of $5.8/kWh and $5.4/kWh, respectively. Specifically, EUs tend to consume more power under lower retail electricity prices and less power under higher prices, resulting in the same profits for the retailer with two different retail

electricity prices, thereby maximizing profits. At the same time, the retailer is indifferent to these two retail electricity prices, leading to the execution of mixed strategies containing both of them. Similarly, EUs may also execute mixed strategies during certain time slots, resulting in different profits for the retailer, such as during 8:00-9:00 and 13:00-14:00. Hence, it is evident that the mixed strategy of one participant may cause profit losses for the other

Scenario 3 of MSNE occurs when both the retailer and EUs adopt mixed strategies. This is reflected in Fig.4.5 (a)-(b), where all participants have different profits in different SPE. For example, during 20:00-21:00, the retailer sets retail electricity prices as \$7.4/kWh, \$6.6/kWh, and \$7.4/kWh in SPE 1, 2, and 3, respectively, resulting in profits of \$5710, \$5700, and \$5820, respectively. Meanwhile, EUs consume power of 1328 kWh, 1629 kWh, and 1353 kWh, resulting in profits of \$1460, \$2710, and \$1470 in SPE 1, 2, and 3, respectively. Thus, there are three different outcomes for each SPE during 20:00-21:00 due to the adoption of mixed strategies by both the retailer and EUs. Specifically, if two retail electricity prices yield similar profits for the retailer, the retailer may play a mixed strategy containing both prices. Similarly, under constant retail electricity prices, EUs may also adopt mixed strategies containing two pure strategies, resulting in similar profits for them. When both the retailer and EUs play mixed strategies simultaneously, it falls under the third scenario.

Table 5.4 Results comparison between three SPE

SPE	Total profit of the RE (\$)	Total profit of EUs (\$)	Total power savings (kWh)
1	146260	68950	8700
2	146080	74320	7360
3	146100	61580	10550

Based on Table 5.4 and the analysis provided, it can be observed that scenario 1, 2, and 3 occur 13 times, 9 times, and 2 times, respectively. This indicates that players (the retailer and EUs) engage in mixed strategy transactions for approximately half of the

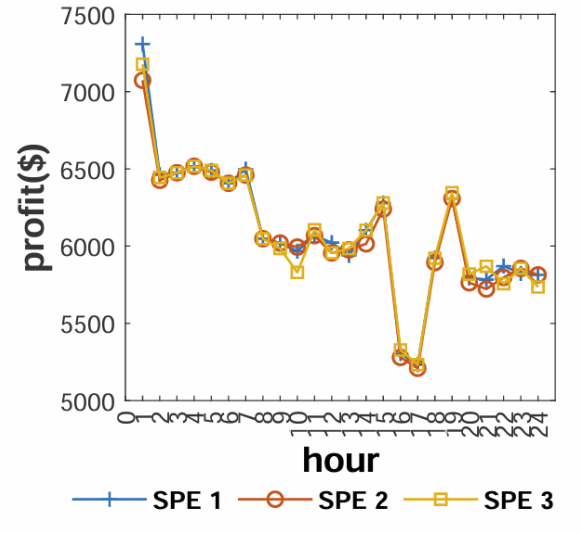
period, while both the retailer and EUs employing mixed strategies in a single time slot is relatively rare.

Furthermore, the retailer's profits are not significantly affected by the mixed strategy of EUs, with fluctuations always remaining below $100. In contrast, the profits of EUs exhibit significant variations, often exceeding $1000, during numerous time slots when the retailer employs a mixed strategy. This disparity is attributed to the retailer's market power as the price-maker in REM, which allows them to enhance profits. Since the retailer has almost the same profits at multiple equilibria while end users do not, it is possible for regulators to propose new market rules for a higher social welfare, for example, dynamic pricing caps across different transaction interval.

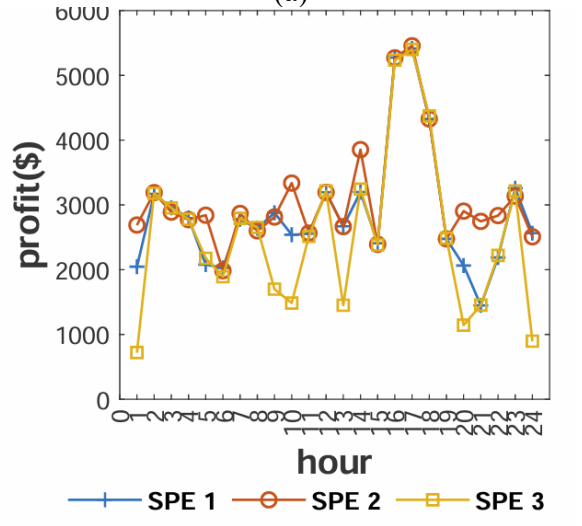
5.4.4 Analysis of the power savings

This subsection focuses on power savings in the SPE to demonstrate the effectiveness of RTP-DR in terms of energy-saving and network power balancing. Fig.4.6 (c) presents the power savings, which reflects the discrepancy between expected power demand and actual consumption. Proper adjustment of power savings is crucial for system operation. In Fig.5.5 (a)-(c), even though EUs have high power consumption during the on-peak period (15:00-19:00), SPE 1 and 2 exhibit high power savings (over 300kWh) in Fig.5.6 (c). This suggests that EUs may display greater demand elasticity during peak hours due to high demand. Hence, implementing the RTP-DR scheme becomes vital to stimulate EUs' demand elasticity and facilitate power balance between the demand and supply sides.

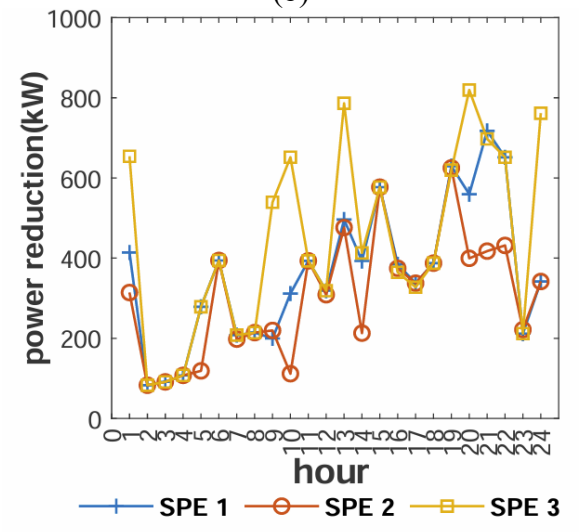
To summarize, the profits and power savings of the three SPE are compared. Table 5.4 presents the total profits of the retailer and 50 EUs, along with the total power savings. SPE 1 and 2 yield higher overall profits for both the retailer and EUs compared to SPE 3. The profit difference for the retailer between SPE 1 and 2 is $680 out of $146,460, while for EUs, it is $5,370. The profit differences for EUs in each SPE appear to be larger than those for the retailer. The retailer's profits remain relatively stable as they have market power as the price-maker in REM. Therefore, improving REM



(a)



(b)



(c)

Fig.5.6 Comparison of each SPE in (a) profits of the RE, (b) profits of all EUs, (c) power savings of EUs.

regulations to protect EU profits is necessary. Furthermore, SPE 1 exhibits curtail 2% higher power savings than SPE 2, leading to lower power and energy consumption. Although SPE 3 has the highest power savings (10,550kWh), market participants are dissatisfied with its lowest EU profits ($61,580) compared to the other SPE.

5.5 Summary

In summary, this chapter formulates the RTP-DR problem between the retailer and EUs into a Bayesian Stackelberg game by considering the incomplete information in REM transactions. The game is non-convex due to the network constraints and temporal-correlated non-linear power usage, thus is analyzed from the view of MSNE. Subsequently, a novel BaS-MAQL is proposed to stimulate the market transaction and estimate the SPE in this game. By representing the numerical Q-value with probability distributions, this algorithm offers solution optimality, and robustness under uncertain market transactions. Additionally, it can be scalable and adaptable to diverse scenarios due to its flexible structure and few hyperparameters. The simulation results demonstrate the existence of MSNE by comparing the SPEs in the transactions.

Nevertheless, there remain several limitations in the present study that point to promising avenues for further investigation: 1) Current framework does not encompass a variety of energy devices—such as electric vehicles, solar panels, and energy storage systems—nor does it factor in potential bounded rationality in end-user decision-making. Incorporating these aspects would significantly alter energy consumption pattern dramatically, lead to different market equilibria and provide deeper insights.

2) By not fully modeling retailer involvement in both wholesale electricity and capacity markets, this work may be overlooking intricate inter-dependencies between various market signals and decision-making processes. Future work could investigate how retailer strategies in these markets affect overall efficiency in retail market and consumer welfare.

Chapter VI
Conclusions and Future Perspectives

6.1 Conclusions

Motivated by the need for smart control of distributed energy demand in the face of growing renewable penetration, energy integration, and dynamic pricing, this thesis focuses on developing RL techniques in demand side energy management across multiple scales. Specifically, novel RL algorithms are designed for profit maximization accounting for external uncertainty, operational safety, and equilibrium estimation for demand side scale ranging from individual buildings to community microgrids and up to REM interactions.

Moreover, the RL designs in this thesis are guided by the dominant challenge at each scale of demand-side energy management. At the building level (BIES), limited dispatchable assets (BESS, micro-CHP, GB) make operations highly sensitive to demand and price uncertainty; accordingly, Chapter 3 proposes a forecast-enhanced RL approach that integrates energy and price forecasting with control. At the community level (ICES), safety precedes economics under multi-energy network constraints; Chapter 4, therefore, develops a safe RL method for constrained operational optimization. In the retail market (REM), RTP-DR introduces strategic interactions among multiple stakeholders; Chapter 5 focuses on equilibrium learning to extract actionable insights. Overall, the thesis tailors RL algorithms to the most critical issue in each scenario across scales.

The key contributions of this work are summarized as follows:

1) *Forecast-Enhanced RL for Operation in Building Integrated Energy System*: An RL-based algorithm is designed for grid connected BIES that leverages load and price forecasting to make proactive energy management decisions. This approach satisfies multi-energy demands and device constraints while reducing the energy cost. The

forecasting method, TFT, is interpretable and provides more information for subsequent decision-making. The proposed approach shows good generalization performance with real-world data in different seasons.

2) *Multi-Network Constrained Integrated Community Energy Systems Model:* A novel MNC-ICES model is proposed to interpret the concept of ICES. The proposed model accounts for the constraints of multiple networks, which captures the physical characteristics of energy flow and imposes security operational constraints for the distribution level energy transmissions. Energy devices are modeled in high fidelity to describe the realistic physical operating attributes in practice. Additionally, the renewable uncertainty and integrated demand elasticity are considered to describe the novel characteristics of modern distribution-level energy systems. Overall, the proposed model can be implemented as a basis for practical network-constrained community operation tools.

3) *Safe Reinforcement Learning for Multi-Network Constrained Integrated Community Energy Systems*: A novel Safe RL algorithm, namely PD-TD3, is proposed to solve the C-MDP and the constrained operational optimization problem in MNC-ICES. In the proposed algorithm, constraints are incorporated directly into the learning process to ensure practical feasibility. Specifically, the PD-TD3 algorithm using double networks reduces the over-estimation problem of the action value for both the reward and cost, and the delayed update stabilizes the training process of policy and its dual variable. With accurate estimation of Q values, the proposed algorithm converges to the optimal solution that balances the maximal profits and the lowest constraint violation.

4) *Game-Theoretic MARL for Real-Time Pricing and Demand Response*: The interaction between EUs and a utility (or retailer) is modeled as a dynamic Stackelberg game and MARL is applied to find multiple behavioral equilibria. Specifically, a pricing strategy and corresponding consumption policies can be learned such that no participant has an incentive to deviate (akin to a Nash equilibrium in demand response). This is one of the first demonstrations of MARL achieving stable market outcomes in a demand

response context, bridging the gap between individual learning agents and system-level economic equilibrium.

The findings of this thesis demonstrate that RL techniques can effectively manage and reduce energy demand on the consumer side, responding adaptively to price incentives and contributing to grid reliability. By fulfilling these objectives, the thesis has demonstrated that RL can serve as an effective tool for energy management on the demand side. The outcomes have several important implications. For demand-side consumers, this suggests that deploying RL algorithms as smart agents in energy management systems or via community aggregators could automate DR at scale, achieving cost savings for EUs and operational benefits for utilities. For the field of power systems, this work provides experimental proof that RL agents, if properly designed, can maximize utilities' profits while coordinating to achieve grid-level goals such as peak shaving and load shifting with only price incentives. Importantly, the multi-level energy management–spanning devices to markets – indicates that coordination at different layers of the power system is feasible with decentralized artificial intelligence controllers, potentially accelerating the adoption of smart grid technologies.

6.2 Future Perspectives

The thesis has proposed several algorithm schemes for multi-scale demand-side energy management problems, from individual BIES to multi-network constrained ICES and RTP-DR problems in REM. To make the current work more comprehensive, the following topics should be investigated in the future.

1) The power system calls for Safe RL algorithms that can enforce safety constraints in both exploration and exploitation because of the severe constraints. Most currently developed Safe RL algorithms, for example, Lagrangian-based Safe RL, enforce soft constraints by penalizing the constraints violation in exploration while not guaranteeing hard constraint enforcement, which may endanger energy system operation

2) Most RL algorithms are mathematically developed to solve dynamic decision-making problems in stationary MDPs. However, the practical operation environment for demand-side energy systems has seasonality and trends inherently due to the seasonal changes and industrial expansion, which makes the environment highly non-stationary so as to reduce the performance of the RL algorithms. Moreover, another important scenario in the demand-side, multi-agent interaction environment is also considered as non-stationary for the evolution and stochastic of others' strategies. Such non-stationarity in the scenario may lead to the failure of the algorithm learning. For the reasons above, developing a novel RL algorithm that can handle the non-stationary environment is a significant step in promoting the real-world implementation of RL techniques.

References

[1] S. Dunn and J. A. Peterson, *Micropower: the next electrical era*. Worldwatch Institute Washington, DC, 2000.

[2] R. Shulga, A. Petrov, and I. Putilova, "The Arctic: Ecology and hydrogen energy," *International Journal of Hydrogen Energy,* vol. 45, no. 11, pp. 7185-7198, 2020.

[3] K. E. Holbert, G. Heydt, and H. Ni, "Use of satellite technologies for power system measurements, command, and control," *Proceedings of the IEEE,* vol. 93, no. 5, pp. 947-955, 2005.

[4] W. N. Association. "Carbon Dioxide Emissions From Electricity." https://world-nuclear.org/information-library/energy-and-the-environment/carbon-dioxide-emissions-from-electricity.aspx (accessed Mar. 24th, 2025).

[5] S. Alfonso, M. Gesto, and B. Sadoul, "Temperature increase and its effects on fish stress physiology in the context of global warming," *Journal of Fish Biology,* vol. 98, no. 6, pp. 1496-1508, 2021.

[6] P. Agreement, "Paris agreement," in *report of the conference of the parties to the United Nations framework convention on climate change (21st session, 2015: Paris). Retrived December*, 2015, vol. 4, no. 2017: HeinOnline, p. 2.

[7] G. D. Kroeger and M. G. Burgess, "Electric utility plans are consistent with Renewable Portfolio Standards and Clean Energy Standards in most US states," *Climatic Change,* vol. 177, no. 1, p. 1, 2024.

[8] E. Sarker *et al.*, "Progress on the demand side management in smart grid and optimization approaches," *International Journal of Energy Research,* vol. 45, no. 1, pp. 36-64, 2021.

[9] G. Strbac, "Demand side management: Benefits and challenges," *Energy policy,* vol. 36, no. 12, pp. 4419-4426, 2008.

[10] X. Cao, X. Dai, and J. Liu, "Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade," *Energy and buildings,* vol. 128, pp. 198-213, 2016.

[11] B. P. Koirala, E. Koliou, J. Friege, R. A. Hakvoort, and P. M. Herder, "Energetic communities for community energy: A review of key issues and trends shaping integrated community energy systems," *Renewable and Sustainable Energy Reviews,* vol. 56, pp. 722-744, 2016.

[12] P. Samadi, A.-H. Mohsenian-Rad, R. Schober, V. W. Wong, and J. Jatskevich, "Optimal real-time pricing algorithm based on utility maximization for smart grid," in *2010 First IEEE international conference on smart grid communications*, 2010: IEEE, pp. 415-420.

[13] T. M. Mitchell and T. M. Mitchell, *Machine learning* (no. 9). McGraw-hill New York, 1997.

[14] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research,* vol. 4, pp. 237-285, 1996.

[15] S. Yang, H. O. Gao, and F. You, "Integrated optimization in operations control and systems design for carbon emission reduction in building electrification with distributed energy resources," *Advances in Applied Energy,* vol. 12, p. 100144, 2023.

[16] Z. Qi, Q. Gao, Y. Liu, Y. Yan, and J. D. Spitler, "Status and development of hybrid energy systems from hybrid ground source heat pump in China and other countries," *Renewable and Sustainable Energy Reviews,* vol. 29, pp. 37-51, 2014.

[17] H. Qiu, V. Veerasamy, C. Ning, Q. Sun, and H. B. Gooi, "Two-Stage Robust Optimization for Assessment of PV Hosting Capacity Based on Decision-Dependent Uncertainty," *Journal of Modern Power Systems and Clean Energy,* 2024.

[18] X. Huang *et al.*, "Heat and power load dispatching considering energy storage of district heating system and electric boilers," *Journal of Modern Power Systems and Clean Energy,* vol. 6, no. 5, pp. 992-1003, 2018.

[19] C. Huang, H. Zhang, L. Wang, X. Luo, and Y. Song, "Mixed deep reinforcement learning considering discrete-continuous hybrid action space for smart home energy management," *Journal of Modern Power Systems and Clean Energy,* vol. 10, no. 3, pp. 743-754, 2022.

[20] Z. Zhu, Z. Hu, K. W. Chan, S. Bu, B. Zhou, and S. Xia, "Reinforcement learning in deregulated energy market: A comprehensive review," *Applied Energy,* vol. 329, 2023, doi: 10.1016/j.apenergy.2022.120212.

[21] H. Zhao *et al.*, "Active dynamic aggregation model for distributed integrated energy system as virtual power plant," *Journal of Modern Power Systems and Clean Energy,* vol. 8, no. 5, pp. 831-840, 2020.

[22] C. Lv *et al.*, "Model predictive control based robust scheduling of community integrated energy system with operational flexibility," *Applied energy,* vol. 243, pp. 250-265, 2019.

[23] N. Li, R. A. Hakvoort, and Z. Lukszo, "Cost allocation in integrated community energy systems-A review," *Renewable and Sustainable Energy Reviews,* vol. 144, p. 111001, 2021.

[24] A. Caramizaru and A. Uihlein, *Energy communities: an overview of energy and social innovation*. Publications Office of the European Union Luxembourg, 2020.

[25] J. Wang, R. El Kontar, X. Jin, and J. King, "Electrifying high-efficiency future communities: impact on energy, emissions, and grid," *Advances in Applied Energy,* vol. 6, p. 100095, 2022.

[26] R. Liu, T. Yang, G. Sun, S. Lin, F. Li, and X. Wang, "Multi-objective optimal scheduling of community integrated energy system considering comprehensive customer dissatisfaction model," *IEEE Transactions on Power Systems,* 2022.

[27] Y. Li, B. Wang, Z. Yang, J. Li, and C. Chen, "Hierarchical stochastic scheduling of multi-community integrated energy systems in uncertain environments via Stackelberg game," *Applied Energy,* vol. 308, 2022, doi: 10.1016/j.apenergy.2021.118392.

[28] T. Jiang, X. Dong, R. Zhang, and X. Li, "Strategic active and reactive power scheduling of integrated community energy systems in day-ahead distribution electricity market," *Applied Energy,* vol. 336, 2023, doi: 10.1016/j.apenergy.2022.120558.

[29] X. Jin, Q. Wu, H. Jia, and N. D. Hatziargyriou, "Optimal integration of building heating loads in integrated heating/electricity community energy systems: A bi-level MPC approach," *IEEE Transactions on Sustainable Energy,* vol. 12, no. 3, pp. 1741-1754, 2021.

[30] W. Lin *et al.*, "Decentralized optimal scheduling for integrated community energy system via consensus-based alternating direction method of multipliers," *Applied Energy,* vol. 302, p. 117448, 2021.

[31] C. Zhou *et al.*, "Two-stage robust optimization for space heating loads of buildings in integrated community energy systems," *Applied Energy,* vol. 331, 2023, doi: 10.1016/j.apenergy.2022.120451.

[32] A. Dolatabadi, H. Abdeltawab, and Y. A.-R. I. Mohamed, "A Novel Model-Free Deep Reinforcement Learning Framework for Energy Management of a PV Integrated Energy Hub," *IEEE Transactions on Power Systems,* vol. 38, no. 5, pp. 4840-4852, 2023, doi: 10.1109/tpwrs.2022.3212938.

[33] T. Ma, W. Pei, H. Xiao, L. Kong, Y. Mu, and T. Pu, "The energy management strategies based on dynamic energy pricing for community integrated energy system considering the interactions between suppliers and users," *Energy,* vol. 211, 2020, doi: 10.1016/j.energy.2020.118677.

[34] Y. Li, M. Han, Z. Yang, and G. Li, "Coordinating flexible demand response and renewable uncertainties for scheduling of community integrated energy systems with an electric vehicle charging station: A bi-level approach," *IEEE Transactions on Sustainable Energy,* vol. 12, no. 4, pp. 2321-2331, 2021.

[35] C. Liu, C. Wang, Y. Yin, P. Yang, and H. Jiang, "Bi-level dispatch and control strategy based on model predictive control for community integrated energy system considering dynamic response performance," *Applied Energy,* vol. 310, p. 118641, 2022.

[36] Y. Li *et al.*, "Optimization of integrated energy system for low-carbon community considering the feasibility and application limitation," *Applied Energy,* vol. 348, p. 121528, 2023.

[37] F. Han, J. Zeng, J. Lin, and C. Gao, "Multi-stage distributionally robust optimization for hybrid energy storage in regional integrated energy system considering robustness and nonanticipativity," *Energy,* vol. 277, 2023, doi: 10.1016/j.energy.2023.127729.

[38] E. A. M. Cesena, E. Loukarakis, N. Good, and P. Mancarella, "Integrated electricity–heat–gas systems: Techno–economic modeling, optimization, and application to multienergy districts," *Proceedings of the IEEE,* vol. 108, no. 9, pp. 1392-1410, 2020.

[39] R. Lu, R. Bai, Y. Huang, Y. Li, J. Jiang, and Y. Ding, "Data-driven real-time price-based demand response for industrial facilities energy management," *Applied Energy,* vol. 283, p. 116291, 2021.

[40] M. Casini, *Construction 4.0: Advanced technology, tools and materials for the digital transformation of the construction industry*. Woodhead Publishing, 2021.

[41] Z. Wang, M. Sun, C. Gao, X. Wang, and B. C. Ampimah, "A new interactive real-time pricing mechanism of demand response based on an evaluation model," *Applied Energy,* vol. 295, p. 117052, 2021.

[42] K. Li, F. Wang, Z. Mi, M. Fotuhi-Firuzabad, N. Duić, and T. Wang, "Capacity and output power estimation approach of individual behind-the-meter distributed photovoltaic system for demand response baseline estimation," *Applied energy,* vol. 253, p. 113595, 2019.

[43] M. Chen, C. Gao, M. Shahidehpour, and Z. Li, "Incentive-compatible demand response for spatially coupled internet data centers in electricity markets," *IEEE Transactions on Smart Grid,* vol. 12, no. 4, pp. 3056-3069, 2021.

[44] P. Yang, G. Tang, and A. Nehorai, "A game-theoretic approach for optimal time-of-use electricity pricing," *IEEE Transactions on Power Systems,* vol. 28, no. 2, pp. 884-892, 2012.

[45] F. A. Wolak, "Residential customer response to real-time pricing: The anaheim critical peak pricing experiment," 2007.

[46] H. Li, Z. Wan, and H. He, "Real-time residential demand response," *IEEE Transactions on Smart Grid,* vol. 11, no. 5, pp. 4144-4154, 2020.

[47] M. Yu and S. H. Hong, "A real-time demand-response algorithm for smart grids: A stackelberg game approach," *IEEE Transactions on smart grid,* vol. 7, no. 2, pp. 879-888, 2015.

[48] L. He, Y. Liu, and J. Zhang, "An occupancy-informed customized price design for consumers: A Stackelberg game approach," *IEEE Transactions on Smart Grid,* vol. 13, no. 3, pp. 1988-1999, 2022.

[49] T. Lu, Z. Wang, J. Wang, Q. Ai, and C. Wang, "A data-driven Stackelberg market strategy for demand response-enabled distribution systems," *IEEE Transactions on Smart Grid,* vol. 10, no. 3, pp. 2345-2357, 2018.

[50] N. Aguiar, A. Dubey, and V. Gupta, "Network-constrained Stackelberg game for pricing demand flexibility in power distribution systems," *IEEE Transactions on Smart Grid,* vol. 12, no. 5, pp. 4049-4058, 2021.

[51] W. Chen, J. Qiu, and Q. Chai, "Customized critical peak rebate pricing mechanism for virtual power plants," *IEEE Transactions on Sustainable Energy,* vol. 12, no. 4, pp. 2169-2183, 2021.

[52] V. C. Pandey, N. Gupta, K. R. Niazi, A. Swarnkar, and R. A. Thokar, "A hierarchical price-based demand response framework in distribution network," *IEEE Transactions on Smart Grid,* vol. 13, no. 2, pp. 1151-1164, 2021.

[53] Z. Liu and L. Wang, "Defense strategy against load redistribution attacks on power systems considering insider threats," *IEEE Transactions on Smart grid,* vol. 12, no. 2, pp. 1529-1540, 2020.

[54] C. Lindig-León, G. Schmid, and D. A. Braun, "Nash equilibria in human sensorimotor interactions explained by Q-learning with intrinsic costs," *Scientific Reports,* vol. 11, no. 1, p. 20779, 2021.

[55] R. B. Myerson, "Refinements of the Nash equilibrium concept," *International journal of game theory,* vol. 7, pp. 73-80, 1978.
[56] M. Samadi, H. Kebriaei, H. Schriemer, and M. Erol-Kantarci, "Stochastic demand response management using mixed-strategy Stackelberg game," *IEEE Systems Journal,* vol. 16, no. 3, pp. 4708-4718, 2022.
[57] D. Jay and K. Swarup, "Game theoretical approach to novel reactive power ancillary service market mechanism," *IEEE Transactions on Power Systems,* vol. 36, no. 2, pp. 1298-1308, 2020.
[58] D. Qiu, Y. Wang, T. Zhang, M. Sun, and G. Strbac, "Hierarchical multi-agent reinforcement learning for repair crews dispatch control towards multi-energy microgrid resilience," *Applied Energy,* vol. 336, p. 120826, 2023.
[59] C. Feng, Y. Wang, Q. Chen, Y. Ding, G. Strbac, and C. Kang, "Smart grid encounters edge computing: Opportunities and applications," *Advances in Applied Energy,* vol. 1, p. 100006, 2021.
[60] Z. Chen, F. Xiao, F. Guo, and J. Yan, "Interpretable machine learning for building energy management: A state-of-the-art review," *Advances in Applied Energy,* vol. 9, p. 100123, 2023.
[61] A. Dolatabadi, H. Abdeltawab, and Y. A.-R. I. Mohamed, "A novel model-free deep reinforcement learning framework for energy management of a PV integrated energy hub," *IEEE Transactions on Power Systems,* 2022.
[62] Y. Li, F. Bu, Y. Li, and C. Long, "Optimal scheduling of island integrated energy systems considering multi-uncertainties and hydrothermal simultaneous transmission: A deep reinforcement learning approach," *Applied Energy,* vol. 333, p. 120540, 2023.
[63] H. Li and H. He, "Learning to Operate Distribution Networks With Safe Deep Reinforcement Learning," *IEEE Transactions on Smart Grid,* vol. 13, no. 3, pp. 1860-1872, 2022, doi: 10.1109/tsg.2022.3142961.
[64] B. Wang, Y. Li, W. Ming, and S. Wang, "Deep reinforcement learning method for demand response management of interruptible load," *IEEE Transactions on Smart Grid,* vol. 11, no. 4, pp. 3146-3155, 2020.
[65] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe Off-Policy Deep Reinforcement Learning Algorithm for Volt-VAR Control in Power Distribution Systems," *IEEE Transactions on Smart Grid,* vol. 11, no. 4, pp. 3008-3018, 2020, doi: 10.1109/tsg.2019.2962625.
[66] Y. Chow, O. Nachum, A. Faust, E. Duenez-Guzman, and M. Ghavamzadeh, "Lyapunov-based safe policy optimization for continuous control," *arXiv preprint arXiv:1901.10031,* 2019.
[67] A. R. Sayed, X. Zhang, Y. Wang, G. Wang, J. Qiu, and C. Wang, "Online operational decision-making for integrated electric-gas systems with safe reinforcement learning," *IEEE Transactions on Power Systems,* 2023.
[68] D. Qiu, Z. Dong, X. Zhang, Y. Wang, and G. Strbac, "Safe reinforcement learning for real-time automatic control in a smart energy-hub," *Applied Energy,* vol. 309, p. 118403, 2022.

[69] Z. Yan and Y. Xu, "A hybrid data-driven method for fast solution of security-constrained optimal power flow," *IEEE Transactions on Power Systems,* vol. 37, no. 6, pp. 4365-4374, 2022.
[70] D. Qiu, J. Wang, Z. Dong, Y. Wang, and G. Strbac, "Mean-field multi-agent reinforcement learning for peer-to-peer multi-energy trading," *IEEE Transactions on Power Systems,* 2022.
[71] D. Qiu, T. Chen, G. Strbac, and S. Bu, "Coordination for multienergy microgrids using multiagent reinforcement learning," *IEEE Transactions on Industrial Informatics,* vol. 19, no. 4, pp. 5689-5700, 2022.
[72] Y. Zhou, Z. Ma, J. Zhang, and S. Zou, "Data-driven stochastic energy management of multi energy system using deep reinforcement learning," *Energy,* vol. 261, p. 125187, 2022.
[73] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting,* vol. 37, no. 4, pp. 1748-1764, 2021.
[74] W. J. Von Eschenbach, "Transparency and the black box problem: Why we do not trust AI," *Philosophy & Technology,* vol. 34, no. 4, pp. 1607-1622, 2021.
[75] G. Pan, W. Gu, Y. Lu, H. Qiu, S. Lu, and S. Yao, "Optimal planning for electricity-hydrogen integrated energy system considering power to hydrogen and heat and seasonal storage," *IEEE Transactions on Sustainable Energy,* vol. 11, no. 4, pp. 2662-2676, 2020.
[76] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka, "A multi-horizon quantile recurrent forecaster," *arXiv preprint arXiv:1711.11053,* 2017.
[77] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.
[78] Q. Zhang, J. Yan, H. O. Gao, and F. You, "A systematic review on power systems planning and operations management with grid integration of transportation electrification at scale," *Advances in Applied Energy,* p. 100147, 2023.
[79] M. Sang, Y. Ding, M. Bao, Y. Song, and P. Wang, "Enhancing resilience of integrated electricity-gas systems: A skeleton-network based strategy," *Advances in Applied Energy,* vol. 7, p. 100101, 2022.
[80] G. Von Wald, K. Sundar, E. Sherwin, A. Zlotnik, and A. Brandt, "Optimal gas-electric energy system decarbonization planning," *Advances in Applied Energy,* vol. 6, p. 100086, 2022.
[81] L. Frölke, T. Sousa, and P. Pinson, "A network-aware market mechanism for decentralized district heating systems," *Applied Energy,* vol. 306, p. 117956, 2022.
[82] M. S. Diéguez, A. Fattahi, J. Sijm, G. M. España, and A. Faaij, "Modelling of decarbonisation transition in national integrated energy system with hourly operational resolution," *Advances in Applied Energy,* vol. 3, p. 100043, 2021.
[83] W. Peng, H. Chen, J. Liu, X. Zhao, and G. Xu, "Techno-economic assessment of a conceptual waste-to-energy CHP system combining plasma gasification, SOFC, gas turbine and supercritical CO2 cycle," *Energy Conversion and Management,* vol. 245, p. 114622, 2021.

[84] J. Hu *et al.*, "Implications of a Paris-proof scenario for future supply of weather-dependent variable renewable energy in Europe," *Advances in Applied Energy,* vol. 10, p. 100134, 2023.

[85] W. He, M. King, X. Luo, M. Dooner, D. Li, and J. Wang, "Technologies and economics of electric energy storages in power systems: Review and perspective," *Advances in Applied Energy,* vol. 4, p. 100060, 2021.

[86] D. Xiao, Z. Lin, H. Chen, W. Hua, and J. Yan, "Windfall profit-aware stochastic scheduling strategy for industrial virtual power plant with integrated risk-seeking/averse preferences," *Applied Energy,* vol. 357, p. 122460, 2024.

[87] S. Yang, H. O. Gao, and F. You, "Building electrification and carbon emissions: Integrated energy management considering the dynamics of the electricity mix and pricing," *Advances in Applied Energy,* vol. 10, p. 100141, 2023.

[88] Q. Liang, F. Que, and E. Modiano, "Accelerated primal-dual policy optimization for safe reinforcement learning," *arXiv preprint arXiv:1802.06480,* 2018.

[89] G. Ruan, D. Qiu, S. Sivaranjani, A. S. Awad, and G. Strbac, "Data-driven energy management of virtual power plants: A review," *Advances in Applied Energy,* p. 100170, 2024.

[90] A. Ajoulabadi, S. N. Ravadanegh, and B. Mohammadi-Ivatloo, "Flexible scheduling of reconfigurable microgrid-based distribution networks considering demand response program," *Energy,* vol. 196, p. 117024, 2020.

[91] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems,* vol. 32, 2019.

[92] J. Lin, B. Xiao, H. Zhang, X. Yang, and P. Zhao, "A novel multitype-users welfare equilibrium based real-time pricing in smart grid," *Future Generation Computer Systems,* vol. 108, pp. 145-160, 2020.

[93] P. J. Reny, "On the existence of pure and mixed strategy Nash equilibria in discontinuous games," *Econometrica,* vol. 67, no. 5, pp. 1029-1056, 1999.

[94] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning,* vol. 8, pp. 279-292, 1992.

[95] F. Charbonnier, T. Morstyn, and M. D. McCulloch, "Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility," *Applied energy,* vol. 314, p. 118825, 2022.

[96] R. Dearden, N. Friedman, and S. Russell, "Bayesian Q-learning," *Aaai/iaai,* vol. 1998, pp. 761-768, 1998.

[97] F. Che, *Bayesian Q-learning from Imperfect Expert Demonstrations*. McGill University (Canada), 2021.

[98] S. Bahrami, Y. C. Chen, and V. W. Wong, "Deep reinforcement learning for demand response in distribution networks," *IEEE Transactions on Smart Grid,* vol. 12, no. 2, pp. 1496-1506, 2020.

[99] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Applied energy,* vol. 220, pp. 220-230, 2018.