RESILIENT POWER SYSTEM OPERATION WITH REINFORCEMENT LEARNING

XIANG WEI

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University

Department of Electrical and Electronic Engineering

Resilient Power System Operation with Reinforcement Learning

Xiang Wei

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)

XIANG WEI (Name of student)

Abstract

The ongoing evolution toward low-carbon and decentralized power systems, driven by high renewable penetration and widespread integration of inverter-based DERs, has raised significant concerns regarding the security, stability, and resilience of both transmission and distribution systems. These systems are now more frequently exposed to high-impact, low-probability events, such as extreme weather conditions, cyber-attacks, multi-component failures, and real-time operational uncertainties. Traditional model-based optimization approaches, such as security-constrained optimal power flow (SCOPF) and contingency-constrained OPF (CCOPF), while mathematically rigorous, often suffer from scalability limitations, long computation times, and a lack of adaptability to rapidly changing conditions. There is an urgent need for new intelligent decision-making methodologies capable of handling uncertainty, maintaining physical constraint feasibility, and enabling fast response in both centralized and decentralized operation frameworks.

This thesis addresses these critical challenges by proposing a deep reinforcement learning (DRL)-based framework for resilient and adaptive power system operation under uncertainty. The thesis spans three interconnected layers of power system control: transmission system scheduling under contingencies, real-time voltage regulation in active distribution networks, and coordinated transmission and distribution (T&D) system load restoration during emergency events. Each layer is investigated through a dedicated contribution, incorporating DRL techniques tailored to the respective operational requirements and system architectures.

The first contribution introduces a novel adversarial learning-based approach to solving the CCOPF problem under worst-case N-*k* contingency conditions. A defender–attacker soft actor-critic (DA-SAC) framework is proposed, in which two non-cooperative agents—representing the system operator and an adversarial uncertainty generator—interact within a reinforcement learning environment. The defender agent learns robust dispatch actions, while the attacker agent identifies the worst-case contingency scenarios in a discrete action

space. The proposed algorithm embeds constraint violation information directly into the reward function and employs dual-timescale policy updates to enhance convergence and learning stability. This approach shifts robust power system operation from static, model-based optimization to a dynamic, game-theoretic learning paradigm.

The second contribution extends the SCOPF model into a two-stage preventive–corrective control framework incorporating fast-response virtual power plants (VPPs). The model is formulated as a constrained Markov decision process (CMDP) and solved using a Lagrangian-based soft actor-critic (L-SAC) algorithm. Preventive and corrective agents are trained to minimize pre-contingency risk and post-contingency recovery costs while satisfying AC power flow constraints. The state-dependent Lagrange multiplier mechanism enables real-time enforcement of safety constraints without relying on static penalty parameters. The inclusion of VPPs in the operational framework enhances flexibility and responsiveness, allowing for dynamic adjustment to unexpected load and generation fluctuations.

The third contribution focuses on voltage regulation in active distribution networks (ADNs), where high penetration of inverter-based DERs results in frequent and unpredictable voltage violations. A hierarchical multi-mode voltage control strategy is proposed, featuring day-ahead dispatch of on-load tap changers (OLTCs) and capacitor banks via single-agent RL, and real-time inverter-based control using a multi-agent SAC (MASAC) algorithm with an embedded attention mechanism. The attention module enables each agent to prioritize relevant local observations, ensuring stable policy learning even in large-scale, multi-agent environments. Additionally, the voltage regulation problem is decomposed into three dynamic operational modes—power loss minimization, under-voltage mitigation, and over-voltage correction—allowing the system to flexibly respond to varying operational conditions.

The fourth contribution addresses the real-time coordination of load restoration across transmission and distribution systems under N-*k* emergency conditions. A distributed DRL architecture is proposed, comprising a centralized SAC controller for the transmission system and a complementary attention-enhanced MASAC controller for the distribution

system. A VPP is introduced as an aggregator to coordinate distributed DERs and reduce communication burdens. The hierarchical architecture enables asynchronous but coherent interaction between system layers, ensuring scalable and rapid recovery under contingency conditions. The integration of an attention mechanism improves inter-agent coordination and decision accuracy during system-wide restoration efforts.

Collectively, the four contributions of this thesis form a comprehensive and integrated framework for enhancing the resilience, adaptability, and operational efficiency of modern power systems under contingencies and uncertainties. By systematically addressing three critical aspects—transmission dispatch against worst-case contingencies, dynamic voltage regulation in active distribution networks, and real-time coordinated restoration across transmission and distribution systems—this work bridges the gap between traditional model-based optimization techniques and data-driven, learning-based control approaches. The proposed reinforcement learning strategies are specifically tailored to overcome key challenges such as computational delays, model inaccuracies, and coordination inefficiencies, which have historically limited the practical deployment of robust control frameworks in real-world systems. Furthermore, by incorporating multi-agent models, adversarial training mechanisms, and hierarchical decision-making structures, the thesis lays the foundation for autonomous, decentralized, and scalable control methodologies that can adapt to evolving system configurations and unforeseen operational scenarios. Extensive case studies on IEEE 30-bus, 118-bus, and modified distribution systems validate the effectiveness and generalizability of the methods, laying a strong foundation for the next generation of learning-augmented decision support systems in modern power networks. Ultimately, this thesis contributes toward realizing the vision of resilient, sustainable, and smart grids capable of ensuring security, stability, and flexibility under the transformative pressures of high renewable integration, decentralization, and digitalization.

Acknowledgments

Pursuing a Ph.D. has been one of the most challenging yet rewarding journeys of my life. I will always cherish this invaluable experience as a defining chapter in both my personal and professional development.

First and foremost, I would like to express my deepest gratitude to my supervisors, Prof. Siqi Bu and Dr. Ka Wing Chan, for their unwavering guidance, trust, and mentorship throughout my Ph.D. studies. I am equally thankful to Prof. Guibin Wang and Prof. Xian Zhang for their valuable academic advice, constructive feedback, and continuous support during my doctoral studies. I also wish to extend my heartfelt thanks to my co-supervisors at Cardiff University, Prof. Jianzhong Wu and Prof. Yue Zhou, for their collaborative guidance during my research attachment. I am especially grateful to Prof. Ting Wu, who served as my earliest mentor in my academic career and introduced me to the world of scientific research.

I would also like to acknowledge the many colleagues and friends who made this journey more meaningful and enjoyable. Special thanks to Dr. Ziqing Zhu, Dr. Zezhou Lin, Dr. Wei Qiu, Dr. Sunhuang Huang, Dr. Rabee Ahmed, Mr. Ze Hu, Mr. Yafeng Zhao, Mr. Xinxiang Zhong, and Mr. Changlei Gu, whose companionship, support, and technical discussions have been invaluable. I am also grateful to many other unnamed friends and colleagues who offered their encouragement and assistance over the years. I wish them all continued success and happiness in their future endeavors.

Finally, I owe my deepest appreciation to my family. I am forever thankful to my parents, Mr. Jianlin Wei and Ms. Wenfeng Li, for their unconditional love, understanding, and constant encouragement. Their belief in me has been my greatest source of strength. I am also profoundly grateful to my wife, Ms. Mengyin Xu, for her endless patience, support, and love throughout this journey. Her presence and unwavering faith in me have been the driving force behind my pursuit of a Ph.D.

Publications Arising from This Thesis

Published papers:

[1] **Xiang Wei**, Ka Wing Chan, Guibin Wang, Ze Hu, Ziqing Zhu, Xian Zhang, Robust preventive and corrective security-constrained OPF for worst contingencies with the adoption of VPP: A safe reinforcement learning approach, Applied Energy, Volume 380, 124970, 15 Feb 2025, DOI: 10.1016/j.apenergy.2024.124970.

[2] **Xiang Wei**, Xian Zhang, Guibin Wang, Ze Hu, Ziqing Zhu and Ka Wing Chan, "Online Voltage Control Strategy: Multi-Mode Based Data-Driven Approach for Active Distribution Networks," in IEEE Transactions on Industry Applications, Vol.61, No.1, p.1569-1580, Jan/Feb 2025, doi: 10.1109/TIA.2024.3462891.

[3] **Xiang Wei**, Ka Wing Chan, Ting Wu, Guibin Wang, Xian Zhang, Junwei Liu, Wasserstein distance-based expansion planning for integrated energy system considering hydrogen fuel cell vehicles, Energy, Volume 272, 2023, 127011, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2023.127011.

[4] Ze Hu, Ziqing Zhu, **Xiang Wei**, Ka Wing Chan, Siqi Bu, Mixed strategy Nash equilibrium analysis in real-time pricing and demand response for future smart retail market, Applied Energy, Volume 391, 2025, 125815.

[5] Ze Hu, Ka Wing Chan, Ziqing Zhu, **Xiang Wei**, Weiye Zheng, Siqi Bu, Techno–Economic Modeling and Safe Operational Optimization of Multi-Network Constrained Integrated Community Energy Systems, Advances in Applied Energy, Volume 15, 2024, 100183, https://doi.org/10.1016/j.adapen.2024.100183.

[6] Wei Qiu, He Yin, Yuqing Dong, **Xiang Wei**, Yilu Liu and Wenxuan Yao, "Synchro-Waveform-Based Event Identification using Multi-task Time-frequency Transform Networks," in IEEE Transactions on Smart Grid, Vol.16, No.3, p.2647-2658, May 2025, DOI: 10.1109/TSG.2025.3546568

[7] Ze Hu, Peijun Zheng, Ka Wing Chan, Siqi Bu, Ziqing Zhu, **Xiang Wei**, "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, Vol.13, No.3, May 2025, DOI: 10.35833/MPCE.2024.000909

Working papers:

[8] **Xiang Wei**, Ka Wing Chan, Ahmed Sayed, Xian Zhang, Guibin Wang, Real-time Resilient Power System Operation with Defender-Attacker Soft Actor-Critic Reinforcement Learning. (Submitted to the "IEEE Transactions on Industrial Informatics")

[9] **Xiang Wei**, Wei Qiu, Ka Wing Chan, Wenxuan Yao, and C.Y. Chung. Distribution System Voltage Regulation: An Attack-Defense Strategy via Visual based Reinforcement Learning. (Submitted to the "IEEE Transactions on Smart Grid")

[10] **Xiang Wei**, Ziqing Zhu, Linghua Zhu, Ze Hu, Xian Zhang, Guibin Wang, Siqi Bu, Ka Wing Chan. Foresight-Seeing Quantum Reinforcement Learning for Two-Stage Unit Commitment with Virtual Power Plants and Renewable Power Integration. (Submitted to the "Journal of Modern Power Systems and Clean Energy")

[11] **Xiang Wei**, Yue Zhou, Ze Hu, Ziqing Zhu, Guibin Wang, Xian Zhang, Ka Wing Chan, Jianzhong Wu. Coordinated Transmission-Distribution Load Restoration under N-*k* Contingencies: A Distributed Optimization and Reinforcement Learning Approach. (Submitted to the "Applied Energy")

[12] **Xiang Wei**, Ziqing Zhu, Ze Hu, Guibin Wang, Xian Zhang, Ka Wing Chan. Enhancing Resilience in Networked Microgrids: A Multi-Agent Deep Reinforcement Learning Approach for Post-Attack Load Restoration. (Submitted to the " IEEE Transactions on Industry Applications ")

[13] **Xiang Wei**, Wei Qiu, Ze Hu, Wenxuan Yao, Yuqing Dong, Ka Wing Chan, and C.Y. Chung. Event-triggered Real Time Voltage Regulation using Wavelet Spectrum and Reinforcement Learning. (Submitted to the " IEEE Transactions on Industry Applications")

Table of contents

List of Figures

List of Tables

List of Acronyms

SCOPF	Security-constrained optimal power flow
CCOPF	Contingency-constrained optimal power flow
OPF	Optimal power flow
DER	Distributed energy resources
DRL	Deep reinforcement learning
T&D	Transmission and distribution
TSO	Transmission system operator
DSO	Distribution system operator
SAC	Soft actor-critic
DA-SAC	Defender–attacker soft actor-critic
VPP	Virtual power plants
MDP	Markov decision process
CMDP	Constrained Markov decision process
L-SAC	Lagrangian-based soft actor-critic
ADN	Active distribution networks
MASAC	Multi-agent soft actor-critic
EV	Electric vehicle
OLTC	On-Load Tap Changer
DA	Defender agent
AA	Attacker agent
DCV	Degree of constraint violation
PCSCOPF	Preventive-corrective security-constrained OPF
CA	Corrective agent
PA	Preventive agent
MADRL	Multi-agent DRL
CMS	Complementary attention for multi-agent SAC
CB	Capacitor bank
C&CG	Column-and-constraint generation

PV	Photovoltaic
BESS	Battery energy storage systems
CVR	Conservation voltage reduction
VVC	Voltage-var control
DDPG	Deep deterministic policy gradient
MADDPG	Multi-agent deep deterministic policy gradient
KKT	Karush–Kuhn–Tucker
DGs	Distributed generators
DSAC	Discrete action-based SAC
DQN	Deep Q network
PPO	Proximal Policy Optimization
A3C	Asynchronous Advantage Actor Critic
UE	Unserved electricity
LS	Load shedding
FST	Short-term emergency rating
FLT	Long-term emergency limits
DRPs	Demand response programs
TOU	Time of use
RTP	Real-time pricing
CPP	Critical peak pricing
PVR	Peak-to-valley ratio
SOC	State of charge
TA	Transmission system agent
DA	Distribution system agent
SDU	State dividing unit
AIU	Attention improvement unit
ACU	Attention complement unit

Chapter 1 Introduction

1.1 Background

The secure and stable operation of power systems under uncertain and unexpected conditions is one of the most critical and fundamental requirements in modern power systems [1]. With the increasing complexity and interconnection of power systems, ensuring that both transmission and distribution systems can operate reliably in the presence of external disturbances has become essential [2]. Power systems are vulnerable to various uncertainties and contingencies, including extreme weather events, abrupt equipment failures, cyber-attacks, and human-induced operational errors [3], [4]. These events can significantly disturb the normal balance between electricity supply and demand, thereby threatening the integrity of the entire power infrastructure. Particularly, high-impact and low-probability disruptions can trigger widespread outages, posing threats to social stability, economic activity, and the stability of essential services [5]. Therefore, maintaining operational robustness, rapid fault recovery, and adaptive response capacity across the transmission and distribution layers is crucial to ensuring the reliability of the power supply. Moreover, with the transition towards low-carbon and renewable energy systems [6], the ability of these networks to withstand, absorb, and recover from unexpected events has become not only a reliability issue but also a core component of modern power system design. The coordinated secure operation of both transmission and distribution systems under uncertainty must therefore be prioritized in power system research, planning, and real-time control.

As the primary infrastructure of the power system, the transmission network is responsible for transporting electricity generated from centralized generation units over long distances to major consumption regions and distribution networks [7]. Its role in ensuring the large-scale balance of power and maintaining grid-wide voltage and frequency stability is indispensable. However, the secure operation of transmission systems has become increasingly vulnerable to high-impact, low-probability events [3], [4]. These include multiple simultaneous equipment failures, line overloads, substation faults, and natural disasters such as wildfires, floods, or hurricanes. Such incidents are often modeled as N-*k* contingencies, reflecting the potential disconnection of multiple transmission elements and leading to severe congestion, voltage fluctuation, or even cascading blackouts. Due to the highly interconnected structure of modern transmission systems, a disruption in the network can result in a widespread outage that impacts large areas of the power system. Traditional protection and control strategies, although effective in managing localized issues, often lack the flexibility and speed required to handle large-scale system-wide disturbances [8]. Furthermore, the increasing complexity of market mechanisms and renewable power injections into transmission networks exacerbates these challenges. As a result, enhancing the resilience of transmission systems through preventive-corrective scheduling, fast response reserves, and intelligent decision-making has become a top research priority for system operators and policymakers worldwide.

In parallel with transmission networks, distribution systems play a vital role in ensuring the final delivery of electricity to end-users, ranging from residential loads to industrial customers [9]. Traditionally, distribution systems functioned passively, relying on predictable single-direction power flows from the transmission level to local loads. However, this operational model has shifted dramatically with the increasing penetration of distributed energy resources (DERs), especially renewable sources such as rooftop photovoltaics and community wind turbines. These resources introduce significant variability and stochasticity into the distribution system, altering voltage profiles and disturbing the load-generation balance at the local level [10]. Moreover, many of these DERs are inverter-based, meaning they lack the inertia that conventional synchronous generators provide, making the system more sensitive to rapid transients [11]. The intermittent characteristics of these resources, along with load uncertainty and electric vehicle (EV) charging behavior, make real-time operation and planning of distribution networks more complex than ever before. Conventional control devices such as on-load tap changers (OLTCs) and capacitor banks (CBs) operate on slower time scales and cannot respond effectively to rapid fluctuations [12]. Consequently, voltage violations, reverse power flows, and protection coordination issues have become more prevalent. These challenges highlight the urgent need for advanced voltage regulation techniques, real-time DER coordination, and predictive management strategies to ensure secure and reliable distribution system operation under increasing uncertainty. Given the parallel challenges facing both transmission and distribution systems, it becomes imperative to consider how these two subsystems interact — particularly under conditions where disturbances in one can propagate to the other. This brings us to the critical need for coordinated operation between transmission and distribution layers.

The need for coordinated operation between transmission and distribution systems has attracted increasing attention due to the growing complexity and interdependence of modern power system operations [13]. Historically, transmission and distribution networks were operated independently, with insufficient real-time interaction [14]. However, the growing complexity of the power system, characterized by the bidirectional flow of power and information, requires a fundamental shift in this operational model. Transmission-level decisions, such as generator redispatch or emergency load shedding, have immediate impacts on downstream distribution systems, potentially causing voltage instability or unexpected disconnection of critical DERs and loads. At the same time, distribution networks are increasingly equipped with controllable DERs and flexible demand-side resources that can provide support to the bulk power system during contingencies if properly coordinated. The traditional top-down control approach is no longer sufficient for ensuring power system stability in such an environment [15]. The absence of synchronized data exchange, integrated modeling frameworks, and real-time control interfaces between transmission system operators (TSOs) and distribution system operators (DSOs) creates significant observability limitations, especially during high-impact events [16]. Achieving effective transmission and distribution (T&D) system coordination requires advanced communication infrastructure, shared situational awareness, and jointly optimized control actions. This coordination is not only essential for improving system resilience but also for enabling new services such as local energy markets, distributed ancillary services, and enhanced system restoration capabilities after major disturbances.

The modernization and restructuring of power systems—under the motivation of decarbonization policies, digital innovation, and widespread electrification—has resulted in growing operational complexity and uncertainty [17]. Unlike traditional power system structures where power flowed predictably from large generation plants to passive loads, the modern power system is a dynamic and interactive system where generation, consumption, and storage are distributed across all voltage levels [18]. This transformation challenges the existing planning and operational strategies, particularly when considering the need to ensure security and reliability under dynamic scenarios. One of the most pressing issues is the dual exposure to stochastic renewable generation on both the transmission and distribution sides, coupled with unexpected contingencies such as multiple devices outages, cyberattack, or natural hazards [19]. These events can induce cascading failures across the entire power system without effective preventive and corrective control mechanisms. Previous research has made considerable progress in tackling transmission system security through contingency-constrained optimal power flow (CCOPF) and security-constrained economic dispatch. Similarly, in the distribution level, work on local voltage control, DER capacity regulations, and microgrid resilience has advanced significantly [20]. However, these efforts are often developed independently and fail to fully capture the range of interactions between the transmission and distribution network levels. In many real-world incidents, it has become apparent that local issues in distribution systems—such as reverse flows, islanding, or sudden DER tripping—can exacerbate transmission-level stresses, and vice versa [21]. This interdependence necessitates the development of joint resilience strategies that simultaneously consider operational flexibility, inter-layer uncertainty interactions, and shared resource utilization. Furthermore, time-scale challenges complicate coordination. While transmission-level decisions are typically made at slower intervals (e.g., 5–15 minutes) [22], distribution systems may require fast responsiveness, especially when integrating fast-reacting DERs and loads [23]. Thus, there is an increasing need for unified frameworks that bridge these temporal and spatial gaps, enabling the joint optimization of T&D systems under uncertainty. Such frameworks must integrate robust preventive planning, real-time corrective actions, and adaptive learning from historical operational data.

Only by treating the transmission and distribution layers as a single, dynamic, and interactive system can power system operators effectively respond to the multifaceted challenges posed by the development of modern power systems.

While emerging technologies such as distributed renewable energy, smart sensors, electric vehicles, and demand response systems offer tremendous opportunities for power system modernization, they also result in numerous new operational challenges [24]. One of the most prominent issues is the spatially dispersed and highly dynamic characteristics of distributed resources, which lack centralized control and exhibit insufficient communication. These resources can fluctuate rapidly, making it difficult to forecast aggregate behavior or respond uniformly during disturbances. In the absence of a centralized control center, the lack of coordination among decentralized units significantly hinders the power system's ability to act swiftly in the face of unexpected events. Moreover, the supporting communication infrastructure is subject to latency, unstable data links, and a lack of standardized protocols across distributed components, all of which complicate timely and reliable system coordination [25]. In addition, many DERs participate in power system operations using inverter-based interfaces that, while fast, are highly sensitive to control errors and disturbances. Without well-coordinated control strategies, these systems can worsen power system instabilities. These technical barriers underscore the need for novel architectures that incorporate real-time communication, hierarchical control, and distributed intelligence. Designing such architectures is crucial for enabling reliable decision-making and fast response across both transmission and distribution systems when confronted with increasing operational uncertainty and complexity.

In response to the increasingly dynamic and uncertain environment in power systems operation, artificial intelligence—particularly deep reinforcement learning (DRL)—has emerged as a promising tool for achieving intelligent decision-making [26]. DRL is capable of addressing the complexity of power system operations due to its ability to learn optimal control strategies from high-dimensional and stochastic environments. Unlike traditional optimization approaches that rely on model-based formulations and are computationally intensive in real-time [27], DRL agents can be trained offline using historical and simulated data, and then deployed online for fast and adaptive decision-making. This paradigm is especially valuable when coordinating transmission and distribution resources under uncertainty, where the need for speed and accuracy is critical. By learning from interaction with the system environment, DRL algorithms can handle nonlinear, time-dependent constraints and unknown disturbances. They can make decisions that are adaptive to real-time system conditions and forward-looking, enabling more robust and flexible operational strategies [28]. In the context of contingency response, load balancing, or voltage regulation, DRL-based controllers can outperform conventional control methods or heuristic methods [29]. Therefore, integrating DRL into the operational framework of power systems has the potential to enhance resilience, reduce response times, and improve overall system efficiency, particularly in scenarios where traditional methods are infeasible or too slow.

Alongside intelligent decision-making techniques, the concept of the virtual power plant (VPP) offers a scalable and practical solution to the challenge of coordinating highly distributed energy resources [30]. A VPP aggregates diverse DERs—including PV, wind, battery storage, controllable loads, and electric vehicles—into a single, flexible, dispatchable entity [31]. Through the use of cloud-based control platforms and advanced communication technologies, VPPs enable real-time coordination and optimization of decentralized resources, effectively functioning as a centralized management center for distributed systems [32]. This capability is crucial for supporting both transmission and distribution systems during high-impact or uncertain conditions. VPPs enhance operational visibility, enable aggregated participation in electricity markets, and support ancillary services such as frequency regulation, ramping, and voltage support. Most importantly, during contingency scenarios, a well-designed VPP can provide rapid load/generation rebalancing and contribute to the restoration of affected areas. By establishing bi-directional communication channels with both TSOs and DSOs, VPPs serve as a bridge between system-level coordination and local flexibility [33], [34]. This approach not only improves overall power system resilience but also facilitates the integration of renewable energy and demand-side resources in a controllable and efficient manner. Therefore, VPPs are expected to play an increasingly central role in the realization of a smart, flexible, and low-carbon power system.

1.2 Research Objectives

This thesis aims to enhance the secure and stable operation of power systems under high-impact contingency events and various uncertainties, which are increasingly frequent due to the integration of renewable energy and extreme external disturbances. As conventional optimization and control methods often fall short in providing real-time, scalable, and adaptive responses to such complex conditions, this work proposes intelligent and robust control frameworks to fill these gaps. Specifically, the thesis addresses these challenges from three key perspectives: the transmission system, the distribution system, and the coordinated operation of transmission and distribution (T&D) systems. For the transmission system, a DRL-based approach is developed to enhance the robustness of CCOPF solutions under the worst-case N-*k* contingencies. In the distribution system, a multi-mode DRL strategy is introduced to manage fast voltage violations in the presence of DER uncertainties. Lastly, the thesis presents a reinforcement learning-enhanced T&D coordination scheme to facilitate intelligent, system-wide response to cascading failures. These contributions aim to improve operational resilience, situational awareness, and real-time decision-making across the entire power network. Each of these aspects is explored in detail in the following sections.

- *Robust transmission system operation under N-k contingencies*: Power system resilience and optimal decision-making under contingency scenarios have become central to ensuring secure operation. Among existing approaches, two-stage decision-making frameworks such as CCOPF are widely adopted, though they present significant computational and modeling challenges due to their large-scale, nonconvex, and discrete decision characteristics. To address this, this thesis proposes a novel DRL-based robust optimization framework, specifically tailored for CCOPF problems under N-*k* security criteria. The method leverages a multi-agent learning architecture that enables the system operator to identify the worst-case contingency scenarios, thereby enhancing the robustness of the resulting operational strategy. This DRL-enhanced CCOPF model improves computational tractability and adaptive response, making it a promising tool for real-time contingency analysis in large-scale transmission networks. The proposed method fills an important research gap by integrating AI-based decision-making with traditional CCOPF, contributing to both the theory and practice of resilient transmission system operation.

- *Multi-mode real-time voltage regulation in active distribution networks*: In distribution systems, real-time operation is increasingly affected by uncertainties such as fluctuating renewable generation and stochastic load demand behaviors. Conventional devices such as OLTCs and capacitor banks, operating at slower timescales, are insufficient for mitigating fast voltage violations. Moreover, single-mode voltage control strategies often fail to satisfy the complex economic and security constraints associated with voltage and reactive power margins in active distribution networks (ADNs). To address these challenges, this thesis develops a two-stage DRL-based multi-modal voltage regulation strategy. The proposed framework combines fast inverter-based reactive power control with traditional device coordination, allowing for real-time adaptation to system uncertainties. The objective is to minimize total power losses while maintaining voltage profiles within secure limits. By introducing adaptive multi-mode control, the strategy enhances voltage stability and distribution-level operational efficiency. This study contributes significantly to the field by offering a scalable, intelligent, and real-time voltage regulation framework tailored to the dynamics of modern ADNs.
- *Coordinated T&D load restoration under N-k contingencies*: Ensuring rapid and coordinated load recovery in T&D systems under emergency conditions is critical for maintaining overall power system stability. With the growing presence of active distribution systems and DERs, there is increasing potential to utilize flexible resources in the distribution layer to support transmission-level operations. In particular, during transmission system contingencies, distribution networks can assist in relieving congestion and mitigating voltage support deficiencies. To achieve this, an effective T&D coordination strategy must facilitate bidirectional information exchange and joint

decision-making under uncertainty. This thesis introduces a reinforcement learning-based control strategy for optimizing load restoration during N-*k* contingency events, leveraging the flexibility of distribution systems. The proposed framework improves global system resilience by allowing T&D subsystems to respond jointly and intelligently to critical disruptions. This contribution is particularly relevant in light of the ongoing transition toward decentralized and distributed grid architectures, providing a pathway for integrated emergency response strategies in coupled T&D environments.

1.3 Contributions of the Thesis

This thesis presents four original contributions aimed at improving the resilient operation of power systems under high-impact contingencies and uncertainties. These contributions address key limitations in traditional optimization and control methods by integrating advanced reinforcement learning algorithms, distributed optimization models, and hybrid control frameworks. Specifically, the proposed solutions tackle challenges in robust transmission system operation, real-time voltage control in active distribution networks, and coordinated transmission and distribution system restoration under emergency scenarios. The technical novelty of this thesis lies in the customized design of DRL algorithms, multi-agent architectures, and the integration of physical constraints into decision-making frameworks. Each research effort targets a specific gap in existing literature and collectively contributes to enhancing the robustness, scalability, and intelligence of modern power system operation. The main contributions are summarized below.

1) Robust real-time transmission operation via defender-attacker reinforcement learning

- *Research gaps*: Traditional approaches to solving two-stage robust optimization problems in power systems, such as CCOPF under contingency constraints, are hindered by their high computational complexity and difficulty in handling real-time uncertainty. Moreover, most DRL applications lack a formal structure to model adversarial uncertainty scenarios dynamically, limiting their effectiveness in ensuring worst-case performance. There is also a lack of multi-agent formulations that explicitly model adversarial interactions between decision-makers and uncertainty realizations, especially under nonconvex constraints such as AC power flow. Thus, there is a pressing need for a DRL framework that supports real-time, robust, and scalable contingency management in power system operations.

- *Contributions*: This study proposes a novel DRL-based method, defender-attacker soft actor-critic (DA-SAC), to solve robust two-stage optimization problems for real-time power system operations under uncertainty. The formulation introduces a Markov decision process (MDP) incorporating two non-cooperative agents: a defender agent (DA) that generates robust control actions, and an attacker agent (AA) that identifies the worst contingency scenarios. A model-free entropy-regularized soft actor-critic (SAC) algorithm is used for the DA in a continuous action space, while a discrete SAC algorithm is designed for the AA. To stabilize learning, the most recent DA action is used as the input state for the AA, and a dual-timescale learning rate mechanism is introduced. Moreover, the degree of constraint violation (DCV) is integrated into the reward function to enhance the feasibility of the final CCOPF solutions. This adversarial DRL framework enables efficient, online learning of robust operational strategies, significantly improving grid reliability during worst-case events.

2) Safe preventive-corrective SCOPF with VPPs under deep reinforcement learning

- *Research gaps*: Preventive-corrective security-constrained optimal power flow (PCSCOPF) models have been widely used to manage N-k contingencies. However, most traditional formulations are limited by their reliance on deterministic scenarios and their inability to integrate the fast-response capabilities of VPPs. Moreover, the inclusion of AC power flow constraints, time-dependent dynamics, and cumulative operational costs increases the complexity and scalability issues of such models. While DRL has emerged as a promising solution for complex control tasks, its application to two-stage PCSCOPF problems under uncertainty, particularly with virtual resources, has been limited. Additionally, existing SAC-based approaches do not explicitly enforce constraint satisfaction, leading to sub-optimal and potentially infeasible solutions during real-time operation.

- *Contributions*: This study presents a robust DRL-based two-stage PCSCOPF framework that integrates fast-response VPPs into AC power systems under N-*k* contingencies. The proposed approach formulates the problem as a constrained Markov decision process (CMDP), where two agents, a preventive agent (PA) and a corrective agent (CA), are designed to minimize unmet demand, constraint violations, and adjustment costs. To solve this CMDP efficiently, a Lagrangian-based SAC (L-SAC) algorithm is developed. The algorithm dynamically tunes state-dependent Lagrange multipliers, ensuring both optimality and constraint satisfaction. This structure captures the full complexity of AC power flows while maintaining computational efficiency through agent decomposition. The proposed framework outperforms existing methods in scalability, constraint handling, and convergence speed, offering a safe and effective control strategy for real-time preventive-corrective dispatch with VPPs under high-impact contingencies.

3) Multi-mode voltage regulation in active distribution networks using MADRL

- *Research gaps*: The increasing penetration of rooftop PV and inverter-based DERs introduces rapid voltage fluctuations in ADNs. Traditional devices such as OLTCs and capacitor banks operate on slower timescales and are unable to provide adequate voltage support in real time. Existing voltage control strategies often rely on single-mode regulation, which fails to account for varying grid conditions such as under-voltage and over-voltage scenarios. Additionally, most control strategies are either centralized, incurring high communication burdens, or fully decentralized, lacking coordination. There is a lack of scalable control architectures that can balance global coordination and local responsiveness under high-dimensional uncertainties in distribution networks.
- *Contributions*: This work introduces a two-stage, multi-mode voltage regulation strategy that coordinates slow-response traditional devices and fast-response PV inverters to optimize voltage control across timescales. A single-agent DRL algorithm performs day-ahead control of OLTCs and CBs, while a MADRL algorithm is employed for real-time local voltage control by distributed PV inverters. Each inverter acts as an agent in a decentralized framework trained using centralized training with decentralized execution. An attention mechanism is integrated to allow each agent to focus on reward-relevant

information, improving learning efficiency and robustness under communication constraints. The strategy supports three operating modes, power loss minimization, under-voltage mitigation, and over-voltage mitigation, enabling dynamic adaptation to network conditions. The framework reduces energy consumption, enhances voltage stability, and minimizes communication overhead, providing a scalable and adaptive solution for voltage regulation in ADNs.

4) Distributed load restoration for T&D systems under *N*-*k* contingency

- *Research gaps*: Traditional load restoration strategies under N-*k* emergencies often treat transmission and distribution systems separately, leading to sub-optimal recovery actions. Moreover, centralized restoration approaches face scalability issues and heavy communication burdens, particularly in distribution networks with high DER penetration. The complexity of coordinating DERs during emergencies presents a major challenge. While VPPs have been proposed as aggregators, their role in coordinated restoration strategies under uncertainty has not been fully explored. Additionally, there is a need for learning-based optimization techniques that can manage large-scale dynamic problems across system layers with real-time performance.
- *Contributions*: This study proposes a distributed optimization and multi-agent DRL framework for coordinated load restoration in T&D systems under N-*k* contingencies. The transmission and distribution layers are modeled as two coupled MDPs, addressed by a SAC and a MASAC algorithm, respectively. A VPP serves as an aggregator in the distribution system, mediating between the DSO and DERs to reduce communication burdens and facilitate coordinated recovery. To enhance system-wide cooperation, a complementary attention mechanism is introduced in the MASAC framework, improving the ability of agents to prioritize relevant information and align decisions with shared objectives. This complementary attention for MASAC (CMS) structure enables scalable, communication-efficient, and effective restoration of loads across T&D boundaries. The proposed approach demonstrates superior performance in terms of convergence speed, adaptability, and restoration coverage under extreme event conditions, making it highly applicable to future resilient grid architectures.

1.4 Organization of the Thesis

This thesis is organized into seven chapters, as illustrated in Fig. 1.1. The overall structure reflects the layered approach of this research, which systematically addresses the secure and resilient operation of power systems under uncertainties and contingencies from three interconnected perspectives: transmission system operation, distribution system control, and coordinated T&D system restoration. Each core contribution is aligned with one of these layers and builds upon a reinforcement learning–based algorithmic framework tailored to the operational characteristics of each subsystem.

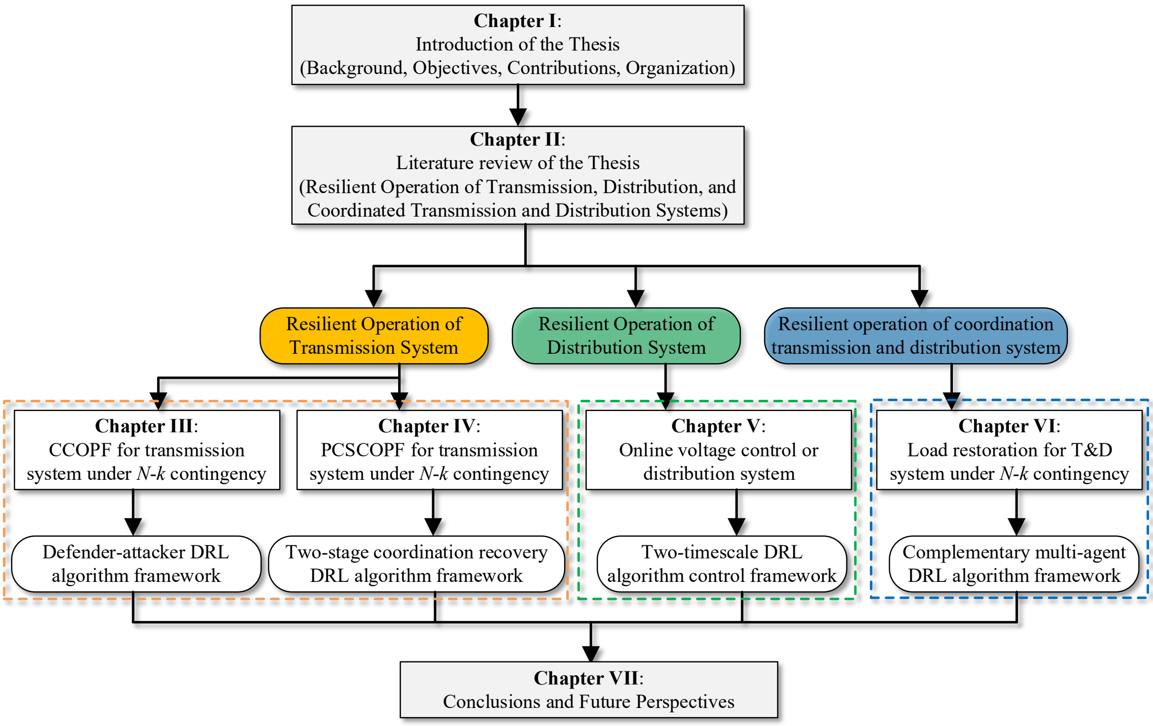

Chapter I:
Introduction of the Thesis
(Background, Objectives, Contributions, Organization)
Chapter II:
Literature review of the Thesis
(Resilient Operation of Transmission, Distribution, and Coordinated Transmission and Distribution Systems)
Resilient Operation of Transmission System
Resilient Operation of Distribution System
Resilient operation of coordination transmission and distribution system
Chapter III:
CCOPF for transmission system under N-k contingency
Chapter IV:
PCSCOPF for transmission system under N-k contingency
Chapter V:
Online voltage control or distribution system
Chapter VI:
Load restoration for T&D system under N-k contingency
Defender-attacker DRL algorithm framework
Two-stage coordination recovery DRL algorithm framework
Two-timescale DRL algorithm control framework
Complementary multi-agent DRL algorithm framework
Chapter VII:
Conclusions and Future Perspectives

Fig. 1.1 Overall Organization of the Thesis and the Structure of Resilient Power System Operation Framework.

- Chapter I introduces the thesis background, research motivations, objectives, main contributions, and the overall organization of the thesis. It outlines the increasing challenges posed by uncertainties and N-*k* contingencies in modern power systems and highlights the need for intelligent, robust, and scalable operational strategies.

- Chapter II presents a comprehensive literature review, categorizing existing works into three domains: resilient operation of transmission systems, distribution networks, and coordinated T&D systems. It identifies key research gaps in traditional optimization-based methods and establishes the rationale for adopting DRL methodologies.
- Chapter III focuses on the resilient operation of transmission systems by addressing the CCOPF (contingency-constrained optimal power flow) problem under N-*k* contingencies. A novel defender–attacker DRL algorithm is proposed, in which a two-agent adversarial learning structure is developed to identify worst-case contingencies and derive robust control strategies.
- Chapter IV continues with the transmission layer and proposes a two-stage PCSCOPF model. By leveraging fast-response VPPs and a Lagrangian-based DRL formulation, the chapter introduces a two-stage coordination recovery DRL framework that enables safe and adaptive control under complex operational constraints.
- Chapter V transitions to the distribution system, addressing the real-time voltage regulation problem under renewable energy uncertainty. A two-timescale DRL control architecture is developed, where traditional devices (OLTCs, CBs) and inverter-based DERs are coordinated using single-agent and multi-agent learning mechanisms. A multi-mode voltage control strategy is proposed to balance power loss minimization and voltage constraint satisfaction.
- Chapter VI explores the coordinated operation of transmission and distribution systems for load restoration under N-*k* contingencies. A distributed, complementary multi-agent DRL (MASAC) algorithm is designed to support real-time decision-making in both TSO and DSO domains. A VPP is employed as an aggregator to reduce communication burdens and enhance coordination between distributed agents.
- Chapter VII concludes the thesis with a summary of key findings, practical implications, and outlines several promising directions for future work. These include extending adversarial learning to broader uncertainty models, integrating formal safe-

RL techniques, developing plug-and-play multi-agent control architectures, and validating learning-based restoration strategies in real-time test environments.

Chapter 2 Literature Review

2.1 Literature Review on Resilient Operation of Power Systems under Contingencies and Uncertainties

2.1.1 Review of Robust SCOPF for Transmission Network Resilience

The secure operation of transmission systems under uncertain and high-impact contingency events remains a fundamental challenge in modern power system operation. To maintain grid stability and avoid cascading failures, SCOPF models have been widely adopted. Among them, the CCOPF problem plays a central role, especially under N-*k* security criteria where multiple simultaneous component failures must be considered. The primary objective of SCOPF is to determine generation dispatch schedules that satisfy all operational constraints in both pre-contingency and post-contingency states.

Early efforts on SCOPF modeling focused on deterministic, single-level formulations under N-1 security assumptions. These models were commonly solved using interior point methods [35], Newton methods [36], projected sub-gradient algorithms [36], sequential linear programming [4], and conic programming [4]. Although effective for relatively small systems, these methods suffer from scalability issues and long convergence times, making them unsuitable for real-time implementation in large-scale networks [37], [38]. To accelerate solution times, simplified models such as DC-OPF approximations [39], sparse tableau formulations[40], and compensation-based approaches [39] have been introduced. However, while these approximations provide computational benefits, they often result in solutions that are not AC-feasible and may be suboptimal in terms of system security and constraint satisfaction.

As power system operation evolves toward resilience-oriented planning, robust optimization techniques have gained significant attention in CCOPF studies. These approaches consider the worst-case realization of uncertainties within a predefined uncertainty set [41], thus enabling system operators to obtain dispatch decisions that remain feasible even under extreme conditions. Typically, robust CCOPF models are formulated as two-stage bi-level [40] or tri-level [42] optimization problems. Decomposition methods such as Benders decomposition [43] and the column-and-constraint generation (C&CG) algorithm [44] are often employed to solve these models. For example, in [43], a robust nonconvex AC OPF problem is dualized and solved via primal Benders decomposition, with feasibility and optimality cuts iteratively refined. However, the presence of AC power flow constraints and numerous mixed-integer variables introduces substantial computational overhead and convergence challenges [45].

More recently, the SCOPF framework has been expanded to incorporate fast-response resources, such as VPPs and battery energy storage systems (BESS), to improve operational flexibility. The inclusion of BESS in CSCOPF models has demonstrated enhanced post-contingency corrective capabilities [46], [47], although performance depends heavily on state-of-charge limitations. To overcome this constraint, [48] proposes integrating controllable loads and DERs within VPPs to support corrective dispatch. Further extensions include coordinated control schemes for VPPs [49], showing promise in improving power system robustness under stochastic contingencies. Nonetheless, existing works often focus solely on post-contingency corrective actions and neglect pre-contingency preventive strategies or the probabilistic nature of the contingencies [46].

In parallel, researchers have explored DRL as an alternative to traditional optimization techniques, particularly for real-time OPF solutions. Unlike supervised learning methods, which require large-scale labeled datasets [50], DRL methods learn directly through interaction with the environment, thereby enabling adaptive control without explicit system modeling [51]. Actor-critic structures and policy gradient algorithms such as DDPG and proximal policy optimization (PPO) have shown strong potential in deriving near-optimal policies in dynamic and uncertain environments [52]. However, most conventional DRL methods ignore hard physical constraints (e.g., voltage and thermal limits), and use reward penalties instead, which leads to difficulties in guaranteeing safety and feasibility during real-time deployment [53], [54]. Moreover, as the number and scope of system constraints grow, tuning appropriate penalty parameters becomes increasingly complex [55], and convergence to safe solutions is not always ensured [56]. To address these issues, recent studies have explored CMDPs, projection-based techniques [57], and robust DRL formulations that incorporate physical constraints more directly. However, the majority of DRL-based methods are still in early stages of application and require further development in terms of scalability, training stability, and integration with existing OPF solvers.

In conclusion, existing solution methodologies for SCOPF and CCOPF can be classified into two main categories: (i) model-based optimization techniques [35], [36],[4], which provide theoretical guarantees but face difficulties in handling uncertainty, dimensionality, and real-time constraints, and (ii) model-free learning-based approaches [58], [59], which offer adaptability and computational speed but often suffer from feasibility and interpretability issues. While supervised learning approaches require extensive datasets and retraining for system changes [60], reinforcement learning methods provide a promising path toward adaptive and robust decision-making. Nevertheless, many existing DRL-based CCOPF applications focus only on N-1 criteria [35], or ignore system security constraints entirely [58],[59], limiting their ability to ensure resilience in the face of worst-case contingencies. Thus, there remains a critical need to develop DRL-based frameworks that are explicitly tailored for robust SCOPF solutions under N-*k* security standards, while maintaining feasibility, adaptability, and computational efficiency in large-scale transmission systems.

2.1.2 Review of Voltage Control Strategies in Distribution Networks under Uncertainty

With the increasing penetration of DERs, particularly inverter-based PV systems, distribution networks are facing unprecedented challenges in maintaining voltage stability. The intermittent and stochastic nature of DER output introduces significant fluctuations in local voltage profiles, often leading to both under-voltage and over-voltage violations. Additionally, the growth of EV charging and flexible loads further increases uncertainty in distribution system operation. Traditional voltage regulation mechanisms, including OLTCs OLTCs, shunt capacitors, and voltage regulators, operate at relatively slow timescales and are not well-suited to managing rapid fluctuations caused by high-frequency PV variability [61].

Early research in voltage regulation has predominantly focused on alleviating under-voltage issues and minimizing power losses. For instance, a deep reinforcement learning (DRL)-based dispatch strategy was proposed in [62] to mitigate under-voltage problems in low-voltage distribution systems. Similarly, a P-Q adjustment strategy for PV inverters was introduced in [63] to provide local voltage support. Although these methods have shown effectiveness in managing specific voltage issues, most of them operate under a single-mode control paradigm, which limits their ability to adapt to dynamically varying operational conditions. In contrast, modern distribution networks often require multi-objective voltage regulation that can simultaneously address over-voltage, under-voltage, and energy efficiency concerns.

To improve adaptability, recent works have introduced mode-switching control strategies for voltage regulation. For example, [64] proposed a scheme for PV inverter control under unbalanced voltage sag conditions, while [65] explored multi-mode control strategies for voltage support in high-voltage DC transmission systems. Further, [66] introduced a voltage-var control (VVC) and conservation voltage reduction (CVR) strategy that switches between control modes to handle fluctuations and optimize energy consumption. However, few of these studies simultaneously consider the three major challenges in distribution networks, over-voltage, under-voltage, and high energy losses, in an integrated and dynamic regulation framework.

From a control architecture perspective, voltage regulation strategies are generally categorized as centralized, decentralized, or distributed. Centralized methods require global system information for real-time decision-making [67], which entails high communication costs and computation burdens. Moreover, these centralized approaches are typically limited in their ability to track fast voltage deviations caused by volatile DER output. On the other hand, decentralized strategies rely solely on local measurements [68], leading to a lack of system-wide coordination and limited performance under complex network conditions. To overcome the limitations of both paradigms, distributed voltage regulation frameworks have been proposed. These strategies often use two-timescale structures to coordinate fast inverter-based regulation and slow mechanical control devices. For instance, [69] developed a two-timescale voltage control strategy for managing smart inverters and capacitors, while [70] proposed a distributed coordination mechanism that aligns the control schedules of OLTCs and DERs. A hybrid hierarchical framework was also introduced in [71] to simultaneously minimize power loss and regulate real-time PV output using both centralized and distributed elements.

Despite their practical value, most model-based distributed voltage regulation methods depend heavily on accurate system models and reliable communication infrastructure [72]. However, in reality, the acquisition of real-time topology and system state information is often constrained by communication bandwidth and measurement accuracy [73]. To overcome this challenge, data-driven methods based on DRL have emerged as promising alternatives. These methods learn voltage control policies through interaction with simulation environments, without requiring detailed system models [74]. An agent-based volt–var control strategy was proposed in [75] to optimize energy dispatch in integrated energy systems. Meanwhile, [59] introduced a decentralized voltage control strategy for active distribution networks using the DDPG algorithm. In [76], a collaborative multi-agent DDPG framework was developed for volt–var control in the presence of high DER penetration.

However, many DRL-based strategies, particularly those based on DDPG and its multi-agent variants, face critical limitations in practice. These include instability during training, sensitivity to hyperparameter tuning, and performance degradation in high-dimensional environments with a large number of agents [77]. In particular, the standard multi-agent deep deterministic policy gradient (MADDPG) algorithm becomes increasingly ineffective as agent count grows, making it difficult to scale DRL-based voltage control to large distribution systems.

To address these issues, recent works have attempted to integrate attention mechanisms into MADRL frameworks to enhance control performance in multi-agent environments. Such mechanisms allow each agent to selectively focus on relevant local or global observations, improving coordination and learning stability. While these methods show promise, most existing studies have yet to fully incorporate multi-mode voltage regulation objectives and dynamic adaptation to diverse network conditions. In addition, practical considerations such as energy efficiency, real-time responsiveness, and communication constraints remain insufficiently addressed.

In summary, while traditional voltage regulation methods offer valuable foundations, they fall short in addressing the full range of operational challenges introduced by high DER penetration and real-time uncertainty. Centralized approaches are often impractical for real-time control, and decentralized strategies lack coordination. Distributed frameworks improve scalability but remain heavily reliant on system models and communications. Reinforcement learning–based methods provide model-free adaptability and fast decision-making, but their scalability and robustness must be further enhanced. Therefore, there is a strong research need to develop scalable, stable, and multi-objective DRL-based voltage regulation strategies that can coordinate both traditional and inverter-based devices across multiple timescales under uncertainty.

2.1.3 Review on T&D System Coordination for Emergency Load Restoration

The coordinated restoration of loads across T&D systems during large-scale contingencies has emerged as a critical area of research in power system resilience. Traditionally, TSOs and DSOs have managed their respective networks independently, with limited information sharing or control coordination. This lack of integration can lead to conflicting operational decisions, suboptimal load recovery strategies, and even exacerbation of system stress during emergencies [78].

Centralized restoration frameworks have been proposed to coordinate T&D system operation. These approaches typically rely on comprehensive system models and global data exchange, resulting in significant communication overhead and computational complexity. In large-scale systems with extensive DER deployment, centralized models become increasingly impractical due to the combinatorial explosion of variables and the latency involved in data acquisition and optimization. To mitigate this, distributed optimization frameworks have gained popularity for coordinating T&D restoration in emergency scenarios [79] , [80].

Among distributed optimization techniques, the Lagrangian relaxation method is one of the most widely used. It allows TSOs and DSOs to solve their respective subproblems independently, exchanging boundary conditions iteratively. Variants of this method include the alternating direction method of multipliers [78], analytical target cascading [79], proximal message passing [80], and mixed-integer boundary-compatible approaches [81]. Other techniques leverage Karush–Kuhn–Tucker (KKT) conditions to facilitate primal–dual decentralized optimization, enabling distributed solutions for economic dispatch and AC OPF [82], [83], [84]. For example, Benders decomposition has been applied to decentralized reactive power dispatch, utilizing transmission-level voltage regulation to mitigate distribution-level overvoltage conditions [85].

Despite these advances, the majority of distributed coordination frameworks adopt sequential update mechanisms. In such schemes, each subsystem must solve its local problem in sequence, based on the latest received boundary conditions. This sequential dependency significantly limits the scalability and responsiveness of restoration algorithms, particularly when rapid recovery is required during N-*k* contingencies. To address this issue, parallel computing approaches have been proposed to enhance the efficiency of distributed optimization in T&D coordination. However, these methods often face trade-offs between convergence speed and solution quality, especially when strict operational constraints are imposed across system layers.

In addition to optimization-based frameworks, recent research has focused on improving the responsiveness and scalability of distribution-level restoration using DERs. Strategies have been developed that utilize distributed generators (DGs) and mobile energy storage to accelerate localized recovery [86], [87]. For instance, DG scheduling algorithms have been designed to rapidly match local demand following outages [88]. However, as the penetration of DERs increases, the complexity of managing these resources in a coordinated manner across the entire distribution network also grows [89].

Traditional centralized scheduling strategies for DERs suffer from high communication costs, limited scalability, and computational bottlenecks. To alleviate these issues, decentralized DER coordination schemes have been introduced. These include approaches based on local terminal measurements [90] or real-time local control without relying on centralized data [49]. However, such methods either require sophisticated communication infrastructure or are limited in scope to small-scale or localized restoration tasks. As an alternative, aggregator-based approaches, such as the VPP concept, have been proposed to manage large populations of DERs through a hierarchical structure [91]. By serving as an intermediary between DSOs and DERs, the VPP reduces communication burdens, enhances controllability, and provides a scalable platform for coordinated response during emergencies[92], [93]. Despite their advantages, aggregator-driven strategies for emergency restoration have received limited attention and remain underexplored in current literature.

Beyond model-based optimization, DRL has been increasingly adopted for real-time decision-making in complex power system restoration tasks. DRL methods offer the advantage of learning optimal control policies through interaction with the environment, without requiring precise system modeling or extensive pre-defined datasets. For example, a DRL-based SCOPF strategy was proposed in [94] to enhance the robustness of transmission systems, while a hybrid DRL approach for preventive control under uncertainty was introduced in [35]. These methods demonstrate the ability of DRL to replace conventional control logic with adaptive, data-driven strategies that better handle nonlinear system behavior and uncertain disturbances [95].

MADRL has also been explored for large-scale power system control tasks. For instance, [96] presented a MADDPG framework for voltage control in transmission systems. While this framework improves local autonomy, its learning performance degrades in high-dimensional settings with many agents. To alleviate this issue, attention mechanisms have been embedded into MADRL algorithms to improve scalability and coordination [97]. However, standard MADRL frameworks often rely heavily on local observations and fail to incorporate system-wide information, which limits the agents' ability to optimize a shared objective function.

In summary, existing literature highlights a clear trajectory toward integrating distributed optimization and learning-based approaches for T&D coordinated restoration. Model-based distributed methods are theoretically sound but computationally intensive and hard to scale for real-time application. DRL-based methods offer adaptive control and reduce reliance on explicit models, but face challenges related to coordination, information sharing, and learning stability. Aggregator-driven architectures such as VPPs offer a promising solution by reducing communication overhead and enabling DER coordination across the distribution layer. To fully realize resilient T&D restoration, future research must focus on hybrid frameworks that combine distributed optimization, DRL, and scalable communication infrastructures to ensure rapid, reliable, and coordinated recovery during extreme events.

Chapter 3 Real-time Resilient Power System Operation with Defender-Attacker Soft Actor-Critic Reinforcement Learning

Threatened by weather disasters and operational uncertainties, resilient and economic decision-making in power systems has garnered significant attention for maintaining system security. Consequently, formulating operational models has become crucial, particularly with the adoption of two-stage decision-making frameworks such as contingency-constrained optimal power flow (CCOPF), a complex, large-scale, nonconvex problem. This paper introduces a novel robust deep reinforcement learning approach named defender-attacker soft actor-critic (DA-SAC), tailored for CCOPF with N-*k* security criteria. Initially, a specialized Markov decision process (MDP) model is standardized for the nested two-agent system. The primary agent generates resilient control actions, while the adversarial agent identifies the worst-contingency scenarios to maximize regulation costs. A power flow-based best response procedure is developed in a computationally efficient uncertain environment to minimize load shedding during attack scenarios. To enhance the feasibility and stability of the foundational soft actor-critic (SAC) algorithm, the degree of constraint violation (DCV) is introduced along with two-timescale learning rates. The effectiveness of the proposed DA-SAC algorithm is validated on two benchmark systems, demonstrating its capability to generate rapid, resilient, and feasible control actions while maintaining stable learning performance.

3.1 Framework

This work addresses the challenge of real-time resilient power system operation under uncertain and high-impact N-*k* contingencies by formulating a two-stage contingency-constrained optimal power flow (CCOPF) problem. The first stage involves pre-contingency planning, where robust control actions are determined to minimize operating costs while accounting for possible future disruptions. The second stage simulates post-contingency conditions, evaluating the system's performance under the worst-case scenarios that result in load shedding, reserve violations, and constraint breaches. This forms a nested max-min optimization problem in which the inner layer represents the attacker's objective of maximizing operational losses, while the outer layer seeks to minimize total system cost and violations. Due to the complexity and nonconvexity of AC power flow equations, traditional decomposition-based optimization methods are computationally intensive and unsuitable for real-time applications. Therefore, this work reformulates the CCOPF as a dynamic decision-making process to enable rapid and robust responses.

To solve this problem efficiently, a novel defender-attacker soft actor-critic (DA-SAC) algorithm is proposed, grounded in a competitive Markov decision process (MDP) framework. Two agents are defined: the defender agent (DA), which produces resilient control strategies under uncertainties, and the attacker agent (AA), which identifies the most disruptive contingencies to test the robustness of these strategies. The DA uses a continuous SAC algorithm to generate optimal power dispatch and load shedding actions, while the AA employs a discrete SAC variant to select attack scenarios. The reward function integrates operational cost and a normalized degree of constraint violation (DCV) to ensure feasibility and system security. To stabilize the adversarial learning process, a non-cooperative strategy is adopted where the DA receives auxiliary information from the AA's Q-values, improving learning stability and convergence. Additionally, a power flow-based best response mechanism is integrated into the environment to simulate realistic post-contingency responses. Experimental validation on IEEE test systems demonstrates that the proposed DA-SAC framework effectively minimizes unserved energy and constraint violations, achieving fast, reliable solutions suitable for real-time grid operation.

3.2 Problem Formulation

To ensure continuous operation under uncertain contingencies such as extreme weather and equipment failures, system operators seek to determine the optimal economic and resilient operational strategy with minimal computational burden. Consequently, this decision-making problem is framed as a two-stage robust optimization model. The goal is to minimize operating costs in the pre-contingency stage while accounting for the worst-case scenarios that maximize post-contingency operational losses in the second stage. The objective function can be expressed as follows:

$$\max_{u\in\mathcal{U}} \min_{\Omega'} \sum_{\forall t}\Big[\sum_{\forall g\in\mathcal{G}} C_g^{+}\Delta r_{g,t}^{+} + C_g^{-}\Delta r_{g,t}^{-} + \sum_{\forall d\in\mathcal{D}} C_d^{'}\Delta p_{d,t}^{'}\Big] + \min_{\Omega} \sum_{\forall t}\Big[\sum_{\forall g\in\mathcal{G}} C_g p_{g,t} + \sum_{\forall d\in\mathcal{D}} C_d \Delta p_{d,t}\Big], \tag{3.1}$$

where the first three terms represent post-contingency operating costs, and the last two represent pre-contingency costs. G and D are the sets of generators and power demands, respectively. The prime symbol (′) indicates post-contingency variables. $p_{g,t}$ and $\Delta p_{d,t}$ denote the active power output from generator *g* and load shedding from demand *d* at time *t*, respectively.

In the proposed framework, power demands are randomly generated according to their stochastic profiles. Consequently, power demands can be high, potentially making the problem infeasible under normal operating conditions. Therefore, load shedding is considered in the first stage to relax constraints and improve stability. Additional load shedding, $\Delta p'_{d,t}$, which occurs due to contingencies and is referred to as unserved electricity, is penalized in the second stage with a penalty C'_d significantly more significant than C_d. After a contingency, some generators automatically adjust their outputs to maintain system stability according to their reserves $r_{g,t}$. Thus, reserve violations, $\Delta r_{g,t}^{+}$ and $\Delta r_{g,t}^{-}$, are penalized in the objective of the second stage.

Operational constraints for the CCOPF in the pre-contingency stage are defined in (3.2)-(3.9). Linear constraints (3.2)-(3.8) represent generation capacities, voltage security constraints, power flow limits, and logical limits of load shedding. $p_{g,t}/q_{g,t}$, $v_{i,t}/\theta_{i,t}$, and $s_{ij,t}/s_{ji,t}$ represent the active/reactive power outputs from generator g, voltage magnitude/ angle at bus i, and sending/receiving power flow between buses i and j, respectively. $\underline{X}_x/\overline{X}_x$ denote the minimum/maximum values of parameter X_x. $P_{d,t}$ is the total power demand.

RD_g/RU_g are the ramping down/up limits of generator g. Finally, (3.9) represents all non-convex AC power flow equations, where $f^{Pre}(\cdot)$ is a nonlinear function [41], [98].

$$\underline{P}_g \le p_{g,t} \le \overline{P}_g, \forall g, t, \tag{3.2}$$

$$\underline{Q}_g \le q_{g,t} \le \overline{Q}_g, \forall g, t, \tag{3.3}$$

$$-RD_g \le p_{g,t} - p_{g,t-1} \le RU_g, \forall g, t, \tag{3.4}$$

$$\underline{V}_i \le v_{i,t} \le \overline{V}_i, \forall i, t, \tag{3.5}$$

$$\underline{\Theta}_i \le \theta_{i,t} \le \overline{\Theta}_i, \forall i, t, \tag{3.6}$$

$$\underline{S}_{ij} \le s_{ij,t} \le \overline{S}_{ij}, \forall ij, ji, t, \tag{3.7}$$

$$0 \le \Delta p_{d,t} \le P_{d,t}, \forall d, t, \tag{3.8}$$

$$[p_{g,t}, s_{ij,t}] = f^{\text{Pre}}(v_{i,t}, \theta_{i,t}, P_{d,t} - \Delta p_{d,t}) \tag{3.9}$$

In the post-contingency stage, the operational constraints consider the attacker's actions, denoted by u, where the objective is to maximize operational losses in this stage. The attacker operates within an uncertainty set $\mathcal{U}$. Various forms of uncertainty sets have been proposed in relevant studies to encompass different electric power system components, such as generation units, transformers, power lines, and reactive power injections [99]. Additionally, these sets can model extreme storm behaviors with specific time and geographical constraints [98].

For simplicity, the uncertainty set in this study considers only the availability of transmission lines and power units. Nevertheless, other uncertainty sets can be integrated into the proposed approach without modification. Consequently, $\mathcal{U}$ is defined as follows:

$$\mathcal{U} := \left\{ u \in \{0,1\} \middle| \sum_{\forall ij} u_{ij,t} + \sum_{\forall g} u_{g,t} \le k, h_{ij,t} = 1 - u_{ij,t}, h_{g,t} = 1 - u_{g,t}, \forall ij, g, t \right\} \tag{3.10}$$

where $u_{ij,t}/u_{g,t}$ indicates the attacker status of the component, 1 if it is attacked and 0 otherwise; $h_{ij,t}/h_{g,t}$ represents the availability of the power component. Once the attack status is realized, the system will try to maintain stability given the robust control actions in the first stage. Therefore, the post-contingency constraints are defined as

$$h_{g,t}(p_{g,t}-r_{g,t}^{-}) \le p_{g,t}^{'} \le h_{g,t}(p_{g,t}+r_{g,t}^{+}), \forall g,t, \tag{3.11}$$

$$0 \le \Delta r_{g,t}^{+} \le p_{g,t}^{'} - (p_{g,t}+r_{g,t}^{+}), \forall g,t, \tag{3.12}$$

$$0 \le \Delta r_{g,t}^{-} \le (p_{g,t}-r_{g,t}^{-}) - p_{g,t}^{'}, \forall g,t, \tag{3.13}$$

$$h_{g,t}\underline{Q}_g \le q_{g,t}^{'} \le h_{g,t}\bar{Q}_g, \forall g,t, \tag{3.14}$$

$$-h_{g,t}RD_g - (1-h_{g,t})\bar{P}_g \le p_{g,t}^{'} - p_{g,t-1}^{'} \le h_{g,t}RU_g + (1-h_{g,t})\bar{P}_g, \forall g,t, \tag{3.15}$$

$$\underline{V}_i \le v_{i,t}^{'} \le \bar{V}_i, \forall i,t, \tag{3.16}$$

$$\underline{\Theta}_i \le \theta_{i,t}^{'} \le \bar{\Theta}_i, \forall i,t, \tag{3.17}$$

$$h_{ij,t}\underline{S}_{ij} \le s_{ij,t}^{'} \le h_{ij,t}\bar{S}_{ij}, \forall ij, ji, t, \tag{3.18}$$

$$0 \le \Delta p_{d,t}^{'} \le P_{d,t} - \Delta p_{d,t}, \forall d,t, \tag{3.19}$$

$$[p_{g,t}^{'}, s_{ij,t}^{'}] = f^{\text{Post}}(p_{g,t}, v_{i,t}^{'}, \theta_{i,t}^{'}, P_{d,t} - \Delta p_{d,t} - \Delta p_{d,t}^{'}, h_{g,t}, h_{ij,t}) \tag{3.20}$$

where (3.11) ensures that regulated power outputs comply with generation reserves based on their availability. The violations in reserves are calculated through (3.12)-(3.13) and minimized in the objective function (3.1). It is important to note that this work focuses on solving the two-stage CCOPF problem, so generation reserves are considered predefined and not optimized in the first stage [42]. However, the proposed model can incorporate generation reserves in the robust action of DA without additional modifications. Ramping capacities, voltage magnitudes, and angles are defined in (3.15)-(3.17), respectively. Depending on the availability of power lines, their flows are restricted by (3.18), and additional power shedding is defined in (3.19). Finally, f^{Post} encompasses all nonlinear AC power flow equations in the post-contingency stage.

The CCOPF model is formulated as a two-stage robust optimization problem to identify optimal robust control actions while considering the worst-case scenarios under N-*k* security criteria. This model can be expressed as follows:

$$\begin{aligned} &\min\nolimits_{\Omega} \text{Cost}^{\text{Pre}} + \text{Cost}^{\text{Post}} \\ &s.t. \quad (3.2)-(3.9), \end{aligned} \tag{3.21}$$

$$\begin{aligned} & u \in \arg\max_u \min_{\Omega,\Omega'} \text{Cost}^{\text{Post}} \\ & s.t. \quad (3.11)-(3.20) \end{aligned} \tag{3.22}$$

The objectives are defined in (3.1). This model cannot be directly solved with traditional solvers. It can be reformulated as a single-level problem by introducing optimality and feasibility cuts for all possible worst-case contingency scenarios. The resulting model can then be solved using a nonlinear solver. However, this approach results in a high-dimensional problem with extensive nonconvex constraints, making it challenging to find an optimal resilient solution for real-time operations. Another approach is to use decomposition or nested algorithms to handle the large number of contingencies. However, this method is time-consuming due to the increasing constraints per iteration. In this work, we employ advanced deep learning technology to solve the problem effectively and quickly, ensuring high reliability for real-time operation.

However, traditional optimization techniques, such as mixed-integer linear programming for unit commitment and nonlinear programming for optimal power flow, guarantee strict adherence to all physical and operational constraints. These methods are particularly suitable for day-ahead scheduling or planning problems, where solution feasibility and optimality are of paramount importance, even at the expense of long computation times. In contrast, deep reinforcement learning (DRL) shifts the heavy computation to the offline training phase, enabling real-time decision-making with negligible inference cost. This makes DRL attractive for real-time operation under high uncertainty, such as corrective dispatch after contingencies or fast voltage regulation with high renewable penetration. Nevertheless, DRL frameworks may generate unsafe or infeasible control actions if constraints are not properly embedded, and thus require careful design.

3.3 Methodology

3.3.1 Markov decision process formulation

To apply a reinforcement learning approach, optimization problems or control tasks are reformulated as a MDP model [50]. In this model, one or more agents interact with an uncertain environment to gradually improve their control policy through exploration. Unlike commonly adopted simple MDP models, which typically involve a single agent [35] or multiple agents cooperating on the same task [100], this work develops a specialized MDP for competitive agents. Specifically, the power system operator, DA, aims to minimize operational costs by implementing robust and resilient control actions a_t^d against all possible contingency scenarios. Conversely, the attacker, AA, seeks to maximize post-contingency costs by determining the attack action a_t^a. To identify the worst contingency scenario, the predicted actions of the DA should be considered in the state of the AA to expedite policy exploration. Fig. 3.1 illustrates the interactions between the two agents and the environment. The DA predicts robust actions based on the latest states of power demands and renewable energy outputs, denoted as s_t^d. The AA predicts the worst attack given the predicted action a_t^d and other environmental states s_t^a. The environment then generates rewards for each agent, r_t^d for the DA and r_t^a for the AA, along with the new states s_{t+1}^d and s_{t+1}^a.

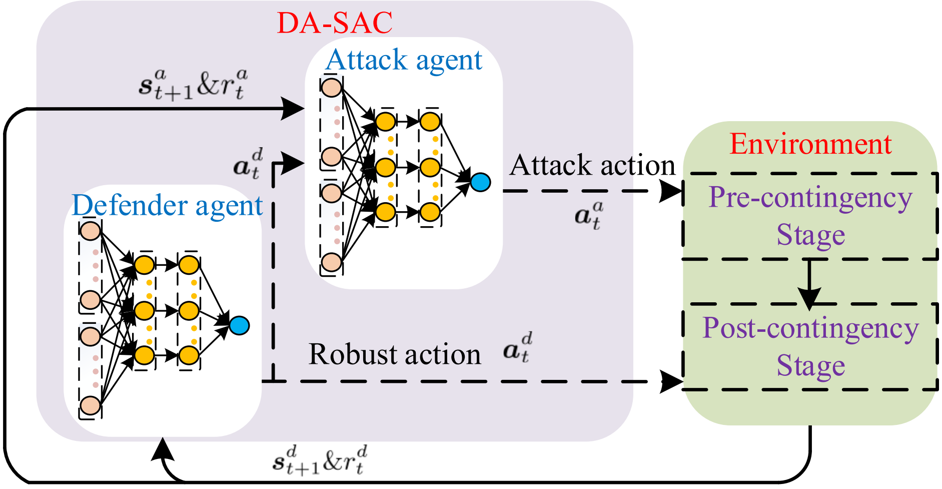


Fig. 3.1 The developed MDP model for the nested agents.

To formulate the MDP model, the main components of the DA, the AA, and the environment are defined as follows. The DA generates robust actions a_t^d using the control policy $\pi^d(s_t^d)$ to maximize the cumulative discounted reward $\sum_{k=1}^{N}(\gamma^d)^{k-1}r_k^d$. Thus, it can be defined by the tuple $(s_t^d, a_t^d, r_t^d, \gamma^d, \mathbb{P}^d)$. s_t^d represents the input states, including active and reactive power demands as defined in (3.23). It is important to note that renewable energy outputs are considered uncertain in this setting and are therefore included in $P_{d,t}$ with a negative sign. The predicted action a_t^d is defined in (3.24), where $\mathcal{I}_v$ and $\mathcal{I}_g$ are subsets of power buses that include reactive power injections and active power injections (excluding the slack bus), respectively. Instead of considering all decision variables from (3.21)-(3.22), the selected actions in (3.24) are controllable and include the minimum necessary actions to improve learning convergence and stability. The defense action chosen here is continuous action, as it involves regulating the adjustment of generators and the load shedding in the demand buses, which is inherently a continuous action. This paper leverages recent advancements in AC power flow solvers [41], [101] to derive the full decision vector from this action space. The reward value per time step r_t^d should reflect the action value taken by the DA. It is defined in (3.25) to include all operational costs, i.e., pre- and post-contingency costs, and the DCV value of violated constraints in the two stages. *K* represents a penalty value. Finally, γ^d is the discount rate for the cumulative reward, and $\mathbb{P}^d$ is the transition function, which the reinforcement learning algorithm will learn.

$$s_t^d = \left(P_{d,t}, Q_{d,t}, \forall d\right), \forall t, \tag{3.23}$$

$$a_t^d = \left(v_{i,t}, \forall i \in \mathcal{I}_v, p_{i,t}, \forall i \in \mathcal{I}_g, \Delta p_{d,t}, \forall d\right), \forall t, \tag{3.24}$$

$$r_t^d = -\text{Cost}^{\text{Pre}} - \text{Cost}^{\text{Post}} - K(\text{DCV}^{\text{Pre}} + \text{DCV}^{\text{Post}}) \tag{3.25}$$

In the reward function, two degrees of constraint violation are incorporated to account for both the pre-contingency and post-contingency system conditions. Specifically, the defense agent applies its control actions in the pre-contingency stage. Since reinforcement learning agents may produce unsafe actions that could violate operational constraints, the degree of constraint violation is explicitly evaluated during the pre-contingency power flow calculation and included in the reward to discourage infeasible preventive actions. At the same time, the defense agent's actions also propagate into the post-contingency environment, where the system is subjected to the worst contingency scenarios. In this stage, the resulting system state reflects how the SCOPF solution performs under stressed conditions, and constraint violations may arise due to line overloads, voltage deviations, or generation limits. Therefore, the degree of constraint violation is also computed in the post-contingency stage and incorporated into the reward function.

Similarly, the AA is defined by the tuple $(s_t^a, a_t^a, r_t^a, \gamma^a, \mathbb{P}^a)$. The states s_t^a include active and reactive power demands as well as the robust action from the pre-contingency stage, as defined in (3.26). The predicted attack action a_t^a is a discrete action space, represented in (3.27), where $\mathcal{G}$ and $\mathcal{L}$ are sets of generation units and power lines, respectively. In other words, the AA selects an attack from a list that considers all possible combinations of equipment failures according to the adopted security criterion *k*. In contrast to the defense action a_t^d, the chosen attack action is a discrete action. This is due to the fact that deciding to disconnect generators or transmission lines is a binary decision, which makes it fundamentally discrete in nature. The reward value per time step r_t^a is defined in (3.28) to encompass post-contingency costs and the DCV value of violated constraints in the second stage.

$$s_t^a = \left(a_t^d, P_{d,t}, Q_{d,t}, \forall d,\right), \forall t, \tag{3.26}$$

$$a_t^a = \left\{1, 2, \ldots (|\mathcal{G}| + |\mathcal{L}|)^k\right\}, \forall t, \tag{3.27}$$

$$r_t^a = \sum C_d^{'} \Delta p_{d,t}^{'} + K \cdot \mathrm{DCV}^{\mathrm{Post}}, \forall d \in \mathcal{D} \tag{3.28}$$

This formalism is modeled as a two-player zero-sum MDP with one-sided incomplete information. In particular, the attacker knows the actions of the operator while the operator does not. The game proceeds as follows. Initially, a state-reward pair $(s_0,\ r_0)$ is sampled from the prior distribution $\mathcal{P}_0$. The state s_0 is publicly observed by both players, while the attacker observes the operator's action $a^d \in \mathcal{D}$ and chooses action $a^a \in \mathcal{D}$. Given both actions, the current state s_t transitions to a successor state s_{t+1} according to the transition model $\mathbb{P}(s_{t+1}|s_t, a^d, a^a)$. The attacker receives a reward $\mathcal{R}\left(s_t^a; a_t^a; a_t^d; s_{t+1}^a\right)$; the operator receives the reward $\mathcal{R}\left(s_t^d; a_t^d; s_{t+1}^a\right)$. The competition results in the operator possessing incomplete information, requiring it to maintain a belief of enhancing robustness over the worst contingency scenario. The attacker and the operator aim to minimize operating costs while maximizing post-contingency operational losses.

$$\min_{\pi^d} \max_{\pi^a} \mathbb{E}\left[R(\pi^d, \pi^a)\right], \tag{3.29}$$

$$s.t. \quad (3.21)-(3.22) \tag{3.30}$$

To ensure the secure operation of the power system, all constraint violations are normalized in one number called the degree of constraint violations (DCV) to be included in the reward function as defined in (3.25) and (3.28). It is defined as

$$\mathrm{DCV} = \sqrt{\frac{1}{|\mathcal{X}|}\sum_{\forall x_n} \zeta_n \left(\frac{[x_n - \overline{x}_n]^+ + [\underline{x}_n - x_n]^+}{\overline{x}_n - \underline{x}_n}\right)^2} \tag{3.31}$$

where x_n collects all uncontrolled constraints in (3.21)-(3.22). The number of constraints is $|\mathcal{X}|$, with minimum $\underline{x}_n$ and maximum $\overline{x}_n$ limits. These limits are obtained from (3.2)-(3.8) and (3.11)-(3.19) for DCV$^{\mathrm{Pre}}$ and DCV$^{\mathrm{Post}}$, respectively. $[\cdot]^+$ indicates $\max\{0,\cdot\}$. Finally, ζ_n is an optional factor to increase the weight of the constraint *n* compared with others.

Because the post-contingency stage can result in disconnected subsystems (zones *z*) within the power system, a new procedure is required to calculate rewards and generate new states in the simulation environment under these disconnections. Recent advances in power flow (PF) solvers [102] have demonstrated their capability to find fast and robust solutions. Building on these advances, **Algorithm 1** is developed to execute the environment with a few straightforward steps.

Algorithm 1: Power Flow-based Best Response Procedure

```
Input: $s_t^d, a_t^d, v_{i,t}$ and.$a_t^a$.
Solve: pre-contingency PF problem with $a_t^d$ and $s_t^d$ . Get Cost$^{\mathrm{Pre}}$
    by (3.1) and DCV$^{\mathrm{Pre}}$ by (3.31).
Apply: attack actions $a_t^a$, update system topology and get $\mathcal{B}_z, \forall z$.
For each zone z:
    If slack bus $k \in \mathcal{B}_z$,
        Solve PF with $a_t^d(z)$ and $s_t^d(z)$.
        $\Delta r_k^+ = [p_k' - (p_k + r_k^+)]^+, \Delta r_k^- = [(p_k - r_k^-) - p_k']^+, \Delta p_d'(z)$
            $= 0$.
    Else if $\exists g \in \mathcal{G}_i, i \in \mathcal{B}_z$,
        Select a slack bus $k$, where $k = arg\,max_k \sum_{g\in\mathcal{G}_k} \Delta r_g^+ + \sum_{g\in\mathcal{G}_k} \Delta r_g^-$
        Solve PF with $a_t^d(z)$ and $s_t^d(z)$.
      $\Delta r_k^+ = [p_k' - (p_k + r_k^+)]^+, \Delta r_k^- = [(p_k - r_k^-) - p_k']^+, \Delta p_d'(z) = 0.$
7:      Else if $\nexists g \in \mathcal{G}_i, i \in \mathcal{B}_z$,
            $\Delta p_d'(z) = P_d(z) - \Delta p_d(z)$.
8: Calculate: $\mathrm{Cost}^{\mathrm{Post}}$ using (3.1), $\mathrm{DCV}^{\mathrm{Post}}$ using (3.31), $r_t^d$ using
      (3.25), $r_t^a$ using (3.28).
```

When attack actions a_t^a are applied, the system topology changes, grouping the buses into sets $\mathcal{B}_z, \forall z$. In some zones, such as those involving lines 5 and 6 in the procedure, the PF solver updates the active power generation from the slack bus to mitigate the attack. Power regulation is penalized if it exceeds generator reserves. Conversely, in zones without generators (as in line 7), the PF solver is unnecessary, and the power demands of that zone remain unmet.

Considering these dynamics, the environment calculates rewards and generates new states accordingly. The simulated power system environment accurately reflects the impact of disconnections and ensures that the power system's response is realistic and robust.

3.3.2 Defender-attacker soft actor-critic DRL algorithm

Traditional model-free DRL algorithms often need help with low sampling efficiency and weak convergence. While practical, on-policy algorithms like A3C and PPO suffer from lower sampling efficiency due to their reliance on data directly related to the current policy, limiting their data utilization scope [103]. On the other hand, off-policy algorithms such as DDPG utilize a broader data set by sampling from an experience buffer, which can potentially increase sampling efficiency [104]. However, the increased sampling scope does not inherently translate to higher efficiency, as both algorithm types must effectively manage the relevance and diversity of sampled experiences. To address these challenges, the soft actor-critic (SAC) algorithm, an off-policy method, was proposed to enhance sampling efficiency by maximizing both the information entropy of system states and discounted cumulative rewards, ensuring robust and efficient learning [103]. Meanwhile, to enhance the operational performance of power systems under pre and post-contingency stages, a robust DRL algorithm is developed based on the fundamental SAC, namely the defender-attacker SAC (DA-SAC) algorithm, where two independent policies for DA and AA are learned in a competitive scenario. Because the action spaces of DA and AA are continuous and discrete, respectively, the fundamental SAC (continuous) and its modified version [52] (discrete) are employed in the developed DA-SAC algorithm.

3.3.2.1 SAC algorithm with continuous action space

The prominent features of the SAC algorithm are due to several essential mathematical and technical techniques. First, the SAC algorithm utilizes a replay buffer to reuse prior experiences for an off-policy formulation, improving sample utilization efficiency. During each gradient step, the actor and critic networks are updated based on a mini-batch of prior experiences sampled from the replay buffer $\mathcal{M} = [(s, a, r, s')]$, where s' represents the new states after applying action a.

Second, SAC is the state-of-the-art entropy maximization-based deep reinforcement learning (DRL) algorithm, where the entropy of the policy is augmented in the policy objective to balance the exploration process, as shown below [105]:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau \sim \pi} \sum_{t=0}^{T-1} \gamma^t [r - \alpha \log \pi(a \mid s)], \tag{3.32}$$

where τ indicates one trajectory. The policy $\pi(s|a)$ maps the system's states to control actions. α represents the entropy temperature, tuning the stochasticity of the optimal policy, i.e., the weight of the entropy term. $\gamma \in [0,1)$ denotes the discounting coefficient.

Third, the SAC algorithm is based on an actor-critic architecture with stochastic actors, where the optimal maximum entropy policies are updated by alternating between critic update (policy evaluation) and actor update (policy improvement). The critic network receives the states and actions and outputs the action value $Q_{\theta}(s, a)$, where θ are the parameters. Using the modified Bellman backup operator, the soft Q-value is given by

$Q_\theta(s,a) = E_{(a',s')\sim\pi}[r_t + \gamma V_\pi(s')]$, where $V_\pi(s) = E_{a\sim\pi}[Q_\theta(s,a) - \alpha\, log(\pi(a|s))]$ is the soft state-value function.

The proposed model: i) Implements two critic networks with different parameters θ_1 and θ_2, taking their minimum values to avoid overestimation issues [99]. ii) Uses a target network for each critic with parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ to improve learning stability [99]. iii) Neglects the state-value network V_π and uses its exact equivalent.

Thus, in the policy evaluation step, the critics are updated by the following loss function:

$$J_Q(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{M}}\Big[\frac{1}{2}\big(Q_{\theta_i}(s,a) - \big(r + \gamma\big\{Q_{\hat{\theta}_i}(s',a') \;-\alpha\log(\pi^*_\varphi(a'|s'))\big\}\big)\big)^2\Big], \forall i \in \{1,2\}, \tag{3.33}$$

where φ are the parameters of the actor network. a' is the control action predicted from the latest updated policy π^*_φ given states s'. Note that target networks are smoothly and periodically updated by

$$\hat{\theta}_i = (1-\tau)\hat{\theta}_i + \tau\theta_i, \forall i \in \{1,2\}, \tag{3.34}$$

where τ is the target update factor. In the policy improvement step, the policy is optimized to maximize the soft Q-function by minimizing the KL-divergence as [105]:

$$J_\pi(\varphi) = \mathbb{E}_{s\sim\mathcal{M}}\big[\mathbb{E}_{a\sim\pi}\big[\alpha\log(\pi_\varphi(a|s)) - \min_{i\in\{1,2\}} Q_{\hat{\theta}_i}(s,a)\big]\big], \tag{3.35}$$

which can be minimized using a reparameterization trick. Given the system states, the policy is modified to predict the mean and standard deviation of the actions' probability distribution (spherical Gaussian). Additionally, policy entropy is maximized to enhance the exploration-exploitation balance and improve learning stability. However, the effectiveness of exploration and learning stability depends on the entropy temperature, which varies across different tasks. Consequently, automating entropy adjustment is proposed in [105] by computing the objective in (35), where $\underline{\mathcal{H}}$ denotes the expected minimum entropy.

$$J(\alpha) = \mathbb{E}_{a\sim\pi}\big[-\alpha\log(\pi(a|s)) - \alpha\underline{\mathcal{H}}\big] \tag{3.36}$$

3.3.2.2 SAC algorithm with disrete action space

The attacker's actions are modeled as discrete to reflect the binary nature of decisions, such as disconnecting specific transmission lines or generators in the power system, as defined by the N-k contingency criteria in (3.27). According to statistical theory, a discrete action A follows a categorical distribution, represented by a probability vector $\mathcal{P} = [p_1, p_2, \ldots, p_k]$. The probability of selecting a specific action $a^* \in A$ is given by $\mathcal{P}(a^* = a_i) = p_i$, where a_i denotes an action in the action space. This distribution enables the attack agent to assign probabilities to potential actions, facilitating exploration of the action space and selection of worst-case contingency scenarios based on the learned policy. This approach ensures the learning process aligns with power system operations while maintaining computational efficiency and stability during policy training [106]. Modeling the attacker's policy with a categorical distribution provides a practical framework for addressing contingencies in CCOPF.

To derive the discrete action-based SAC (DSAC) algorithm, four essential modifications are required in the SAC algorithm: i) Instead of implementing the policy network with outputs representing the mean and variance of control actions, DSAC directly predicts the probability of discrete actions. The softmax function is employed in the output layer to ensure an accurate probability distribution for the outputs, thereby transforming the policy space from continuous to discrete. ii) Because the expectation of the discrete actions can be directly calculated from the probability of discrete distributions, the soft state-action function can be expressed as $V_\pi(s) = \pi_\varphi(s)^\intercal$ [30]. The Q-function loss can then be calculated as follows:

$$J_Q(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{M}}\Big[\frac{1}{2}\Big(Q_{\theta_i}(s,a) \; -\Big(r+\gamma\Big\{\pi_\varphi(s')^\top\Big[Q_{\theta_i}(s',a')-\alpha\log(\pi(a'|s'))\Big]\Big\}\Big)\Big)^2\Big], \forall i, \tag{3.37}$$

where $\pi_\varphi(s')^\intercal$ indicates the expectation value of the discrete action; iii) Similarly, the automating entropy adjustment can be changed to

$$J(\alpha) = \pi_\varphi(s')^\top\Big[-\alpha\log(\pi(s)) - \alpha\underline{\mathcal{H}}\Big], \tag{3.38}$$

Finally, iv) there is no need for the reparameterization trick because the policy predicts the exact action distribution and the new objective is changed to

$$J_\pi(\varphi)=\mathbb{E}_{s\sim\mathcal{M}}\left[\pi_\varphi(s)^{\top}\left[\alpha\log(\pi_\varphi(a\,|\,s))-\min\nolimits_{i\in\{1,2\}}Q_{\hat{\theta}_i}(s,a)\right]\right] \tag{3.39}$$

3.3.2.3 Practical implementation of DA-SAC

The DA and AA are constructed within a min-max framework, sharing the same neural network architecture, except for the policy network, but with independently updated parameters. Each agent generates different actions and interacts with the same environment to obtain distinct rewards. The replay buffer $\mathcal{M}$ stores their prior experiences and randomly samples a mini-batch to update the parameters of DA and AA according to (3.32)-(3.39). The proposed DA-SAC is summarized in **Algorithm 2**.

Algorithm 2: Defender-Attacker SAC Algorithm

```
Initialize: Defender agent networks $\varphi^d$, $\theta_i^d$, $\hat{\theta}_i^d$, and attacker agent networks $\varphi^a$, $\theta_i^a$, $\hat{\theta}_i^a$
For each episode do
    For each time step do
        $a^d \sim \pi_{\varphi^d}(\cdot\,|s^d = s); a^a \sim \pi_{\varphi^a}(\cdot\,|s^a = [a^d, s])$.
        $r^d, r^a, s' \leftarrow$ call Algorithm 1 to execute $a^d, a^a$.
        $\mathcal{M} \leftarrow \mathcal{M} \cup (s, a^d, a^a, r^d, r^a, s')$.
    End For
    For each gradient step do
        Sample random $N$ experiences from $\mathcal{M}$.
        Update soft Q-value parameters $\theta_i^d$ and $\theta_i^a$ by (3.33) and (3.37).
        Update policy parameters $\varphi^d$ and $\varphi^a$ by (3.35) and (3.39).
        Adjust temperature $\alpha^d$ and $\alpha^a$ by (3.36) and (3.38).
        Update targets $\hat{\theta}_i^d$ and $\hat{\theta}_i^a$ by (3.34).
    End For
End For
```

The architecture of the DA-SAC algorithm consists of two sets of networks, as discussed above. Each set can be represented as shown in Fig. 3.2. It includes the three components: the actor, critics/targets, and replay buffer. The actor network comprises a fully connected layer with several hidden layers and outputs a two-dimensional vector that characterizes a

Gaussian distribution of the predicted control actions. In other words, it maps the states s to actions $\mathcal{N}(\mu,\sigma)$. The action determined by the policy $\pi_{\varphi}(\cdot|s)$ and the latest state s are fed into the environment, which produces the reward r and the next state s'. These are then stored in the replay buffer $\mathcal{M}$.

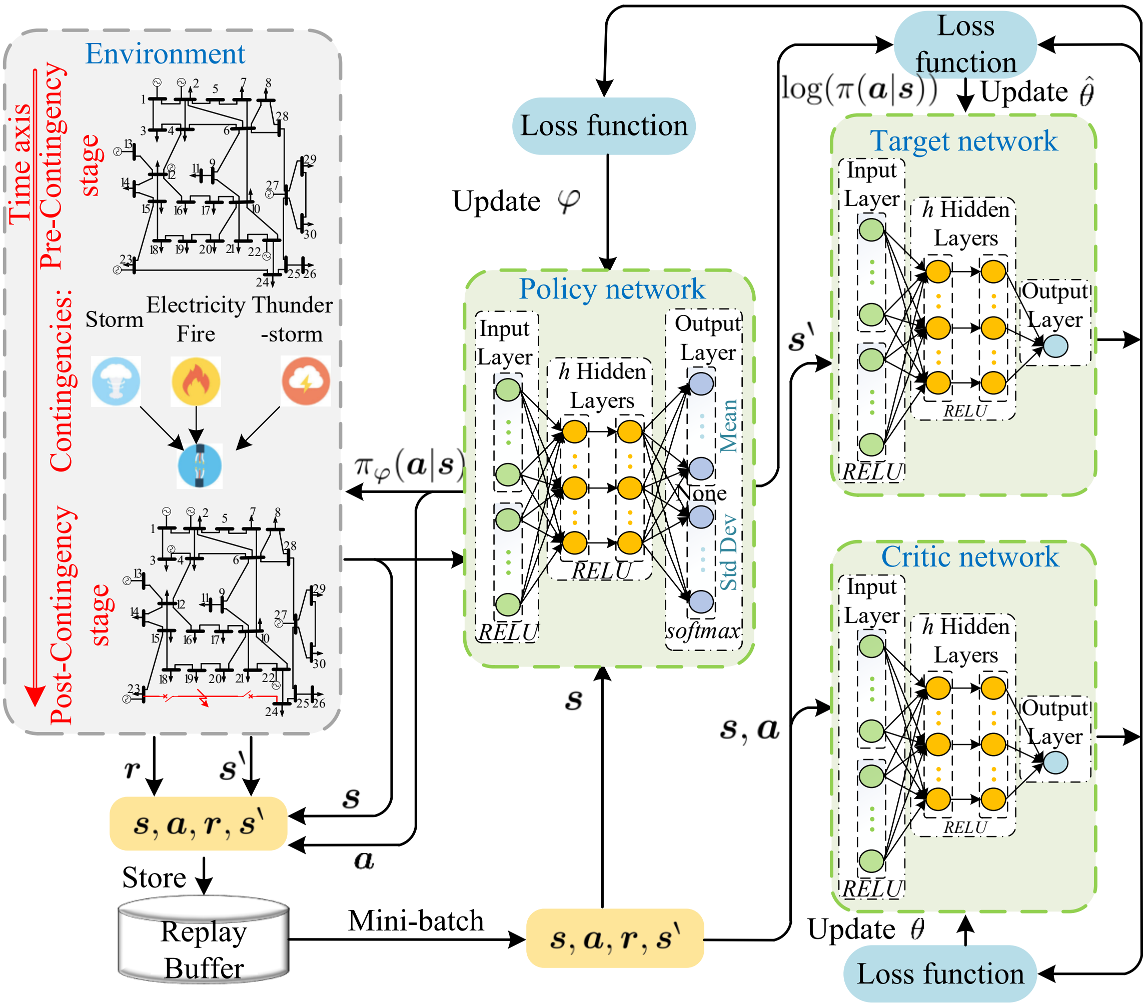


Fig. 3.2 Structure of a set of networks in the DA-SAC agent.

In our study, the training of DRL algorithms is primarily conducted in a simulated environment rather than relying on real-world historical records of equipment failures. During the early stage of training, experiences are generated through random sampling of actions. Each action is applied to the environment, where the resulting system state transition and the corresponding reward value are calculated. These state–action–reward samples are then stored in the replay buffer to serve as the initial learning experience for the DRL agent. This simulation-driven setup ensures that the agent can learn from a wide variety of contingency scenarios, including rare but severe failures, thereby improving the robustness of the trained policy.

The critic network predicts the Q-value of the cumulative reward using an output layer with a single neuron and multiple hidden layers. As previously discussed, the target network shares the same structure as its associated critic. A mini-batch of N experiences, *i.e*, $\{s, a^d, a^a, r^d, r^a, s'\} \sim \mathcal{M}$, is sampled from the replay buffer to calculate the loss functions for each critic and the gradient descent for the actor network to update their parameters. The updating process alternates between collecting prior experiences and updating the parameters of the actor-critic components until the termination criteria are met, such as reaching the maximum number of episodes or achieving the local optimal policy.

3.3.2.4 Stability enhancement for noncooperative agents

The DA and AA compete iteratively in the MDP framework to find optimal actions that maximize their rewards. However, the DA is disadvantaged since the AA can access additional information from the robust action. The states of the AA are designed to include the robust action predicted by the DA's policy to ensure the identification of worst-contingency scenarios, which aligns with practical operation. This essentially forms an embedded Stackelberg game [107] between the attacker and the defender and helps develop robust defense mechanisms [108]. These recent studies [109] underline the validity and necessity of assuming complete information in attackers to realistically prepare and defend against potential sophisticated attacks on power systems.

To address this issue, this work introduces a noncooperative strategy. This strategy involves two competitive agents with completely misaligned objectives, where only one agent has perfect information. The less informed agent receives auxiliary information from the environment rather than cooperating with its competitor. In this setup, the latest improvement of the informed agent's policy parameters and the action value function are transferred to the state of the less informed agent as auxiliary information for the next step. The action value function evaluates the expected future performance of the informed agent's action a_t under state s_t. Consequently, based on the action value function, the less informed agent (DA) can generate effective actions to compete with the informed agent

(AA). This approach is underpinned by existing literature on multi-agent systems where the sharing of strategic information has proven to enhance system robustness and reliability significantly [110]. This proactive strategy not only improves the robustness of the DA's solutions but also reinforces the overall resilience of the electrical power systems against a wide range of threats and disturbances, thereby enhancing reliability and security.

In the proposed setting, the AA agent has two critics with two targets, and the minimum operator is used to ensure solution stability. The input state for the DA during training is given by $min_{i\in\{1,2\}} Q_{\hat{\theta}_i^a}(s,a)$. Therefore, the states of the DA agent can be written as follows:

$$s_{t+1}^{d} = \begin{cases} [P_{d,t+1}, Q_{d,t+1}, \forall d, 0], t = 0, \\ [P_{d,t+1}, Q_{d,t+1}, \forall d, \min_{i\in\{1,2\}} Q_{\hat{\theta}_i^a}(s_t^a, a_t^a)], t > 0 \end{cases} \tag{3.40}$$

Fig. 3.3 illustrates the procedure of the noncooperative strategy within the MDP framework. In this process, the environment produces immediate rewards (r^m, r^a) and the next state s' as a result of the prior actions, which are then stored in the replay buffer. The informed agent (AA) receives complete information from its state space, whereas the less informed agent (DA) only has access to limited information.

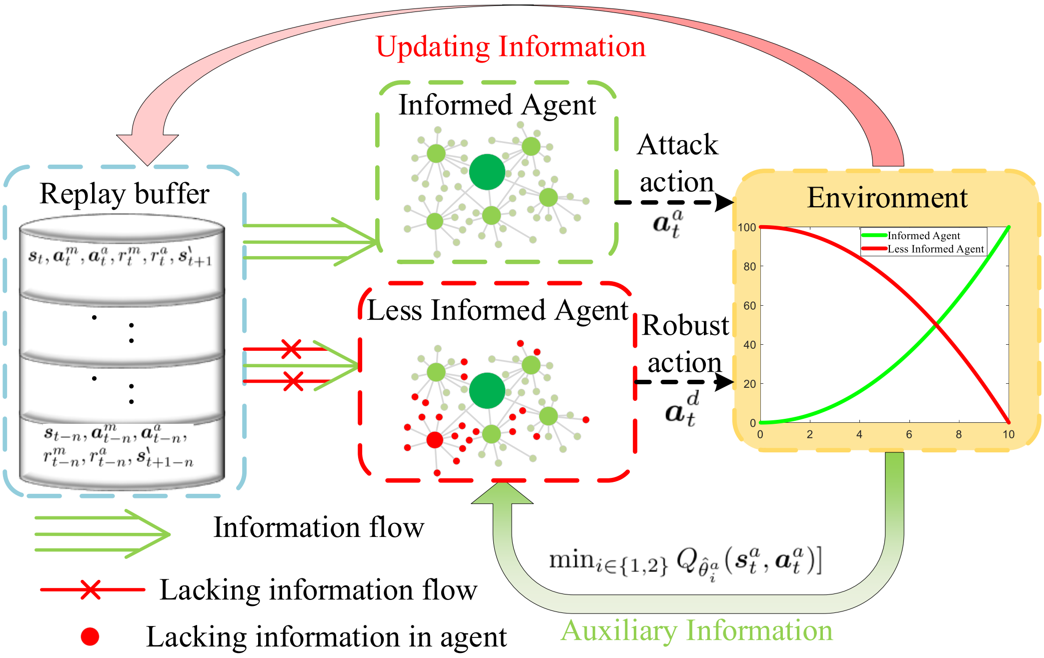


Fig. 3.3 The process of the noncooperative strategy.

Imperfect information games can lead to different learning rate scales and significant fluctuations while the DA and AA compete during the training process. It can result in instability, as most actor-critic approaches with explicit parameterization of π are particularly sensitive to large fluctuations. To enhance the stability and robustness of the DA, the noncooperative strategy provides auxiliary information to the DA, as defined in (3.40). This auxiliary information enables the DA to compete effectively with the AA and generate robust actions against worst-case scenarios during exploration.

This design of the noncooperative strategy ensures that the defense agent is not persistently placed in a weak or dominated position, which is critical for maintaining its learning efficiency. While the interaction between the defense agent and the attack agent may introduce oscillatory dynamics due to their competitive objectives, the use of noncooperative training prevents these oscillations from becoming excessive or destabilizing. As a result, the defender–attacker competition converges to a stable learning process where the defense agent is able to consistently improve its policy, thereby enhancing the robustness of the overall SCOPF solution under worst-case contingencies.

3.4 Case Study

3.4.1 Experimental Setup

The following experimental results and simulations are programmed using Python language with Pycharm as an IDE, and the learning process of the multilayer neural networks in the DRL algorithm is formulated using PyTorch. Numerical tests are implemented on a computer with Intel i7−10700 CPU and 16 GB of RAM. The hyperparameters of the SAC algorithm are presented in Table 3.1. The system parameters, including system topology, generation capacities, and line parameters, are obtained from PYPOWER. Two test systems, namely IEEE 30-Bus and IEEE 118-Bus, are selected for this work. Additional modeling data are generalized as follows. Power reserves are set as $r_{g,t}^{+} = r_{g,t}^{-} = 0.05 \times \overline{P}_g, \forall g, t$, penalties of (3.1) are set as $C_d = 2 \times max_g C_g, C_g^{+} = C_g^{-} = C_d' = 5 \times C_a$, and the violation penalty K in (3.25) and (3.28) is set as 1 × 10^3. Finally, power demands are randomly generated, where maximum and minimum values are set at 120% and 80% of the normal operating point in the data set PYPOWER.

Table 3.1 Main hyper-parameters and data setting.

Parameters	Value	Parameters	Value
Optimizer	Adam	Discount factor	0.99
Critics learning rate	1e-2	Minibatch size	128
Actor learning rate	1e-3	Neurons number	512
Target learning rate	1e-3	Time step	1 hour
Entropy learning rate	1e-4	Max steps	24 hours
Initial temperature	1	Activation	RELU

3.4.2 Training performance of the DA-SAC algorithm

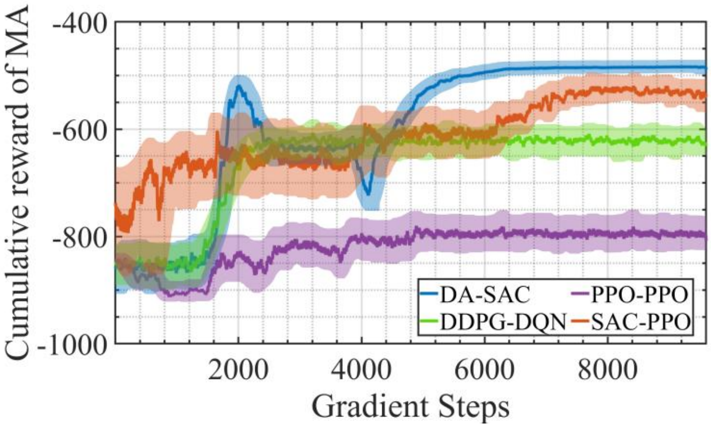
This subsection investigates the training performance of the proposed algorithm with two advanced DRL algorithms. The first benchmark algorithm, DDPG, struggles with discrete actions during offline training. To overcome this limitation, we combined the deep Q-network (DQN) with DDPG, creating the DDPG-DQN algorithm, which effectively generates discrete actions for AA. The second benchmark, PPO, is an on-policy algorithm capable of handling both DA and AA actions, termed the PPO-PPO method. An analysis of the experimental results for the three DRL algorithms reveals that the PPO-based AA and the SAC-based DA agents demonstrate the best convergence performances in attack and defense, respectively. To further validate the effectiveness of the proposed DA-SAC algorithm, a new scheme, SAC-PPO, is introduced, where the DA is trained using SAC and the AA using PPO. All DRL algorithms were tested on the IEEE 30-bus system using the same dataset. Ten independent experiments with different initial seeds and training datasets were conducted for each algorithm to illustrate the DA and AA cumulative reward curves in Fig. 3.4 under the N-1 criterion. The solid curve represents the average of the ten experiments. It should be noted that the DA actors (continuous actions) and AA actors (discrete actions) share identical structures across all algorithms (DDPG, PPO, SAC for DA;

DQN, PPO, SAC for AA) to ensure a fair comparison based solely on algorithmic differences.

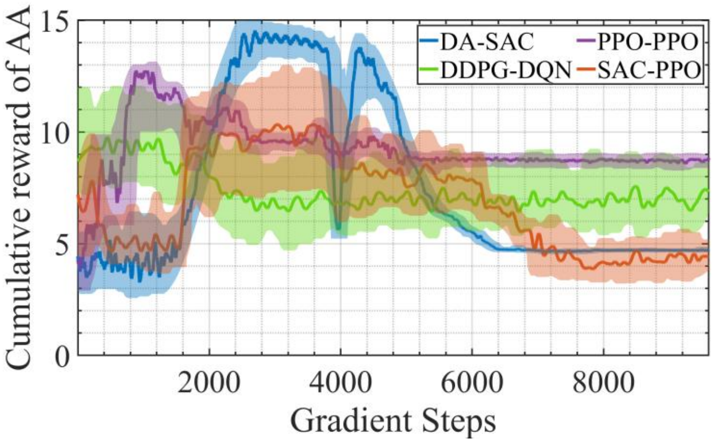
The proposed algorithm demonstrates noticeable reward oscillations during the initial 1800 steps in Fig. 3.4. This oscillation is due to the stochastic exploration by the nested agent to fill the replay buffer. The policy network parameters are updated as the agent interacts with the environment, increasing the reward. After approximately 2000 steps, the DA-SAC reaches an optimal local solution, where the DA achieves high rewards and generates robust actions with less than 10 kWh of unserved electricity. However, as the AA gains an advantage, the reward of the DA quickly drops to 650, and the AA generates the worst contingency scenario with over 15 kWh of unserved electricity. During steps 2400 to 3600, the noncooperative strategy shifts the stable action-value function from AA to DA, promoting competition for robust actions against the worst scenarios. As depicted in Fig. 3.4, the reward curve fluctuates significantly from the 3600th to the 6000th step, ultimately resulting in consistently high cumulative rewards for the DA and low cumulative rewards for the AA. The proposed DA-SAC algorithm converges after 6000 steps, progressing through three stages: policy exploration (first 1800 steps), policy training (1800 to 4800 steps), and policy convergence (after 4800 steps). During the first stage, the algorithm randomly selects actions to collect sufficient initial experience in the replay buffer, resulting in significant reward fluctuations. As the replay buffer accumulates enough experience, the policy network learns and updates according to the principles outlined in Algorithm 2. These sequential updates enable the proposed algorithm to find the optimal policy, resulting in cumulative reward convergence in the final stage. The SAC-PPO algorithm demonstrates strong learning capabilities. While its convergence performance for the AA is weaker compared to the PPO-PPO algorithm, its DA achieves convergence results comparable to those of the proposed DA-SAC algorithm. This advantage stems from SAC's entropy-based offline learning strategy, which provides the DA with a learning edge during adversarial training against the PPO-based AA. However, the interaction between the PPO and SAC algorithms introduces larger fluctuations in the convergence outcomes of both agents, compared to other benchmark algorithms. Overall, the results confirm that the proposed

DA-SAC algorithm outperforms the benchmark DRL algorithms in terms of cumulative rewards for both DA and AA and demonstrates superior convergence stability compared to alternative approaches

Furthermore, to demonstrate the superiority of the proposed algorithm, Table 3.2 presents the average computational results over the last 1e2 episodes across ten independent experiments for the including unserved electricity, load shedding, and DCV per hour, over the last 1e2 episodes across ten independent experiments. The performance comparison of the DA-SAC, DDPG-DQN, PPO-PPO, and SAC-PPO algorithms reveals distinct differences influenced by their unique optimization strategies. DA-SAC achieves a balanced performance with an offline computation time of 1807.29 seconds and an online response time of 152.59 milliseconds. This results in a load shedding of 27.78 kWh, an unserved electricity of 5.46 kWh, and a degree of constraint violation of 0.0636. This indicates that DA-SAC effectively manages contingencies by preemptively reducing demand, thus enhancing system robustness without excessive penalties.



(a) Cumulative reward of DA



(b) Cumulative reward of AA

Fig. 3.4 Cumulative reward of the proposed and benchmark algorithms.

Table 3.2 Training performance comparison of the different algorithms.

Algorithm	CT(s)	OT(ms)	LS(kWh)	UE(kWh)	DCV
DA-SAC	1807.29	152.59	27.7811	4.46	0.0636
DDPG-DQN	2048.94	152.59	42.7158	6.83	0.1053
PPO-PPO	1571.14	152.59	66.3634	8.67	0.1703
SAC-PPO	1728.25	152.59	42.2552	4.19	0.0712

CT: Offline computation time; OT: Online response time; LS: Load shedding; UE: Unserved electricity; DCV: Degree of constraint violation.

In contrast, DDPG-DQN, which has the highest offline computation time of 2048.94 seconds, performs moderately. It has a load shedding of 42.72 kWh, unserved electricity of 6.83 kWh, and a degree of constraint violation of 0.1053. This reflects a reasonable balance but less efficiency compared to DA-SAC. Despite having the fastest computation times with 1571.14 seconds offline, PPO-PPO exhibits significant load shedding of 66.36 kWh. This results in the highest unserved electricity of 8.67 kWh and a constraint violation of 0.1703, suggesting a less optimal approach to maintaining system stability. It should be noted that all DA's policies have identical structures; therefore, online operating times are the same for all algorithms. The SAC-PPO algorithm, leveraging SAC's entropy-based strategy, outperforms the other two benchmarks but still lags behind the proposed DA-SAC. This is attributed to the increased volatility introduced by the adversarial interaction between its

DA and AA components. In conclusion, DA-SAC emerges as the most effective algorithm for generating robust CCOPF solutions, striking a superior balance between computational efficiency and performance in the face of contingencies.

Table 3.3 Online performance comparison of DA policies based on contingencies generated by the SAC-based AA.

Algorithm	OC	LS(kWh)	UE*(kWh)	DCV*	TC
DA-SAC	241.23	27.78	4.46	0.0636	488.38
DDPG-DQN	221.86	42.42	11.26	0.1244	826.76
PPO-PPO	201.13	67.01	9.38	0.2115	811.48
SAC-PPO	208.91	43.20	6.24	0.10696	614.27

∗ Worst cases generated by the same AA learned by DSAC; OC: Operational cost; LS: Load shedding; UE: Unserved electricity; DCV: Degree of constraint violation; TC: Total cost.

To evaluate the learned agents further, an additional performance analysis was conducted to assess their behavior during online operation. A fair comparison was ensured by excluding the stochastic elements of the policies, such as Gaussian noise in PPO and SAC or Ornstein-Uhlenbeck process noise in DDPG. All learned DA policies were tested under the same SAC-based AA, which exhibited the best performance among all AA policies. Table 3.3 presents the operational costs and loadshedding values for the four algorithms, where the worst-case contingencies were generated using the SAC-based AA. Metrics such as unserved electricity and degree of constraint violation were recorded. The results demonstrate that the proposed DASAC algorithm achieves the lowest total operational costs and constraint violations, outperforming the benchmark algorithms. Combining these results with those in Table 3.2 (training performance) confirms that the proposed DA-SAC excels in both training and online operations, offering a robust and cost effective solution for CCOPF.

3.4.3 Training performance in N-*k* outage contingencies

Table 3.4 presents the numerical results for the abovementioned algorithms under N-2 and N-3 criteria, including operational cost, pre-contingency stage load shedding, unserved electricity, and DCV. This table shows that the proposed DASAC algorithm outperforms the other alternative algorithms in terms of CCOPF solution quality. It minimizes total cost, pre-contingency shedding, and post-contingency unserved power, thanks to the combined coordination of continuous and discrete SAC. This approach yields an optimal policy by leveraging auxiliary information that enables DA to generate robust actions against the worst contingency scenarios.

Table 3.4 Training performance comparison of the proposed and benchmark algorithms in N-*k* situation.

Case	Algorithm	CT(s)	OT(ms)	LS	UE	DCV	TC
N-2	DA-SAC	1961.61	152.59	209.60	14.30	0.0640	589.70
	DDPG-DQN	2341.72	152.59	188.95	18.51	0.1062	796.85
	PPO-PPO	1620.93	152.59	206.11	16.22	0.1709	944.71
N-3	DA-SAC	2062.01	152.59	198.91	17.85	0.0629	644.31
	DDPG-DQN	2342.55	152.59	176.81	27.48	0.1251	942.36
	PPO-PPO	1813.16	152.59	184.28	25.05	0.2534	1112.58

OC: Operational cost (k$); LS: load shedding (kWh); UE: Unserved electricity (kWh); DCV: Degree of constraint violation; CT: Offline computation time; OT: Online response time; TC: Total cost (k$).

The proposed algorithm exhibits the highest CCOPF solution quality, with an average improvement of 82.42% and 234.95% in constraint violations compared to the DDPG and PPO algorithms, respectively. In contrast, DDPG produces the worst CCOPF solutions, with the highest unserved electricity under N-2 and N-3 situations, due to competition between the DDPG and DQN algorithms, which fail to find an optimal CCOPF solution. The PPO algorithm, on the other hand, requires more pre-contingency stage load shedding to reduce unserved electricity under N-2 and N-3 situations compared to the other two algorithms.

Furthermore, the PPO-PPO algorithm excels in computational efficiency with the smallest offline computation time of 1620.93s in the N-2 situation. However, despite these computational advantages, PPO-PPO suffers from significantly higher load shedding of 81.10 kWh and unserved electricity of 16.22 kWh, leading to the highest total cost of 944.71. While PPO-PPO can quickly compute and respond, its inability to manage robustness effectively increases penalties from unmet demand and operational constraint violations. On the other hand, although not the fastest in computation, with 1961.61s in the N-2 scenario, the DA-SAC algorithm strikes a better balance between computational time and system robustness. DA-SAC achieves a lower load shedding of 31.76 kWh and unserved electricity of 14.30 kWh, resulting in a total cost of 589.70. In conclusion, the proposed DA-SAC's nested structure, which allows for more comprehensive planning and contingency handling, offsets its slower learning times by minimizing the penalties associated with load shedding and unserved electricity.

3.4.4 Effectiveness of auxiliary information

A numerical comparison between the proposed method and the model is conducted without considering auxiliary information (woAI) in the training process. The MDP for both methods is depicted in Fig. 3.5, whereas the cumulative reward of DA and AA is presented in Fig. 3.6. In the convergence curves 6, the proposed method for the DA shows higher cumulative rewards, indicating better learning efficiency and more robust performance. Meanwhile, the AA achieves lower cumulative rewards, reflecting effective mitigation of adverse actions. Conversely, the woAI method exhibits more fluctuations and lower cumulative rewards for the defender agent and higher cumulative rewards for the attacker agent, signaling less effective learning and increased vulnerability. This is mainly caused by the two learning rate scales caused by the different strategies, as shown in Fig. 3.5, leading to faster exploration in the AA and significant fluctuations between the DA and AA reward curves. In contrast, the proposed method leverages the auxiliary information, allowing DA to compete with AA and stabilize its solution at the 6000th gradient step. Although both two methods achieve the worst scenario nearly at 2500th gradient step, the proposed method leverages the $Q_\pi(s_t, a_t)$from AA as auxiliary information to improve the DA reward after the 2500th gradient step. The high reward of the DA and the lower reward of the AA indicate that the proposed method learns an optimal policy to compete with AA and optimize the cumulative reward under the worst contingency scenario. This disparity is further substantiated by the data in Table 3.5, where proposed method results in significantly lower load shedding (27.7811 kWh vs. 33.0019 kWh), reduced unserved electric load (6.46 kWh vs. 7.65 kWh), and a lower damage cost value (0.0636 vs. 0.1283), illustrating its superior performance in maintaining power system stability and reducing operational risks compared to woAI method.

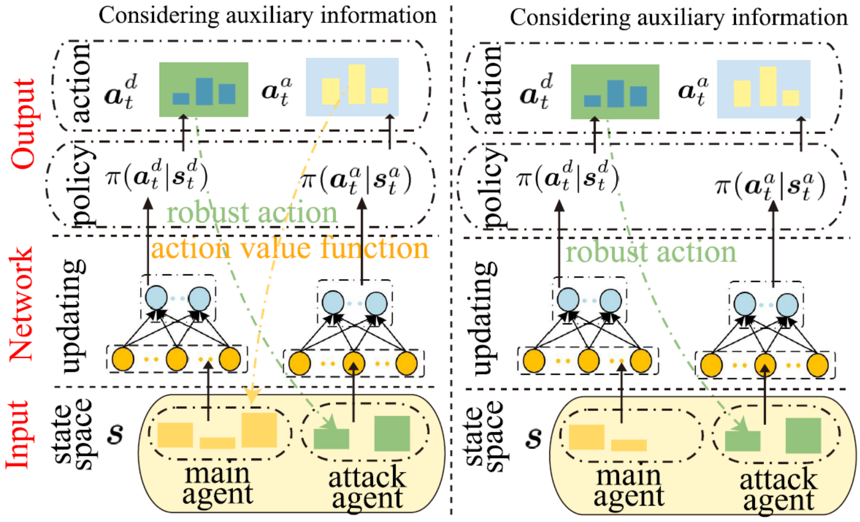
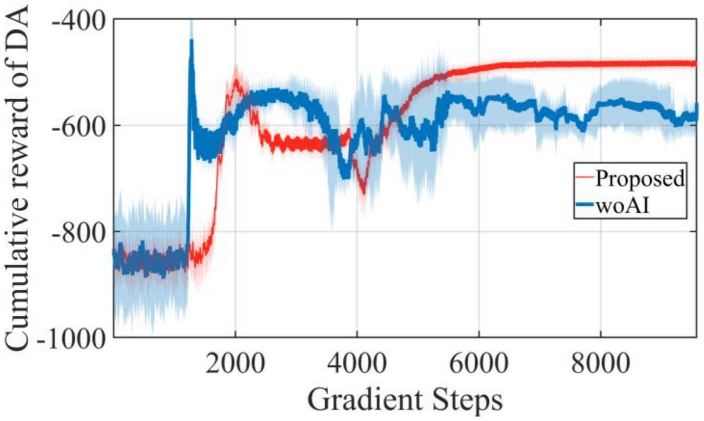
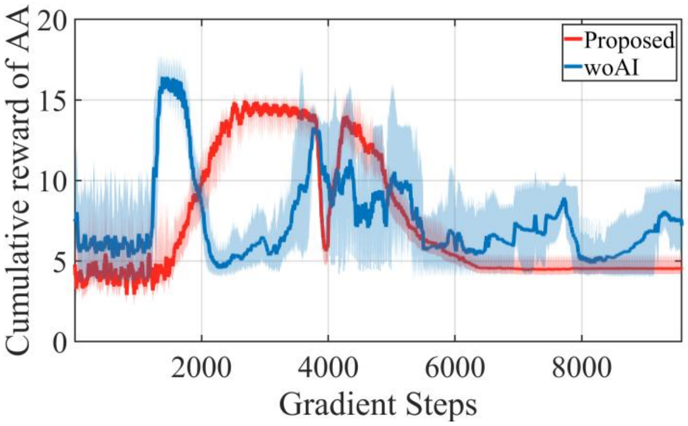


Fig. 3.5 Different settings in MDP with or without considering auxiliary information.



(a) Cumulative reward of DA



(b) Cumulative reward of AA

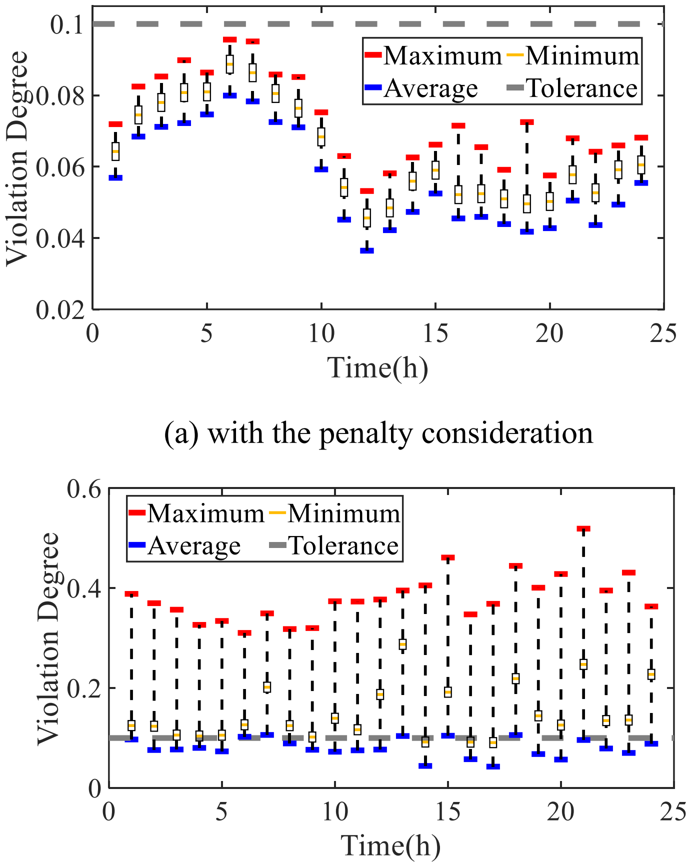
Fig. 3.6 Convergence comparison with and without auxiliary information.

Table 3.5 Performance comparison of the different strategies.

	Load shedding	Unserved electricity	DCV
DA-SAC	27.7811 kWh	6.46 kWh	0.0636
woAI	33.0019 kWh	7.65 kWh	0.1283

3.4.5 Performance evaluation

In terms of the solution quality of CCOPF, Figs. 3.7 depict the hourly distribution of DCVs. Fig. 3.7(a) illustrates the DCV when the DCV penalty is considered in the reward function (scenario 1), while Fig. 3.7(b) shows the DCV without the DCV penalty (scenario 2). As shown in Fig. 3.7(a), the DCVs are restricted within the upper tolerance value of 0.1 when the DCV penalty is included in the CCOPF operation. Introducing the DCV penalty incentivizes the DA to avoid violations, minimize total operational costs, and ensure the minimum degree of constraint violations.



(a) with the penalty consideration

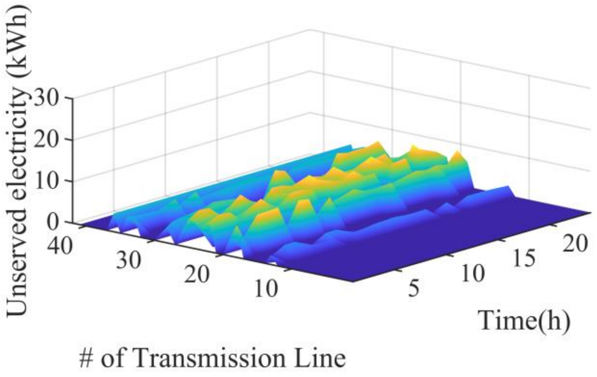
(b) without the penalty consideration

Fig. 3.7 Distribution of DCV with or without the penalty consideration.

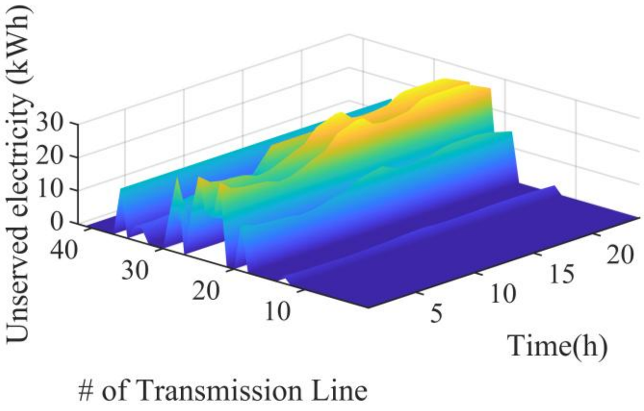
On the other hand, while most of the minimum DCVs in scenario 2 are below the upper tolerance value, the average DCV exceeds the tolerance limit, especially during peak electricity demand hours, leading to significant constraint violations. Consequently, the DCV penalty plays a critical role in maintaining the stability of the power system and improving the solution quality of CCOPF under contingency scenarios.

3.4.6 Robustness analysis

To evaluate the robustness of the proposed CCOPF solution using the DA-SAC algorithm, the unserved electricity of the CCOPF and OPF solutions when different transmission lines are tripped hourly is examined. The corresponding results are presented in Figs. 3.8. As shown, the CCOPF solution demonstrates superior performance in reducing unserved electricity compared to the OPF solution, with a maximum of less than 10kWh of unserved electricity in the CCOPF solution, as opposed to more than 30kWh in the OPF solution.



(b) Unserved electricity of the proposed CCOPF



(b) Unserved electricity of the OPF

Fig. 3.8 Unserved electricity(UE) of the CCOPF or OPF after contingency.

By training the nested agents to compete with each other, the DA-SAC algorithm generates robust actions for the CCOPF solution against the worst contingency scenarios. Consequently, when transmission lines are out of service, the CCOPF solution effectively mitigates the unserved electricity caused by such incidents. In contrast, the OPF solution struggles to mitigate unserved electricity and maintain solution quality when different transmission lines are tripped, making it vulnerable to contingencies without a defensive strategy. Therefore, the DA-SAC algorithm effectively generates robust CCOPF solutions under contingency scenarios.

3.4.6 Computational performance test

To verify the computational performance of the proposed DASAC algorithm, the interior point optimizer (IPOPT) is utilized as a baseline solver for the CCOPF problem examined for the IEEE 30-Bus and 118-Bus systems under N-1 criteria. The task can be formulated as a single-stage optimization problem with an additional N set of constraints. We considered 100 random profiles of power demands and recorded the average results per time step. Table 3.6 summarizes the simulation results of the IPOPT and proposed DA-SAC algorithms in terms of average operation cost, DCV, and computation time. Notably, the proposed algorithm achieves comparable average operation costs to IPOPT, with improvements of approximately 0.7% and 0.96% in the IEEE 30-Bus and IEEE 118-Bus systems, respectively. These values are reliable from an economic perspective. Note that the RMS value of DCV is below 0.1% of the range of the physical quantities of voltages at demand buses and line flows under all cases of contingencies. Although the proposed DRL method results in near-zero violation degrees, this level of violation is acceptable, especially under the worst contingency scenario, and our future work is to generate zero-DCV decisions. The average computation times of these two methods differ significantly, with IPOPT taking 8.624s and 15.087s due to the large number of constraints and variables. The proposed DRL algorithm requires only 0.153s and 0.841s for the IEEE 30-Bus and IEEE 118-Bus systems, respectively. Due to the simple mathematical operation in predicting actions through the learned policy, the proposed DRL method is much faster than the IPOPT method, with nearly 98.22% and 94.4% time savings, respectively, about 56x and 18x speedup.

Table 3.6 Computational performance.

Method	Test system	Operation cost	DCV	Computation time
IPOPT	IEEE 30	216.33	0	8.624s
Proposed DRL	IEEE 30	217.94	0.0636%	0.153s
IPOPT	IEEE 118	4552.27	0	15.087s
Proposed DRL	IEEE 118	4596.64	0.0736%	0.841s

3.5 Summary

This paper proposes a novel robust deep reinforcement learning algorithm, DA-SAC, for solving the CCOPF problem and generating a robust solution against the worst contingency scenario while satisfying system constraints. The proposed optimization process has several unique features: (i) the design of two competitive agents, a DA, and an AA, as nested agents to obtain robust actions through iterative interaction with the environment; (ii) the inclusion of DCV penalties in power system operations to ensure the feasibility of the CCOPF solution; and (iii) the enhancement of stability and robustness in the noncooperative learning strategy to find optimal CCOPF solutions. Specifically, the DA and AA utilize continuous and discrete SAC algorithms to generate robust decisions and attack actions.

To evaluate the proposed algorithm, numerical simulations are conducted on IEEE 30-bus and 118-bus systems. The results demonstrate that the proposed algorithm generates a robust solution for the CCOPF problem under the worst contingency scenario, achieving significant time savings compared to other state-of-the-art optimization approaches and learning techniques.

Although the proposed DRL-based frameworks demonstrate strong empirical performance in enhancing system resilience, the interpretability of learned policies remains an important consideration for practical deployment. DRL decisions are often viewed as opaque, which may reduce operator confidence in automated control. One promising direction is to conduct sensitivity analysis, for example by examining how agent control actions shift under variations in distributed energy resource (DER) outputs or load conditions. Such analysis would help clarify the behavioral patterns of the learned agents and provide operators with a better understanding of the underlying decision-making process. In addition, explainable AI techniques could be integrated with DRL to further enhance transparency and trust. While a detailed sensitivity study is beyond the scope of this work, this discussion highlights interpretability as a promising avenue for future research.

Chapter 4 Robust preventive and corrective security-constrained OPF for worst contingencies with the adoption of VPP: A safe reinforcement learning approach

The rising frequency of extreme weather events calls for urgent measures to improve the resilience and reliability of power systems. This paper, therefore, presents a robust preventive-corrective security-constrained optimal power flow (PCSCOPF) model designed to strengthen power system reliability during N-*k* outages. The model integrates fast-response virtual power plants (VPPs), dynamically adjusting their injections to mitigate post-contingency overloads and maintain branch flows within emergency limits. Additionally, a novel approach combining deep reinforcement learning (DRL) with Lagrangian relaxation is introduced to efficiently solve the PCSCOPF decision-making problem. By framing the problem as a constrained Markov decision process (CMDP), the proposed Lagrangian-based soft actor-critic (L-SAC) algorithm optimizes control actions while ensuring constraint satisfaction during the exploration process. Extensive investigations have been conducted on the IEEE 30-bus and 118-bus systems to evaluate their computational efficiency and reliability.

4.1 Framework

This study aims to tackle the issue of robust PCSCOPF by incorporating the ACPF constraints and dividing it into PSCOPF and CSCOPF in pre- and post-contingency stages, respectively. The PSCOPF is responsible for enhancing the robustness of the power system during the pre-contingency stage; hence, load shedding is implemented to alleviate constraint violations. The solution to the PSCOPF involves finding a balance between the amount of load shedding prior to the contingency and the degree of constraint violation following the contingency. A min-max optimization problem is specifically formulated to determine the minimum quantity of load shedding required in the pre-contingency stage, considering the worst-case contingency scenario. In the CSCOPF problem, power flow may exceed short-term emergency ratings [111], potentially leading to cascading line outages in the post-contingency stage. However, due to the limitation of ramping rate constraints and large inertia, it is hard for conventional generators to respond immediately to contingencies. Therefore, the introduction of the flexible VPP with the capacity to rapidly dispatch generation and absorb overflow from the system can quickly restore the power flow back to the long-term emergency rating. To gain a better understanding of the implementation of PCSCOPF with VPP involvement in dispatching, a timeline-based illustration is depicted in Fig. 4.1.

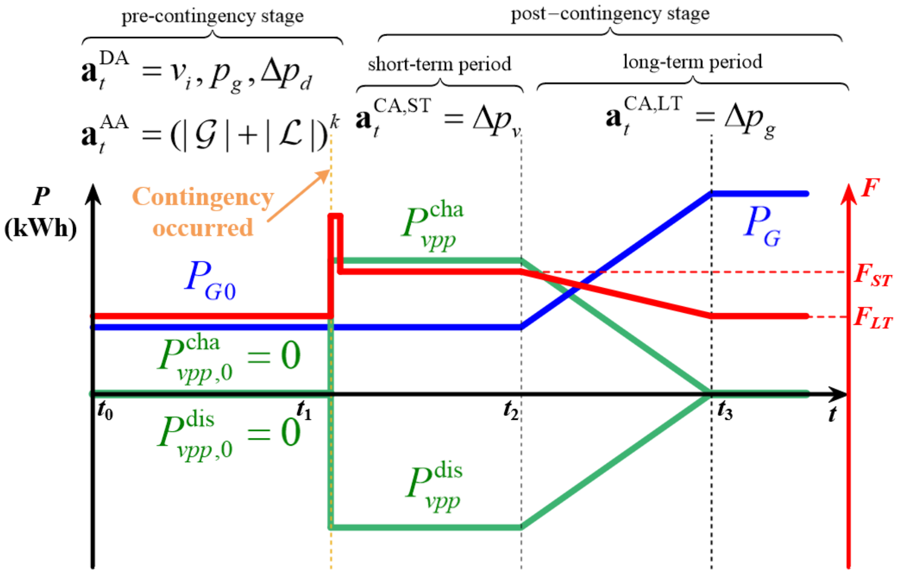

pre-contingency stage
post−contingency stage
short-term period
long-term period
$\mathbf{a}_t^{\mathrm{DA}} = v_i, p_g, \Delta p_d$
$\mathbf{a}_t^{\mathrm{AA}} = (|\mathcal{G}|+|\mathcal{L}|)^k$
$\mathbf{a}_t^{\mathrm{CA,ST}} = \Delta p_v$
$\mathbf{a}_t^{\mathrm{CA,LT}} = \Delta p_g$
P (kWh)
Contingency occurred
P_{G0}
P_{vpp}^{cha}
P_G
F
F_{ST}
F_{LT}
$P_{vpp,0}^{\mathrm{cha}} = 0$
t_0
t_1
t_2
t_3
t
$P_{vpp,0}^{\mathrm{dis}} = 0$
P_{vpp}^{dis}

Fig. 4.1 Timeline-based illustration of the PCSCOPF implementation.

The process can be effectively delineated into two distinct stages based on the time axis. The first stage, known as the pre-contingency stage, is resolved through the utilization of PSCOPF. The load shedding and stochastic contingencies are considered in this stage to identify a robust action against the worst contingency scenario. Subsequently, the second stage, referred to as the post-contingency stage, is comprised of two distinct periods: a rapid short-term emergency period and a gradual long-term emergency period. As indicated in the figure, the stochastic contingencies lead to transmission lines exceeding their short-term emergency rating (FST). However, during the fast short-term period, conventional generators encounter limitations in their ability to promptly respond owing to the constraints imposed by their ramping rates and substantial inertia. Thus, VPPs swiftly dispatch their active and reactive power output, enabling them to discharge or charge power and effectively bring the branch flow back down with short-term emergency violations. Throughout the long-term period, VPPs consistently decrease their power output until it reaches zero, while conventional generators commence the process of redistributing the power flow within the confines of long-term emergency limits (FLT). By integrating the two aforementioned stages, a comprehensive framework known as PCSCOPF is established. This framework ensures the uninterrupted functionality of the power system when confronted with various contingency scenarios. For clarity and readability, the PSCSOPF optimization problem is divided into PSCOPF and CSCOPF and defined separately as follows.

4.2 Problem Formulation

4.2.1 Problem Formulation of PSCOPF

The PSCOPF optimization objective function can be formulated as follows:

$$\min_{\Omega} \sum_{\forall t} \left[\sum_{\forall g \in \mathcal{G}} C_g p_{g,t} + \sum_{\forall d \in \mathcal{D}} C_d \Delta p_{d,t} + \max_{w \in \mathcal{W}} \left[\sum_{\forall d \in \mathcal{D}} C_d^o \Delta p_{d,t}^o \right] \right] \tag{4.1}$$

where Ω Feasible region for primary variables. C_g, C_d and C_d^o are operation cost of generator, load shedding cost, and unserved electricity cost, respectively. $p_{g,t}$ is active power from generator g. Δp_d, Δp_d^o are Load-shedding and unserved electricity of bus d. $\forall g \in \mathcal{G}$, $\forall d \in \mathcal{D}$, and $w \in \mathcal{W}$ are generators set, power demand buses, and uncertainty set.

The first two terms correspond to the operational costs and load shedding penalty during the pre-contingency stage; the last term pertains to the penalties imposed for unserved electricity following the occurrence of contingencies. The superscript symbol (o) designates variables after the contingency event. In this work, power demands are generated randomly in accordance with their stochastic profiles. Consequently, power demands may be heavy, potentially rendering the problem infeasible under normal operating conditions. Hence, load shedding is incorporated in the first stage to relax constraints and enhance the stability of the learning process. Furthermore, any additional load shedding that may occur as a result of a contingency event, also known as unserved electricity, is penalized in the second stage with a penalty of c_d^o.

The operational constraints for PSCOPF during the pre-contingency stage are explicitly outlined in equations (4.2)-(4.10). Equation (4.2) denotes the set of equality constraints that pertain to the active and reactive power balance equations. Equation (4.3)-(4.9) represents the inequality set of constraints encompassing generation capacities, voltage security constraints, and power flow boundaries. Equation (4.10) defines the operational constraint governing the behavior of the attacker, which is devised to maximize the constraint violation and the magnitude of unserved electricity during the post-contingency stage. Different forms of sets are presented in the previous studies to include different types of electric power system components, such as generation units, transformers, power lines, and reactive power injections, or to model extreme storm behavior with additional time and geographical constraints [112]. In this work, the behavior of the attacker only considers the availability of the transmission lines and generation units.

$$[p_{g,t}, s_{ij,t}] = g(v_{i,t}, \theta_{i,t}, p_t^{net,L}, p_t^{PV,net}, p_t^{BESS,net}, P_{d,t} - \Delta p_{d,t}) \tag{4.2}$$

$$\underline{P}_g \le p_{g,t} \le \overline{P}_g, \forall g,t \tag{4.3}$$

$$\underline{Q}_g \le q_{g,t} \le \overline{Q}_g, \forall g,t \tag{4.4}$$

$$-RD_g \le p_{g,t} - p_{g,t-1} \le RU_g, \forall g,t \tag{4.5}$$

$$\underline{V}_i \le v_{i,t} \le \overline{V}_i, \forall i,t \tag{4.6}$$

$$\underline{\Theta}_i \le \theta_{i,t} \le \overline{\Theta}_i, \forall i,t \tag{4.7}$$

$$\underline{S}_{ij} \le s_{ij,t} \le \overline{S}_{ij}, \forall ij, ji, t \tag{4.8}$$

$$0 \le \Delta p_{d,t} \le P_{d,t}, \forall d,t \tag{4.9}$$

$$\mathcal{W} = \{\mathbf{w} \in \{0,1\} \mid \sum_{\forall ij} w_{ij,t} + \sum_{\forall g} w_{g,t} \le k, l_{ij,t} = 1 - w_{ij,t}, l_{g,t} = 1 - w_{g,t}, \forall ij, g, t\} \tag{4.10}$$

where s_{ij} is receiving power flow between buses i and j. v_i , θ_i are voltage magnitude/angle at bus i. $p^{\text{net,L}}$, $p^{\text{PV,met}}$, and $p^{\text{BESS,net}}$ are the consumption from the network of controllable load in the VPP, PV and BESS generation injected into the network. $\underline{P_g}$, $\overline{P_g}$ are min/max active power limit of generator. $\underline{Q_g}$, $\overline{Q_g}$ are min/max reactive power limit of generator. RD_g, RU_g are ramping up/down limit of generator. $\underline{\mathrm{V_i}}$, $\overline{\mathrm{V_i}}$ are min/max voltage limit of bus. $\underline{\Theta_i}$, $\overline{\Theta_i}$ are min/max angle phase limit of bus. $\underline{S_{ij}}$, $\overline{S_{ij}}$ are min/max power flow limit of line. w_{ij}, w_g are attacker status of transmission line *ij*/ generator *g*, 1 if it is attacked and 0 otherwise. l_{ij}, l_g are availability of the transmission line *ij*/generator *g*.

4.2.2 Problem Formulation of CSCOPF

The CSCOPF optimization problem considering the VPP fast-response control action can be defined as follows:

$$\min_{\Omega} \sum_{\forall t} \{ \sum_{\forall v \in vpp} C_{vpp} \Delta p_v + \sum_{\forall g \in \mathcal{G}} C_g \Delta p_{g,t} \} \tag{4.11}$$

$$[p^s_{g,t}, s^s_{ij,t}] = g^s(v^s_{i,t}, \theta^s_{i,t}, p_t^{PV,net,s}, p_t^{BESS,net,s}, p_t^{net,L,s}, P_{d,t} - \Delta p_{d,t} - \Delta p^o_{d,t}) \tag{4.12}$$

$$h^s\left(p^s_{g,t}, q^s_{g,t}, s^s_{ij,t}, v^s_{i,t}\right) \le h^{\max} \tag{4.13}$$

$$\left|p^s_v - p_v\right| \le \Delta P_v^{Max} \tag{4.14}$$

$$[p^l_{g,t}, s^l_{ij,t}] = g^l(v^l_{i,t}, \theta^l_{i,t}, p_t^{PV,net,l}, p_t^{BESS,net,l}, p_t^{net,L,l}, P_{d,t} - \Delta p_{d,t} - \Delta p^o_{d,t}) \tag{4.15}$$

$$h^l\left(p^l_{g,t}, q^l_{g,t}, s^l_{ij,t}, v^l_{i,t}\right) \le h^{\max} \tag{4.16}$$

where C_{vpp} is adjustment cost of VPP. $\forall v \in vpp$ is VPP set. Superscript primes *s*, *l*, *o* indicate the short-term, long-term emergency period, and occurrence of the contingencies.

In (4.11), the first term of the objective is to minimize VPP adjustments during the short-term period, and the second term is to minimize the adjustment of generators during the long-term period. Equations (4.12) and (4.15) are the equality set of constraints, incorporating the active and reactive power balance equations during short-term and long-term periods, respectively. Equations (4.13) and (4.16) are the inequality set of constraints, encompassing generation capacities, voltage security constraints, and power flow limits during short-term and long-term periods, respectively. Equation (4.14) aims at avoiding unrealistic variations of VPPs, ensuring that their responses remain within reasonable bounds.

4.2.3 Fast Response Model of VPP

The VPP is introduced to recover quickly and ensure the continuous operation of the power system in the presence of contingencies. Fig. 4.2 shows the structure of the VPP profile, which illustrates all internal energy flows between the VPP elements and the power system.

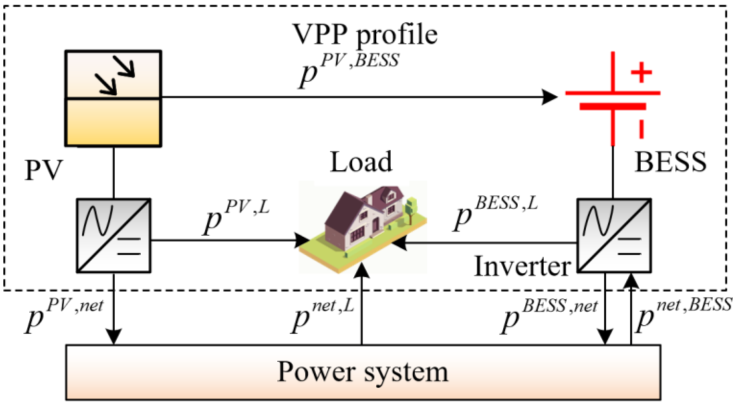


Fig. 4.2 The structure of the VPP profile.

Constraint (4.17) ensures that the total PV production is equal to the summation of the directly injected into the power system, the power consumption by the load, and the power consumption by the BESS during the charging phase. Constraint (4.18) restricts the hourly production of the PV production during each hour *t*. Constraint (4.19) ensures that the total load consumption is equal to the summation of the load that is directly absorbed from the power system, the load fed by the PV, and the load fed by the BESS. Constraint (4.20) ensures that the total BESS production is equal to the summation of the directly injected into the power system and the power consumption by the load. Similarly, constraint (4.21) ensures that the total BESS load that is consumed during the charging phase is equal to the summation of the load directly absorbed from the power system and the load fed by the PV. Constraints (4.22) restrict the energy generation and consumption of the BESS to their discharging and charging limits, respectively, while constraint (4.23) requires that the BESS may not operate in discharging and charging mode simultaneously in a given hour. Constraint (4.24) represents the hourly BESS energy balance.

$$p_t^{PV} = p_t^{PV,net} + p_t^{PV,BESS} + p_t^{PV,L} \tag{4.17}$$

$$p_t^{PV} \le \overline{P^{PV}} \tag{4.18}$$

$$p_t^{L} = p_t^{net,L} + p_t^{BESS,L} + p_t^{PV,L} \tag{4.19}$$

$$p_t^{BESS,p} \cdot \chi^{p} = p_t^{BESS,net} + p_t^{BESS,L} \tag{4.20}$$

$$p_t^{BESS,c} \cdot \chi^{c} = p_t^{net,BESS} + p_t^{PV,BESS} \tag{4.21}$$

$$p_t^{BESS} \le \overline{P^{BESS}} \cdot \eta^{BESS} \tag{4.22}$$

$$\chi^{p} + \chi^{c} \le 1 \tag{4.23}$$

$$\mathrm{SOC}_{i,t}^{\mathrm{BESS}} = \begin{cases} \mathrm{SOC}_{i,t-1}^{\mathrm{BESS}} + \eta^{\mathrm{BESS}} p_t^{BESS,c} \Delta t \\ \mathrm{SOC}_{i,t-1}^{\mathrm{BESS}} + p_t^{BESS,p} \Delta t / \eta^{BESS} \end{cases} \tag{4.24}$$

where $p^{\mathrm{net,L}}$, $p^{\mathrm{PV,L}}$, $p^{\mathrm{BESS,L}}$ are the consumption from the network/ PV/ BESS of controllable load in the VPP. $p^{\mathrm{BESS,p}}$, $p^{\mathrm{BESS,c}}$ are the power production/consumption of the BESS. χ^{p}, χ^{c} are discharge/charge status of the BESS. p^{BESS} is operation power of the BESS. $p^{\mathrm{net,L}}$, $p^{\mathrm{PV,L}}$, $p^{\mathrm{BESS,L}}$ are the consumption from the network/ PV/ BESS of controllable load in the VPP.

4.2.4 Comprehensive Model of the PCSCOPF with VPP

The presented CSCOPF problem is integrated with the PSCOPF problem, resulting in the formulation of a min-max optimization framework known as PCSCOPF. This comprehensive formulation aims to determine the optimal robust control actions by considering worst-case scenarios based on N-*k* security criteria. Mathematically, the PCSCOPF formulation can be expressed as follows:

$$\min_{\Omega} \sum_{\forall t} \left[\sum_{\forall g \in \mathcal{G}} C_g p_{g,t} + \sum_{\forall d \in \mathcal{D}} C_d \Delta p_{d,t} + \max_{u \in \mathcal{U}} \left[\sum_{\forall d \in \mathcal{D}} C_d^o \Delta p_{d,t}^o \right] \right] + \min_{\Omega^{s,l}} \left[\sum_{\forall v \in vpp} C_{vpp} \Delta p_v^s + \sum_{\forall g \in \mathcal{G}} C_g \Delta p_{g,t} \right] \tag{4.25}$$

$$\text{s.t.} \qquad (4.2)\text{-}(4.10), (4.12)\text{-}(4.16), (4.19)\text{-}(4.24) \tag{4.26}$$

To surmount this optimization problem, cutting-edge DRL technology is adopted to effectively address it in a time-efficient manner, facilitating real-time operation with a heightened level of robustness. In the following section, a CMDP model is introduced, which encapsulates the problems above into two intelligent agents.

4.2.5 Contingency Filtering Approach

To enhance scalability and reduce the computation time of the PCSCOPF solution, a contingency filter [47] is employed to filter non-dominated contingencies. In this work, the contingency filter leverages the constraint violations observed after simulating all contingencies using a Newton-Raphson power flow program. The contingency filter selects a critical contingency set, where the scenarios in this set exhibit greater violations compared to others. Two constraint limits—branch flow and voltage—are considered critical violation parameters. Consequently, the Pareto set [113] is used to determine the contingency set, which is defined as:

$$PS = \left\{ w' \in \mathcal{W} \mid F(w) \prec F(w') \right\} \tag{4.27}$$

The corresponding Pareto front is defined as:

$$PF = \left\{ F(w') = \left[f_1(w'), f_2(w') \right] \mid w' \in PS \right\} \tag{4.28}$$

where $F(w) \prec F(w')$ implies that any contingency scenario $\mathrm{w} \in \mathrm{W}$ is dominated by contingency scenario w', mathematically, $f_1(w) \leq f_1(w'), w \in \mathcal{W}, w' \in PS$, and $f_2(w) \leq f_2(w'), w \in \mathcal{W}, w' \in PS$; $f_1(w')$ and $f_2(w')$ represent the magnitude of branch flow and voltage violations under contingency scenario w', respectively. A contingency scenario belonging to the Pareto front is able to form a critical contingency set. This ensures that the contingency scenarios w' on the Pareto front result in greater violations than other scenarios. By eliminating redundant contingencies, the Pareto set-based contingency filtering approach accelerates the PCSCOPF computation and guarantees the scalability of the proposed model.

4.3 Methodology

In accordance with the established framework, the PCSCOPF problem is sequentially addressed by solving the PSCOPF problem, followed by the CSCOPF problem. The PSCOPF problem adopts a min-max optimization formulation where the power system operator engages in competition with an attacker, aiming to identify a robust solution against contingency scenarios. Therefore, the PSCOPF problem can be standardized as an MDP with two adversarial agents, which is an important way to build the RL framework. On the other hand, the CSCOPF problem focuses on generating a resilient solution to restore power system operation in the presence of contingencies, which can be effectively formulated as an MDP. Notably, unlike previous study [114], this work addresses the issue of constraint violation in the MDP framework and formulates the process as a CMDP. The proposed CMDP framework is employed for decision-making to maximize rewards within the constraint-satisfying regime. This approach offers enhanced clarity and readability by separately defining the CMDP within the PSCOPF and CSCOPF domains, which will be elaborated upon in the subsequent subsection.

4.3.1 CMDP characteristics in preventive agent

In the robust PSCOPF problem, two competitive agents, a defense agent (DA) and an attack agent (AA), are designed in CMDP as preventive agents (PA). Simultaneously, a corrective agent (CA) is designed to make sequential decisions during the post-contingency stage while minimizing the cumulative reward. The elements of those agents are summarized in Table 4.1.

Table 4.1 Elements of Constraint Markov Decision Process in Different Agents.

Elements of CMDP	Pre-contingency stage		Post-contingency stage
	Preventive agent		Corrective agent
	Defense agent	Attack agent	
State	s^d	s^a	s^c
Action	a^d	a^a	a^c
Reward	r^d	r^a	r^c
Constraint cost function	C^d	-	C^c

Specifically, the state space of DA can be defined as the active and reactive load of each bus, which is defined in (4.29). The actions of the DA are represented by the active outputs and voltage magnitudes of the non-slack buses, and load shedding in demand buses, which can be formulated in (4.30). Noting that, instead of considering all decision variables of (4.29), the selected actions in (4.30) are controllable and include fewer actions to improve learning convergence and stability. This paper adopts the recent progress in AC power flow solvers [52] to extract the full decision vector from this action space. The reward function of DA is to evaluate the action performance to facilitate the update process of the policy network, which is defined in (4.31).

$$s^d = \left(P_d, Q_d\right), \forall d \tag{4.29}$$

$$a^d = \left(v_g, p_g, \Delta p_d\right), \forall g \in U(pv), d \tag{4.30}$$

$$r_t^d = -\left(\sum C_g p_{g,t} + \sum C_d \Delta p_{d,t} + \sum C_d^o \Delta p_{d,t}^o\right), \forall g, d \tag{4.31}$$

where *U*(*pv*) denote the set of PV buses in the generator set. To enhance the robustness of the PSCOPF solution by identifying the worst-case contingency scenario, it is essential to incorporate the action of the DA a^d into the state of the AA. Consequently, the state space of the AA is mathematically expressed in (4.32). Meanwhile, the AA is responsible for attacking transmission lines and generation units. As a result, the action space of the AA is discrete and can be precisely defined in (4.33). On the other hand, the reward function for the AA aims to maximize the amount of unserved electricity. This can be formulated in (4.34).

$$s^a = \left(P_d, Q_d, a^d\right), d \in DP \tag{4.32}$$

$$a^a = \left(w_{ij}, w_g\right), \forall ij, g \tag{4.33}$$

$$r_t^a = \sum C_d^o \Delta p_{d,t}^o, \forall d \tag{4.34}$$

In this formulation, the state of the corrective agent is defined in (4.35), which consists of active, reactive power demands and action of the PA to describe the state of the post-contingency environment. To quickly determine the corrective action to recover the power system operation, corrective action a^c is defined to determine the adjustment of the output of the VPPs and the generator, as defined in (4.36). The reward function r^c assesses the action value taken by the CA, which is defined in (4.37).

$$s^c = \left(P_d, Q_d, a^d, a^a\right), d \in DP \tag{4.35}$$

$$a^c = \left(\Delta p^{BESS}, \Delta p^{net,L}, \Delta p_g\right), \forall g \in U(pv), d \in DP \tag{4.36}$$

$$r_t^c = -\left(C_{vpp} \Delta p_t^{BESS} + C_g \Delta p_{g,t}\right), \forall g \tag{4.37}$$

The interactions between the PA and CA in the environment are depicted in Fig. 4.3. The PA generates preventive action to address the robust PSCOPF problem, while the CA predicts corrective actions to efficiently control the VPPs and adjust the generators, ensuring a swift recovery of power system operations. The environment provides feedback to each agent in the form of rewards (r^d, r^a, and r^c) and transitions to the next state, thereby shaping the learning process. Fig. 4.3 visually illustrates these interactions between the agents and the environment, highlighting their collaborative efforts in achieving the desired power system performance.

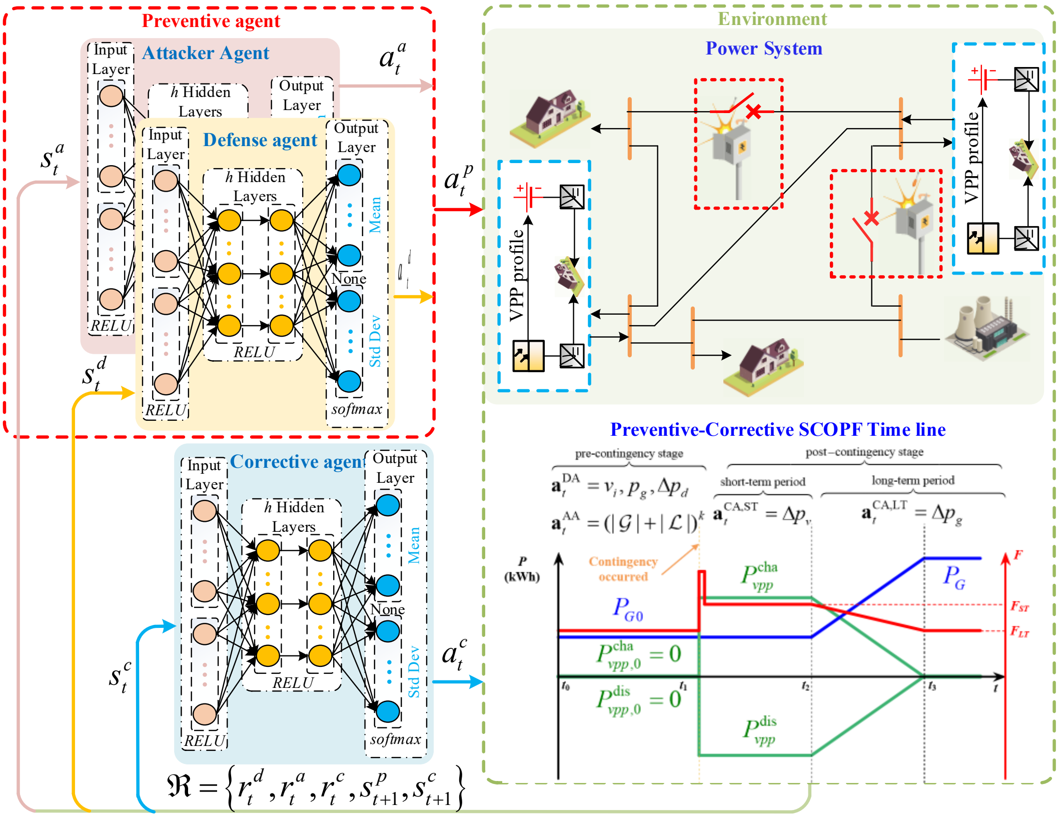


Fig. 4.3 The developed CMDP model for the PCSCOPF problem.

4.3.2 Constraint cost function

The CMDP incorporates a cost function that enforces constraints, ensuring that selected actions must satisfy prescribed conditions at each exploration step. Deviation from these constraints incurs a substantial penalty imposed by the constraint cost function, impacting the overall reward. This mechanism drives the CMDP framework to prioritize the exploration of action policies that adhere to the specified constraints, fostering the selection of actions that prioritize constraint satisfaction throughout the decision-making process. Within the CMDP framework, the constraint cost function, as presented in equation (4.38), quantifies the extent of constraint violation.

$$C = \sqrt{\frac{1}{|\mathcal{X}|}\sum_{\forall x_n} \zeta_n \left(\frac{[x_n - \overline{x}_n]^+ + [\underline{x}_n - x_n]^+}{\overline{x}_n - \underline{x}_n}\right)^2} \tag{4.38}$$

where x_n indicates all uncontrollable variations in PCSCOPF, such as branch flows and voltage magnitudes at demand buses, with minimum $\underline{x_n}$ and maximum $\overline{x_n}$ which are obtained from (4.2)-(4.6), (4.8)-(4.12). The total number of various is $|\mathcal{X}|$. $[\cdot]^+$ denotes max{0, .} function. In contrast to simply summing the constraint violations with their varying scales, the proposed constraint cost function normalizes these values before summation. This normalization step ensures that the constraint violations are treated on an equal scale, thereby facilitating a more accurate evaluation of the overall constraint violation degree within the CMDP framework.

4.3.3 Soft Actor-Critic Algorithm for PSCSOPF Problem

The DRL algorithm is responsible for determining the optimal control actions that maximize the expected cumulative reward. Typically, this is achieved within the actor-critic framework. However, on-policy approaches like asynchronous advantage actor-critic (A3C) and PPO algorithms often face challenges related to updating contradictions and efficiency. In contrast, off-policy algorithms such as DDPG have been introduced to enhance exploration capability. Nevertheless, DDPG suffers from issues, such as hyper-parameter sensitivity, which can hinder training performance. In this work, these limitations were addressed by adopting the off-policy algorithm known as SAC. SAC combines the actor-critic framework with entropy maximization to promote exploration and ensure learning stability. However, this paper highlights a critical issue prevalent in existing DRL algorithms, wherein the unrestricted ability of the agent to select control actions through trial and error can lead to violations of operational constraints [115]. Such violations can result in equipment failures and instability in the power system operation. Therefore, it is crucial to ensure zero-constraint violations during the RL training process, not only at convergence but also throughout the exploration and learning phases. To tackle this challenge, this section introduces the L-SAC algorithm, which effectively manages the CMDP while operating within a constraint-satisfying regime.

The SAC algorithm trains a stochastic policy to maximize not only the cumulative reward but also the entropy of the policy, and the policy function π can be expressed as follows:

$$\pi^* = \arg\max_{\pi} \sum_{t} E_{(s_t,a_t)\sim\tau_\pi}\left[\gamma^t (r(s_t,a_t) + \alpha H(\pi(\cdot|s_t)))\right] \tag{4.39}$$

where t is the time step; τ denotes a trajectory; r is a reward under state s_t and action a_t; α is an entropy temperature which regulates the stochastic degree of the policy; $H(\pi(\cdot|s_t))$ represents the entropy of the policy under state s_t. The prominent feature of the SAC algorithm is that the hyperparameter entropy temperature is learned by an automated entropy adjustment, which is presented in (4.40).

$$J(\alpha) = E_{a_t\sim\pi}\left[-\alpha\left(\log\pi\left(a_t|\ s_t\right) + \bar{H}\right)\right] \tag{4.40}$$

where $\overline{H}$ indicates the target entropy. Based on the maximum entropy framework, the soft iteration is employed to maximize the objective by alternating between policy estimation and amendment. Thus, the soft state value function can be defined as:

$$V(s_t) = E_{a_t\sim\pi}[Q(s_t,a_t) - \alpha\log(\pi(a_t|s_t))] \tag{4.41}$$

where the value function $Q(s_t, a_t)$ estimates the performance of the action a_t at state s_t. Since the action space of the PA is hybrid with continuous and discrete action, the soft state value function in discrete action can be transferred to:

$$V(s_t) = \pi(s_t)^T[Q(s_t) - \alpha\log(\pi(s_t))] \tag{4.42}$$

Consequently, the optimization task is transferred to identify the optimal policy based on the state-value function *V*, which is defined as follows:

$$\pi^* = \arg\max_{\pi} E_{s_t \sim D}[V_\pi(s_t)] \tag{4.43}$$

where *D* is the minibatch prior sample from the replay buffer. Soft policy iteration is to learn an optimal maximum entropy policy that alternates between policy evaluation and policy improvement in the maximum entropy framework. To satisfy large continuous domain requirements, instead of alternating between the soft policy evaluation and improvement, the approximator functions are introduced to derive a practical approximation to soft policy iteration in subsection 4.3.5.

4.3.4 Lagrangian-Based Soft Actor-Critic Algorithm

Automatically turning the Lagrange multipliers for each power constraint during the exploration process of the DRL algorithm is the key point in limiting the violation of constraints. In this way, constraints can be imposed on not only expected reward or cost but also their instantaneous values. In this subsection, the Lagrangian-based SAC algorithm is formulated in the CMDP. The objective of a CMDP is to select a feasible action to satisfy all of its necessary constraints within the feasibility budget. Mathematically, the discounted cumulative constraint within the feasibility budget is of the form:

$$J_c^\pi = E_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t C(s_t, a_t, s_{t+1})\right] \le \overline{J}_c \tag{4.44}$$

where $\overline{J_c}$ indicates the upper bound for constraint violation cost. Finally, the goal of a CMDP is recast as a constrained optimization problem as expressed in (4.45):

$$\begin{aligned} &\max_{\pi} J_r^\pi = E_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, s_{t+1})\right] \\ &s.t. \quad E_{\tau \sim \pi}\left[\sum_{t=0}^{T} \gamma^t C(s_t, a_t, s_{t+1})\right] \le \overline{J}_c \end{aligned} \tag{4.45}$$

where J_r^π is the expected discounted cumulative reward. Therefore, the SAC maximization task (4.39) in CMDP can be shifted to

$$\pi^* = \arg\max_{\pi} E_{s_t \sim D}[V_\pi(s_t)] \qquad s.t. E_{s_t \sim D}[V_\pi^c(s_t)] \le \overline{V_c} \tag{4.46}$$

where $V_\pi^c(s_t)$ is the state-value function of constraint violation C; $\overline{V_c} = (1-\lambda^T)/(1-\lambda)\overline{C_t}$ pertains to the limit for state value associated with the operation constraint; $\overline{C_t}$ is the maximum violation in each time step. This inequality-constrained problem can be solved by the Lagrangian relaxation approach, wherein the hard constraint is relaxed into a soft constraint. Specifically, the Lagrangian function for the CMDP problem can be written as:

$$\min_{\lambda \geq 0} \max_{\pi} E_{s_t \sim D}[V_\pi(s_t)] + \lambda(\overline{V_c} - E_{s_t \sim D}[V_\pi^c(s_t)]) \tag{4.47}$$

where λ is the Lagrange multiplier, and the Lagrangian function can be converted into

$$\mathcal{L}(\pi,\lambda) = E_{s_t \sim D}\left[V_\pi^*\left(s_t\right)\right] + \lambda\overline{V_c} \tag{4.48}$$

where $V_\pi^*\left(s_t\right) = E_{(s_t,a_t)\sim\tau_\pi}\left[\left(\gamma^t\left(r(s_t,a_t) - \lambda c_t\right) + \alpha H(\pi(\cdot|s_t))\right)\right]$. In the proposed Lagrangian-SAC framework, the partitioning of constraints into soft and hard is guided by both power system operational priorities and algorithmic considerations. From an operational perspective, hard constraints correspond to safety-critical limits whose violations may immediately compromise system security, such as transmission line ratings, bus voltage bounds, and generation capacity limits. These constraints must be strictly satisfied and are therefore explicitly modeled in the CMDP formulation through Lagrangian multipliers. In contrast, soft constraints are associated with objectives of economic efficiency or service quality, such as minimizing active power losses or reducing load shedding costs. Since occasional deviations in these objectives are tolerable, they are incorporated directly into the cost function, where they can be traded off against operational costs during optimization. This partitioning strategy ensures that the algorithm focuses on strictly preserving system security while maintaining flexibility in optimizing system performance.

Note that as λ increases, the solution of (4.45) converges to that of (4.44). However, a larger λ results in a higher penalty for violating the constraint. Therefore, a slower timescale solution [54] is needed to iteratively update λ by gradient descent on the state-value function and alternate with policy optimization until the constraint is satisfied.

$$\lambda^{(k+1)} = \Gamma_\lambda\left[\lambda^k + \eta_\lambda\left(\overline{V_c} - E_{s_t \sim D}[V_\pi^c(s_t)]\right)\right] \tag{4.49}$$

where η_λ is the step size for updating λ, Γ_λ projects λ into its logical range [0, $\lambda^{\max}$]. However, at the beginning of the training process, as the constraints are typically not satisfied, λ will increase to surpass the cost $E_{s_t \sim D}[V_\pi(s_t)]$ and focus the optimization on maximizing $E_{s_t \sim D}[V_\pi^c(s_t)]$. This can result in unstable learning as most actor-critic methods that have an explicit parameterization of π are especially sensitive to large (swings in) values. To improve stability, a change of variable $\lambda' = log\ (\lambda)$ is performed to obtain the following dual optimization problem (4.47). Therefore, the weight of constraint cost in the Lagrangian-based state-value function is mitigated.

$$\min_{\lambda \geq 0} \max_{\pi} \frac{E_{s_t \sim D}[V_\pi(s_t)] + \exp(\lambda')(\overline{V_c} - E_{s_t \sim D}[V_\pi^c(s_t)])}{\exp(\lambda') + 1} \tag{4.50}$$

In the proposed Lagrangian-SAC algorithm, the stability and feasibility of constraint handling are achieved through the careful design of the Lagrange multiplier update scheme. First, stability is maintained through soft updates: as formulated in Eq. (3.49)–(3.50), the multipliers are updated using a projected gradient descent scheme on a slower timescale than the policy updates. This mechanism ensures that the values of λ gradually converge rather than oscillate, preventing instability and avoiding over-penalization during the early stage of training. Second, feasibility is guaranteed in expectation: the multiplier update is driven by the difference between the allowed violation budget and the observed violation under the current policy. When constraint violations persist above the threshold, λ increases and shifts the optimization emphasis toward satisfying constraints. Conversely, once the violations fall within the acceptable region, λ stabilizes, allowing the algorithm to continue optimizing operational costs without compromising feasibility.

4.3.5 Practical implementation in Lagrangian-based SAC algorithm

This subsection presents the structure of the off-policy L-SAC algorithm and provides an overview of the overall updating procedure of the proposed algorithm. The L-SAC algorithm implementation comprises the following sets of DNNs: (i) Two critic DNNs, characterized by distinct parameters, are employed to accurately represent the value function $Q(s,a)$ and mitigate the problem of overestimation. (ii) A safety network, with parameters ϑ, is utilized to update the Lagrange multiplier, ensuring the convergence of the algorithm. (iii) To enhance learning stability, two target networks with parameters are adopted. These target networks share the same task and construction as the critics, promoting improved learning efficiency. (iv) The policy, referred to as the actor-network, utilizes parameters φ. It accepts environmental states as input and generates a probability density function, parameterizing a Gaussian distribution for the control action. Fig. 4.4 illustrates the comprehensive workflow of the L-SAC algorithm, depicting the interactions among the components above and highlighting the necessary loss functions. Further elaboration on these loss functions can be found in subsequent sections within this subsection.

The parameters of the networks are updated by performing updates on the critics, actor, and safety networks using an experience buffer. Firstly, the target networks undergo periodic and gradual adjustments derived from the relative critics and safety utilizing equations (4.51) and (4.52) with $\mu \in (0,1)$ [55].

$$\theta_r^* = \mu\theta_r + (1-\mu)\theta_r^*, \forall r \tag{4.51}$$

$$\vartheta_c^* = \mu\vartheta_c + (1-\mu)\vartheta_c^*, \forall r \tag{4.52}$$

Secondly, the safety network parameters ϑ are updated through the utilization of the loss function outlined in equation (4.53), where the mini-batch size M is introduced.

$$J_c = \frac{1}{M}\sum_{n=1}^{M}\frac{1}{2}\left[Q_{\vartheta_c}\left(\mathbf{s}_n,\mathbf{a}_n\right) - c_n - Q_{\vartheta_c^*}\left(\mathbf{s}'_n,\mathbf{a}'_n\right) + \lambda\log\left(\pi^*\left(\mathbf{a}'_n \mid \mathbf{s}'_n\right)\right)\right] \tag{4.53}$$

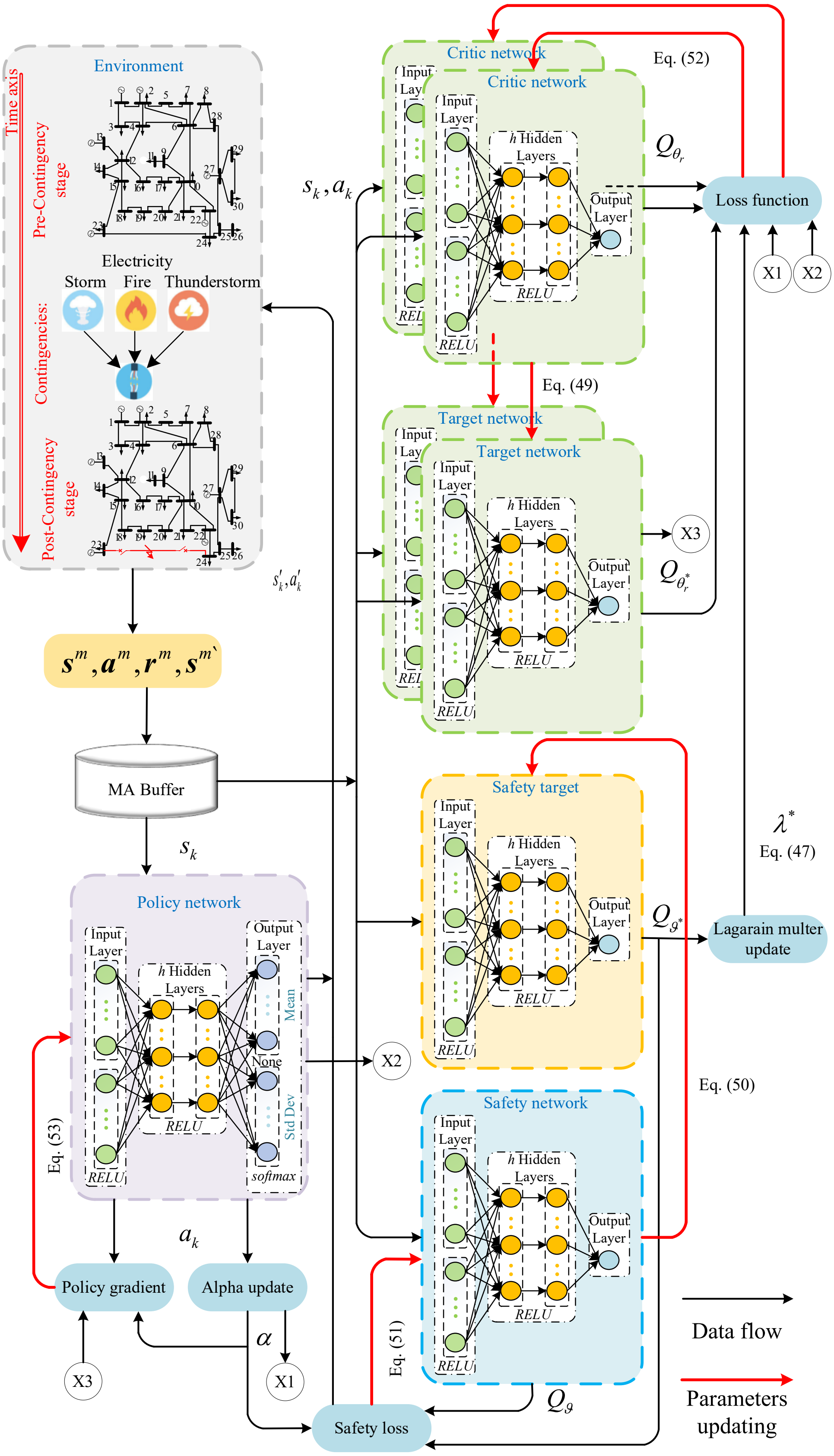


Fig. 4.4 The workflow of the L-SAC algorithm.

Thirdly, the actor parameters φ are updated by employing the policy gradient defined in equation (5.54). Finally, the critics undergo updates by minimizing the loss function defined in equation (5.55).

$$J_\pi = \frac{1}{M}\sum_{n=1}^{M}\left(-\min_{r\in\{1,2\}} Q_{\theta_r^*}\left(\mathbf{s}_n,\mathbf{a}_n\right) + \lambda\log\left(\pi\left(\mathbf{a}_n | \ \mathbf{s}_n\right)\right)\right) \tag{4.54}$$

$$J_r = \frac{1}{M}\sum_{n=1}^{M}\frac{1}{2}\left[Q_{\theta_r}\left(\mathbf{s}_n,\mathbf{a}_n\right) - r_n + \lambda^* c_n - Q_{\vartheta_r^*}\left(\mathbf{s}'_n,\mathbf{a}'_n\right) + \lambda\log\left(\pi^*\left(\mathbf{a}'_n | \ \mathbf{s}'_n\right)\right)\right], \forall r\in\{1,2\} \tag{4.55}$$

The proposed L-SAC algorithm is summarized in Algorithm 3. In each iteration, the parameters of the networks are updated using stochastic gradient descent. This process involves performing gradient updates to optimize the network parameters and improve the algorithm's performance.

Algorithm 3 L-SAC Algorithm

```
Initialize: Preventive agent networks $\varphi^d$, $\theta_i^d$, $\theta_i^{d*}$, $\varphi^a$, $\theta_i^a$,
    and $\theta_i^{a*}$, corrective agent networks $\varphi^c$, $\theta_i^c$, and $\theta_i^{c*}$, and
    safety network $\vartheta$
For each episode do
    For each time step do
        $a^d \sim \pi_{\varphi^d}(\cdot | s^d = s)$.
        $a^a \sim \pi_{\varphi^a}(\cdot | s^a = [a^d, s^d])$.
        $a^c \sim \pi_{\varphi^c}(\cdot | s^c = [a^d, a^a, s^d])$.
        Apply $a^d$, $a^a$, and $a^c$ in eq. (4.25) and observe $r^d$, $r^a$,
    $r^c$, $s'$, and $c$.
        $\mathbb{R} \leftarrow \mathbb{R} \cup \left(s, a^d, a^a, a^c, r^d, r^a, r^c, c, s'\right)$
End For
For each gradient step do
        Sample random mini-batch $N$ experiences from $\mathbb{R}$
    Update soft Q-value parameters $\theta_i^d$, $\theta_i^a$, and $\theta_i^c$ using (4.55)
        Update policy parameters $\varphi^d$, $\varphi^a$, and $\varphi^c$ using (4.54)
        Update safety network parameter $\vartheta$ using (4.53)
        Adjust temperature $\alpha^d$, $\alpha^a$, and $\alpha^c$ using (4.40)
        Update targets $\theta_i^{d*}$, $\theta_i^{a*}$, $\theta_i^{c*}$, and $\vartheta^*$ using (4.51)-(4.52)
```

4.4 Case Study

4.4.1 Experiment Setting

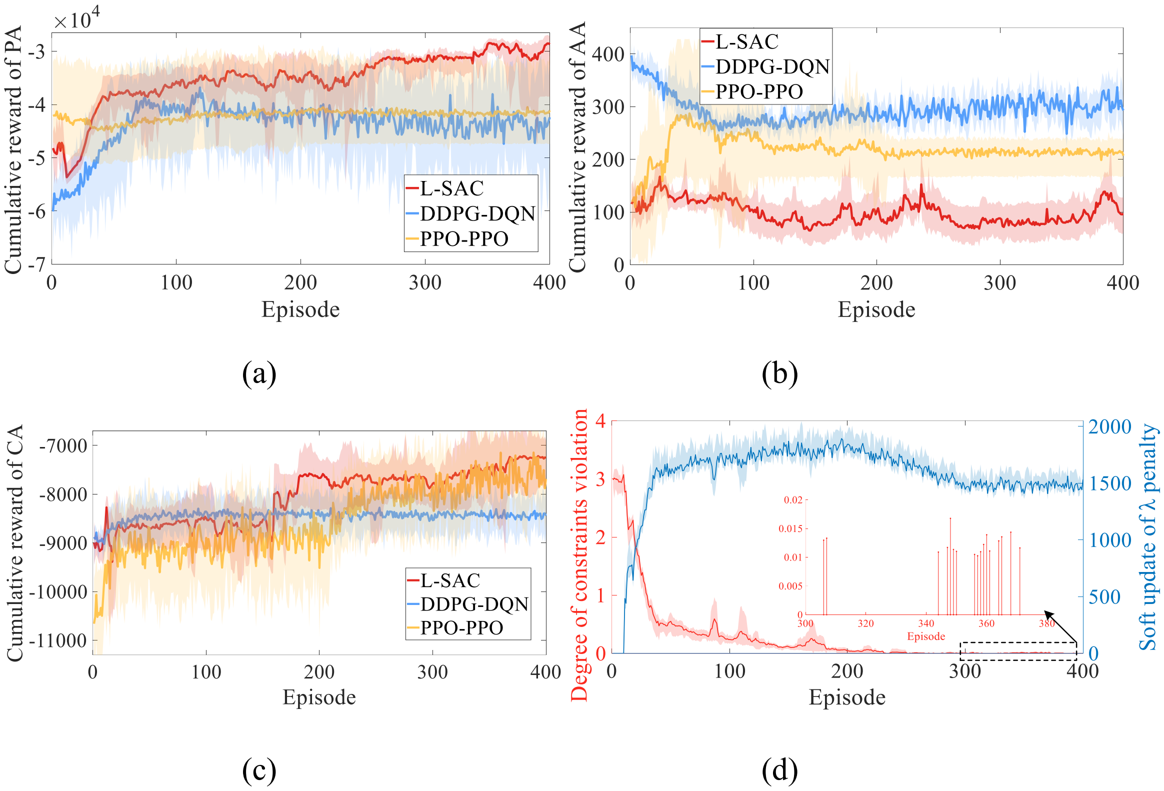
In this section, the formulated CMDP model and the proposed DRL solution approach are evaluated by examining two test systems, IEEE 30-bus and IEEE 118-bus. The system parameters, including system topology, generation capacities, and line parameters, are directly handled in its standard format as in PYPOWER. The numerical results listed below are conducted on a computer with an Intel i7−10700 CPU and 16 GB of RAM. The hyperparameters of the SAC algorithm are presented in Table 4.2. Additional modeling data are generalized as follows. Load shedding penalties of Eq. (4.25) are set as $C_d = 10 \times C_g$ and $C_d^o = 100 \times C_g$. The PV generation profile data are from pvoutput.org, whose generation power capacity is 6 kW, and the BESS power/energy capacity is 10 kW/30 kWh. Finally, power demands are randomly generated, where maximum and minimum values are set at 120% and 80% of the normal operating point in the data set PYPOWER. The response time (*t*2 - *t*1 as shown in Fig. 4.1) and ramping time (*t*3 – *t*2) of the generators are assumed to be 5 and 10 min, respectively.

Table 4.2 Main Hyper-Parameters and Data Setting.

Parameters	Value	Parameters	Value
Optimizer	Adam	Activation	RELU
Actor learning rate	1e-3	Critics learning rate	1e-2
Entropy learning rate	1e-4	Targets learning rate	1e-3
Discount factor	0.99	Initial temperature	1
Neurons number	512	Time step	1
Max steps	24	Minibatch size	128

4.4.2 Case 1: 30-Bus System

The modified IEEE 30-bus system is used to test the proposed algorithm. This system has 30 buses, six generators, 41 branches, and 24 loads. Six VPPs are connected at buses 4, 7, 10, 15, 24, and 30. The maximum allowed adjustment of long-term branch flow than the continuous ratings is 1.2.



(a)

(b)

(c)

(d)

Fig. 4.5 Convergence performance of the proposed and benchmark algorithms (a) reward of the DA; (b) reward of the AA; (c) reward of the CA; (d) constraint penalty.

In this subsection, the training performance comparison of the proposed algorithm with two state-of-the-art DRL algorithms is investigated. The first benchmark algorithm, DDPG, performs offline training but struggles to handle discrete action spaces. To overcome this limitation, the deep Q-network (DQN) is introduced to combine with DDPG to form the DDPG-DQN algorithm, which generates the discrete action for AA. The second benchmark algorithm, PPO, is an on-policy algorithm. The PPO algorithm is responsible for generating action for PA and CA and is referred to as the PPO-PPO method. All DRL algorithms are implemented in the IEEE 30-bus system with the same data set. Ten independent experiment results of each algorithm with different initial seeds and training datasets are collected to depict the DA, AA, and CA cumulative reward curves in Figs. 4.5 (a), (b), and (c) under the N-1 criterion. Fig. 4.5 (d) demonstrates the degree of constraint violations and the soft update in the penalty λ per episode. It is worth noting that in Fig. 4.5, the solid curve in each algorithm represents the average value of ten experiments, while the light-colored shadow area is bounded by the minimum and maximum rewards obtained over ten experiments.

It is observed from Figs. 4.5 (a), (b), and (c) that the proposed algorithm exhibits significant oscillations in reward during the first ten episodes as the SAC agent stochastically explores to fill the replay buffer. After the first ten episodes, the reward of the DA begins decreasing as the reward of the AA increases. Nevertheless, this situation is reversed when the DA updates robust action against the attack and improves the reward after the 50th episode. In the end, the reward of the DA continues to grow and reaches 2.8×10^4. This is owing to the DA policy learned from the prior experience to generate optimal action against the AA. As seen in Fig. 4.5 (d), the power system generates a high degree of violation in the first ten episodes. This is mainly because the Lagrange multiplier λ starts with a low value, 0, and the updating process of the network begins after filling enough prior experience in the replay buffer. However, the dramatic increase in the penalty value encourages the agent to find feasible control decisions and reduce the degree of the constraint violation during 10 - 50 episodes. Then, the agent changes its focus from avoiding violations to minimizing the total operation costs while guaranteeing the minimum violation degree, and the penalty value moves to its saturation value. Therefore, the safety network generates small violation degrees and encourages a reduction in the penalty value. Based on the results presented in Fig. 4.5, the DDPG algorithm fails to learn a good policy for the DA and CA to achieve a steady reward, and there are big fluctuations in the cumulative reward of the AA as well. On the other hand, PPO demonstrates considerable fluctuation in the start-up training phase, with the reward of DA decreasing due to the increased reward of the AA. Numerical results show that the proposed algorithm outperforms the benchmark DRL algorithms in terms of the cumulative reward of the DA and AA and exhibits better convergence.

Furthermore, to further demonstrate the superiority of the proposed algorithm, Table 4.3 illustrates the numeric results for the aforementioned algorithms under N-1, N-2, and N-3 criteria, wherein the operational cost, pre-contingency stage load shedding, unserved electricity, short-term and long-term adjustment are included in each hour, over the last 100 episodes over ten independent experiments. Additionally, the computation time of each DRL algorithm is presented to compare the training efficiency of the algorithm.

Table 4.3 Computational Results of the Proposed and Benchmark Algorithm for the N-*k* PCSCOPF Problem.

	Method	OC	LS	UE	ST	LT	CT
N-1	L-SAC	293.73	29.52	7.33	121.19	179.29	1850.58
	DDPG-DQN	297.54	28.28	12.60	140.93	224.52	2172.39
	PPO-PPO	275.44	49.91	9.351	102.11	236.04	2059.36
N-2	L-SAC	271.93	47.82	28.19	149.09	239.87	1902.71
	DDPG-DQN	260.12	51.43	27.64	103.83	218.53	2341.72
	PPO-PPO	254.84	55.92	17.08	127.99	184.72	2252.45
N-3	L-SAC	245.95	64.62	30.19	154.70	239.58	2143.00
	DDPG-DQN	247.80	59.46	32.27	134.22	228.79	2466.21
	PPO-PPO	271.93	47.82	28.19	149.09	239.87	2477.67

OC: Operational cost; LS: Load shedding (Unit: MW); UE: Unserved electricity (Unit: MW); ST: Short-term corrective adjustment (Unit: MW); LT: Long-term corrective adjustment (Unit: MW); CT: Computation time (Unit: s).

As shown in this table, the proposed L-SAC algorithm outperforms the other alternative algorithms in terms of the CCOPF solution quality. The proposed algorithm attains the minimum total cost with the lowest pre-contingency load shedding and post-contingency unserved electricity, which is achieved through the coordination of the continuous action SAC and discrete SAC. Simultaneously, the proposed L-SAC generates an optimal policy to reduce the adjustment of the VPP and the generators after the contingencies to redispatch the power system. Additionally, the proposed L-SAC algorithm outperforms both DDPG and PPO in terms of computation time. This is obvious because the proposed L-SAC algorithm inherits such an advantage from the traditional SAC algorithm. Therefore, the proposed algorithm demonstrates an advantage in training efficiency, with an average of 18.51% and 15.09% time savings compared to the DDPG and PPO algorithms, respectively.

Table 4.4 Corrective Actions in Different Contingencies Scenario (Unit: MW).

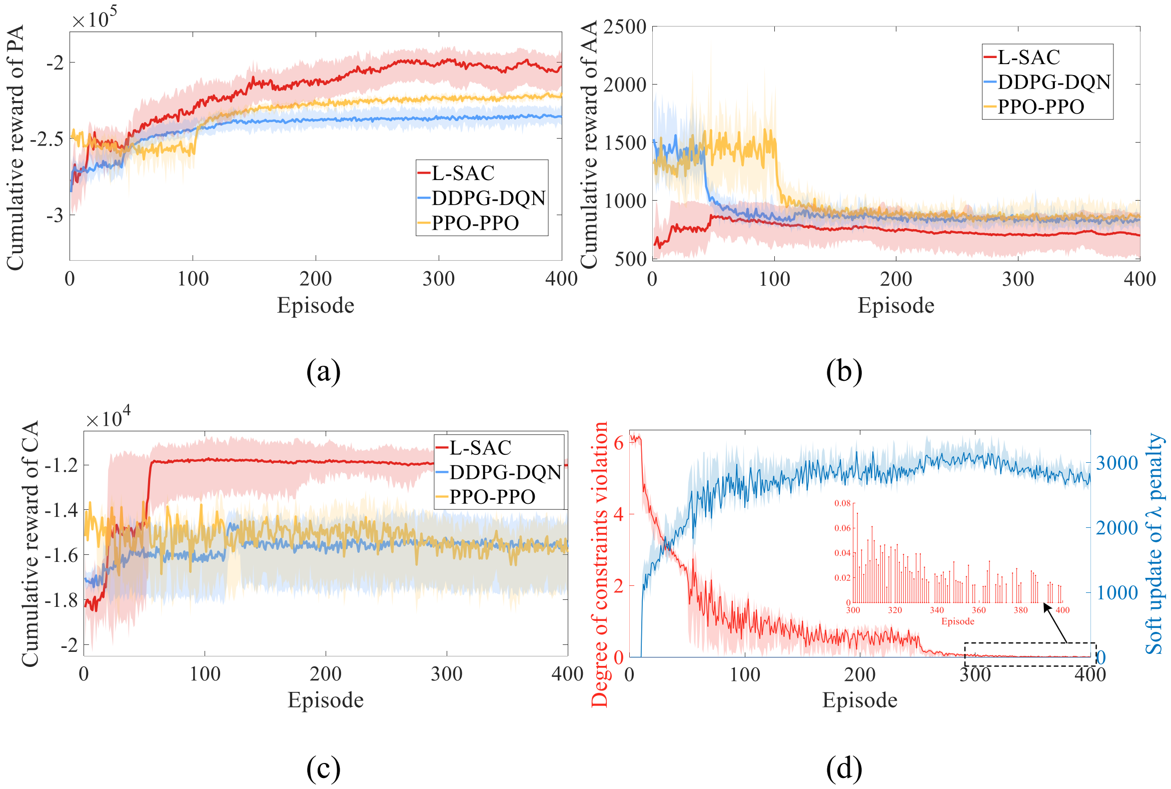
		L10	L36	L25
Short term	P_{vpp1}	-29.62	-29.96	-12.69
	P_{vpp2}	-29.80	-29.88	-21.12
	P_{vpp3}	8.54	24.29	15.32
	P_{vpp4}	29.22	28.36	6.01
	P_{vpp5}	4.11	20.26	0.31
	P_{vpp6}	19.90	-11.30	11.60
Long term	ΔP_{G1}	-36.17	-1.41	-21.11
	ΔP_{G2}	-54.15	-55.23	-10.51
	ΔP_{G3}	29.11	27.41	3.58
	ΔP_{G4}	18.56	-4.51	14.75
	ΔP_{G5}	16.03	10.73	2.11
	ΔP_{G6}	25.27	23.81	10.77
DCV		0.6043	0.3371	0.2070

Table 4.4 demonstrates the corrective control actions during short- and long-term emergency periods when part transmission lines are disconnected. The stochastic contingencies will cause overflow in the power system, which results in constraints violation during the post-contingencies stage. These corrective actions are provided by the VPPs immediately to remedy the branch flow above their short-term emergency rating after each contingency and the long-term generation adjustment for the same contingencies. This demonstrates that, after solving the pre-contingency problems of the PSCOPF, suitable corrective actions by the VPPs and generators are required to relieve overloads.

4.4.3 Case 2: 118-Bus System

The scalability of the proposed CCOPF is tested on the modified IEEE 118-bus system. This system has 118 buses, 54 generators, 186 branches, and 99 loads. Six VPPs are connected at buses 5, 30, 37, 64, 82, and 94. The maximum allowed adjustment of long-term branch flow compared to the continuous ratings is 1.2. The convergence curves of the cumulative reward of the DA, AA, and CA based on different DRL algorithms for the IEEE

118-bus system in the N-1 criterion are presented in Figs. 4.6 (a), (b), and (c). Fig. 4.6 (d) demonstrates the degree of constraint violations and the soft update in the penalty λ per episode. Similarly, the solid curve in each algorithm corresponds to the average value of ten independent experiments, and the light-colored shadow area is bounded by the minimum and maximum rewards over the experiments.



(a)

(b)

(c)

(d)

Fig. 4.6 Convergence performance of the proposed and benchmark algorithms based on the IEEE 118-bus system (a) reward of the DA; (b) reward of the AA; (c) reward of the CA; (d) constraint penalty.

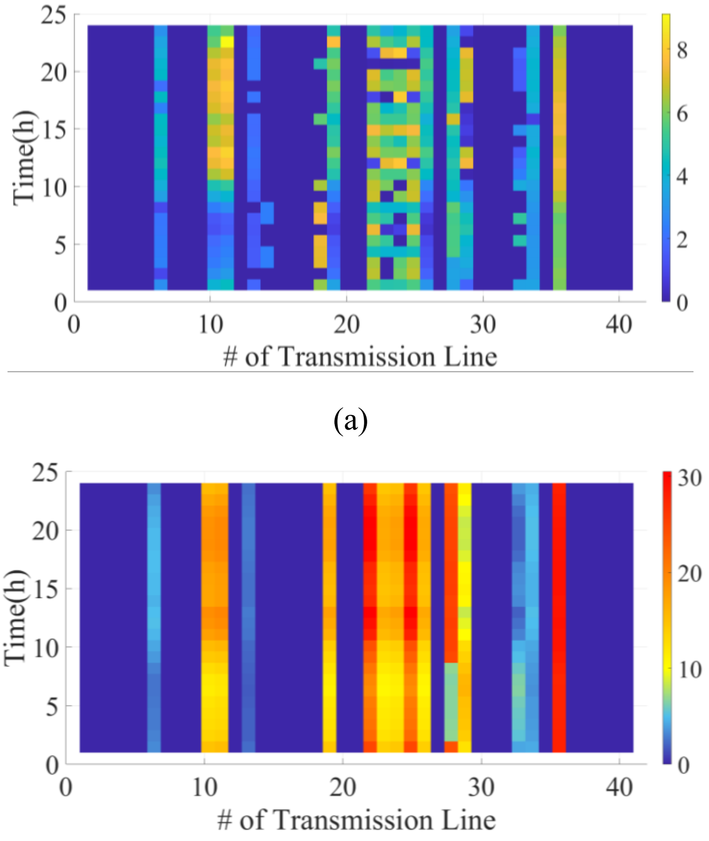
As shown in the figures, the training process of the proposed algorithm can be divided into three continuous stages: policy exploration (first ten episodes), policy training (from 10 to 150 episodes), and policy convergence (after 150 episodes). During the initial stage, the proposed algorithm collects initial experiences, which demonstrate slight fluctuations in the cumulative reward of DA and AA. With the training process preceding, DA and AA start to learn the optimum policy and generate satisfactory actions to compete with each other. Therefore, the cumulative rewards of DA and AA exhibit significant fluctuations during this stage. The competition facilitates DA in generating robust action against the worst contingency scenario. Hence, the cumulative reward of the proposed algorithm converges to minimize operational costs. On the other hand, the PPO algorithm demonstrates fast update progress in the cumulative reward of AA, but the cumulative reward of DA suffers from obvious competition from AA. Even after the training process ends, the cumulative reward for the DA is at a lower level. DDPG-DQN provides AA with a gradually growing cumulative reward due to DQN's on-policy training. However, when the replay buffer of the DDPG algorithm stores enough prior experience, the cumulative reward of DA starts to rise.

Table 4.5 Computational Results based on the IEEE 118-Bus System.

	Method	OC	LS	UE	ST	LT	CT
N-1	L-SAC	4715.23	109.88	28.61	79.93	426.16	6946.48
	DDPG-DQN	4614.75	214.04	33.92	88.17	557.43	8157.34
	PPO-PPO	4919.29	108.42	35.42	130.34	465.94	7966.76
N-2	L-SAC	4768.45	133.19	34.19	119.98	486.07	7210.43
	DDPG-DQN	4544.411	235.15	44.63	84.07	512.80	8537.56
	PPO-PPO	4871.24	137.42	42.10	144.70	494.32	8014.78
N-3	L-SAC	4628.78	213.33	41.20	141.97	525.97	7679.51
	DDPG-DQN	4400.83	294.62	49.46	86.26	599.54	8849.11
	PPO-PPO	4744.89	221.09	52.11	159.0	532.36	8511.75

To demonstrate the superiority of the proposed algorithm, Table 4.5 illustrates the detailed computational results of different DRL algorithms in the N-*k* security criterion in the IEEE 118-bus system. The table includes key performance metrics such as operational cost, load shedding, unserved electricity, computation time, and total cost. Although the solutions of the proposed algorithm result in a high operational cost, they outperform the two benchmark algorithms in terms of load shedding and unserved electricity under different N-*k* contingency scenarios. Meanwhile, the proposed algorithm demonstrates high sample efficiency and requires less computation time to train the PCSCOPF problem while curtailing fewer pre-contingency load shedding to reduce unserved electricity. As a result, the total cost of the solution generated by the L-SAC algorithm is the lowest, with an average improvement of 17.02% and 12.23% compared to the DDPG and PPO algorithms, respectively.

4.4.4 Case 3: Robustness Effectiveness of the PCSCOPF model



(a)

(b)

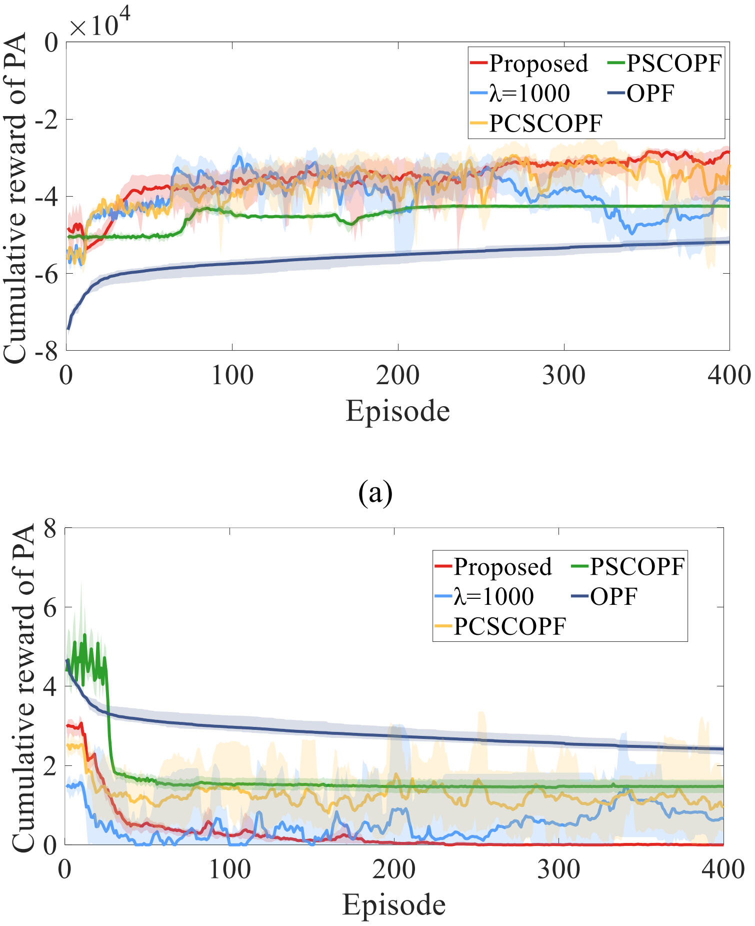
Fig. 4.7 (a) Unserved electricity in the proposed method and (b) in the OPF method.

The stochastic contingencies will cause not only constraint violation but also unserved electricity on demand buses. The subsection verified the effectiveness of the preventive action in enhancing the robustness of the power system. Fig. 4.7 demonstrates the unserved electricity of the PCSCOPF solution and the OPF solution when different transmission lines are out of service hourly. As shown in the figures, the PCSCOPF solution exhibits superior performance in reducing unserved electricity as compared to the OPF solution, with less than 10kWh maximum unserved electricity in the PCSCOPF solution compared to more than 30kWh in the OPF solution. The L-SAC algorithm, by training the defense and attack agents to compete with each other in the pre-contingency stage, can generate robust actions for the PCSCOPF solution against the worst contingency scenarios. Therefore, when the transmission lines are out of service, the PCSCOPF solution can effectively mitigate the unserved electricity caused by attacks. On the contrary, the OPF solution struggles to mitigate the unserved electricity and maintain solution quality when different transmission lines are tripped, leaving it vulnerable to contingencies without a defensive strategy. Hence, the L-SAC algorithm is effective in generating robust PCSCOPF solutions under contingency scenarios.

4.4.5 Case 4: Soft Update of Lagrange multiplier

The power system is the most important infrastructure, and it must be ensured that any decision taken is safe and does not violate any crucial operating constraints. However, the stochastic contingencies will cause significant constraint violations. Numerical comparisons of the security effectiveness of the soft update of the Lagrange multiplier with the existing methods of optimal power flow, PSCOPF, PSCOPF methods, and the fixed penalty methods in the IEEE 30-bus and IEEE 118-bus systems are conducted. Fig. 4.8 (a) demonstrates the operation cost of the IEEE 30-bus system under the N-1 contingency scenario, and the degree of constraints violation will multiply by 1×10^4 as part of the operation cost. Fig. 4.8 (b) illustrates the cumulative constraint violation degree of the five methods. It should be noticed that the solid curve in each method represents the average value of ten experiments, while the light-colored shadow area is bounded by the minimum and maximum values obtained over ten experiments. The OPF and PSCOPF generate more conservative decisions than the proposed method because of the lack of safe action in the post-contingency stage to avoid violations. As indicated by the figures, the PCSCOPF provides high reward values because the agent focuses on the operational costs, neglecting the low values of penalty terms. However, it suffers from high values of constraint violations. It generates unsafe actions during the training process, resulting in an unpromising method for the power system operation. The degree of constraint violation is low when setting the λ=1000.

However, its operation cost suffers from big fluctuation due to the unstable degree of constraint violation.



(a)

(b)

Fig. 4.8 (a) Optimality performance of the proposed safety method; (b) Safety performance of the proposed safety method.

To investigate the feasibility of the five methods after the stochastic contingency, the numerical data of the last 100 episodes, i.e., 2400 time steps, are collected. Table 4.6 presents the statistical results of the five methods in the security action and the cumulative DCV over the 2400 time steps. The Scur (%) is defined in (4.56), where the limitation of the DCV, C_t, is set as 0.05 in the IEEE 30-bus system and 1 in the IEEE 118-bus system. From the table, it is obvious that the proposed method outperforms both existing methods and the fixed penalty method by predicting high-quality nonconservative control actions, promoting safety and economical operation.

$$\text{Scur\%} = \sum_{n=1}^{N}\left[\frac{1}{|\mathcal{T}|}\left(\sum_{\forall x_n}(\mathbf{x}-\bar{\mathbf{x}}) > \bar{C}_t + (\underline{x}-\mathbf{x}) > \bar{C}_t\right) = 0\right] \tag{4.56}$$

Table 4.6 Security Performance of Different Safety OPF Methods.

Method	IEEE 30-Bus system		IEEE 118-Bus system	
	Scur (%)	Cumulative C_t	Scur (%)	Cumulative C_t
OPF	29.9%	248.2038	16.6%	1050.26
PSCOPF	37.0%	147.3566	32.5%	918.69
PCSCOPF	68.9%	114.7575	50.0%	836.09
λ=1000	79.1%	87.2114	71.9%	609.82
Proposed	99.99%	0.4424	99.78%	18.44

4.4.6 Case 5: Demand Response Programs Analysis

Demand response programs (DRPs) contributes to the power system by offering a flexible model that can enhance both the security and economic performance of the SCOPOF model [47]. This subsection discusses the simulation results of the proposed method for handling DRPs on the IEEE 30-bus system. Four different time-based DRP scenarios—flat rates, time of use (TOU), real-time pricing (RTP), and critical peak pricing (CPP)—are considered to evaluate their performance in terms of operational costs, total load shedding, and the peak-to-valley ratio (PVR) of demand. Furthermore, an economic model for DRPs is introduced to simulate the economic behavior of responsive loads, with their corresponding electricity tariff ratios provided in [47]. Based on the four different DRP scenarios, Fig. 4.9 illustrates their daily demand curves based on demand ratios in [47].

Based on the load demand from Fig. 4.9, the safe RL method solves the PCSCOPF problem for each of the four DRP scenarios, with the hourly load shedding for each scenario shown in Fig. 4.10. In the flat rate scenario, which remains constant throughout the day, load shedding is relatively stable, reflecting the absence of price incentives to shift demand. In contrast, TOU pricing leads to more dynamic load shedding, with reduced shedding during cheaper peak hours (1–9) and increased shedding during more expensive off-peak and valley periods (10–24). RTP and CPP demonstrate greater responsiveness, with significantly higher load shedding during periods of elevated pricing, especially between hours 20–22 when RTP and CPP prices spike. These two schemes show the system's sensitivity to real-time or critical price changes, leading to the highest load shedding during peak price periods. Overall, while the flat rate results in a uniform shedding pattern, TOU, RTP, and CPP pricing strategies encourage more targeted load reductions, with RTP and CPP proving most effective during critical peak times.

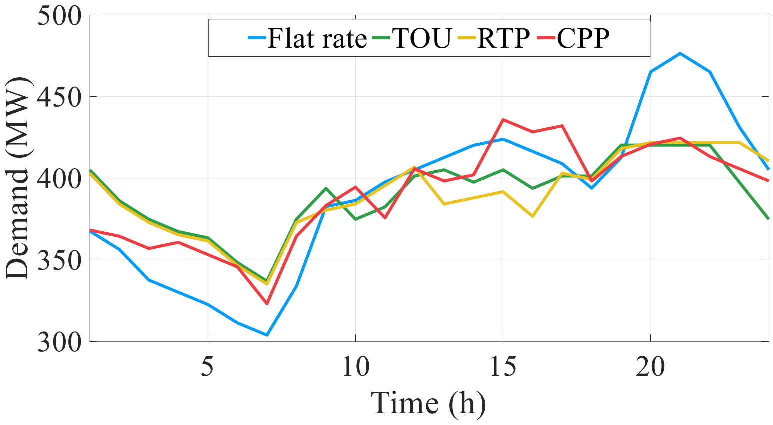


Fig. 4.9 Power demand curve based on four different scenarios.

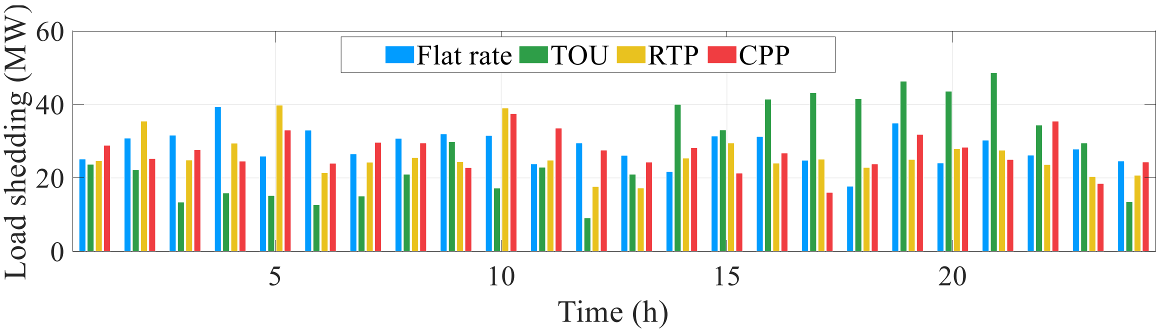


Fig. 4.10 Load shedding based on four different scenarios.

Table 4.7 evaluates the performance of power systems under four DRP scenarios based on the safe RL solution for the PCSCOPF model. The results highlight that dynamic pricing schemes, such as RTP and TOU, generally outperform flat rate and CPP across various performance metrics. Although the flat rate achieves the lowest operational cost, it also results in the highest load shedding and the greatest demand imbalance, as indicated by its PVR. In contrast, RTP and TOU provide a better balance by reducing both load shedding and PVR, reflecting more efficient load management and a smoother demand curve. Meanwhile, CPP lowers PVR and load shedding compared to the flat rate, incurs the highest operational costs due to the sharp price spikes during peak periods. Overall, RTP demonstrates as the most effective strategy for balancing operational costs, reducing load shedding, and flattening demand, whereas CPP offers similar benefits in demand smoothing but with reduced cost efficiency. The DRPs contribute to reducing the system's operational costs and improving security performance by lowering the load shedding.

Table 4.7 Comprehensive assessment of performance of four different scenarios.

DRPs Indices	Flat rate	TOU	RTP	CPP
Operation costs	7.67e+03	7.69e+03	7.92e+03	1.17e+04
PVR	1.5679	1.2472	1.2584	1.3488
Load shedding (MW)	681.0367	654.5655	620.8196	647.9215

To clarify the effect of the contingency filtering approach in the proposed model, the solution times of the problem in scenarios of IEEE 30-bus system are given in Table 4.8.

Table 4.8 Computation time performance of four different scenarios.

DRPs Indices	Flat rate	TOU	RTP	CPP
Computation time without contingency filtering (s)	2029.73	2053.14	2058.44	2042.22
Computation time with contingency filtering (s)	1376.63	1371.51	1365.55	1370.99

This clearly demonstrates that applying the contingency filtering approach effectively reduces the computational burden, enabling faster solutions to the PCSCOPF problem under various DRPs. The filtering approach filters unnecessary contingency scenarios, accelerates the calculation process, making it a practical method to enhance computational efficiency, especially for real-time power system operations.

4.5 Summary

This paper presents a novel, fast, and safe solution method for PCSCOPF problem that uses a combination of a robust DRL algorithm and the Lagrangian relaxation methods. By modeling the problem as a CMDP with two DRL agents, the approach ensures robust and efficient solutions to prevent and correct N-*k* outages. The enhanced L-SAC algorithm, featuring soft Lagrange multiplier updates, guarantees the safe exploration of control actions, improving policy robustness. Meanwhile, the incorporation of the VPPs with the power system in this work enables a fast response to stochastic contingencies, thereby avoiding short-term violations of the operating constraints. Finally, test results on IEEE 30-bus and 118-bus systems verify the computational efficiency and reliability of the proposed method, outperforming traditional OPF approaches in handling stochastic contingencies.

Chapter 5 Online Voltage Control Strategy: Multi-Mode Based Data-Driven Approach for Active Distribution Networks

Active distribution network (ADN) is faced with significant challenges, including frequent and fast voltage violations, due to the increased integration of intermittent renewable energy resources. This paper proposes a two-stage multi-mode voltage control strategy based on a deep reinforcement learning (DRL) algorithm, designed to alleviate voltage violations in ADN and minimize network power loss. In the first stage, a DRL algorithm, the soft actor-critic (SAC), is introduced to determine the hourly dispatch of on-load tap changers and capacitor banks, ensuring voltage security during the day-ahead stage. A multi-mode voltage regulation strategy is then proposed to obtain real-time dispatch of PV inverters, aiming to save energy and enforce voltage constraints under various conditions. The real-time voltage regulation problem is formulated as a Markov decision process and solved using a multi-agent SAC integrated with an attention mechanism. All agents undergo centralized offline training to learn the optimal coordinated voltage control strategy, then make decentralized online decisions based on locally available information only. The effectiveness of the proposed approach is confirmed through extensive testing on the IEEE 33-bus distribution system, with simulation results conclusively demonstrating its ability to address voltage violation challenges.

5.1 Framework

This study addresses the challenge of optimal voltage regulation within a two-stage framework, as illustrated in Fig. 5.1. The framework coordinates the collaboration of both traditional and innovative voltage control devices across two timescales, aiming to ensure secure operation in ADN using the DRL algorithm. Two intelligent agents, designed as day-

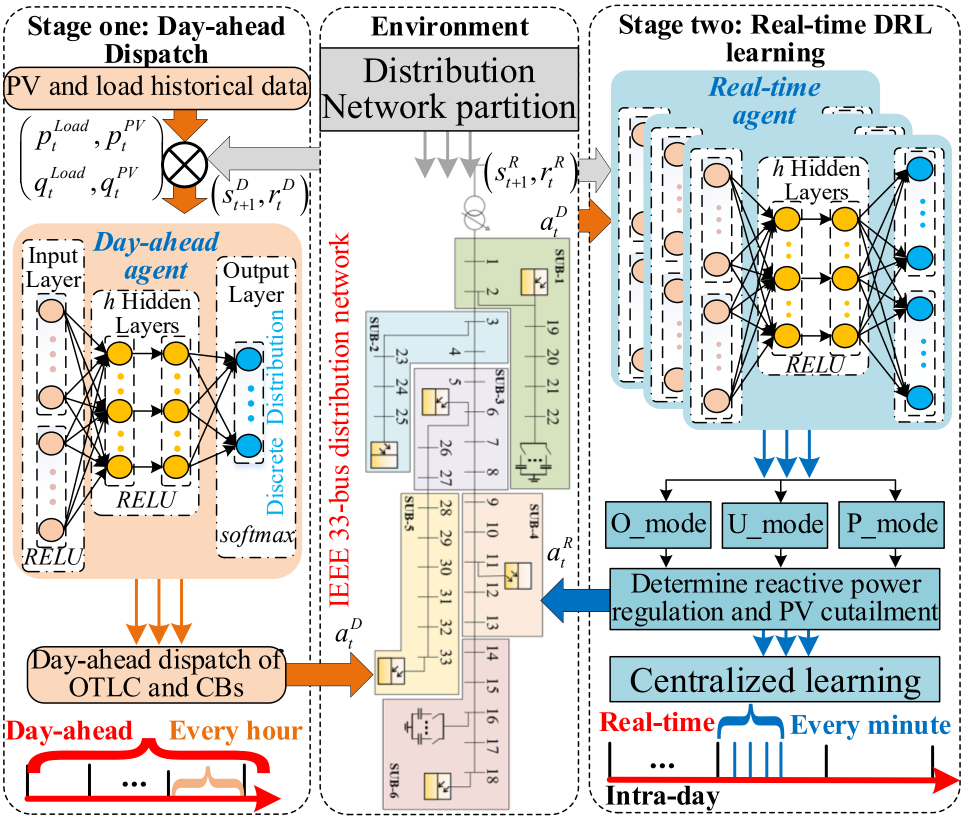


Fig. 5.1 Proposed multi-mode based data-driven voltage control framework.

ahead and real-time agents, are specifically tailored for voltage regulation at different timescales. In the first stage, forecasts of PV generation and power demand for the upcoming day are generated and communicated as observations to the day-ahead agent. This agent undergoes training to learn the optimal control policy for voltage regulation within the MDP framework. Subsequently, to address the slower timescale control, the day-ahead agent executes optimal power flow within the environment, yielding day-ahead dispatch schedules for OLTC and CBs. In the second stage, system information such as voltages and switch statuses of OLTC and CBs from the first stage are recorded as observations for the real-time agent. In order to mitigate power loss and voltage violations, the well-trained real-time agent, operating within the MDP framework, adjusts the output of PV inverters to achieve fast timescale control. To further guide the selection of distinct voltage regulation modes in the real-time stage, two security operation margins are designed. Specifically, three operation modes – power loss minimization mode (P_Mode), under- voltage optimization mode (U_Mode), and over-voltage optimization mode (O_Mode) – are formulated for the real-time agent. This multi-mode control strategy is designed to ensure both economic and secure operation in ADN.

This study notably incorporates multiple agents for optimal voltage regulation. A multi-agent soft actor-critic (MASAC) algorithm is introduced to address the formulated MDP, representing an off-policy entropy maximization-based DRL algorithm. Power flow calculations are then executed in the modeling environment, incorporating injection actions and the dispatch of PV, OLTC, and CBs. The actor and critic networks of the SAC algorithm are further augmented with an attention mechanism to extract pertinent information from extensive state-action spaces, thereby mitigating potential issues related to local observations. Compared to single-based DRL algorithms, the attention-based MADRL requires less information to generate optimal voltage control, a notable departure from the MADRL algorithm, which experiences performance degradation when handling numerous agents. Considering the longevity concerns and sluggish response of traditional voltage control devices, OLTC and CBs are adjusted on an hourly basis during day-ahead dispatch. In the real-time stage, the output of PV inverters is regulated with a 1-minute time interval between adjacent control steps in each agent to address rapid voltage changes. Leveraging the offline training characteristic of the SAC algorithm, all agents in the MADRL algorithm undergo centralized training to learn the coordination voltage regulation strategy. Upon completion of the exploring process, the parameters of the DNN stabilized and subsequently transitioned to online implementation for each agent, enabling real-time voltage control based on local observations. This approach significantly mitigates the degradation of control performance caused by communication delays within the entire system.

5.2 Problem Formulation

5.2.1 Day-ahead Stage for voltage regulation

The aim of the two-stage voltage control is to determine the optimal dispatch of OLTC and CBs at each time step. This ensures that both the cumulative voltage violation and the long-term switching operations of mechanical devices are reduced. Consequently, the mathematical optimization formulation for the first stage is articulated as follows:

$$\min \sum\nolimits_{t\in T}\left(\sum\nolimits_{ij\in B} l_{ij,t} r_{ij} + \sum\nolimits_{i\in N} v_{i,t}^{D}\right) \tag{5.1}$$

$$p_{j,t}^{PV} + \sum_{ij\in B} p_{ij,t} = \sum_{jk\in B}\left(l_{jk,t} r_{jk} + p_{jk,t}\right) + p_{j,t}^{Load}, \forall i, j \tag{5.2}$$

$$q_{j,t}^{PV} + \sum_{ij\in B} q_{ij,t} + q_{j,t}^{CB} = \sum_{jk\in B}\left(l_{jk,t} x_{jk} + q_{jk,t}\right) + q_{j,t}^{Load}, \forall i, j \tag{5.3}$$

$$v_{i,t} = v_{j,t} - 2\left(r_{ij} p_{ij,t} + x_{ij} q_{ij,t}\right) + (r_{ij}^2 + x_{ij}^2) l_{ij,t}, \forall ij, t \tag{5.4}$$

$$v_{1,t} = \left(V_s + tap_t \cdot \Delta V_T\right)^2, \forall t \tag{5.5}$$

$$p_{ij,t}^2 + q_{ij,t}^2 = v_{i,t} l_{ij,t}, \forall i, j, ij, t \tag{5.6}$$

$$\left|q_{i,t}^{PV}\right| \le \sqrt{\left(p_{i,t}^{PV}\right)^2 + \left(s_{i,t}^{PV}\right)^2}, \forall i \tag{5.7}$$

$$q_{i,t}^{CB} = \sum\nolimits_n u_{i,n,t}^{CB} q_{i,n,t}^{CB} \tag{5.8}$$

$$\underline{V} \le v_{i,t} \le \overline{V}, \forall i, t \tag{5.9}$$

$$\sum_{t=1}^{T}\left|tap_{t+1} - tap_t\right| \le tap_{\max} \tag{5.10}$$

$$\sum_{t=1}^{T}\left|u_{i,n,t+1} - u_{i,n,t}\right| \le cap_{i,\max} \tag{5.11}$$

where *B* and *N* are the sets of transmission lines and power buses; l_{ij}, p_{ij}, and q_{ij} are current active, and reactive power flow of transmission line ij; $v_{i,t}$ and $v_{i,t}^{D}$ are voltage and voltage deviation of bus *i* at time *t*; $p_{i,t}^{PV}$ and $q_{i,t}^{PV}$ are active and reactive of PV of bus *i* at time t; $p_{i,t}^{Load}$ and $q_{i,t}^{Load}$ are active and reactive demand of bus *i* at time *t*; $q_{i,t}^{CB}$ and $u_{i,n,t}^{CB}$ are reactive power and status of *n*th capacitor of CB of bus i at time *t*; tap_t is the status of OLTC at time *t*; r_{ij} and x_{ij} are resistance and reactance of transmission line ij; ΔV_T is voltage regulation of OLTC for one-tap step; V_s is the primary voltage of transformer at the slack bus; S_i^{PV} is power capacitor of the PV inverter at bus *i*; $\overline{V}$ and $\underline{V}$ are max and min bus voltage limit; $q_{i,n,t}^{CB}$ is reactive power of one capacitor at bus *i*; tap_{max} and $cap_{i,max}$ are maximum operation number of OLCT and nth capacitor at bus *i*.

The objective function (5.1), consisting of two terms, minimizes the total cost. The first term represents the power loss costs, which are crucial for efficient energy distribution. The second term represents the voltage deviation costs, penalizing deviations from the desired voltage levels to ensure system stability. Equations (5.2)-(5.3) delineate the nodal power balance constraints. Equation (5.4) describes the constraints related to bus voltages. Equation (5.5) calculates the substation voltage based on the OLTC positioning. Equation (5.6) denotes the power flow constraint, ensuring that the power transmitted through each branch does not exceed its limits. Equation (5.7) specifies the reactive power constraint for the PV inverter, ensuring that the inverter operates within its reactive power capability limits. Equation (5.8) computes the reactive power injections facilitated by the CBs, which provide necessary reactive power support to maintain voltage levels and improve power factor. Equations (5.9)-(5.11) establish the constraints concerning bus voltage limits, branch flow, and switch time for both OLTC and CBs.

5.2.2 Multi-mode for real-time stage voltage regulation

In the context of real-time voltage regulation, the distribution system operator endeavors to minimize energy consumption while ensuring that bus voltages are maintained within predefined acceptable thresholds within power systems. In scenarios where fast reactive power resources are scarce and voltage margins are tight, the system operator will focus on maintaining voltage levels to mitigate security concerns. Conversely, when reactive power availability or voltage margins are sufficient, a single mode of voltage regulation may neglect economic factors [20], [23], [116]. In such scenarios, a multi-mode voltage regulation approach offers appropriate power support across varied conditions, thereby enhancing practical flexibility. To illustrate the proposed multi-mode voltage control methodology comprehensively, we introduce two distinct criteria: namely, the voltage margin (VM) and the flexible PV margin (PVM) for bus *i* at time *t*, as described as follows.

$$\overline{VM_{i,t}} = (\overline{V} - v_{i,t}) \tag{5.12}$$

$$\underline{VM_{i,t}} = (v_{i,t} - \underline{V}) \tag{5.13}$$

$$PVM_{i,t} = \sqrt{\left(s_{i,t}^{PV}\right)^2 - \left(p_{i,t}^{PV}\right)^2} - \left(q_{i,t-1}^{PV}\right)^2 \tag{5.14}$$

where $\overline{VM_{i,t}}$ and $\underline{VM_{i,t}}$ denote voltage margins when the voltage is close to the upper limit and lower limit, respectively; $PVM_{i,t}$ represents the reactive power margin of PV inverters.

1) Mode 1: Power-Loss Minimization Mode (P_Mode): The P_Mode is developed to curtail the overall power loss of the distribution network while ensuring requisite voltage limits. If the proposed two margins satisfy security levels, the optimization problem will exclusively prioritize power loss minimization through the subsequent objective:

$$\min \sum_{t\in T} \sum_{ij\in B} l_{ij,t} r_{ij} \tag{5.15}$$

$$\text{s.t. (5.2)-(5.9)} \tag{5.16}$$

This objective function seeks to minimize the total power loss by summing the product of the branch current and resistance over all branches and time periods.

2) Mode 2: Under-Voltage Minimization Mode (U_Mode): The U_Mode is structured to guarantee voltage levels when both VM and PVM fall below security thresholds within the distribution system. This implies that the available reactive power resources are approaching exhaustion, increasing the risk of voltage descending below acceptable limits. In this operational mode, the reserved reactive power in the PV inverters will be regulated to sustain voltage levels above a predetermined threshold, as governed by the following model:

$$\min \sum_{t\in T} \left(\sum_{ij\in B} l_{ij,t} r_{ij} + \sum_{i\in N} v_{i,t}^{D} \right) \tag{5.17}$$

$$q_{i,t}^{PV} = q_{i,t}^{PV,r} - \Delta q_{i,t}^{PV} \tag{5.18}$$

$$\text{s.t. (5.2)-(5.9)} \tag{5.19}$$

3) Mode 3: Over-Voltage Minimization Mode (O_Mode): The O_Mode is devised to uphold secure voltage levels, particularly when the VM diminishes and the reactive power reserves of PV inverters are fully depleted. Under such circumstances, the O_Mode orchestrates active power curtailment of PV systems to mitigate voltage escalation issues. To optimize and minimize the overall PV curtailment, the optimization formulation is defined as follows:

$$\min \sum_{t\in T} \left(\sum_{ij\in B} l_{ij,t} r_{ij} + \sum_{i\in N} v_{i,t}^{D} + \sum_{i\in PV} p_{i,t}^{curt} \right) \tag{5.20}$$

$$p_{i,t}^{PV} = p_{i,t}^{PV,r} - p_{i,t}^{curt} \tag{5.21}$$

$$p_{i,t}^{curt} \le p_{i,t}^{curt,\max} = P_{i}^{PV,\max} \left(\frac{\overline{V} - v_{i,t}}{\sum_{n\in PV} \delta v_i / \delta p_n \cdot p_{n,t}^{PV}} + 1 \right) \tag{5.22}$$

$$\text{s.t. (5.2)-(5.9)} \tag{5.23}$$

where the maximum curtailment of each PV inverter, $p_{i,t}^{curt,max}$, in equation (5.22) is given by fairness control among the PV systems based on voltage sensitivities [117]. The detailed shift algorithm of the proposed multi-mode voltage regulation strategy is demonstrated in **Algorithm 4.**

Algorithm 4: Multi-mode voltage control strategy

```
Input: $v_i, p_{i,t-1}^{PV}$ $v_{i,t}$ and.$q_{i,t-1}^{PV}$.
Calculate: $\overline{VM_{i,t}}$, $\underline{VM_{i,t}}$, and $PVM_{i,t}$ using (5.13)-(5.15).
For time step t do
  If $\overline{VM_{i,t}} \leq \gamma_1$ and $PVM_{i,t} \leq \gamma_2$
    Perform O_Mode to address voltage rise problems.
  Else $\underline{VM_{i,t}} \leq \gamma_1$, and $PVM_{i,t} \leq \gamma_2$
    Perform U_Mode to maintain the voltage above a
certain level.
  Else $\overline{VM_{i,t}} > \gamma_1$, $\underline{VM_{i,t}} > \gamma_1$, and $PVM_{i,t} > \gamma_2$
    Perform P_Mode to minimize power loss.
  End If
End For
```

5.3 Methodology

To address the objectives of mitigating rapid voltage violations and minimizing power losses via the developed DRL framework, the two-stage voltage regulation framework is divided into two timescale tasks based on distinct objectives. Subsequently, the two timescale tasks are conceptualized as a single-agent MDP and a multi-agent MDP respectively. The MDP stands as a quintessential paradigm within DRL methodologies, wherein an agent, or potentially multiple agents, engages with an inherently uncertain environment, iteratively refining their control policies through exploration. The orchestrated coordination of this two-stage voltage regulation is realized by concurrently training the two-stage agents, facilitated by information interchange grounded in the reward signals deduced from a data-driven surrogate model. For the slower timescale control, the OLTC and CBs are harmonized through a single agent-driven SAC algorithm, leveraging comprehensive system information. Conversely, the optimization of PV inverters, treated as a fast timescale control, is addressed and resolved by employing the MASAC algorithm.

5.3.1 Day-ahead Voltage Control based on Soft Actor-critic Algorithm

1) MDP Formulation of Day-Ahead Agent

To formulate MDP of the day-ahead voltage control problem, key components for the day-ahead agent (DA) encompass the state set *S*, action set *A*, and reward function *R*. The DA determines the schedule of OLTC and CBs according to the forecasting power demand and PV generation. Therefore, the state set holds comprehensive information regarding the distribution network and is explicitly defined in (5.24). The predicted action a_t^d is defined in (5.25), encapsulating the statuses of OLTC and CBs. Note that, instead of encompassing all decision variables in (5.2)-(5.9), the selected actions in (5.25) are controllable and include the minimum actions to improve the learning convergence and stability. The reward value per time step r_t^d endeavors to reflect the efficacy of actions undertaken by the DA, as specified in equation (5.26), taking into account metrics such as power loss and voltage deviations.

$$s_t^d = \left(p_{i,t}^{Load}, q_{i,t}^{Load}, p_{i,t}^{PV}, q_{i,t}^{PV}\right), \forall d,t \tag{5.24}$$

$$a_t^d = \left(tap_t, u_{i,n,t}\right), \forall t \tag{5.25}$$

$$r_t^d = -\left(\lambda_1 l_{ij,t} r_{ij} + \lambda_2 v_{Di,t}\right), \forall t, ij \tag{5.26}$$

2) SAC Algorithm for Day-Ahead Voltage Control

SAC is an off-policy, actor-critic algorithm in maximum entropy reinforcement learning [105], which concurrently enhances the expected reward and the entropy of the policy to facilitate exploration, i.e.:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau\sim\pi} \sum_{t=0}^{T-1} \gamma^t \left[r + \alpha \mathcal{H}\left(\pi\left(\cdot\middle|s_t\right)\right)\right] \tag{5.27}$$

where τ indicates one trajectory. $\pi(\cdot\,|s_t)$ is a categorical distribution indicating the probability of taking any action under state s_t; α represents the entropy temperature that tunes the stochasticity of the optimal policy; $\mathcal{H}\big(\pi(\cdot\,|s)\big) = -\,log\big(\pi(a_t|s_t)\big)$ denotes entropy term. $\gamma \in [0,1]$ indicates the discounting coefficient. Besides, the exploration and learning stability of the policy is related to the value of entropy temperature. Therefore, one of the technical essential tricks in the SAC algorithm is to automatically adjust the entropy temperature by:

$$J_{\pi}(\alpha) = \mathbb{E}_{a\sim\pi}\left[-\alpha \log\left(\pi(a|s)\right) - \alpha\underline{\mathcal{H}}\right] \tag{5.28}$$

where $\underline{\mathcal{H}}$ denotes the expected minimum entropy. Noteworthy that the action in the day-ahead agent is discrete. Therefore, the discrete SAC algorithm is applied to train our day-ahead control optimization problem. Accordingly, the policy evaluation relies on the actor-critic architecture, wherein the Bellman backup operator is applied for soft Q-function $Q_\theta(s) = r(s_t, a_t) + \gamma\pi_t(s)^T V_\pi(s_{t+1})$ where $\pi_t(s)^T$ indicates the expectation value of the discrete action; $V_\pi(s) = \pi_t(s)^T\big[Q_\theta(s) - \alpha\; log\big(\pi(a|s)\big)\big]$ is the soft state-value function. SAC refines the critic through temporal-difference (TD) learning by minimizing the loss function, in which two critic networks with different parameters θ_1 and θ_2 are implemented to avoid overestimation issues:

$$J_Q(\theta_i) = E_{(s,a,r,s')\sim\mathcal{M}}\left[\frac{1}{2}\Big(Q_{\theta_i}(s) - \big(r + \gamma * \pi_t(s)^T (Q_{\hat{\theta}_i}(s', a') + \alpha\mathcal{H}\;(\pi_\varphi^*(\cdot\,|s'))\big)\Big)^2\right] \tag{5.29}$$

Significantly, SAC utilizes a soft Q-function augmented with an entropy term. The policy is learned through the gradient ascent optimizer, where two target networks for each critic with parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ are used to improve the learning stability

$$J_\pi(\varphi) = \mathbb{E}_{s\sim\mathcal{M}}\left[\pi_t(s)^T\left[\alpha\log\left(\pi_\varphi(a\,|\,s)\right) - \min_{i\in\{1,2\}} Q_{\hat{\theta}_i}(s)\right]\right] \tag{5.30}$$

where φ are the parameters of the actor-network.

5.3.2 Real-Time Voltage Control via Attention-based MASA

1) MDP Formulation of Real-Time Agent

To mitigate the communication burden associated with information exchange, the distribution network is segmented into distinct regional sub-networks based on inherent geographic attributes. Subsequently, each agent is designated to oversee a specific sub-network. The coordination of PV inverters across multiple sub-networks is conceptualized within the framework of MDPs, representing a multi-agent extension thereof. While all agents undergo centralized training to learn a coordinated control strategy, their operational deployment is decentralized, enabling robust decisions grounded in real-time sub-network information. This approach markedly diminishes communication demands and avoids adverse effects on control efficacy stemming from temporal delays. Within the MDP paradigm, each sub-network is formulated as an agent to dispatch PV inverters within its designated domain. The main constituents of this MDP framework are defined subsequently.

$$s_{i,t}^{r}=\left(VM_{i,t},PVM_{j,t},p_{j,t}^{PV},q_{j,t}^{PV},p_{j,t}^{Load},q_{j,t}^{Load},v_{i,t},tap_{t},c_{i,t}\right) \tag{5.31}$$

$$a_{i,t}^{r}=\left(p_{i,t}^{curt},\Delta q_{i,t}^{PV}\right),\forall t \tag{5.32}$$

$$r_{t}^{r}=-\left(\sum\nolimits_{ij\in B}l_{ij,t}r_{ij}+\sum\nolimits_{i\in N}\kappa_{1}v_{i,t}^{D}+\sum\nolimits_{i\in PV}\kappa_{2}p_{i,t}^{curt}\right) \tag{5.33}$$

where s_t^r denotes the state of the agent i in time t. The state s_t^r includes the local observation of sub-network i, which is composed of voltage margin, flexible PV margin, PV output, load demand, voltage in the sub-network i, and the operational statuses of OLTC and CBs across the distribution network. The action $a_{i,t}^r$ represents the strategic control undertaken by agent i at the time t. Specifically, $a_{i,t}^r$ is designed to regulate reactive power and curtail active power across PV inverters situated within sub-network i. The reward r_t^r is the immediate reward subsequent to action execution within the operational environment. Notably, all the agents concur upon a unified reward r_t^r, wherein κ_1 and κ_2 serve as penalty coefficients, addressing deviations in voltage and PV curtailment, respectively.

2) Attention Based MASAC Algorithm for Real-Time Voltage Control

To address the intricacies of the multi-agent MDP problem, this section introduces the MASAC algorithm. Nonetheless, the performance of the multi-agent algorithm suffers from degradation with the increasing number of agents. To ameliorate this issue, an attention mechanism is incorporated into the MADRL framework, enabling each agent to selectively focus on information most pertinent to its corresponding reward structure. The structure of the proposed methodology is illustrated in Fig. 5.2, in which $Q_i^{\psi}(s,a) = f_i(g_i(s_i,a_i),\upsilon_i)$ denotes a function encapsulating the state and action of agent i, augmented by contributions from other agents. Herein, f_i signifies a two-layer multi-layer perceptron, $g_i(\cdot)$ indicates the embedding function pertinent to agent i, and υ_{i} presents the output processed by the attention mechanism, signifying the weighted aggregation of values extracting from other agents:

$$\upsilon_i = \sum_{i\neq j} \ell_i \cdot \mathrm{ReLU}(V \cdot g_i(s_i,a_i)) \tag{5.34}$$

where ReLU denotes the activation function; V stand as the linear transformation matrix.

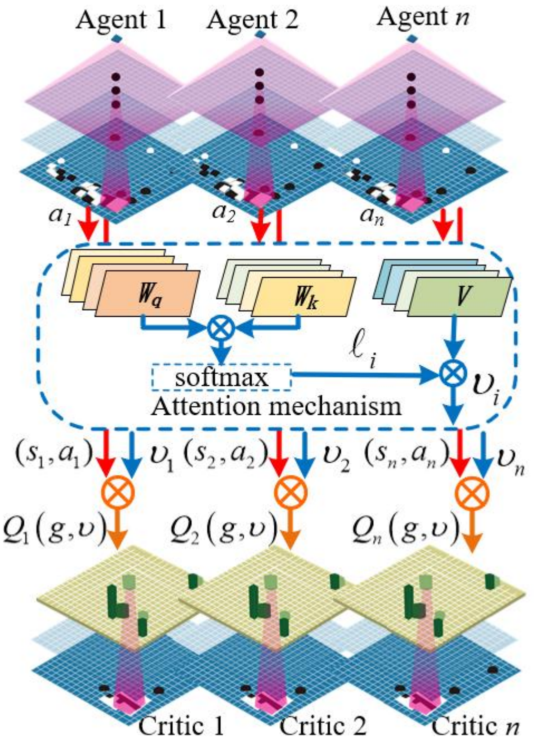


Fig. 5.2 The framework of the proposed attention based MADRL algorithm.

The attention weight ℓ_i evaluates the embedding $g_i(s_i,a_i)$ with $g_j(s_j,a_j)$ through a query-key mechanism:

$$\ell_i \propto \exp(g_j^{\mathrm{T}} W_k^{\mathrm{T}} W_q g_i) \tag{5.35}$$

where W_k and W_q denote the transformation matrices. The computed similarity between two embeddings subsequently undergoes a softmax operation to derive the attention weight ℓ_i. The parameters associated with the attention model, represented as $\langle W_k, W_q, V\rangle$, facilitates a weighted aggregation of contributions from all other agents pertinent to a specific agent. Consequently, the parameters of attention-critic framework comprise both the parameters of the critic function $Q_\theta(s,a)$ and those of attention model $\langle W_k, W_q, V\rangle$. These parameters are refined through optimization techniques aimed at minimizing the ensuing loss function as follows:

$$J_Q(\theta_i) = E_{(s,a,r,s')\sim\mathcal{M}}\left[\frac{1}{2}\left(Q_{\theta_i}(g(s,a),\upsilon) - \left(r + \gamma(Q_{\hat{\theta}_i}(g(s',a'),\upsilon) - \alpha log\,(\pi^*_\varphi(a'|s'))))\right)^2\right] \tag{5.36}$$

The critic function is optimized through the minimization loss among $Q_\theta(s,a)$ and the target. In the policy improvement step, the policy is optimized to maximize the soft Q-function by minimizing the KL-divergence as

$$J_\pi(\varphi) = \mathop{\mathbb{E}}_{s\sim\mathcal{M}}\left[\mathop{\mathbb{E}}_{a\sim\pi}\left[\alpha\log\left(\pi_\varphi(a\,|\,s)\right) - \min_{i\in\{1,2\}} Q_{\theta_i}\left(g(s,a),\upsilon\right)\right]\right] \tag{5.37}$$

which can be minimized by a reparameterization trick. The policy is modified to predict the mean and standard deviation of actions' probability distribution given system states.

Due to the inherent offline training characteristic of the SAC algorithm, the integration of the attention-based MASAC can separately execute centralized training for coordinated strategy and decentralized implementation for voltage regulation. The procedural details of this practical implementation are summarized in **Algorithm 5**.

Algorithm 5: Attention based MASAC

```
Input: the power demand, $p^{Load}, q^{Load}$, PV output
       $p^{PV}, q^{PV}$, and the result from the day-ahead agent.
Initialize: actor network $\varphi_n$, and attention-critic
       network $\theta_{i,n}, \hat{\theta}_{i,n}$ for each agent n.
For each episode do
  For each time step do
```

```
        Generate action $a_n \sim \pi_{\varphi^n}(\cdot\,|s_n = s)$ for each
        agent n, and execute joint action $\mathrm{a} = (\mathrm{a})_1, \cdots, \mathrm{a_n}$
        to obtain reward and next state→ $r, s_n'$
        Store transition $\mathcal{M} \leftarrow \mathcal{M} \cup (s_n, a_n, r, s_n')$ in the
        experience buffer
    **End For**
    **For** each gradient step **do**
        Sample random m experiences from $\mathcal{M}$
        Update soft-Q value parameter by $\theta_{i,n}$ (5.36)
        Update policy parameter $\varphi_n$ by (5.37)
        Adjust temperature $\alpha$ by (5.28)
        Update target $\hat{\theta}_{i,n}$ by $\hat{\theta}_{i,n} = (1-\rho)\hat{\theta}_{i,n} + \rho\theta_{i,n}$
    **End For**
**End For**
```

Note: ρ is the target update factor.

Furthermore, the trained network parameters then are transformed to the real-time stage for voltage regulation. Each agent receives local observation from the sub-network and then executes the voltage regulation in a decentralized manner.

5.4 Case Study

5.4.1 Setting of the Test System

In this section, the proposed two-stage multi-mode voltage regulation strategy is evaluated on a modified IEEE 33-bus distribution network, where six PV inverters were installed at bus 2, 6, 11, 18, 25, 33, respectively, to provide distributed generation and reactive power support and two CBs were added at bus 16 and 22 to help manage reactive power and voltage control. The scalability of the proposed attention-based MASAC framework is an important consideration for its deployment in larger-scale distribution systems. From a training perspective, the framework adopts a centralized training with decentralized execution (CTDE) paradigm: during training, global information is available to stabilize the learning process, while in the inference stage each agent makes decisions solely based on local states and selectively attended neighbor information. This design ensures that the computational complexity of online inference remains manageable even as the number of agents increases. From a communication perspective, the attention mechanism naturally enhances scalability by allowing each agent to focus only on the most relevant neighbors rather than requiring system-wide communication. Such selective information exchange significantly reduces communication overhead, which is particularly critical for large distribution networks where full communication among all controllable units, such as PV inverters, would be impractical.

Firstly, the distribution network is divided into several regional sub-networks according to the default geographic location parameters, with each agent assigned to a specific sub-networks. It's important to note that geographic partition does not inherently ensure voltage control for every bus through local PV inverter adjustments. To address this, an offline evaluation mechanism is established to identify uncontrollable buses following the geographic partition [96]. These uncontrollable buses are then reassigned to an alternative sub-network that has the necessary electrical interconnections. This iterative post-partition adjustment process continues until all buses can be effectively regulated by local resources, as illustrated in Fig. 5.3. The computational analyses presented here were executed on a system equipped with an Intel i7-10700 CPU and 16 GB of RAM. The hyperparameters of the SAC algorithm are presented in Table 5.1. Fig. 5.4 illustrates PV output and load demand across various periods, sourced from online resources. In particular, Fig. 5.4 (a) displays curves of the day-ahead forecast of PV output and load demand for hourly data. This forecast is calculated from the hourly averages of actual data and serves as the foundation for the day-ahead optimization process. Fig. 5.4 (b) depicts real-time minute-by-minute data during a peak PV generation between 13:00 and 14:00, showcasing the variability and dynamics of PV output and load on a finer timescale. Meanwhile Fig. 5.4 (c) presents minute-by-minute real-time data during a period of diminished PV generation and increased load demand from 18:00 to 19:00, highlighting the daily fluctuations in PV output and load demand.

Table 5.1 Main Hyper-Parameters and Data Setting.

Parameters	Value	Parameters	Value
Optimizer	Adam	Activation	RELU
Actor learning rate	1e-3	Critics learning rate	1e-3
Entropy learning rate	1e-3	Targets learning rate	1e-3
Discount factor	0.99	Initial temperature	1
Neurons number	512	Time step	1
Max steps	24	Minibatch size	128
Penalty coefficient λ_1	1	Penalty coefficient λ_2	1e3
Voltage margin limit γ_1	0.005 p.u.	PV margin limit γ_2	70 kVar
Penalty coefficient κ_1	1e3	Penalty coefficient κ_2	1

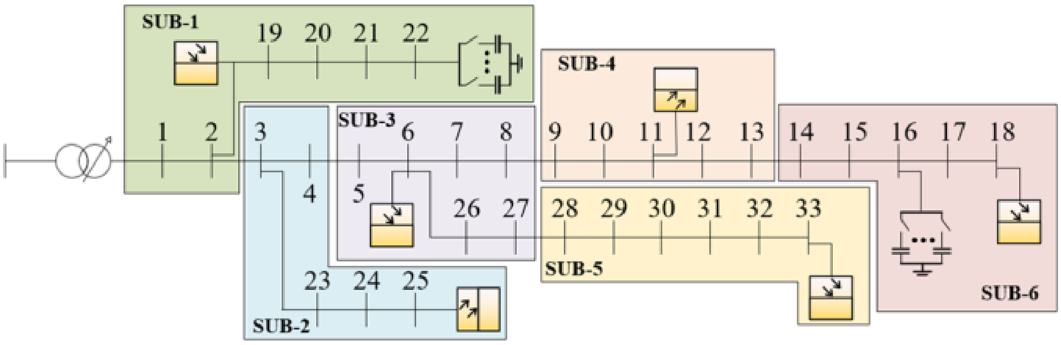
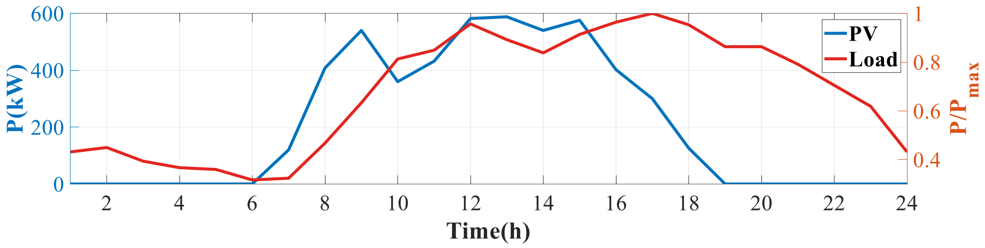
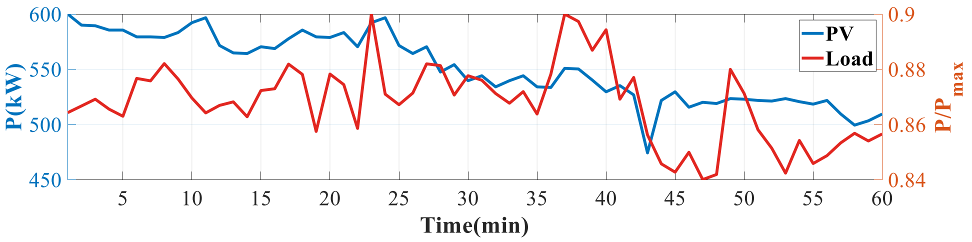


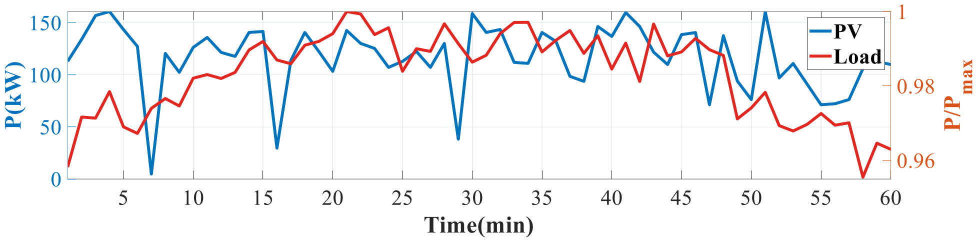
Fig. 5.3 Partition and topology results of the IEEE 33-bus system.



(a) Prediction of PV and load day-ahead hourly data.



(b) Real-time minutely of PV and load data during 13:00-14:00.



(c) Real-time minutely of PV and load data during 18:00-19:00.

Fig. 5.4 PV and load data in distribution network in different timescales.

5.4.2 Numerical Results of Day-Ahead Agent Voltage Regulation

To validate the effectiveness of the day-ahead agent in addressing voltage control problems, voltage profiles following the day-ahead optimization via the DRL algorithm are shown in Fig. 5.5. It is evident that all bus voltages adhere to the security operational range (0.95 p.u. to 1.05 p.u.) throughout the entire day. Notably, at the start and end of the day, the system's voltage distribution progressively converges towards the lower operational threshold, making the system particularly susceptible to fluctuations inherent to real-time operations. Moreover, at midday, the voltage profiles show a reduced margin, approaching the upper operational limit, due to the significant PV output during this time. These observations underscore the need for a real-time voltage control strategy to mitigate voltage infringements arising from system uncertainties and to strengthen the voltage margin during these vulnerable periods.

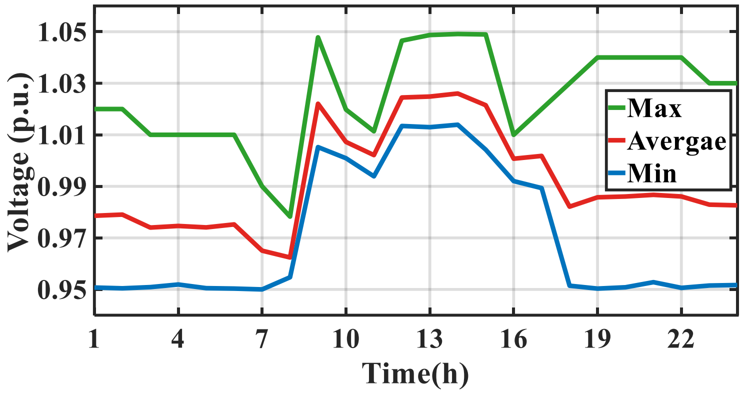


Fig. 5.5 Voltage profiles obtained by the day-ahead agent in the distribution network.

5.4.3 Comparison Results of Real-Time Voltage Regulation with Other Alternative Strategies

To illustrate the effectiveness of the proposed multi-mode voltage control strategy during the real-time stage, comparative analyses were conducted with three benchmark voltage control methodologies for contextual evaluation. The first benchmark control method, referred to as the conventional centralized control method (Method #1), adopts only one agent within the real-time agent for voltage regulation. Following the day-ahead agent's dispatch, this method relies on a single agent to centrally determine the minutely dispatch of PV inverters [118]. The second benchmark control method, referred to as the optimal local control method (Method #2), performs a singular voltage control mode for regulatory purposes. Specifically, this method centralizes the adjustment and curtailment of PV reactive and active power within a single control mode to mitigate voltage violations based on day-ahead optimal dispatch schedules [119]. Lastly, the third benchmark control method, referred to as the original two-stage control method (Method #3), regulates voltage regulation devices with a slow timescale and maintains the day-ahead dispatch constant during the real-time stage.

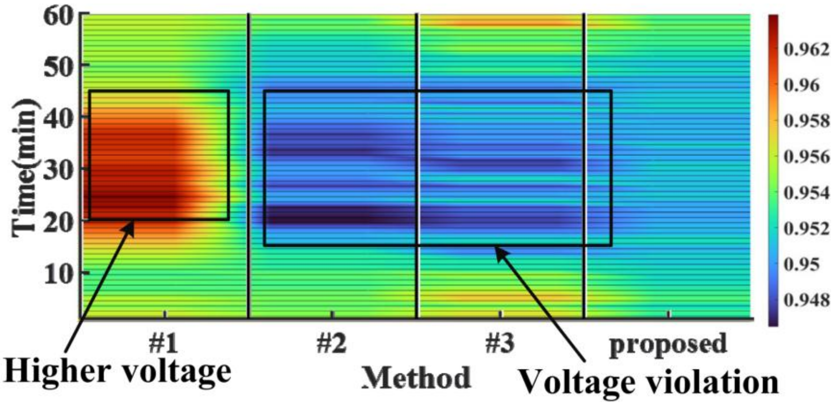


Fig. 5.6 Voltage profiles obtained by different methods during 18:00-19:00.

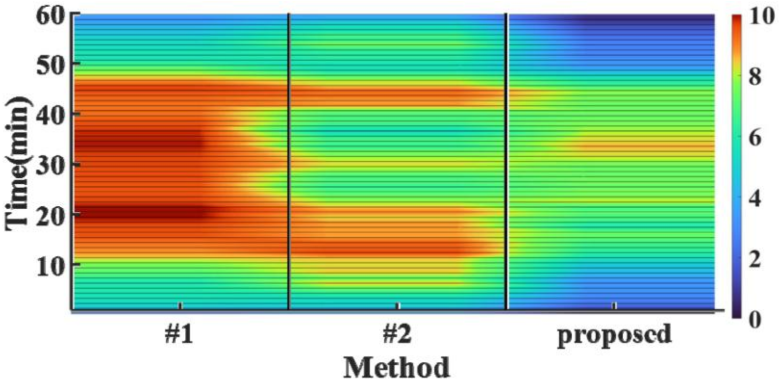


Fig. 5.7 Adjustment of reactive power in PVs with different methods during 18:00-19:00.

Fig. 5.6 presents voltage profiles at bus 18 across distinct voltage regulation methods during the typical time interval of 18:00 to 19:00. Notably, Method #2 shows severe voltage violations, whereas Method #3 displays more moderate violations. These discrepancies arise from Method #2's single mode, which lacks a clear strategy for voltage regulation when voltage margins are constrained. Specifically, when confronted with fewer voltage margins at lower levels, this method may adjust both the active and reactive without a distinct operational mode. In contrast, the proposed method adopts a clear classification of voltage regulation mode, adeptly adjusting the PV reactive power as the voltage margin approaches its lower limit. Furthermore, the persistence of day-ahead dispatch remaining within Method #3 proves inadequate in mitigating voltage violation challenges, primarily due to uncertainties in the real-time stage. On the other hand, Method #1 consistently results in higher voltage levels relative to the proposed method. This discrepancy arises from overoptimization tendencies when a single agent attempts to centrally regulate PV dispatches across different sub-networks. This single agent requires comprehensive knowledge of the entire distribution system, demanding highly communicative capabilities. In contrast, the proposed method only requires local information for the real-time agent, reducing communication requirements and avoiding negative impacts on control performance caused by time delays.

Fig. 5.7 shows the reactive power adjustment of PV across various voltage control methods. Owing to the static strategy inherent in Method #3, adjustments to the reactive power profile for this approach are not considered. As shown in Fig. 5.7, Method #1 exhibits high levels of reactive power adjustments, a result of its inherent tendency towards over-optimization. In contrast, when compared with Method #2, the proposed method exhibits more precise and accurate adjustments to PV reactive power. The clear classification of the voltage control mode facilitates timely voltage support. A comprehensive comparison of outcomes across different voltage regulation methods is further illustrated in Table 5.2. Notably, both Method #2 and Method #3 engender higher power losses, accompanied by unacceptable voltage violations of 0.00931 p.u. and 0.00816 p.u., respectively. Furthermore, Method #3 records a maximum voltage variance of 4.31%, attributable to its real-time control strategy. Compared to Method #1, the proposed approach effectively curtails voltage violations while minimizing power losses.

Table 5.2 Comparison Results for Different Methods.

Methods	Power loss	Voltage Vio.	Voltage
	(kWh)	(p.u.)	Var.
Method #1	27.41	0	3.02%
Method #2	27.70	0.00931	3.19%
Method #3	27.94	0.00817	4.31%
Proposed	26.56	0	3.11%

where the degree of the voltage violation is defined as DCV = $\sqrt{\frac{1}{\mathcal{B}}\sum_{i\in\mathcal{B}}(v_{i,t}^{D})^2}$, $\mathcal{B}$ is the total number of power buses.

To assess the effectiveness of the proposed method in mitigating over-voltage issues, simulations are conducted during the period from 13:00 to 14:00, characterized by heightened PV generation. In Fig. 5.8, the voltage profiles at bus 18 are presented, showcasing the performance of various voltage control methods. While voltage levels and trends differ among methods, both Method #2 and Method #3 exhibit inadequacies in addressing voltage violations. Notably, voltage fluctuations are observed in the profiles of Method #3 between 13:00 and 13:40, underscoring the inherent limitations of relying solely on day-ahead dispatch to manage real-time power system uncertainties. Method #2, while showing improved performance compared to Method #3, falls short in fully mitigating voltage violations due to PV curtailment operations. Conversely, both Method #1 and the proposed method demonstrate superior performance in effectively addressing voltage rise problems. However, Method #1 exhibits lower voltage levels attributed to overoptimization, despite substantial PV curtailment, as depicted in Fig. 5.9, illustrating curtailment of PV generation across distinct voltage control methods. Notably, the adjustment profiles of Method #3 are not presented in this figure. In Fig. 5.9, it is evident that the active power curtailment tendencies of Method #2 and the proposed method align closely. Yet, the proposed method excels in enhancing voltage quality through its clear classification of the voltage control mode. Complementary numerical results are summarized in Table 5.3, where Method #3 incurs a maximum power loss of 22.51 kWh alongside pronounced voltage violations of 0.00907 p.u. In comparison, despite Method #1 curtailing a higher PV active power, resulting in a maximum voltage variance, the proposed method outperforms in terms of minimizing PV active power curtailment and associated power losses, facilitated by its precise classification of operational control modes.

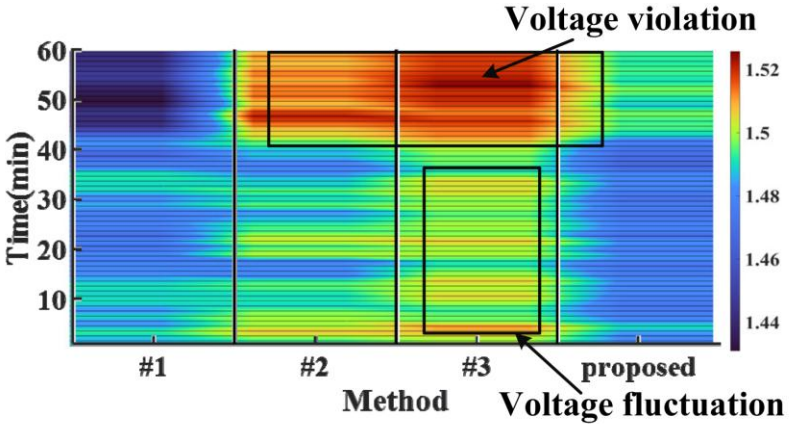


Fig. 5.8 Voltage profiles obtained by different methods during 13:00-14:00.

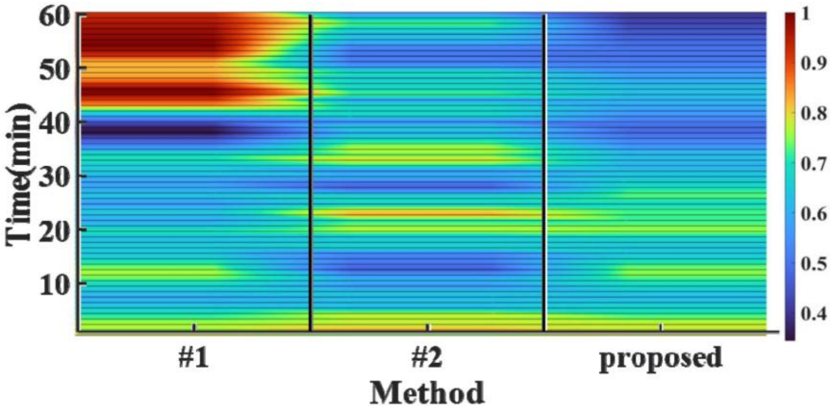


Fig. 5.9 Adjustment of active power in PVs with different methods during 13:00-14:00.

Table 5.3 Comparison Results for Different Methods.

Methods	Power loss (kWh)	Voltage Voi. (p.u.)	Voltage Var.	PV Curtailment (kWh)
Method #1	19.77	0	10.36%	8.2
Method #2	21.66	0.00795	9.86%	6.51
Method #3	22.51	0.00907	9.61%	0
Proposed	19.20	0	9.09%	5.21

5.4.4 Comparison Results with Other Alternative Algorithms

A comparative analysis is conducted on four distinct algorithms, namely: 1) the MASAC algorithm; 2) multi-agent proximal policy optimization (MAPPO), an on-policy algorithm; 3) multi-agent deep deterministic policy gradient (MADDPG), which involves offline training followed by online testing; and 4) attention-based MADDPG (AMADDPG), which integrates the attention mechanism with the MADDPG framework. To ensure robustness and reliability, each algorithm is subjected to ten independent experimental runs using varied initial seeds. The cumulative reward curves resulting from these experiments are illustrated in Fig. 5.10, in which each algorithm is represented by a solid curve denoting the average value across the ten experimental iterations. The shaded region surrounding each curve represents the range between the minimum and maximum rewards obtained across the ten experiments, providing a comprehensive visualization of the performance variability of the algorithms.

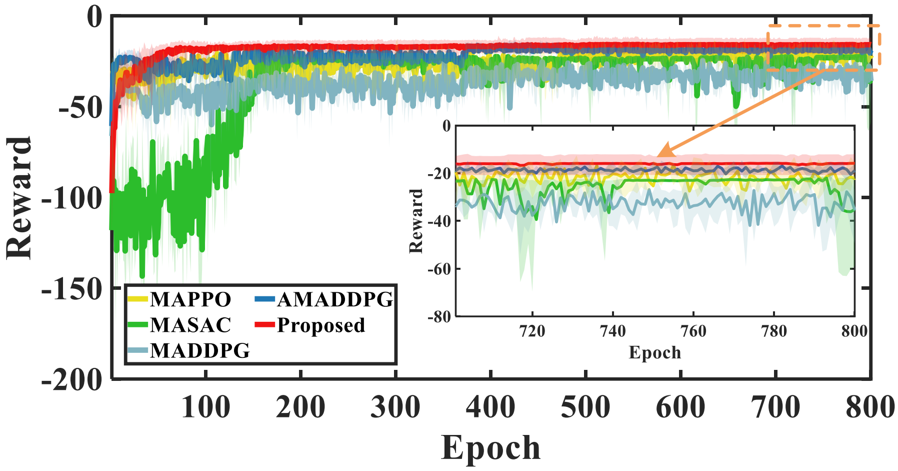


Fig. 5.10 Training process of different algorithms.

From Fig. 5.10, it can be observed that while the cumulative reward reaches high levels with different algorithms, their tendencies vary. Initially, we observe that MASAC and MADDPG exhibit significant oscillations, reflecting the inherent challenges in stabilizing the learning process in a multi-agent environment using off-policy methods. However, as training progresses, these oscillations diminish, with MASAC converging around 350 epochs and MADDPG around 400 epochs, though MADDPG's final reward stabilizes at a higher value of -50 compared to MASAC's -30. In contrast, the on-policy MAPPO demonstrates early convergence around 300 epochs, maintaining relatively low oscillations and a final reward close to -10, showcasing its stability and efficiency. The attention-based variations, AMD and AMS, show marked improvements; AMD converges around 300 epochs with reduced oscillations and a final reward of -40, while AMS displays superior performance with minimal oscillations, early convergence around 250 epochs, and the lowest final reward close to -5. This superior performance of AMS can be attributed to the enhanced capability of attention mechanisms in handling complex interactions, leading to more effective optimization in power system dispatch.

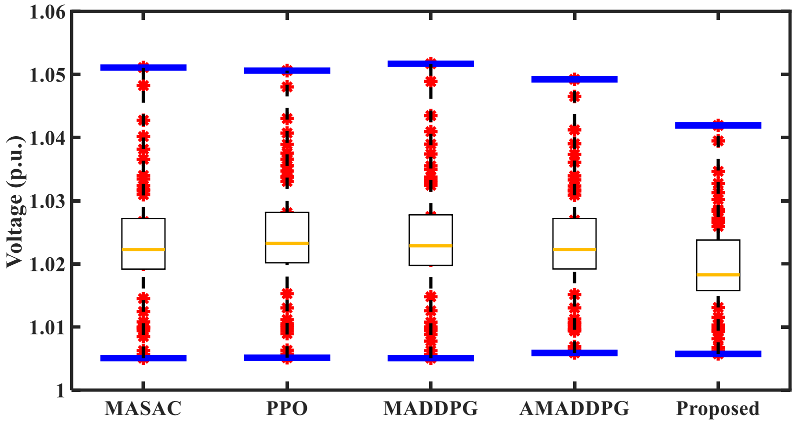


Fig. 5.11 Voltage distribution obtained by different algorithms when t=9:00.

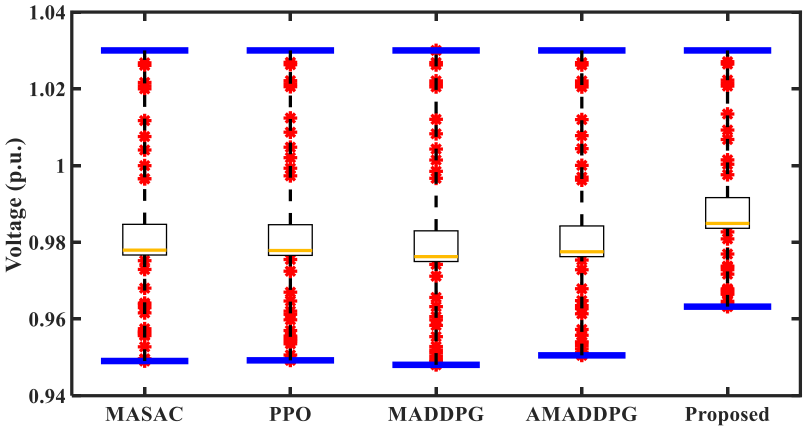


Fig. 5.12 Voltage distribution obtained by different algorithms when t=23:00.

Figs. 5.11 and 5.12 demonstrate the voltage distributions across all buses as generated by the proposed algorithm at specific time instances: t=9:00 and t=23:00, respectively. Notably, the MADDPG algorithm exhibits pronounced voltage violations across both temporal scenarios, attributable to challenges due to hyperparameter calibration. Conversely, while the AMADDPG algorithm adeptly mitigates voltage violation, it slightly lags in voltage margin efficacy relative to the proposed method. This superior performance of the proposed algorithm can be attributed to its integrated attention mechanism during the training process, coupled with stability attributes inherited from the foundational SAC algorithm.

Table 5.4 Training Results with Different Algorithms.

Methods	Power loss	Voltage Vio. (p.u.)	Computation time (s)
MASAC	22.9376	0.00329	2456.72
MAPPO	22.0715	0.00435	1628.16
MADDPG	34.2096	0.01094	2347.01
AMADDPG	17.8841	0.00092	2591.93
Proposed	17.5993	0	2752.01

Table 5.4 offers a numerical exposition of training outcomes across varied algorithms. The MAPPO algorithm, leveraging an on-policy approach, achieves fast computational time; however, this strategy concurrently engenders unacceptable voltage infractions. In contrast, the MADDPG algorithm causes a maximum voltage violation of 0.01094 p.u., accompanied by a maximum power loss. Relative to the AMADDPG algorithm, the proposed algorithm adeptly curtails voltage violations while minimizing power losses. Although the proposed algorithm leads to the lengthiest computational time, it is acceptable, especially when leveraging offline training paradigms for optimizing voltage control strategies.

5.4.5 Scalability of the Proposed Method

The scalability of the proposed voltage regulation strategy is tested on the IEEE 123-bus system, with parameters data obtained from [120]. The convergence curves of the cumulative reward based on different DRL algorithms for the IEEE 123-bus system are presented in Figs. 5.13. Similarly, the solid curve in each algorithm corresponds to the average value of ten independent experiments, and the light-colored shadow area is bounded by the minimum and maximum rewards over the experiments. Initially, MAPPO exhibits significant oscillations but stabilizes around a -145 reward after approximately 100 epochs, showing moderate stability. Similarly, MASAC, with comparable initial oscillations, stabilizes slightly later, around 150 epochs, with a final reward of around -147. On the other hand, MADDPG starts with high oscillations but stabilizes around 200 epochs, converging to a -148 reward and indicating less stability than both MAPPO and MASAC. Furthermore, AMADDPG shows the highest initial oscillations and stabilizes around 300 epochs, with a final reward of approximately -149, suggesting slower convergence and less stability. In contrast, the proposed method stands out with rapid convergence, stabilizing around -145 reward within the first 50 epochs, and maintains the lowest and most stable reward, thus indicating superior performance. These differences can be attributed to the complexity of each algorithm. Specifically, the proposed method incorporates advanced techniques, such as attention mechanisms leading to faster and more stable convergence. Algorithms that balance exploration and exploitation effectively, such as the proposed method and MAPPO, tend to achieve better performance, while stability mechanisms like clipping and entropy regularization further contribute to the superior results observed. Consequently, the proposed method outperforms the others, making it the most effective for minimizing power loss and voltage violation in this scenario.

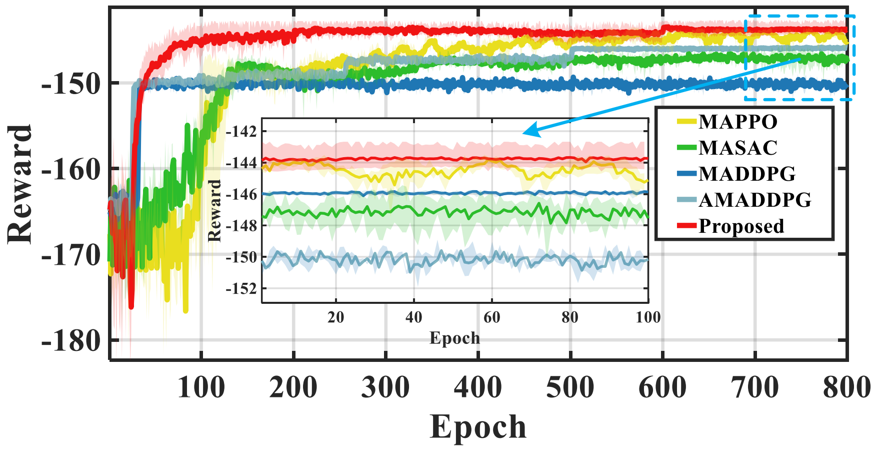


Fig. 5.13 Training process of different algorithms.

To demonstrate the superiority of the proposed algorithm, Table 5.5 illustrates the detailed computational results of different DRL algorithms in the IEEE 123-bus system. The proposed method shows the best overall performance in terms of voltage violation, achieving a perfect score of 0, which indicates a complete elimination of voltage violations. Although it has a slightly higher power loss (110.28) compared to MAPPO (107.43) and MADDPG (108.82), the significant advantage of eliminating voltage violations cannot be overlooked. Meanwhile, MAPPO demonstrates the lowest power loss at 107.43 and a minimal voltage violation of 0.00043, making it a strong contender, though it falls short of the proposed method in completely eliminating voltage violations.

On the other hand, MASAC and AMADDPG exhibit higher power losses (111.32 and 110.77, respectively) and more significant voltage violations (0.00205 and 0.00143, respectively). These results suggest that while they are somewhat effective in reducing power loss, their ability to minimize voltage violations is less effective compared to the proposed method. MADDPG shows a balanced performance with a power loss of 108.82 and a voltage violation of 0.00197. It performs better than MASAC and AMADDPG in terms of voltage violation but still falls short when compared to the proposed method.

Table 5.5 Training Results with Different Algorithms.

Methods	Power loss	Voltage Vio. (p.u.)	Computation time (s)
MASAC	111.32	0.00205	4422.10
MAPPO	107.43	0.00043	3988.99
MADDPG	108.82	0.00197	4224.62
AMADDPG	110.77	0.00143	4665.47
Proposed	110.28	0	4985.62

5.5 Summary

This paper proposes a two-stage voltage control strategy to alleviate fast voltage violations in ADN by coordinating PV inverters and traditional voltage control devices, including OLTC and CBs. In the first stage, the dispatches of OLTC and CBs are determined by a discrete SAC algorithm. In the second stage, a novel multi-mode voltage control method is designed to dispatch the output of PV inverters in real-time operation, achieving a minimized power loss and secure voltage profile simultaneously. An attention-based MASAC algorithm is then proposed to optimize the real-time dispatch of multiple PV resources, which enables each PV inverter to regulate the voltage with only local information. This algorithm helps alleviate the performance degradation associated with a large number of agents in typical MARL algorithms. In case studies, the proposed control strategy is compared with benchmark control strategies. The simulation results show that the proposed multi-mode voltage control method can more precisely dispatch the output of PV inverters and achieve the balance between voltage violation mitigation and power loss minimization. Moreover, the performance of the attention-based MASAC is demonstrated by comparison with benchmark MARL algorithms. It shows that MASAC addresses the performance degradation by facilitating information exchange during off-line training, and the voltage regulation strategy generation by the proposed algorithm mitigates the power loss by 13.53% and reduces the voltage constraint violation by 7.07% compared with benchmark MARL algorithms.

Chapter 6 Coordinated Transmission-Distribution Load Restoration under N-k Contingencies: A Distributed Optimization and Reinforcement Learning Approach

Ensuring the rapid restoration of loads in transmission and distribution (T&D) systems under emergency conditions is crucial for maintaining grid stability. This study addresses the challenge of load restoration when contingencies, such as the disconnection of transmission lines and generators, disrupt the power supply. To address this issue, a coordinated T&D system operation strategy is introduced in this work, where virtual power plants (VPPs) within the distribution system are leveraged to compensate for the curtailed loads, thereby supporting the transmission system's load-shedding efforts. The coordination process involves bidirectional information exchange: the transmission system communicates load-shedding decisions to the distribution system, while the distribution system provides the available maximum curtailment capacity through VPPs. This interaction enhances the system's ability to respond to N-*k* contingency events in an optimized manner, improving overall resilience. To achieve efficient decision-making in this coordinated framework, reinforcement learning techniques are employed to optimize load restoration under N-*k* contingencies. The transmission system is modeled using the soft actor-critic (SAC) algorithm, which determines optimal load-shedding and generator dispatch strategies for rapid system recovery. Meanwhile, the distribution system, responsible for managing multiple VPPs, is controlled using the complementary attention for multi-agent SAC (CMS) algorithm. This approach mitigates the common attention dispersion problem in multi-agent SAC implementations, ensuring optimal decision-making in dynamic multi-agent environments. Simulation results demonstrate that the proposed reinforcement learning-based framework effectively reduces constraints violation in the transmission system while maintaining load supply and voltage stability in the distribution network.

6.1 Framework

The proposed framework integrates the coordination of T&D systems under N-*k* contingency scenarios to enhance power system resilience, as illustrated in Fig. 6.1. A key aspect of this coordination lies in the dynamic exchange of information between the two systems. The distribution system first provides the transmission system with its maximum potential load curtailment capacity, allowing the transmission system to make informed load-shedding decisions. Once the transmission system determines the curtailment strategy, the distribution system immediately adjusts the output of VPPs to compensate for the curtailed loads, ensuring stable power supply within the distribution network. While VPPs play a crucial role in this process, the core of the coordination mechanism is the bidirectional interaction between the transmission and distribution systems, enabling adaptive, efficient, and resilient power dispatch in response to N-*k* contingency scenarios.

As shown in Fig. 6.1, on the transmission side, the load restoration problem for handling N-*k* contingency scenarios is formulated as a Markov decision process (MDP) and solved using a DRL-based approach. The agent receives predicted power demand as the state and generates control actions to adjust the power dispatch and load curtailment, thereby improving system robustness under N-*k* contingency scenarios. These actions involve load-shedding at specific buses, satisfying constraints and load restoration. A simulation environment models the transmission network's operation, providing the cost of the solution to update the agent's policy through offline training. By iteratively refining the policy with state-action-reward feedback, the framework achieves a stability power flow operation and power supply for the transmission system under N-*k* contingency scenarios.

On the distribution side, after coordinating with the transmission system to handle emergency scenarios and implement load curtailments, the framework enables the distribution system to work in cooperation with VPP centers to restore its load supply. The activate distribution network mechanism coordinates individual VPPs, collecting flexible

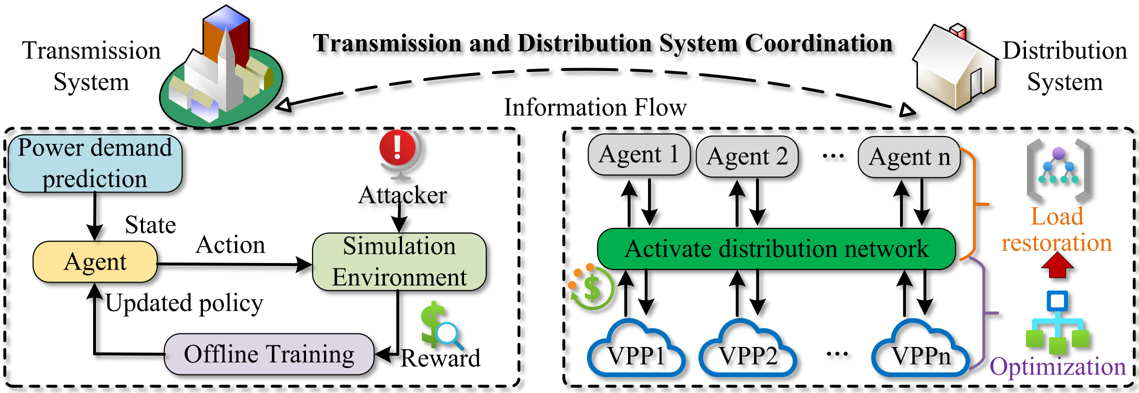


Fig. 6.1 Decentralized Coordination Framework for Transmission and Distribution System.

load supply capacities from VPP centers while offering economic incentives. Based on the available load supply capacity from VPP centers and the load curtailment decisions transmitted from information flow, a multi-agent DRL algorithm is then employed to dynamically adjust the distributed generators within each VPP center, ensuring an optimal power supply to the distribution system. To address the limitations of traditional MADRL algorithms with increasing agent numbers, this framework introduces the CMS algorithm, enhancing learning efficiency. This enables the distribution network to meet load demands effectively during emergencies while optimizing economic returns for VPPs.

The proposed framework establishes a coordinated mechanism between the T&D systems by facilitating efficient information exchange. Through this information flow, both systems can access critical decision-making information, enabling the transmission system to implement informed load curtailment strategies while allowing the distribution system to respond effectively. In particular, the distribution system collaborates with VPP centers, utilizing flexible distributed generators to dynamically adjust power output in response to the transmission system's curtailment decisions. This coordinated approach enhances the system's ability to mitigate N-k contingency scenarios, ensuring adaptive, resilient, and efficient load restoration. Additionally, the adoption of the CMS algorithm addresses the limitations of traditional multi-agent reinforcement learning by improving scalability and learning efficiency. The complementary attention mechanism ensures optimal decision-making across multiple VPP control centers, resulting in robust and efficient distribution system operations during emergencies. Overall, the proposed TSO framework combines robust power flow optimization for the transmission system with flexible load restoration in the distribution system, providing a comprehensive, resilient, and economically viable solution for maintaining grid stability under N-k contingency scenarios.

6.2 Problem Formulation

6.2.1 Mathematical Model of the Transmission-Level System

This section focuses on constructing the objective function and associated constraints for the transmission-level system to ensure robust operation during emergency scenarios. The aim is to optimize the system's performance by minimizing the objective function under emergency condition S. The objective function consists of three key components: the generation cost of power in the transmission network, the penalty for load shedding at critical load buses, and the penalty for providing insufficient power to load buses during emergencies.

$$\min_{\mathcal{S}\in\Xi}\sum_{\forall t}\left[\sum_{\forall g\in\mathcal{G}} C_g p_{g,t}^T + \sum_{\forall l\in\mathcal{L}} C_l \Delta p_{l,t}^T + \sum_{\forall l\in\mathcal{L}} C_l^* \Delta p_{l,t}^{T*}\right] \tag{6.1}$$

where $p_{g,t}^T$ is the output of generator at time t and C_g is its operation cost; $\Delta p_{l,t}^T$ is the load shedding at time t and C_l is its penalty cost; $\Delta p_{l,t}^{T*}$ is the unserved electricity at time t and C_l^* is its penalty cost. The objective function is subject to the following constraints to ensure that the transmission network operates within its physical and operational limits during emergency scenarios. Equations (6.2)-(6.3) ensure the nodal power balance constraints. Equation (6.4) used to calculate the bus voltages. Equation (6.5) calculates power flow. Equations (6.6)-(6.8) limit the power output of generators to their operational limits. Equations (6.9)-(6.10) are restricted the voltage magnitude and phase angle at each bus. The maximum amount of load shedding is constrained at Equations (6.11). Equations (6.12) constraints model the disconnection of k transmission lines and generators from the network under N-k contingency scenarios.

$$p_{g,n,t}^T + \sum_{mn\in\mathcal{D}} p_{mn,t}^T = \sum_{nk\in\mathcal{D}}\left(l_{nk,t}^T r_{nk}^T + p_{nk,t}^T\right) + (p_{l,n,t}^T - \Delta p_{l,n,t}^T - \Delta p_{l,n,t}^{T*}), \forall m,n,t \tag{6.2}$$

$$q_{g,n,t}^{T} + \sum_{mn \in \mathcal{D}} q_{mn,t}^{T} = \sum_{nk \in \mathcal{D}} \left(l_{nk,t}^{T} x_{nk}^{T} + q_{nk,t}^{T} \right) + q_{l,n,t}^{T}, \forall m,n,t \tag{6.3}$$

$$v_{n,t}^{T} = v_{m,t}^{T} - 2\left(r_{mn}^{T} p_{mn,t}^{T} + x_{mn}^{T} q_{mn,t}^{T} \right) + \left((r_{mn}^{T})^{2} + (x_{mn}^{T})^{2} \right) l_{mn,t}^{T}, \quad \forall mn,t \tag{6.4}$$

$$(p_{mn,t}^{T})^{2} + (q_{mn,t}^{T})^{2} = v_{m,t}^{T} l_{mn,t}^{T}, \quad \forall m,n,mn,t \tag{6.5}$$

$$\underline{P}_{g}^{T} \le p_{g,t}^{T} \le \overline{P}_{g}^{T}, \forall g,t \tag{6.6}$$

$$\underline{Q}_{g}^{T} \le q_{g,t}^{T} \le \overline{Q}_{g}^{T}, \forall g,t \tag{6.7}$$

$$-RD_{g} \le p_{g,t}^{T} - p_{g,t-1}^{T} \le RU_{g}, \forall g,t \tag{6.8}$$

$$\underline{V}_{m}^{T} \le v_{m,t}^{T} \le \overline{V}_{m}^{T}, \forall m,t \tag{6.9}$$

$$\underline{\Theta}_{m}^{T} \le \theta_{m,t}^{T} \le \overline{\Theta}_{m}^{T}, \forall m,t \tag{6.10}$$

$$0 \le \Delta p_{l,t}^{T} \le p_{l,t}^{T}, \forall l,t \tag{6.11}$$

$$\mathcal{S} = \{ s \in \{0,1\} \mid \sum_{\forall ij} s_{mn,t} + \sum_{\forall g} s_{g,t} \le k, I_{mn,t} = 1 - s_{mn,t}, I_{g,t} = 1 - s_{g,t}, \forall mn, g, t \} \tag{6.12}$$

where $p_{g,n,t}^{T}$ and $p_{mn,t}^{T}$ are the power input into bus *n* at time *t* from generator *g* and tranmission line mn, respecitvely, and $q_{g,n,t}^{T}$ and $q_{mn,t}^{T}$ are the reactive power input; l_{nk}^{T} and l_{nk}^{T} is square of bus voltage and current; r_{nk}^{T} and x_{nk}^{T} are resistance and reactance of transmission line mn; ($\underline{P_g^T}$ $\overline{P_g^T}$) and ($\underline{Q_g^T}$ $\overline{Q_g^T}$) are active and reactive power limits of generator, repsecitvely; *RD*$_g$ and *RU*$_g$ are ramping up and down limits of generator, repsecitvely; ($\underline{V_m^T}$ $\overline{V_m^T}$) and ($\underline{\Theta_m^T}$ $\overline{\Theta_m^T}$) are voltage and angle phase limits, repsecitvely; s_g and s_{mn} are status of generator and transmission line, repsecitvely; I_g and I_{mn} are availability of generator and transmission line, repsecitvely. To accelerate the solution process for the robust optimal power flow (OPF) problem, a worst contingency scenario is generated by previous preventive security-constrained method to simulate the disconnection of *k* components [111]. This method captures a wide range of possible contingencies without the computational burden of worst contingency scenario identification, significantly enhancing the efficiency of solving the robust OPF problem while maintaining robustness against common and severe contingencies.

6.2.2 Mathematical Model of the Distribution-Level System

This section focuses on constructing the objective function and corresponding constraints for the distribution-level system to ensure efficient and cost-effective operation, particularly during emergency scenarios. The objective function is designed to minimize the overall operational cost of the distribution network, which consists of two key components: the power losses within the distribution network and the operational cost of purchasing power from the VPP centers.

$$\min \sum_{\forall t} \left(\sum_{mn \in \mathcal{D}} l_{mn,t}^{D} r_{mn}^{D} + \sum_{\forall v \in \mathcal{V}} C_v p_{v,t} \right) \tag{6.13}$$

where superscript represents the variable in the distribution system; $p_{v,t}$ is the output of the VPP at time *t* and C_v is its operation cost. The objective function for the distribution-level system is subject to the following constraints, ensuring that the distribution network operates within its physical and operational limits while providing reliable power supply and minimizing costs. Equations (6.14)-(6.15) ensure the nodal power balance constraints. Equation (6.16) is used to calculate the bus voltages. Equation (6.17) calculates power flow. Equation (6.18) is restricted the voltage magnitude at each bus. Equation (6.19) calculates the substation voltage based on the OLTC positioning.

$$p_{v,n,t} + \sum_{mn \in \mathcal{D}} p_{mn,t}^{D} = \sum_{nk \in \mathcal{D}} \left(l_{nk,t}^{D} r_{nk}^{D} + p_{nk,t}^{D} \right) + p_{l,n,t}^{D}, \forall m,n,t \tag{6.14}$$

$$q_{v,n,t} + \sum_{mn \in \mathcal{D}} q_{mn,t}^{D} = \sum_{nk \in \mathcal{D}} \left(l_{nk,t}^{D} x_{nk}^{D} + q_{nk,t}^{D} \right) + q_{l,n,t}^{D}, \forall m,n,t \tag{6.15}$$

$$v_{n,t}^{D} = v_{m,t}^{D} - 2\left(r_{mn}^{D} p_{mn,t}^{D} + x_{mn}^{D} q_{mn,t}^{D} \right) + \left((r_{mn}^{D})^2 + (x_{mn}^{D})^2 \right) l_{mn,t}^{D}, \quad \forall mn,t \tag{6.16}$$

$$(p_{mn,t}^{D})^2 + (q_{mn,t}^{D})^2 = v_{m,t}^{D} l_{mn,t}^{D}, \quad \forall m,n,mn,t \tag{6.17}$$

$$\underline{V}_{m}^{D} \le v_{m,t}^{D} \le \overline{V}_{m}^{D}, \forall m,t \tag{6.18}$$

$$v_{1,t}^{D} = \left(V_s + tap_t \cdot \Delta V_T \right)^2, \forall t \tag{6.19}$$

where V_s are the primary voltage of transformer at the slack bus; tap_t is the status of OLTC at time *t*; ΔV_T is voltage regulation of OLTC for one-tap step.

6.2.2 Optimization Model for Virtual Power Plants

The introduction of VPPs into the distribution network enhances system flexibility and resilience, particularly during emergency scenarios. As illustrated in Fig. 6.2, VPPs act as aggregators of DERs, such as photovoltaic (PV) systems, battery energy storage systems (BESS), and controllable loads (e.g., electric vehicles). These resources are centrally managed through a unified VPP dispatch strategy, which reduces the complexity of directly controlling individual DER responses to distribution system energy dispatch commands.

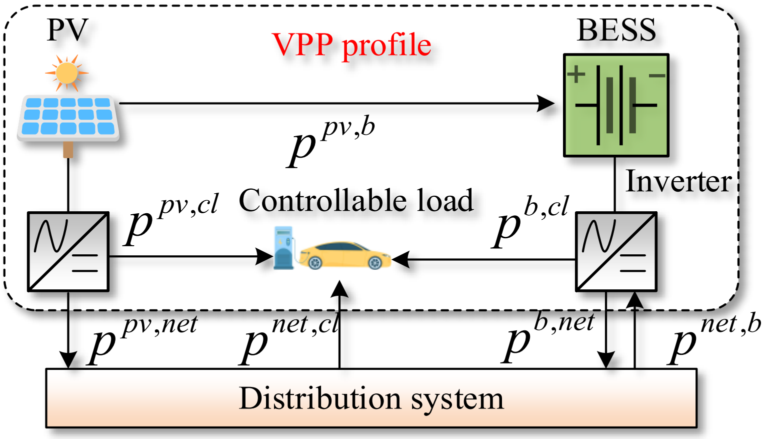


Fig. 6.2 The structure of the VPP profile.

The optimization model for VPPs aims to maximize the economic operating benefits of the VPP centers while ensuring reliable power supply to the distribution network during emergency conditions. The objective function of VPP optimization is designed to maximize the net revenue, defined as the difference between energy sales and operating costs. The first term is revenue from energy sales to the distribution system; the second and third terms are operating cost of PV and BESS systems; the last term is energy consumption of controllable loads.

$$\max \sum_{\forall t} \left(C_v p_{v,t} - C_{pv} p_t^{pv} - C_b p_t^b - C_c p_{cl,t} \right) \tag{6.20}$$

where p_t^{pv} is the output of PV; p_t^b is the output of BESS; C_{pv} and C_b are management cost of PV and BESS, respectively; $p_{cl,t}$ is the power demand of the controllable load and C_c is its cost. The optimization of the VPP is subject to several constraints to ensure that its operation remains within technical and operational limits while achieving the economic objective. Constraint (6.21) calculates the net power output of VPP sells to the distribution system. Constraint (6.22) ensures total PV production. Constraint (6.23) calculates the power consumption of the controllable load. Constraints (6.24)-(6.26) regulate the operation of the BESS and govern the state of charge (SOC) of the BESS.

$$p_{v,t} = p_t^{pv,net} + p_t^{b,net} - p_t^{net,b} - p_t^{net,cl} \tag{6.21}$$

$$p_t^{pv} = p_t^{pv,net} + p_t^{pv,cl} + p_t^{pv,b} \tag{6.22}$$

$$p_t^{cl} = p_t^{net,cl} + p_t^{b,cl} + p_t^{pv,cl} \tag{6.23}$$

$$p_t^{b} = p_t^{b,cl} / \lambda^p + p_t^{b,net} / \lambda^p - \lambda^c p_t^{pv,b} - \lambda^c p_t^{net,b} \tag{6.24}$$

$$\lambda^p + \lambda^c \leq 1 \tag{6.25}$$

$$\mathrm{SOC}_t^b = \begin{cases} \mathrm{SOC}_{t-1}^b + \lambda^c p_t^b \Delta t & p_t^b > 0 \\ \mathrm{SOC}_{t-1}^b + p_t^b \Delta t / \lambda^p & p_t^b < 0 \end{cases} \tag{6.26}$$

where $p_t^{pv,net}$, $p_t^{pv,cl}$, and $p_t^{pv,b}$ are the power that PV sends to distribution system, controllable load and BESS, respectively; $p_t^{net,cl}$ and $p_t^{b,cl}$ are the power consumption of controllable load from distribution system and BESS, respectively; $p_t^{b,net}$ and $p_t^{net,b}$ are the power output and input from BESS to distribution system; λ^p and λ^c are the discharge and charge efficiency of BESS; SOC_t^b is the state of charge of the BESS at time *t*. The VPP optimization model is designed to determine the feasible range of power supply that VPP centers can dispatch within the distribution system during emergency scenarios.

6.3 Methodology

Building upon previous mathematical models for both the T&D systems, traditional optimization algorithms, such as mixed-integer programming and gradient-based methods, are often limited in handling the high-dimensional, nonlinear, and stochastic characteristic of power system coordination, especially under contingency scenarios where uncertainty and rapid decision-making are critical. To overcome these limitations, DRL algorithms are introduced, offering the ability to learn optimal policies directly from interaction with the system without requiring an explicit mathematical model of all uncertainties. To achieve this, the control problems in both the transmission and distribution systems are formulated as MDPs. The following sections present the MDP formulation, the SAC algorithm for the transmission system, and the CMS algorithm for the distribution system in detail.

6.3.1 MDP characteristics in agent

To apply a reinforcement learning approach, the optimization problems and control tasks are reformulated as MDPs, where one or more agents interact with an uncertain environment to gradually improve their control policies while exploring this environment. Unlike commonly adopted simple MDP models, which typically involve a single agent or multiple agents cooperating on the same task, this work develops two specialized MDPs tailored to address the coordination challenges of the T&D systems under contingency scenarios. Specifically, the transmission system agent (TA) aims to minimize the operational cost by providing robust and resilient control actions a_t^T to against all possible contingency scenarios. Meanwhile, the distribution system agents (DA) are formulated as a multi-agent system to enable distributed control of the VPPs dispersed across the distribution network. The DA seeks to minimize operational costs by determining the optimal economic dispatch a_t^D , while ensuring sufficient load restoration to ensure power supply, effectively coordinating the operation of the transmission system under contingency scenarios.

To construct the MDP model, the key components involving the TA, DA, and the environment are defined as follows. The TA generates robust control actions a_t^T based on the policy $\pi^T(s_t^T)$, aiming to maximize the cumulative discounted reward $\sum_t \gamma^{n-1} r_n^T(s^T, a^T)$. Therefore, the main elements of the MDP can be described using the tuple (s_t^T , a_t^T , r_t^T , γ, Γ^T). The state s_t^T consists of the input features, including active and reactive power demands and the maximum load curtailment capacity of load buses connected with distribution system, as defined in (6.27). The determination of this maximum load curtailment capacity is an optimization problem that can be addressed using well-established methods [121]. Then, the predicted action a_t^T , defined in (6.28), is designed to curtail power load at power system nodes to mitigate the impact of emergency events. Instead of using all the decision variables from equations (6.2)-(6.11), the selected actions in (6.28) are chosen to be controllable and minimal, ensuring faster convergence and improved learning stability. The reward value r_t^T at each time step reflects the effectiveness of the action taken by the TA and is defined in (6.29) to include all relevant operational costs, such as generation costs, load curtailment penalties, unserved electricity penalties, and penalties for operational violations. The discount factor γ is used to calculate the cumulative reward over time, while the transition function Γ^T describes how the system evolves based on the current state and action, which will be learned by the reinforcement learning algorithm.

$$s_t^T = \left(p_{l,t}^T, q_{l,t}^T, p_{l,t}^{ca}, \forall l \in \mathcal{L} \right), \forall t \tag{6.27}$$

$$a_t^T = \left(\Delta p_{l,t}^{T*}, \forall l \in \mathcal{L} \right), \forall t \tag{6.28}$$

$$r_t^T = \left(\sum_{\forall g \in \mathcal{G}} C_g p_{g,t}^T + \sum_{\forall l \in \mathcal{L}} C_l \Delta p_{l,t}^T + \sum_{\forall l \in \mathcal{L}} C_l^* \Delta p_{l,t}^{T*} + \kappa v_t^T \right), \forall t \tag{6.29}$$

where $p_{l,t}^{ca}$ represents the maximum load curtailment capability at load nodes; while v_t^T indicates the degree of operational violations in the transmission system, with κ denoting its penalty coefficient. Similarly, the main components of the MDP in the DA can be defined using the tuple (s_t^D, a_t^D, r_t^D, γ, Γ^{D}). Since each VPP center operates relatively independently, the distribution system is managed by multiple DAs, with each DA responsible for controlling and dispatching power from its respective VPP center to supply electricity to load buses in the distribution system. The states s_t^D include the robust action taken by the TA, the active and reactive power demands within the distribution network, and the available power capacity of the controllable VPPs, as defined in (6.30). However, these agents operate under partial state, meaning that each agent can only access state about the bus load conditions in the neighborhood of its assigned VPP center. The extent of this observability is determined by the observation region $\mathcal{R}$, which defines the subset of the distribution network that an agent can observe. While a larger observability range can potentially improve the learning efficiency of agents, an excessive or redundant information can lead to distracted learning and inefficient coordination, as agents struggle to focus on critical decision variables. To address this issue, the CMS algorithm is employed to enhance attention-driven state processing, ensuring that each agent selectively focuses on the most relevant state within its observation region $\mathcal{R}$. The detailed introduction to the CMS algorithm and its implementation is provided in Section 6.3.3. Then the predicted action a_t^D, described in (6.31), is used to adjust the VPP dispatch to provide sufficient power to the load buses in the distribution network. Each VPP center is scheduled by the $a_{i,t}^D$ generated by its corresponding DA. This action compensates for any power deficits caused by load shedding in the transmission system while responding to emergency contingencies. The reward value r_t^D at each time step is defined in (6.32) and incorporates several factors, including power losses in the distribution network, the cost of power supplied by VPP centers, and penalties for voltage violations. These reward components are designed to guide the DA in optimizing the operation of VPPs, ensuring efficient power supply and system stability during emergencies. Notably, all DAs share the same reward function. These DAs collaborate to minimize the operational costs of the distribution system when operating in coordination with the transmission system during emergency situations.

$$s_{i,t}^D = \left(a_t^T, p_{i,l,t}^D, q_{i,l,t}^D, p_{i,v,t}^{\max}, \forall l \in \mathcal{L}\right), \forall t \tag{6.30}$$

$$a_{i,t}^D = \left(p_{i,v,t}, \forall v \in \mathcal{V}\right), \forall t \tag{6.31}$$

$$r_t^D = \left(\sum_{mn \in \mathcal{D}} l_{mn,t}^D r_{mn}^D + \sum_{\forall v \in \mathcal{V}} C_v p_{v,t} + \kappa v_t^D \right), \forall t \tag{6.32}$$

where v_t^D represents the degree of operational violations in the distribution system. To guarantee the secure and reliable operation of the power system, all constraint violations are aggregated and normalized into a single metric known as the violation metric. This metric quantifies the severity of any violations and is incorporated into the reward function to penalize suboptimal actions, as defined in (6.29) for the transmission system and (6.32) for the distribution system. The DCV is formulated as:

$$v = \sqrt{\frac{1}{|\mathcal{N}|}\sum_{\forall s_n}\left(\frac{\left[s_n - \overline{s}_n\right]^{+} + \left[\overline{s}_n - s_n\right]^{+}}{\overline{s}_n - \underline{s}_n}\right)^{2}} \tag{6.33}$$

where s_n represents the collection of all uncontrolled constraints, which include limits on line power flows, active power outputs of the slack generator, reactive power injections, and voltage magnitudes at the load buses. The total number of constraints is $\mathcal{N}$, with the corresponding minimum $\underline{s}_n$ and maximum $\overline{s}_n$ limits. These limits are derived from equations (6.2)-(6.11) for the transmission system and equations (6.14)-(6.19) for the distribution system. The notation [.]+ is defined as $s_n = \max\{0,\cdot\}$, ensuring that only violations beyond the allowed limits are penalized.

6.3.2 Soft actor-critic algorithm for transmission system

To solve the robust optimal power flow and load restoration problem for the transmission system as a single-agent decision-making problem, the SAC algorithm is introduced. SAC is an off-policy, actor-critic method based on maximum entropy reinforcement learning [105], which simultaneously maximizes the expected reward and the policy entropy. The incorporation of entropy encourages sufficient exploration, making the learning process more stable and robust. The objective of the SAC algorithm is to train a policy $\pi(a_t^T, s_t^T)$ that minimizes the operational cost of the transmission system under emergency scenarios, which is defined as:

$$\pi^{*} = \arg\max_{\pi} \mathbb{E}_{(s_t^T, a_t^T)\sim\pi} \sum_{t}\left[r(s_t^T, a_t^T) + \alpha\mathcal{H}(\pi(\cdot|\ s_t^T))\right] \tag{6.34}$$

where $\pi(\cdot\ |s_t^T)$ is a categorical distribution describing the probability of selecting any load curtailment action a_t^T given the power system state s_t^T. The term $H\big(\pi(\cdot\ |s_t^T)\big) = -log\ (\pi(a_t^T, s_t^T))$ represents the policy entropy, while α is the entropy temperature that controls the balance between exploration and exploitation. The value of α significantly influences learning convergence and exploration, and one important aspect of the SAC algorithm is the automatic adjustment of α using the following optimization:

$$\alpha_t^{*} = \arg\min_{\alpha_t} \mathbb{E}_{a_t\sim\pi_t^{*}}\left[-\alpha_t \log\pi_t^{*}(a_t|\ s_t;\alpha_t) - \alpha_t\underline{\mathcal{H}}\right] \tag{6.35}$$

where $\underline{\mathcal{H}}$ is the target minimum entropy. The α_t^* is an optimal dual variable after solving this dual optimization problem, $max\ E_\pi \sum_t r_n^T(s^T,a^T)\ s.t.\ E_{(s_{t+1}^T,a_{t+1}^T)\sim\pi}[-log\ (\pi(a_t^T,s_t^T))] \geq \mathcal{H}$, and π_t^* . This dual optimization problem ensures that the learned policy maintains sufficient stochasticity to explore efficiently while achieving optimal performance. SAC follows an actor-critic framework with stochastic actors, and the policy is iteratively updated by alternating between the critic network and the actor network. The critic network evaluates the action-value function $Q_\theta(s_t^T,a_t^T)$, parameterized by θ, using the soft Bellman backup operator, $Q_\theta(s_t^T,a_t^T) = E_{(s_{t+1}^T,a_{t+1}^T)\sim\pi}[r_t^T(s_t^T,a_t^T) + \gamma V_\pi(s_{t+1}^T)]$, where $V_\pi(s_t^T) = E_{a_t^T\sim\pi}[Q_\theta(s_t^T,a_t^T) - \alpha log\ (\pi(a_t^T,s_t^T))]$ is the soft state-value function. Instead of using an explicit state-value network, SAC calculates this value directly from the Q-function, improving efficiency. During each training iteration, the actor and critic networks are updated using mini-batches of previous experiences stored in a replay buffer $\mathrm{B} = [s_t^T,a_t^T,r_t^T,s_{t+1}^T]$, where s_{t+1}^T represents the next power system state after applying load curtailment action $\mathrm{a}_\mathrm{t}^\mathrm{T}$. The model employs two separate critic networks $Q_{\theta_1}(s_t^T,a_t^T)$ and $Q_{\theta_2}(s_t^T,a_t^T)$ with distinct parameters θ_1 and θ_2. The minimum of the two Q-values is used to mitigate overestimation bias [122]. Additionally, target networks $Q_{\theta_1'}$ and $Q_{\theta_2'}$ are introduced for each critic to improve stability during training [29]. The critic network update is performed by minimizing the following loss function:

$$J_Q(\theta_i) = \mathbb{E}_{(s_t^T,a_t^T,r_t^T,s_{t+1}^T)\sim\mathcal{B}}\Bigg[\frac{1}{2}\Bigg(Q_{\theta_i}(s_t^T,a_t^T) - \Bigg(r(s_t^T,a_t^T) + \Bigg(Q_{\theta_i'}(s_{t+1}^T,a_{t+1}^T) - \alpha log\pi_\varphi^*(a_{t+1}^T|s_{t+1}^T)\Bigg)\Bigg)\Bigg)^2\Bigg], \forall i \in \{1,2\} \qquad (6.36)$$

where φ are the parameters of the actor network. a_{t+1}^T is the load curtailment action predicted from the latest updated policy π_φ^* given states s_{t+1}^T. In the policy improvement step, the policy is optimized to maximize the soft Q-function by minimizing the KL-divergence as [29]:

$$J_{\pi}(\varphi)=\mathbb{E}_{s\sim\mathcal{B}}\left[\mathbb{E}_{a\sim\pi}\left[\alpha\log\pi_{\varphi}(a_t^T\,|\,s_t^T)-\min_{i\in\{1,2\}}Q_{\theta_i}(s_t^T,a_t^T)\right]\right] \tag{6.37}$$

This optimization is implemented using the reparameterization trick, where the policy is designed to predict the mean and standard deviation of a spherical Gaussian distribution, allowing efficient sampling of continuous control actions. By leveraging this structure, the SAC algorithm provides robust and scalable solutions to the robust optimal power flow and load restoration problem, ensuring resilience against contingencies while minimizing system operational costs.

6.3.3 Complementary attention based SAC for distribution system

To effectively address load restoration problem with VPP cooperation in distribution system, each VPP center within the distribution network is modeled as a DA. These DAs can only observe the load information of buses within their local distribution network, which limits their ability to formulate optimal solutions for coordinating the distribution system with the transmission network in response to contingency scenarios. To overcome these challenges, we introduce the complementary attention for multi-agent soft actor-critic (CMS) algorithm, which integrates a multi-agent SAC architecture with a complementary attention mechanism. This mechanism enhances each agent's local focus while supplementing critical global information, improving coordination, robustness, and performance stability.

CMS extends the standard SAC framework by incorporating multiple agents operating in a partially observable environment. Each agent *i* maintains a stochastic policy $\pi_{i,t}\left(a_{i,t}^D|s_{i,t}^D\right)$ and learns its action-value function by integrating local and global distribution system observation. The global value is communicated via a centralized trainer, while attention mechanisms guide local decision-making. The primary objective of CMS is to maximize cumulative rewards while maintaining a balance between exploration and exploitation through entropy regularization.

The objective for each agent *i* is to maximize the cumulative reward while balancing exploration and exploitation through entropy regularization:

$$\pi_i^*=\arg\max_{\pi}\mathbb{E}_{(s_t^D,a_{i,t}^D)\sim\pi}\sum_t\left[\left(r_t(s_t^D,a_{i,t}^D)+\alpha\mathcal{H}[\pi_i(\cdot\,|\,s_{i,t}^D)]\right)\right] \tag{6.38}$$

where r_t is global reward which is defined in equation (6.32); s_t^D and $a_{i,t}^D$ are global distribution system state and VPP dispatch action generated by agent i. Due to the multi-agent structure, policy evaluation based on local agent observations fails to effectively update the critic network. To enhance the evaluation capability of the critic network in a multi-agent framework, the complementary attention mechanism introduces the state dividing unit (SDU). This unit dynamically partitions the observed global distribution system observation s_t^D into two components: attention-enhanced information $s_{i,t}$, which is locally relevant for decision-making, and attention-replenished information $s_{-i,t}$, which represents complementary global insights derived from a local attention mask $M_{i,s}$. The extraction of $s_{i,t}$ and $s_{-i,t}$ from s_t^D is guided by the attention weights Q_iK^T of the multi-head attention module [123], enabling the selection of a limited number of high-relevance entities while filtering out distractions. This state extraction process is formally defined as follows:

$$M_{i,s} = M_i \odot M_T \tag{6.39}$$

$$M_T[T_i] = 1 \tag{6.40}$$

$$s_{i,t} = \mathrm{MHA}(T_i, s_t^D, M_{i,s}) \tag{6.41}$$

$$s_{-i,t} = \mathrm{MHA}(T_i, s_t^D, \neg M_{i,s}) \tag{6.42}$$

where equation (6.39) ensures that attention is focused on key execution-relevant states, while equation (6.40) utilizes a binary mask to retain the indices of states with the highest attention weights, where $T_i = F_\Xi(Q_iK^T)$ is used to select the top Ξ state with the highest attention weights. $M_i \in \{0,1\}^{S,A}$ is a binary mask of agent *i* applied to the state embeddings, generated by the environment to indicate which global state the agent can observe at a given time step. Here, S and A represent the number of lobal state s_t^D and agents, respectively. Additionally, $M_{i,s}$ and M_T denote the enhanced attention mask and the high-attention state mask, respectively, while $\mathrm{T_i}$ represents the set of selected attention indices. The MHA mechanism is employed to compute each agent's attention distribution over all visible entities [124]. The information extraction process for $s_{i,t}$ ensures that only the most relevant entities are selected for further processing, while the extraction process for $s_{-i,t}$ guarantees that it contains global-level insights that the agent would otherwise miss.

After SDU extracts the attention-enhanced feature $s_{i,t}^{D}$, the attention improvement unit (AIU) is introduced to further filter task-relevant information using an inverse model. The inverse model predicts actions $\hat{a}_{i,t}$ based on a probability $\pi(\hat{a}_{i,t})$ and SDU information $s_{i,t}$ and $s_{i,t+1}$:

$$\pi(\hat{a}_{i}^{t}) = \mathrm{IM}(s_{i,t}, s_{i,t+1}; \vartheta) \tag{6.43}$$

where inverse model, IM, is a two-layer multilayer perceptron with parameters ϑ. This model is trained using the cross-entropy loss as follows:

$$\mathcal{L}_{\mathrm{IM}} = \mathrm{CE}(\pi(\hat{a}_{i,t}), a_{i,t}) \tag{6.44}$$

By optimizing $\mathcal{L}_{IM}$, the model encourages the embedding $s_{i,t}$ to encode only the most relevant task-related information, helping the agent filter out irrelevant observations that could lead to attention distraction. Once the attention-enhanced feature $s_{i,t}$ has been refined through the AIU, it is used as an input to compute the local Q-function. The local Q-function quantifies the agent's expected return based on its refined observations and historical information:

$$Q_{i,t}(\mathrm{local}) = \mathbb{E}_{(s_{i,t}^{D}, a_{i,t}^{D}) \sim \mathcal{B}} \left[r_{t}^{D}(s_{i,t}^{D}, a_{i,t}^{D}) + \gamma V_{\pi}(s_{i,t}) \right] \tag{6.45}$$

where the temporal state of the agent and is updated using an experience buffer, $\mathcal{B}$. It helps retain memory of past decisions, making $Q_{i,t}(\mathrm{local})$ more informed about past actions. This formulation ensures that each agent's local Q-function is grounded in the most relevant observations, improving local decision-making while mitigating distractions from irrelevant entities.

Furthermore, to complement local observations, the complementary attention mechanism introduces attention complement unit (ACU) to leverage a centralized trainer that has access to the global state. The trainer generates a global attention $\zeta_{i,t}$ based on $s_{-i,t}$, providing agents with critical out-of-sight information. The global attention $\zeta_{\mathrm{i,t}}$ is computed as:

$$\zeta_{i,t} = \arg\max_{\zeta} \left(MI(\zeta_{i,t}; s_{-i,t}) - \beta MI(\zeta_{i,t}; s_{t}^{D}) \right) \tag{6.46}$$

where $MI(\cdot,\cdot)$ means the mutual information, which ensures that $\zeta_{i,t}$ captures relevant global information without introducing distractions. Maximizing $MI(\zeta_{i,t}, s_{-i,t})$ enables agent i to perceive information beyond its sight region, thereby alleviating the challenges of cooperation caused by partial observability. $MI(\zeta_{i,t}, s_t^D)$ prevents $\zeta_{i,t}$ from being overloaded with unnecessary details. β controls the trade-off between capturing relevant information and preventing distractions. Then the global message $\zeta_{i,t}$ is passed through a fully connected layer $f(\cdot)$ to compute the global Q-function which is ensuring that agents incorporate global coordination information into decision-making as follows:

$$Q_{i,t}(\text{global}) = f(\zeta_t^i) \tag{6.47}$$

Finally, the total Q-function described as:

$$Q_{i,t} = Q_{i,t}(\text{local}) + Q_{i,t}(\text{global}) \tag{6.48}$$

Fig. 6.3 illustrates the process by which the complementary attention mechanism coordinates the states observed by different agents to learn the total Q-function. First, the SDU receives the states observed by each agent in the environment and then divides and embeds the global state s_t^D into two components: $s_{i,t}$ and $s_{-i,t}$, which are fed into the AIU and the ACU, respectively. For AIU, an inverse model is applied to mitigate the issue of distracted attention. The AIU then generates the local Q-value $Q_{i,t}(\text{local})$ based on the attention-enhanced information $s_{i,t}$. For ACU, a mutual information network with global insights is introduced to generate a communication message $\zeta_{i,t}$, facilitating agent coordination. The ACU subsequently generates the global Q-value $Q_{i,t}(\text{global})$ based on the communication message $\zeta_{i,t}$. The local and global Q-values of all agents are summed to obtain the total Q-value $Q_{n,t}(\text{total})$, which is then used in conjunction with the target network to compute the RL loss. This loss is utilized to update the agents' control policies.

During each training iteration, the total Q-function is updated using mini-batches of previous experiences stored in a replay buffer $\mathcal{B} = [s_t^D, a_t^D, r_t^D, s_{t+1}^D]$. The update process of the Q-values is performed by minimizing the overestimation bias using the target network [29]. This recursive loss can be formulated as follows:

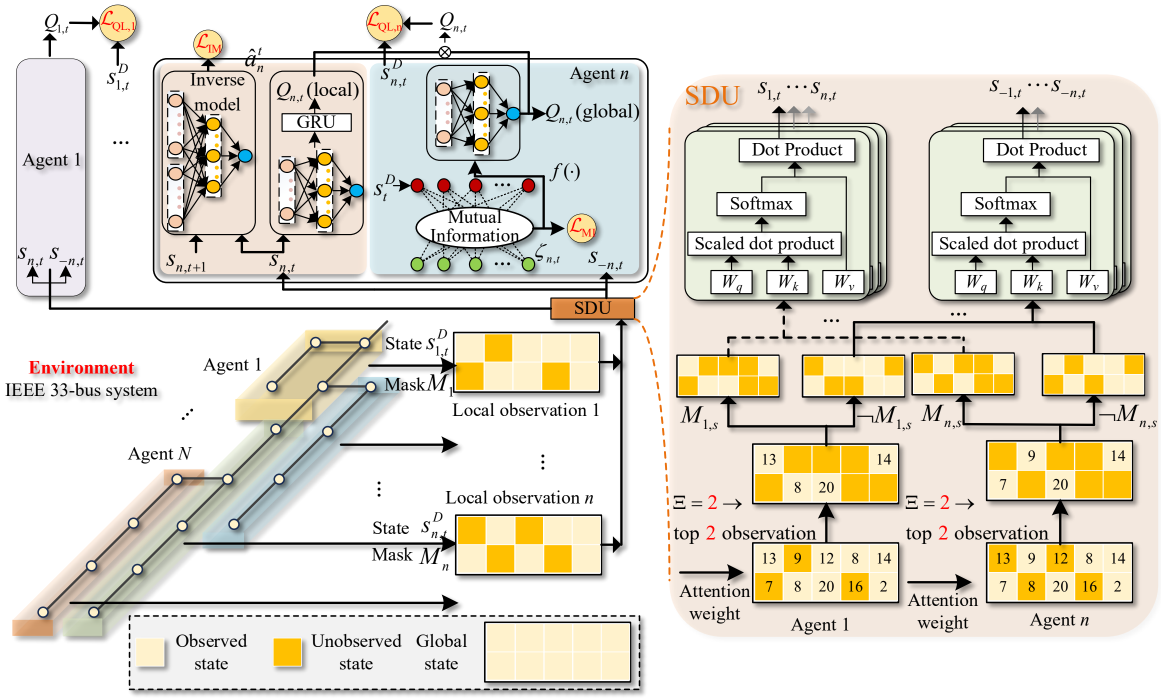


Fig. 6.3 Training structure of the complementary attention mechanism.

$$\mathcal{L}_{Q_{i,t}} = \mathbb{E}_{(s_t^D,a_t^D,r_t^D,s_{t+1}^D)\sim\mathcal{B}}\left[\frac{1}{2}\left(Q_{i,t}(\text{global}) - \left(r(s_t^D,a_t^D) + \left(Q'_{i,t+1}(\text{global}) - \alpha log\pi_{\varphi}^{*}(a_{t+1}^D|s_{t+1}^D)\right)\right)\right)^2\right], \forall i \in \{1,2\} \tag{6.49}$$

where $Q'_{i,t+1}(\text{global})$ is the target network. In the policy improvement step, the input feature of the actor function is the local information of each agent. Its parameters are optimized based on the following equation:

$$J_{\pi}(\varphi) = \mathbb{E}_{s\sim\mathcal{B}}\left[\mathbb{E}_{a\sim\pi}\left[\alpha \log \pi_{\varphi}(a_{i,t}^D | \; s_{i,t}^D) - Q_{i,t}(\text{global})\right]\right] \tag{6.50}$$

By leveraging this structure, the CMS algorithm effectively addresses the multi-VPP coordination challenge in the load restoration problem with VPP cooperation in distribution systems, ensuring efficient and adaptive dispatch of multiple VPP centers.

6.4 Case Study

6.4.1 Experiment setting

In this section, the proposed T&D system coordination model for emergency scenarios is validated using a test network consisting of an IEEE 30-bus system representing the transmission network and an IEEE 33-bus system representing the distribution network, which is connected to the transmission system at bus 8. Within the distribution network, four VPP centers are integrated at nodes 15, 22, 25, and 26, respectively, as illustrated in Fig. 6.4. The system parameters, including network topology, generation capacities, and line characteristics, are directly processed in their standard format as defined in PYPOWER. The numerical experiments were conducted on a computer equipped with an Intel i7-10700 CPU and 16 GB of RAM, with the hyperparameters used in the proposed algorithm summarized in Table 6.1. Additional modeling parameters are set as follows: the penalties for load shedding and unserved electricity, as defined in Eq. (6.29), are set to $C_l = 10 \times C_g$ and $C_l^* = 100 \times C_g$, respectively; the constraint violation penalty κ is set to $\kappa = 1e3$; the PV generation profile data are sourced from pvoutput.org, with a generation capacity of 6 kW; the BESS has a power/energy capacity of 10 kW/30 kWh; and power demand values are randomly generated, with maximum and minimum limits set at 120% and 80% of the normal operating levels as defined in the PYPOWER dataset. This validation framework ensures a comprehensive assessment of the proposed coordination model under emergency scenarios.

Table 6.1 Main Hyper-Parameters and Data Setting.

Parameters	Value	Parameters	Value
Optimizer	Adam	Discount factor	0.99
Critics learning rate	1e-2	Minibatch size	128
Actor learning rate	1e-3	Neurons number	512
Entropy learning rate	1e-4	Top Ξ attention weight	8
Targets learning rate	1e-3	Max steps	10
Initial temperature	1	Activation	RELU, Softmax

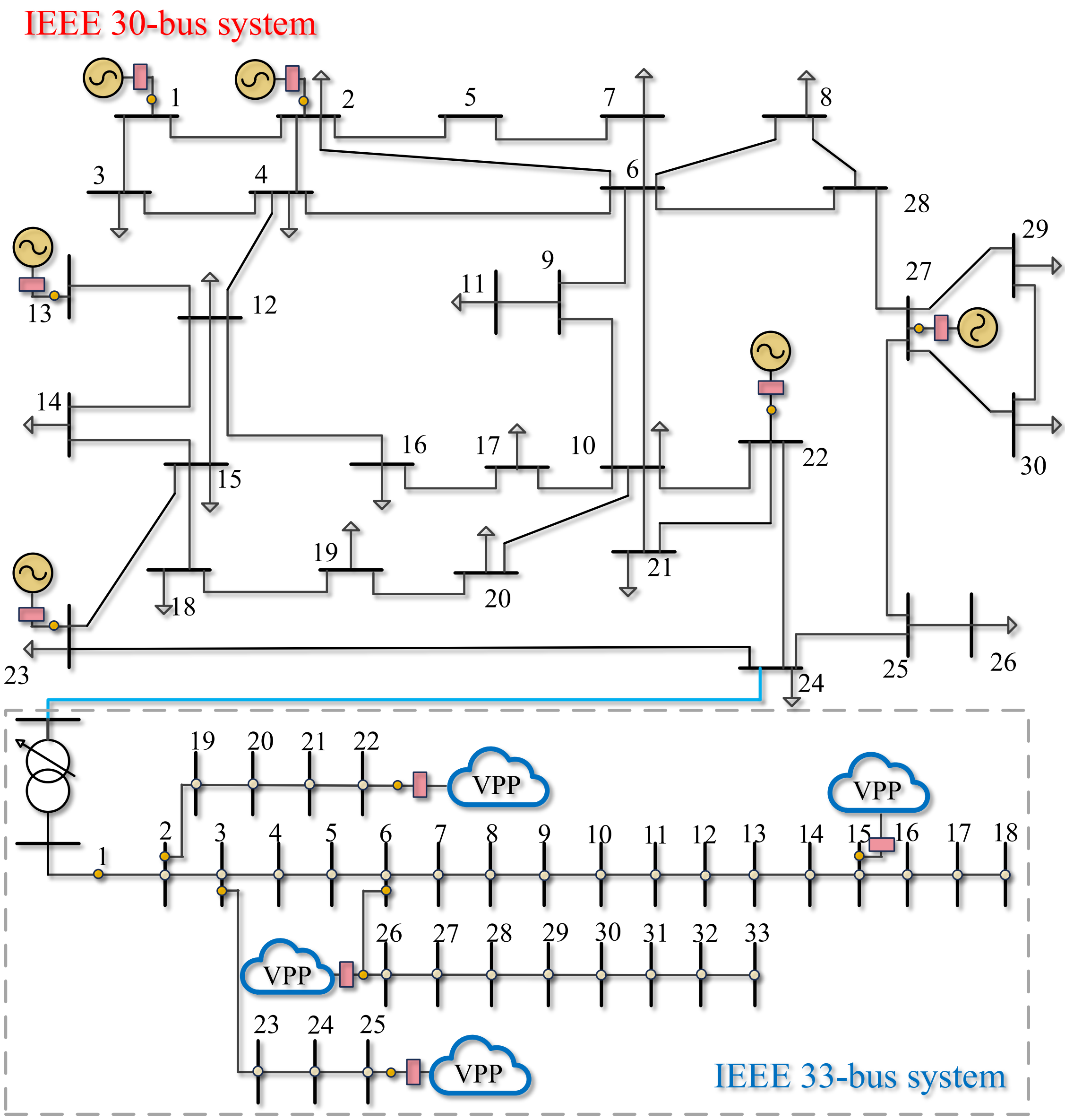
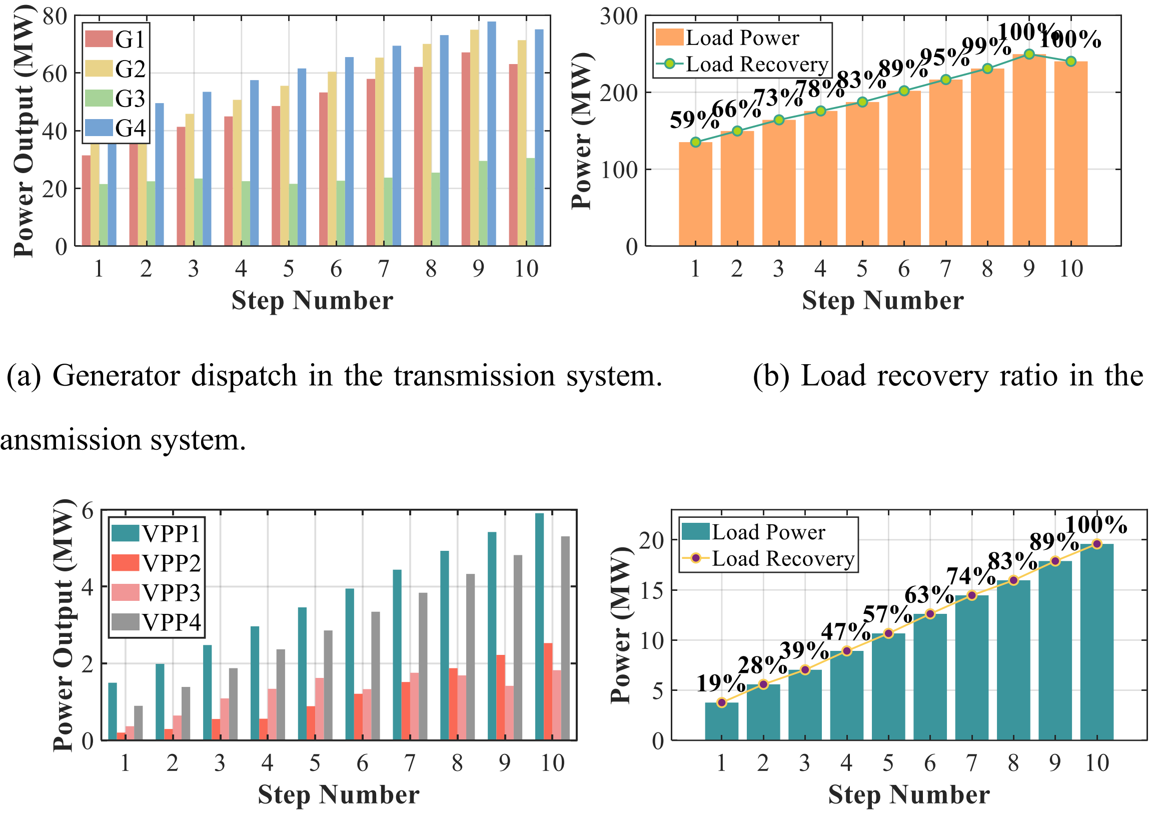


Fig. 6.4 The network topology structure of the coordinated IEEE 30-bus and IEEE 33-bus system.

6.4.2 Simulation Results of Load Restoration Under N-*k* Contingency

To evaluate the effectiveness of the coordinated T&D system in addressing N-*k* contingency scenarios, we conducted simulations over 10 restoration steps, during which the system gradually recovered from an emergency scenario where one transmission line and one generator were disconnected from the grid. The optimization results of generators output and loads restoration in the T&D system throughout the restoration process are depicted in Fig. 6.5.



(a) Generator dispatch in the transmission system.

(b) Load recovery ratio in the transmission system.

(c) VPP dispatch in the transmission system.

(d) Load recovery ratio in the distribution system.

Fig. 6.5 Restoration of generator dispatch and load recovery in the transmission and distribution systems.

As observed in Figs. 6.5(a) and (c), due to ramp rate limitations, the active power output of the generators and VPP centers in both the transmission and distribution systems gradually increases, enabling the progressive restoration of load supply under emergency scenarios. Figs. 6.5(b) and (d) further illustrate that after 9 restoration steps, the load recovery ratio of the transmission system steadily improves from 59% to 100%, while the distribution system, responding to load curtailment from the transmission system, achieves a full recovery from 19% to 100% within 10 restoration steps. To further demonstrate the robustness of the coordinated transmission and distribution system under emergency conditions, Table 6.2 presents a detailed analysis of the operating costs and constraint violations experienced by both systems throughout the restoration process.

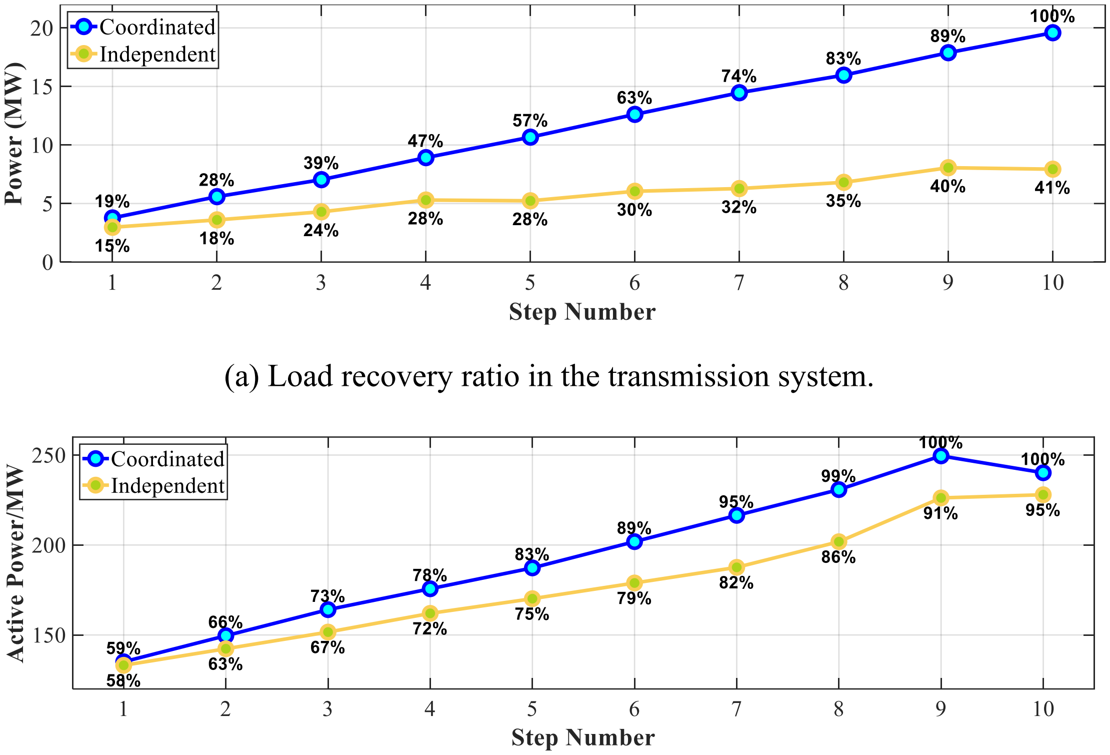
Table 6.2 System performance during the restoration process.

	Operation cost of TSO	Constraints violation of TSO	Operation cost of DSO	Constraints violation of DSO	VPP output
Step 1	228.88	1.3298	6.49	1.3897	296.68
Step 2	225.86	1.1285	7.85	1.2269	432.01
Step 3	226.58	0.9130	9.22	1.0735	600.43
Step 4	228.51	0.7300	10.31	0.9147	723.90
Step 5	233.93	0.5369	11.80	0.8170	883.15
Step 6	242.72	0.4430	12.62	0.6180	983.85
Step 7	251.58	0.1999	14.23	0.5333	1155.55
Step 8	261.22	0.0791	15.30	0.3317	1282.62
Step 9	273.18	0	16.18	0.1571	1388.02
Step 10	267.66	0	17.73	0	1557.19

Initially, both systems encounter severe constraint violations, quantified based on Equation (6.33). This is primarily due to the ramping limitations of the generators, which hinder the immediate reallocation of power dispatch following the transmission line and generator disconnection, leading to congestion in power flow near the affected components. Simultaneously, the distribution system experiences voltage violations as a result of load curtailments imposed by the transmission system, causing an insufficient supply to local consumers. However, as the system progresses through multiple restoration steps, the load restoration ratio of both the transmission and distribution systems improves steadily, while the constraint violations progressively decrease. This demonstrates the effectiveness of the coordinated transmission and distribution response in dynamically mitigating the adverse impacts of N-*k* contingency events.

6.4.3 Comparison of coordinated and independent scheme under N-*k* contingency scenario

To validate the effectiveness of the proposed coordinated transmission and distribution system operation, a comparative analysis is conducted against the independent operation model. Unlike the coordinated model, where the transmission system communicates its load curtailment decisions to the distribution system and receives information about the maximum potential load curtailment capacity in return, the independent model operates without any exchange of information between the two systems. This lack of coordination impacts the system's ability to respond optimally under emergency scenarios.



(a) Load recovery ratio in the transmission system.

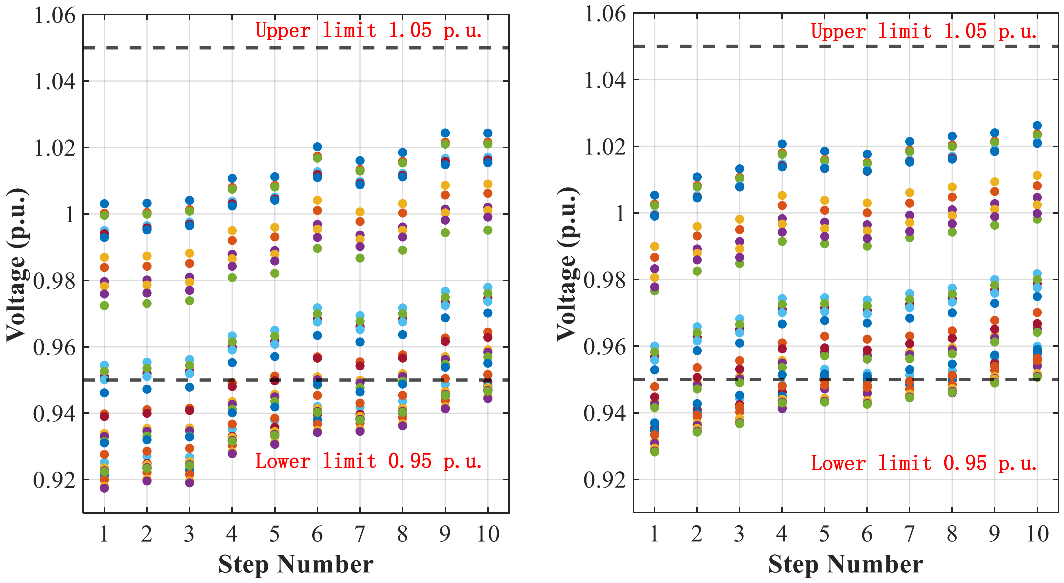
(b) Load recovery ratio in the distribution system.

Fig. 6.6 Load restoration comparison between coordinated and independent scheme.

The optimization results, presented in Fig. 6.6, indicate that both approaches can restore the transmission system's load supply following an emergency. However, the independent operation model exhibits a slower response in restoring load supply compared to the coordinated approach. The absence of information from the distribution system prevents the independent model from accurately formulating a robust security-constrained optimal power flow solution, leading to delays in adjusting dispatch and redistributing power flows efficiently. Additionally, as the transmission system curtails load without knowledge of the distribution system's available flexibility, the distribution network struggles to recover its load supply, resulting in a prolonged period of voltage violations and load deficiency. Overall, the results highlight that the coordinated operation significantly enhances system resilience, accelerates load recovery, and improves stability in the distribution system. By integrating real-time information exchange, the coordinated approach ensures faster decision-making, more precise load curtailment, and reduced power flow congestion, ultimately leading to a more efficient and robust emergency response.

6.4.4 Impact of considering DCV on system constraint satisfaction

This section evaluates the effectiveness of incorporating degree of constraints violation (DCV) in the MDP framework to ensure that the generated actions remain within safe operational limits while minimizing system violations. To highlight the advantages of this approach, we compare the proposed CMS algorithm with its counterpart without DCV consideration. The voltage distribution of the distribution system under both approaches is depicted in Fig. 6.7.



(a) Voltage distribution without DCV consideration.

(b) Voltage distribution with DCV consideration.

Fig. 6.7 Voltage distribution in the distribution system with and without DCV consideration.

As shown in Figs. 6.7(a) and (b), both approaches initially experience voltage violations due to severe load supply shortages in the early restoration steps. However, as the MDP-based optimization progresses, the proposed DCV-integrated approach successfully regulates voltage distribution within the permissible range (0.95–1.05 p.u.), effectively mitigating violations over time. In contrast, the approach that does not consider DCV consistently exhibits lower voltage levels compared to the proposed method and fails to fully eliminate voltage violations by the final MDP step. These results highlight the benefits of incorporating DCV considerations in the optimization framework. By explicitly accounting for system constraints, the proposed approach enhances voltage stability, prevents persistent violations, and ensures secure operation of the distribution system. This contribution strengthens the reliability and resilience of the coordinated transmission and distribution system, particularly in emergency response scenarios.

6.4.5 Performance Evaluation of DRL Algorithms

To validate the effectiveness of the proposed CMS algorithm in handling the coordination of the distribution system with the transmission system under emergency scenarios, a comparative analysis is conducted against four state-of-the-art MADRL algorithms: MASAC [125], MADDPG [126], multi-agent proximal policy optimization (MAPPO) [127], and attention-based MASAC (AMS) [128]. Each of these algorithms is implemented in the IEEE 33-bus distribution system, ensuring a consistent dataset and operational setting for evaluating their performance. The first benchmark algorithm, MADDPG, utilizes an offline training mechanism but struggles with high sensitivity to hyperparameters, making it less robust in practical applications. The second benchmark, MAPPO, operates as an on-policy algorithm, which limits its sample efficiency in complex environments. The third benchmark, MASAC, applies a multi-agent SAC framework without additional mechanisms to enhance coordination, while the fourth benchmark, AMS, integrates an attention mechanism to mitigate the multi-agent attention dispersion problem, thereby improving decision-making efficiency. The training performance of these algorithms is illustrated in Fig. 6.8, which depicts the convergence curves over 400 gradient episodes.

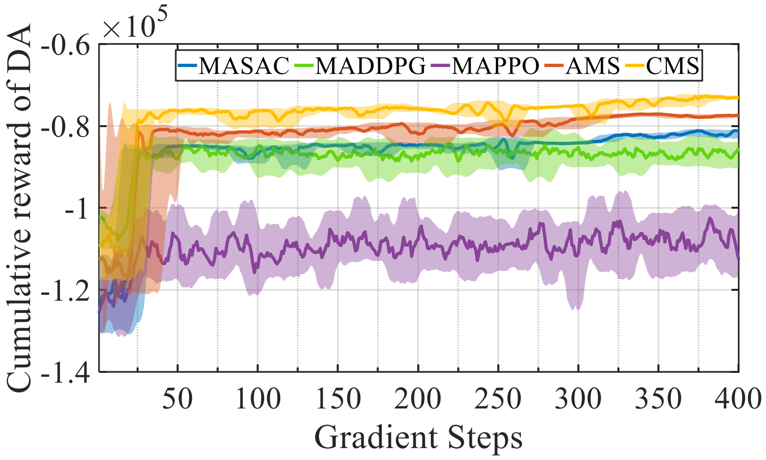


Fig. 6.8 Convergence curves of CMS and benchmark algorithms.

As shown in Fig. 6.8, the CMS algorithm demonstrates superior stability and faster convergence compared to the other four approaches. The solid curves represent the mean cumulative reward across ten independent experiments, while the shaded areas indicate the range between the minimum and maximum rewards obtained. The results show that CMS consistently achieves higher cumulative rewards than the other algorithms, demonstrating better learning efficiency and decision-making capabilities in managing emergency dispatch with VPP coordination. The MASAC, MADDPG, and MAPPO algorithms exhibit slower convergence and lower final rewards, with MAPPO experiencing the most unstable learning process, likely due to its reliance on an on-policy approach. Meanwhile, the attention-based MASAC algorithm performs better than MASAC but still falls short of CMS, indicating that while attention mechanisms enhance coordination, CMS further improves multi-agent cooperation through complementary attention mechanisms. To further analyze the detailed performance of each algorithm in solving this problem, Table 6.3 presents a quantitative comparison of key performance indicators.

Table 6.3 provides a detailed evaluation of five key performance metrics: cumulative reward, constraint violations, VPP adjustment, and unserved electricity at convergence. The results confirm that CMS achieves the lowest reward (7420.58), along with the lowest constraint violations (0.5972) and unserved electricity (23.94 MWh), while also maintaining efficient VPP adjustment (1016.73 MWh). These results demonstrate that CMS effectively balances power dispatch, constraint compliance, and operational efficiency in an emergency dispatch scenario. Among the benchmark algorithms, MAPPO exhibits the highest reward (11091.85), which suggests the worst performance in constraint satisfaction. This is further confirmed by its highest constraint violations (1.0244) and unserved electricity (84.33 MWh), indicating that it struggles to maintain a stable and reliable power dispatch solution. Similarly, MADDPG, while performing better than MAPPO, still shows a relatively high reward (8655.93), along with significant constraint violations (0.7117) and unserved electricity (37.77 MWh), demonstrating suboptimal coordination and resource allocation. MASAC and attention-based MASAC perform better than MAPPO and MADDPG in terms of constraint satisfaction but still exhibit higher reward values and constraint violations compared to CMS. Specifically, attention-based MASAC achieves a reward of 7812.16, with constraint violations of 0.6312 and unserved electricity of 23.68 MWh, showing an improvement over MASAC but still falling short of CMS. These findings further validate that CMS provides the most effective balance between constraint satisfaction, resource allocation, and system stability, reinforcing its superiority in managing the distribution system's response to emergency dispatch scenarios.

Table 6.3 Performance metrics comparison of CMS and benchmark algorithms.

	Reward	Constraints violation	VPP adjustment	Unserved electricity
MASAC	8021.45	0.6668	1369.31	25.55
MADDPG	8655.93	0.7117	1311.48	37.77
MAPPO	11091.85	1.0244	762.78	84.33
AMS	7812.16	0.6312	1268.54	23.68
CMS	7420.58	0.5972	1016.73	23.94

To further assess the effectiveness of AMS and CMS in enhancing agents' focus on global information, we introduce an attention entropy metric. This metric quantifies how uniformly an agent distributes its attention across available entities, thereby reflecting its ability to prioritize crucial information while avoiding distractions [124]. The attention entropy value is defined as follows:

$$\mathcal{M}_{ae} = -\sum_{i}^{N}\sum_{j}^{N} p_{i,j} \log p_{i,j} \tag{6.51}$$

where $p_{i,j} = \frac{e_{i,j}}{\sum_k^N e_{i,k}}$ is the normalized attention weight assigned to agent j; $e_{i,j}$ is the attention weight assigned agent *i* to agent *j*. A lower entropy value indicates higher attention concentration, meaning the agent focuses on fewer but more relevant information [129]. To further analyze the variability of attention distribution across agents and time steps, we compute the standard deviation of the attention distribution:

$$\mathcal{S}_{ae} = \sum_{i}^{N}\sqrt{\frac{1}{N}\sum_{j}^{N}(p_{i,j} - \mu_i)^2} \tag{6.52}$$

where $\mu_i = \frac{1}{N}\sum_j^N p_{i,j}$ is the mean attention weight for agent i. A high standard deviation indicates more variance in attention distribution, suggesting that certain agents receive significantly more attention than others. We conducted simulation experiments to compare the attention entropy and its standard deviation for CMS and AMS under varying sight regions (R $=$ 8,10). The results are summarized in the following table:

Table 6.4 Evaluation of attention entropy in the proposed algorithm and the AMS algorithm.

	$\mathcal{M}_{ae}$, $(\mathcal{R} = 8)$	$\mathcal{M}_{ae}$, $(\mathcal{R} = 10)$	$\mathcal{S}_{ae}$ $(\mathcal{R} = 8)$	$\mathcal{S}_{ae}$ $(\mathcal{R} = 10)$
CMS local	1.3843	1.4208	0.1286	0.09132
CMS global	1.5121	1.5144	0.0922	0.06202
AMS	1.4830	1.5498	0.1018	0.0865

Note: $\mathcal{R}$ represents the size of the sight region that an agent can observe around its VPP center. $\mathcal{M}_{\mathrm{ae}}$ denotes the mean value of attention entropy, while $\mathcal{S}_{ae}$ refers to the standard deviation of the attention distribution.

The results indicate that CMS local exhibits a relatively low attention entropy, which allows agents to focus on critical information efficiently; however, excessive concentration can lead to high training instability due to the lack of sufficient environmental information for decision-making. To counterbalance this, CMS global achieves a higher attention entropy, supplementing the local model by extracting relevant global information, thereby enhancing the algorithm's training stability. This local-global coordination ensures that CMS enables agents to make optimal control decisions even when limited to local observations. The role of sight region R is critical in this analysis, as it defines how much information an agent can perceive at a given time; a larger sight region ($\mathcal{R} = 10$) generally results in a more balanced attention distribution, while a smaller sight region ($\mathcal{R} = 8$) forces agents to rely more on local information. Compared to AMS, CMS local provides a more focused attention mechanism, reducing distractions, while CMS global mitigates excessive attention concentration by incorporating essential global context. This synergy allows CMS to outperform AMS by maintaining both stability and adaptability, making it more robust in dynamic multi-agent environments where agents need to make precise decisions despite partial observability.

6.5 Summary

This study presents a novel coordinated T&D system operation strategy to enhance power system resilience under N-*k* contingencies. The proposed framework leverages VPPs within the distribution network to compensate for loads curtailed by the transmission system, ensuring load restoration and mitigating power flow congestion. Through bidirectional information exchange, the transmission system communicates load-shedding decisions to the distribution network, while the distribution network provides its available maximum curtailment capacity. This coordination optimizes system response to unexpected disruptions, improving power system stability. To achieve efficient real-time decision-making, we adopt reinforcement learning-based optimization. The transmission system is modeled using the SAC algorithm, which determines optimal load-shedding and generator dispatch strategies for rapid restoration. Meanwhile, the distribution system employs the

CMS algorithm to effectively manage multiple VPPs. This approach addresses attention dispersion issues commonly encountered in multi-agent environments, enabling more reliable decision-making. Simulation results validate the effectiveness of the proposed reinforcement learning framework. The coordinated operation of the T&D system successfully reduces power flow congestion in the transmission network while maintaining load supply and voltage stability in the distribution system. These findings demonstrate the potential of reinforcement learning-based methodologies in enhancing the adaptability and resilience of modern power grids.

While the present study has focused on simulation-based validation for transmission, distribution, and coordinated T&D systems, an important direction for future work lies in exploring the engineering validation pathway. One promising step is the adoption of hardware-in-the-loop (HIL) platforms, where the proposed DRL-based controllers can be evaluated against real-time digital simulations of power systems to assess their dynamic performance under realistic operating conditions. At the transmission level, HIL experiments could emulate generator dynamics and contingency responses to provide insights into real-time feasibility. At the distribution level, controller-HIL tests with inverter emulators and OLTCs would allow evaluation of scalability, communication load, and control responsiveness. This staged validation pathway—from simulation to HIL to SCADA-level integration—provides a clear route for bridging the gap between theoretical research and practical deployment.

Chapter 7 Conclusions and Future Perspectives

7.1 General Conclusions

This thesis proposes a unified reinforcement learning–based framework to improve the secure and resilient operation of modern power systems under various contingencies and uncertainties. It addresses a diverse set of operational challenges that arise across the transmission system, distribution system, and their coordinated operation layers. Through the development of four interrelated but methodologically distinct contributions, this thesis highlights how learning-based control, robust optimization, and system-level coordination can jointly address the fundamental limitations of traditional OPF techniques. The contributions extend beyond simulation performance and demonstrate how reinforcement learning can be systematically integrated into power system operation models to handle high-dimensional, dynamic, and safety-critical decision-making. The four major contributions are summarized below from a structural and conceptual perspective.

1) *Multi-Agent Adversarial Learning Architecture for Robust Decision-Making under N-k Contingencies*: This study introduces an innovative multi-agent adversarial reinforcement learning framework for enhancing robustness in CCOPF problems. Unlike prior methods that treat uncertainties as passive stochastic parameters, the proposed model structures the interaction between system control and uncertainty as a game between two learning agents. This modeling shift enables the system operator (defender agent) to proactively generate decisions that are robust to dynamically evolving worst-case contingency scenarios generated by an attacker agent. The novelty lies not just in the application of DRL, but in how uncertainty modeling is internalized within the learning loop through competitive policy interaction. The use of dual-agent SAC—where continuous control spaces for the defender and discrete combinatorial action spaces for the attacker coexist—further introduces a flexible yet computationally efficient framework. Additionally, the approach

supports end-to-end policy learning without dependence on linearized power flow models or surrogate convex approximations. This work sets the foundation for rethinking robust power system operation not just as an optimization problem but as a dynamic adversarial learning task, where resilience emerges from strategic anticipation of system threats.

2) *CMDP-Based Formulation for Safety-Aware Preventive-Corrective Scheduling with VPPs*: The second contribution reformulates the classic preventive–corrective SCOPF problem into a constrained Markov decision process (CMDP) to explicitly address the dual requirements of operational safety and adaptability under uncertainty. By leveraging the control flexibility of VPPs, this formulation builds a temporal structure that separates decision-making into pre-contingency (preventive) and post-contingency (corrective) stages, each governed by a dedicated reinforcement learning agent. A key conceptual advancement is the embedding of constraint satisfaction directly into the policy learning phase via Lagrangian dual variables, which are updated dynamically based on observed system states and actions. This contrasts with conventional DRL approaches that treat safety constraints as soft penalties and often struggle with feasibility. The CMDP-based structure also enables a modular learning approach, where the preventive agent learns to anticipate downstream constraints imposed by corrective actions, creating a feedback-consistent learning environment. Furthermore, the design effectively integrates physical modeling (AC constraints) with policy-based learning, illustrating a hybrid paradigm that balances theoretical rigor with practical adaptability. This methodological innovation is critical for realizing real-time resilient control in large-scale systems where fast, constraint-compliant decision-making is essential.

3) *Hierarchical Multi-Mode Control Design for Data-Driven Voltage Regulation in ADNs*: In addressing voltage regulation in active distribution networks, this thesis presents a hierarchically structured control architecture that operates across two timescales and supports multiple operational objectives. A distinctive contribution of this work is the formalization of a multi-mode voltage regulation model, where the system dynamically transitions between optimization goals such as minimizing power losses, mitigating under-

voltage, or suppressing over-voltage, depending on real-time operating conditions. This contrasts with prior strategies that rely on fixed control objectives or single-mode regulation. The learning framework is structured such that a global agent dispatches slow-acting mechanical devices (OLTCs, CBs) using a discrete SAC approach, while local agents govern inverter-based devices via attention-enabled multi-agent SAC (MASAC). The attention mechanism allows agents to selectively process relevant signals, improving learning efficiency and stability, especially in high-agent-count environments. Beyond algorithmic novelty, the work contributes to the growing field of autonomous grid management by demonstrating how localized inverter agents can collectively realize system-wide control objectives without explicit communication during runtime. This decentralized-yet-coordinated model of voltage regulation provides a scalable pathway for managing uncertainty in high-DER environments without compromising stability or efficiency.

4) Reinforcement Learning–Enhanced Coordination Framework for Distributed T&D Load Restoration: The final contribution addresses the longstanding challenge of real-time T&D coordination during emergency load restoration. Rather than relying on tightly coupled optimization models, the proposed framework decomposes the problem into two separate but communicative RL processes: a centralized SAC controller for the transmission system and a multi-agent MASAC controller for the distribution system. The use of a virtual power plant (VPP) as an intermediate aggregator represents a shift from device-level to resource-cluster coordination, significantly reducing the dimensionality and communication overhead of the distributed control process. What distinguishes this work is not only the agent-level learning design but also the information exchange architecture that supports asynchronous yet coherent decision-making across system layers. Furthermore, the introduction of a complementary attention mechanism into the MASAC framework addresses two pressing MARL challenges: the inability to focus on critical environmental signals and the lack of global coordination among decentralized agents. By combining reinforcement learning, hierarchical aggregation, and cross-layer communication, the

proposed method provides a conceptually unified and computationally efficient approach for restoring power in complex, multi-entity grid environments.

7.2 Future Perspectives

While this thesis introduces several novel and effective strategies for enhancing the secure operation of power systems under uncertainties and emergencies, several limitations and open challenges remain. Future research directions may focus on addressing these limitations, extending current frameworks to broader scenarios, and integrating more practical considerations for deployment in real-world systems. The key future directions are summarized as follows:

1) *Expanding Adversarial RL for Broader Classes of Power System Uncertainties*: The proposed defender–attacker reinforcement learning framework shows promising results in generating robust control strategies under N-*k* contingency scenarios. However, current work mainly considers topological failures and static system parameters. Future research could extend this adversarial framework to include dynamic uncertainties, such as time-varying load demand, renewable forecast errors, and market-driven behaviors. Additionally, incorporating probabilistic forecasting models into the attacker's behavior generation could further enhance the realism and adaptability of the defender's learning strategy. From a theoretical standpoint, there is also a need to explore convergence guarantees and robustness bounds in multi-agent adversarial settings within power system environments.

2) *Integration of Safety-Critical RL Algorithms with Physical System Constraints*: The Lagrangian-based soft actor-critic algorithm developed in this thesis ensures constraint satisfaction through dynamic dual variable adjustment. However, in practice, DRL policies may still suffer from feasibility violations in unseen or extreme scenarios. Future work should explore the integration of formal safe RL techniques, such as Lyapunov-based policy optimization, control barrier function learning, or model predictive safety layers, into the preventive–corrective SCOPF framework. Moreover, extending the CMDP formulation to incorporate chance constraints or distributionally robust optimization (DRO) could better handle uncertainties with known distributions, enhancing safety guarantees during operation.

3) *Toward Plug-and-Play Multi-Agent Voltage Control Architectures*: The proposed multi-mode voltage regulation strategy demonstrates a scalable solution for active distribution networks. However, the MASAC architecture requires pre-defined agent topologies and offline centralized training. For real-world systems with dynamic DER participation, topology reconfigurations, or plug-and-play devices, the control architecture must become more adaptive. Future research should investigate online transferable reinforcement learning, where agents can continuously adapt to evolving network structures. Additionally, adopting graph neural networks (GNNs) as policy encoders could improve coordination under changing connectivity, while reducing the dependency on retraining across different system configurations.

4) *Real-World Implementation of Distributed Load Restoration Frameworks*: While the proposed T&D coordination strategy has shown strong simulation performance, practical implementation still faces challenges in terms of communication delays, scalability, and cybersecurity risks. Future research should explore the deployment of the proposed CMS-based MASAC algorithm in hardware-in-the-loop (HIL) environments or digital twins, integrating SCADA-based data feeds to validate real-time decision-making performance. Additionally, optimizing the communication topology and frequency for aggregator-to-DER coordination is crucial to ensure low-latency control without overloading the system. Finally, considering cyber-attack resilience (e.g., data falsification or denial-of-service attacks) in distributed MARL-based architectures is critical for ensuring the secure operation of restoration strategies under adversarial conditions.

Reference

[1] B. Ti, G. Li, M. Zhou, and J. Wang, "Resilience Assessment and Improvement for Cyber-Physical Power Systems Under Typhoon Disasters," *Ieee Transactions on Smart Grid,* Article vol. 13, no. 1, pp. 783-794, Jan 2022, doi: 10.1109/tsg.2021.3114512.

[2] T. Zhang, Y. Mu, L. Dong, H. Jia, T. Pu, and X. Wang, "Fully parallel decentralized load restoration in coupled transmission and distribution system with soft open points," *Applied Energy,* Article vol. 349, Nov 1 2023, Art no. 121626, doi: 10.1016/j.apenergy.2023.121626.

[3] N. Junnarkar, E. Jensen, X. Wu, S. Gumussoy, and M. Arcak, "Grouping of N-1 Contingencies for Controller Synthesis: A Study for Power Line Failures," *Ieee Transactions on Power Systems,* Article vol. 40, no. 1, pp. 585-596, Jan 2025, doi: 10.1109/tpwrs.2024.3393866.

[4] M. Zhou, C. Liu, A. A. Jahromi, D. Kundur, J. Wu, and C. Long, "Revealing Vulnerability of N-1 Secure Power Systems to Coordinated Cyber-Physical Attacks," *Ieee Transactions on Power Systems,* Article vol. 38, no. 2, pp. 1044-1057, Mar 2023, doi: 10.1109/tpwrs.2022.3169482.

[5] K. Zhou, I. Dobson, and Z. Wang, "The Most Frequent N-k Line Outages Occur in Motifs That Can Improve Contingency Selection," *Ieee Transactions on Power Systems,* Article vol. 39, no. 1, pp. 1785-1796, Jan 2024, doi: 10.1109/tpwrs.2023.3249825.

[6] M. Chen, X. Cao, Z. Zhang, L. Yang, D. Ma, and M. Li, "Risk-averse stochastic scheduling of hydrogen-based flexible loads under 100% renewable energy scenario," *Applied Energy,* Article vol. 370, Sep 15 2024, Art no. 123569, doi: 10.1016/j.apenergy.2024.123569.

[7] Y.-C. Wu, L.-F. Cheung, K.-S. Lui, and P. W. T. Pong, "Efficient Communication of Sensors Monitoring Overhead Transmission Lines," *Ieee Transactions on Smart Grid,* Article vol. 3, no. 3, pp. 1130-1136, Sep 2012, doi: 10.1109/tsg.2012.2186596.

[8] M. Yan, M. Shahidehpour, A. Paaso, L. Zhang, A. Abdulwhab, and A. Abusorrah, "A Convex Three-Stage SCOPF Approach to Power System Flexibility With Unified Power Flow Controllers," *Ieee Transactions on Power Systems,* Article vol. 36, no. 3, pp. 1947-1960, May 2021, doi: 10.1109/tpwrs.2020.3036653.

[9] Y. Xu, C.-C. Liu, K. P. Schneider, F. K. Tuffner, and D. T. Ton, "Microgrids for Service Restoration to Critical Load in a Resilient Distribution System," *Ieee Transactions on Smart Grid,* Article vol. 9, no. 1, pp. 426-437, Jan 2018, doi: 10.1109/tsg.2016.2591531.

[10] W. Liu and F. Ding, "Collaborative Distribution System Restoration Planning and Real-Time Dispatch Considering Behind-the-Meter DERS," *Ieee Transactions on Power Systems,* Article vol. 36, no. 4, pp. 3629-3644, Jul 2021, doi: 10.1109/tpwrs.2020.3048089.

[11] A. Suresh, R. Bisht, and S. Kamalasadan, "A Coordinated Control Architecture With Inverter-Based Resources and Legacy Controllers of Power Distribution System for Voltage Profile Balance," *Ieee Transactions on Industry Applications,* Article;

Proceedings Paper vol. 58, no. 5, pp. 6701-6712, Sep 2022, doi: 10.1109/tia.2022.3183030.

[12] N. Nazir and M. Almassalkhi, "Voltage Positioning Using Co-Optimization of Controllable Grid Assets in Radial Networks," *Ieee Transactions on Power Systems,* Article vol. 36, no. 4, pp. 2761-2770, Jul 2021, doi: 10.1109/tpwrs.2020.3044206.

[13] X. Zhu, J. Wang, N. Lu, N. Samaan, R. Huang, and X. Ke, "A Hierarchical VLSM-Based Demand Response Strategy for Coordinative Voltage Control Between Transmission and Distribution Systems," *Ieee Transactions on Smart Grid,* Article vol. 10, no. 5, pp. 4838-4847, Sep 2019, doi: 10.1109/tsg.2018.2869367.

[14] H. Wang, Z. Liu, Z. Liang, X. Huo, R. Yu, and J. Bian, "Multi-timescale risk scheduling for transmission and distribution networks for highly proportional distributed energy access," *International Journal of Electrical Power & Energy Systems,* Article vol. 155, Jan 2024, Art no. 109598, doi: 10.1016/j.ijepes.2023.109598.

[15] X. Cao, H. Wang, Y. Liu, R. Azizipanah-Abarghooee, and V. Terzija, "Coordinating self-healing control of bulk power transmission system based on a hierarchical top-down strategy," *International Journal of Electrical Power & Energy Systems,* Article vol. 90, pp. 147-157, Sep 2017, doi: 10.1016/j.ijepes.2017.02.004.

[16] C. Zhang, L. Liu, H. Cheng, D. Liu, J. Zhang, and G. Li, "Data-driven distributionally robust transmission expansion planning considering contingency-constrained generation reserve optimization," *International Journal of Electrical Power & Energy Systems,* Article vol. 131, Oct 2021, Art no. 106973, doi: 10.1016/j.ijepes.2021.106973.

[17] S. Rahim and P. Siano, "A survey and comparison of leading-edge uncertainty handling methods for power grid modernization," *Expert Systems with Applications,* Article vol. 204, Oct 15 2022, Art no. 117590, doi: 10.1016/j.eswa.2022.117590.

[18] J. Zhang, N. Zhang, and Y. Ge, "Energy Storage Placements for Renewable Energy Fluctuations: A Practical Study," *Ieee Transactions on Power Systems,* Article vol. 38, no. 5, pp. 4916-4927, Sep 2023, doi: 10.1109/tpwrs.2022.3214983.

[19] G. E. Mejia-Ruiz, M. R. A. Paternina, M. Ramirez-Gonzalez, F. R. S. Sevilla, and P. Korba, "Real-time co-simulation of transmission and distribution networks integrated with distributed energy resources for frequency and voltage support," *Applied Energy,* Article vol. 347, Oct 1 2023, Art no. 121046, doi: 10.1016/j.apenergy.2023.121046.

[20] A. Bedawy, N. Yorino, K. Mahmoud, Y. Zoka, and Y. Sasaki, "Optimal Voltage Control Strategy for Voltage Regulators in Active Unbalanced Distribution Systems Using Multi-Agents," *Ieee Transactions on Power Systems,* Article vol. 35, no. 2, pp. 1023-1035, Mar 2020, doi: 10.1109/tpwrs.2019.2942583.

[21] Q. Wang, S. Lin, Y. Yang, and M. Liu, "A decomposition and coordination algorithm for SVSM interval of integrated transmission and distribution networks considering the uncertainty of renewable energy," *International Journal of Electrical Power & Energy Systems,* Article vol. 136, Mar 2022, Art no. 107761, doi: 10.1016/j.ijepes.2021.107761.

[22] C. Liang, L. Guo, A. Zocca, S. Low, and A. Wierman, "Adaptive Network Response to Line Failures in Power Systems," *Ieee Transactions on Control of Network Systems,* Article vol. 10, no. 1, pp. 333-344, Mar 2023, doi: 10.1109/tcns.2022.3203367.

[23] W. Jiao, J. Chen, Q. Wu, C. Li, B. Zhou, and S. Huang, "Distributed Coordinated Voltage Control for Distribution Networks With DG and OLTC Based on MPC and Gradient Projection," *Ieee Transactions on Power Systems,* Article vol. 37, no. 1, pp. 680-690, Jan 2022, doi: 10.1109/tpwrs.2021.3095523.

[24] S. Wang, C. Zhao, L. Fan, and R. Bo, "Distributionally Robust Unit Commitment With Flexible Generation Resources Considering Renewable Energy Uncertainty," *Ieee Transactions on Power Systems,* Article vol. 37, no. 6, pp. 4179-4190, Nov 2022, doi: 10.1109/tpwrs.2022.3149506.

[25] X. Li, G. Chen, C. Li, Z. Xu, F. Luo, and Z. Y. Dong, "Communication-Efficient Distributed Pricing for Power-Hydrogen Systems With Electric Vehicles and Renewable Energy Integration," *Ieee Transactions on Smart Grid,* Article vol. 16, no. 1, pp. 541-553, Jan 2025, doi: 10.1109/tsg.2024.3413755.

[26] Y. Chen, J. Zhu, Y. Liu, L. Zhang, and J. Zhou, "Distributed Hierarchical Deep Reinforcement Learning for Large-Scale Grid Emergency Control," *Ieee Transactions on Power Systems,* Article vol. 39, no. 2, pp. 4446-4458, Mar 2024, doi: 10.1109/tpwrs.2023.3298486.

[27] N. Sahani and C.-C. Liu, "Model-Based Detection of Coordinated Attacks (DCA) in Distribution Systems," *Ieee Open Access Journal of Power and Energy,* Article vol. 11, pp. 558-570, 2024 2024, doi: 10.1109/oajpe.2024.3489477.

[28] J. Yan, Y. Li, J. Yao, S. Yang, F. Li, and K. Zhu, "Look-Ahead Unit Commitment With Adaptive Horizon Based on Deep Reinforcement Learning," *Ieee Transactions on Power Systems,* Article vol. 39, no. 2, pp. 3673-3684, Mar 2024, doi: 10.1109/tpwrs.2023.3286094.

[29] Y. Zhang, M. Yue, J. Wang, and S. Yoo, "Multi-Agent Graph-Attention Deep Reinforcement Learning for Post-Contingency Grid Emergency Voltage Control," *Ieee Transactions on Neural Networks and Learning Systems,* Article vol. 35, no. 3, pp. 3340-3350, Mar 2024, doi: 10.1109/tnnls.2023.3341334.

[30] H. Gao, T. Jin, C. Feng, C. Li, Q. Chen, and C. Kang, "Review of virtual power plant operations: Resource coordination and multidimensional interaction," *Applied Energy,* Review vol. 357, Mar 1 2024, Art no. 122284, doi: 10.1016/j.apenergy.2023.122284.

[31] Z. Zheng *et al.*, "A De-aggregation strategy based optimal co-scheduling of heterogeneous flexible resources in virtual power plant," *Applied Energy,* Article vol. 383, Apr 1 2025, Art no. 125404, doi: 10.1016/j.apenergy.2025.125404.

[32] Y. Li, W. Chang, and Q. Yang, "Deep reinforcement learning based hierarchical energy management for virtual power plant with aggregated multiple heterogeneous microgrids," *Applied Energy,* Article vol. 382, Mar 15 2025, Art no. 125333, doi: 10.1016/j.apenergy.2025.125333.

[33] Q. Li *et al.*, "Co-optimization of virtual power plants and distribution grids: Emphasizing flexible resource aggregation and battery capacity degradation," *Applied Energy,* Article vol. 377, Jan 1 2025, Art no. 124519, doi: 10.1016/j.apenergy.2024.124519.

[34] J. Liu, Z. Tang, Y. Liu, Y. Zhou, P. P. Zeng, and Q. Wu, "Region-inspired distributed optimal dispatch of flexibility providers in coordinated transmission-distribution framework," *Energy,* Article vol. 319, Mar 15 2025, Art no. 134985, doi: 10.1016/j.energy.2025.134985.

[35] Z. Yan and Y. Xu, "A Hybrid Data-Driven Method for Fast Solution of Security-Constrained Optimal Power Flow," *Ieee Transactions on Power Systems,* Article vol. 37, no. 6, pp. 4365-4374, Nov 2022, doi: 10.1109/tpwrs.2022.3150023.

[36] A. M. Farid, "A Profit-Maximizing Security-Constrained IV-AC Optimal Power Flow Model & Global Solution," *Ieee Access,* Article vol. 10, pp. 2842-2859, 2022 2022, doi: 10.1109/access.2021.3138972.

[37] J. Sliwak, E. D. Andersen, M. F. Anjos, L. Letocart, and E. Traversi, "A Clique Merging Algorithm to Solve Semidefinite Relaxations of Optimal Power Flow Problems," *Ieee Transactions on Power Systems,* Article vol. 36, no. 2, pp. 1641-1644, Mar 2021, doi: 10.1109/tpwrs.2020.3044501.

[38] A. R. Sayed, X. Zhang, G. Wang, C. Wang, and J. Qiu, "Optimal Operable Power Flow: Sample-Efficient Holomorphic Embedding-Based Reinforcement Learning," *Ieee Transactions on Power Systems,* Article vol. 39, no. 1, pp. 1739-1751, Jan 2024, doi: 10.1109/tpwrs.2023.3266773.

[39] X. Pan, T. Zhao, M. Chen, and S. Zhang, "DeepOPF: A Deep Neural Network Approach for Security-Constrained DC Optimal Power Flow," *Ieee Transactions on Power Systems,* Article vol. 36, no. 3, pp. 1725-1735, May 2021, doi: 10.1109/tpwrs.2020.3026379.

[40] A. Velloso and P. Van Hentenryck, "Combining Deep Learning and Optimization for Preventive Security-Constrained DC Optimal Power Flow," *Ieee Transactions on Power Systems,* Article vol. 36, no. 4, pp. 3618-3628, Jul 2021, doi: 10.1109/tpwrs.2021.3054341.

[41] L. You, H. Ma, T. K. Saha, and G. Liu, "Risk-Based Contingency-Constrained Optimal Power Flow With Adjustable Uncertainty Set of Wind Power," *Ieee Transactions on Industrial Informatics,* Article vol. 18, no. 2, pp. 996-1008, Feb 2022, doi: 10.1109/tii.2021.3076801.

[42] A. R. Sayed, C. Wang, and T. Bi, "Resilient operational strategies for power systems considering the interactions with natural gas systems," *Applied Energy,* Article vol. 241, pp. 548-566, May 1 2019, doi: 10.1016/j.apenergy.2019.03.053.

[43] N. Y. Puvvada, A. Mohapatra, and S. C. Srivastava, "Robust AC Transmission Expansion Planning Using a Novel Dual-Based Bi-Level Approach," *Ieee Transactions on Power Systems,* Article vol. 37, no. 4, pp. 2881-2893, Jul 2022, doi: 10.1109/tpwrs.2021.3125719.

[44] S. Zhang, Y. Fang, H. Zhang, H. Cheng, and X. Wang, "Maximum Hosting Capacity of Photovoltaic Generation in SOP-Based Power Distribution Network Integrated With Electric Vehicles," *Ieee Transactions on Industrial Informatics,* Article vol. 18, no. 11, pp. 8213-8224, Nov 2022, doi: 10.1109/tii.2022.3140870.

[45] A. R. Sayed, C. Wang, J. Zhao, and T. Bi, "Distribution-Level Robust Energy Management of Power Systems Considering Bidirectional Interactions With Gas Systems," *Ieee Transactions on Smart Grid,* Article vol. 11, no. 3, pp. 2092-2105, May 2020, doi: 10.1109/tsg.2019.2947219.

[46] Y. Xu, J. Hu, W. Gu, W. Su, and W. Liu, "Real-Time Distributed Control of Battery Energy Storage Systems for Security Constrained DC-OPF," *Ieee Transactions on Smart Grid,* Article vol. 9, no. 3, pp. 1580-1589, May 2018, doi: 10.1109/tsg.2016.2593911.

[47] H. Ebrahimi, A. Yazdaninejadi, and S. Golshannavaz, "Decentralized prioritization of demand response programs in multi-area power grids based on the security

considerations," *Isa Transactions,* Article vol. 134, pp. 396-408, Mar 2023, doi: 10.1016/j.isatra.2022.07.031.

[48] Z. Tan, H. Zhong, Q. Xia, C. Kang, X. S. Wang, and H. Tang, "Estimating the Robust P-Q Capability of a Technical Virtual Power Plant Under Uncertainties," *Ieee Transactions on Power Systems,* Article vol. 35, no. 6, pp. 4285-4296, Nov 2020, doi: 10.1109/tpwrs.2020.2988069.

[49] H. Hui, Y. Chen, S. Yang, H. Zhang, and T. Jiang, "Coordination control of distributed generators and load resources for frequency restoration in isolated urban microgrids," *Applied Energy,* Article vol. 327, Dec 1 2022, Art no. 120116, doi: 10.1016/j.apenergy.2022.120116.

[50] L. Zhu and D. J. Hill, "Data/Model Jointly Driven High-Quality Case Generation for Power System Dynamic Stability Assessment," *Ieee Transactions on Industrial Informatics,* Article vol. 18, no. 8, pp. 5055-5066, Aug 2022, doi: 10.1109/tii.2021.3123823.

[51] F. Wei, Z. Wan, and H. He, "Cyber-Attack Recovery Strategy for Smart Grid Based on Deep Reinforcement Learning," *Ieee Transactions on Smart Grid,* Article vol. 11, no. 3, pp. 2476-2486, May 2020, doi: 10.1109/tsg.2019.2956161.

[52] Y. Zhou, W.-J. Lee, R. Diao, and D. Shi, "Deep Reinforcement Learning Based Real-time AC Optimal Power Flow Considering Uncertainties," *Journal of Modern Power Systems and Clean Energy,* Article vol. 10, no. 5, pp. 1098-1109, Sep 2022, doi: 10.35833/mpce.2020.000885.

[53] 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,* Article vol. 11, no. 4, pp. 3146-3155, Jul 2020, doi: 10.1109/tsg.2020.2967430.

[54] F. Charbonnier, T. Morstyn, and M. D. McCulloch, "Scalable multi-agent reinforcement learning for distributed control of residential energy flexibility," *Applied Energy,* Article vol. 314, May 15 2022, Art no. 118825, doi: 10.1016/j.apenergy.2022.118825.

[55] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward Constrained Policy Optimization," *Arxiv,* preprint Dec 26 2018, doi: arXiv:1805.11074.

[56] T. L. Vu, S. Mukherjee, R. Huang, Q. Huang, and Ieee, "Barrier Function-based Safe Reinforcement Learning for Emergency Control of Power Systems," in *60th IEEE Conference on Decision and Control (CDC)*, Electr Network, 2021, Dec 13-17 2021, in IEEE Conference on Decision and Control, 2021, pp. 3652-3657, doi: 10.1109/cdc45484.2021.9683573. [Online]. Available: <Go to ISI>://WOS:000781990303040

[57] Y. Chow, O. Nachum, A. Faust, E. Duenez-Guzman, and M. Ghavamzadeh, "Lyapunov-based Safe Policy Optimization for Continuous Control," *Arxiv,* preprint Feb 11 2019, doi: arXiv:1901.10031.

[58] J. Lei *et al.*, "A Reinforcement Learning Approach for Defending Against Multiscenario Load Redistribution Attacks," *Ieee Transactions on Smart Grid,* Article vol. 13, no. 5, pp. 3711-3722, Sep 2022, doi: 10.1109/tsg.2022.3175470.

[59] D. Cao *et al.*, "Deep Reinforcement Learning Enabled Physical-Model-Free Two-Timescale Voltage Control Method for Active Distribution Systems," *Ieee Transactions on Smart Grid,* Article vol. 13, no. 1, pp. 149-165, Jan 2022, doi: 10.1109/tsg.2021.3113085.

[60] J. Liu, Y. Zhang, K. Meng, Z. Y. Dong, Y. Xu, and S. Han, "Real-time emergency load shedding for power system transient stability control: A risk-averse deep learning method," *Applied Energy,* Article vol. 307, Feb 1 2022, Art no. 118221, doi: 10.1016/j.apenergy.2021.118221.

[61] X. Sun, J. Qiu, Y. Tao, Y. Ma, and J. Zhao, "Coordinated Real-Time Voltage Control in Active Distribution Networks: An Incentive-Based Fairness Approach," *Ieee Transactions on Smart Grid,* Article vol. 13, no. 4, pp. 2650-2663, Jul 2022, doi: 10.1109/tsg.2022.3162909.

[62] S. Wang, L. Du, X. Fan, and Q. Huang, "Deep Reinforcement Scheduling of Energy Storage Systems for Real-Time Voltage Regulation in Unbalanced LV Networks With High PV Penetration," *Ieee Transactions on Sustainable Energy,* Article vol. 12, no. 4, pp. 2342-2352, Oct 2021, doi: 10.1109/tste.2021.3092961.

[63] A. Das, E. I. Batzelis, S. Anand, and S. R. Sahoo, "Network-Agnostic Adaptive PQ Adjustment Control for Grid Voltage Regulation in PV Systems," *Ieee Transactions on Industry Applications,* Article vol. 58, no. 5, pp. 5792-5804, Sep 2022, doi: 10.1109/tia.2022.3180280.

[64] M. Castilla, J. Miret, J. Luis Sosa, J. Matas, and L. Garcia de Vicuna, "Grid-Fault Control Scheme for Three-Phase Photovoltaic Inverters With Adjustable Power Quality Characteristics," *Ieee Transactions on Power Electronics,* Article vol. 25, no. 12, pp. 2930-2940, Dec 2010, doi: 10.1109/tpel.2010.2070081.

[65] Z. Liu, X. Lv, F. Wu, and Z. Li, "Multi-Mode Active Inertia Support Strategy for MMC-HVDC Systems Considering the Constraint of DC Voltage Fluctuations," *Ieee Transactions on Power Delivery,* Article vol. 38, no. 4, pp. 2767-2781, Aug 2023, doi: 10.1109/tpwrd.2023.3259039.

[66] X. Sun, J. Qiu, Y. Tao, Y. Ma, and J. Zhao, "A Multi-Mode Data-Driven Volt/Var Control Strategy With Conservation Voltage Reduction in Active Distribution Networks," *Ieee Transactions on Sustainable Energy,* Article vol. 13, no. 2, pp. 1073-1085, Apr 2022, doi: 10.1109/tste.2022.3149267.

[67] S. Huang *et al.*, "Distributed Predefined-Time Control for Power System With Time Delay and Input Saturation," *Ieee Transactions on Power Systems,* Article vol. 40, no. 1, pp. 151-165, Jan 2025, doi: 10.1109/tpwrs.2024.3402233.

[68] D. Cao, W. Hu, J. Zhao, Q. Huang, Z. Chen, and F. Blaabjerg, "A Multi-Agent Deep Reinforcement Learning Based Voltage Regulation Using Coordinated PV Inverters," *Ieee Transactions on Power Systems,* Article vol. 35, no. 5, pp. 4120-4123, Sept 2020, doi: 10.1109/tpwrs.2020.3000652.

[69] Q. Yang, G. Wang, A. Sadeghi, G. B. Giannakis, and J. Sun, "Two-Timescale Voltage Control in Distribution Grids Using Deep Reinforcement Learning," *Ieee Transactions on Smart Grid,* Article vol. 11, no. 3, pp. 2313-2323, May 2020, doi: 10.1109/tsg.2019.2951769.

[70] Z. Tang, D. J. Hill, and T. Liu, "Distributed Coordinated Reactive Power Control for Voltage Regulation in Distribution Networks," *Ieee Transactions on Smart Grid,* Article vol. 12, no. 1, pp. 312-323, Jan 2021, doi: 10.1109/tsg.2020.3018633.

[71] A. R. Malekpour, A. M. Annaswamy, and J. Shah, "Hierarchical Hybrid Architecture for Volt/Var Control of Power Distribution Grids," *Ieee Transactions on Power Systems,* Article vol. 35, no. 2, pp. 854-863, Mar 2020, doi: 10.1109/tpwrs.2019.2941969.

[72] 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,* Article vol. 11, no. 4, pp. 3008-3018, Jul 2020, doi: 10.1109/tsg.2019.2962625.

[73] H. Xu, A. D. Dominguez-Garcia, V. V. Veeravalli, and P. W. Sauer, "Data-Driven Voltage Regulation in Radial Power Distribution Systems," *Ieee Transactions on Power Systems,* Article vol. 35, no. 3, pp. 2133-2143, May 2020, doi: 10.1109/tpwrs.2019.2948138.

[74] H. Liu and W. Wu, "Two-stage Deep Reinforcement Learning for Inverter-based Volt-VAR Control in Active Distribution Networks," *Arxiv,* preprint May 20 2020, doi: arXiv:2005.11142.

[75] B. Xu *et al.*, "Var-Voltage Control Capability Constrained Economic Scheduling of Integrated Energy Systems," *Ieee Transactions on Industry Applications,* Article; Proceedings Paper vol. 58, no. 6, pp. 6899-6908, Nov 2022, doi: 10.1109/tia.2022.3199675.

[76] X. Sun and J. Qiu, "Two-Stage Volt/Var Control in Active Distribution Networks With Multi-Agent Deep Reinforcement Learning Method," *Ieee Transactions on Smart Grid,* Article vol. 12, no. 4, pp. 2903-2912, Jul 2021, doi: 10.1109/tsg.2021.3052998.

[77] 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,* Article vol. 61, no. 1, pp. 1569-1580, Jan 2025, doi: 10.1109/tia.2024.3462891.

[78] D. Xu, Q. Wu, B. Zhou, C. Li, L. Bai, and S. Huang, "Distributed Multi-Energy Operation of Coupled Electricity, Heating, and Natural Gas Networks," *Ieee Transactions on Sustainable Energy,* Article vol. 11, no. 4, pp. 2457-2469, Oct 2020, doi: 10.1109/tste.2019.2961432.

[79] A. Mohammadi, M. Mehrtash, and A. Kargarian, "Diagonal Quadratic Approximation for Decentralized Collaborative TSO+DSO Optimal Power Flow," *IEEE Transactions on Smart Grid,* vol. 10, no. 3, pp. 2358-70, May 2019, doi: 10.1109/tsg.2018.2796034.

[80] S. Chakrabarti and R. Baldick, "Look-Ahead SCOPF (LASCOPF) for Tracking Demand Variation via Auxiliary Proximal Message Passing (APMP) Algorithm," *International Journal of Electrical Power & Energy Systems,* Article vol. 116, Mar 2020, Art no. 105533, doi: 10.1016/j.ijepes.2019.105533.

[81] S. Sharma and Q. Li, "Decentralized optimization of energy-water nexus based on a mixed-integer boundary compatible algorithm," *Applied Energy,* Article vol. 359, Apr 1 2024, Art no. 122588, doi: 10.1016/j.apenergy.2023.122588.

[82] A. Ravi, L. Bai, V. Cecchi, and F. Ding, "Stochastic Strategic Participation of Active Distribution Networks With High-Penetration DERs in Wholesale Electricity Markets," *Ieee Transactions on Smart Grid,* Article vol. 14, no. 2, pp. 1515-1527, Mar 2023, doi: 10.1109/tsg.2022.3196682.

[83] N. Pourghaderi, M. Fotuhi-Firuzabad, M. Moeini-Aghtaie, M. Kabirifar, and M. Lehtonen, "Exploiting DERs' Flexibility Provision in Distribution and Transmission Systems Interface," *Ieee Transactions on Power Systems,* Article vol. 38, no. 2, pp. 1961-1975, Mar 2023, doi: 10.1109/tpwrs.2022.3209132.

[84] W. Lu, K. Xie, M. Liu, X. Wang, and L. Cheng, "Online Decentralized Tracking for Nonlinear Time-Varying Optimal Power Flow of Coupled Transmission-Distribution Grids," *Ieee Transactions on Power Systems,* Article vol. 39, no. 2, pp. 2706-2722, Mar 2024, doi: 10.1109/tpwrs.2023.3276049.

[85] C. Lin, W. Wu, B. Zhang, B. Wang, W. Zheng, and Z. Li, "Decentralized Reactive Power Optimization Method for Transmission and Distribution Networks Accommodating Large-Scale DG Integration," *Ieee Transactions on Sustainable Energy,* Article vol. 8, no. 1, pp. 363-373, Jan 2017, doi: 10.1109/tste.2016.2599848.

[86] J. Zhao, H. Wang, Y. Liu, Q. Wu, Z. Wang, and Y. Liu, "Coordinated Restoration of Transmission and Distribution System Using Decentralized Scheme," *Ieee Transactions on Power Systems,* Article vol. 34, no. 5, pp. 3428-3442, Sep 2019, doi: 10.1109/tpwrs.2019.2908449.

[87] W. Wang, X. Xiong, Y. He, J. Hu, and H. Chen, "Scheduling of Separable Mobile Energy Storage Systems With Mobile Generators and Fuel Tankers to Boost Distribution System Resilience," *Ieee Transactions on Smart Grid,* Article vol. 13, no. 1, pp. 443-457, Jan 2022, doi: 10.1109/tsg.2021.3114303.

[88] X. Liu, X. Lin, H. Qiu, Y. Li, and T. Huang, "Optimal aggregation and disaggregation for coordinated operation of virtual power plant with distribution network operator," *Applied Energy,* Article vol. 376, Dec 15 2024, Art no. 124142, doi: 10.1016/j.apenergy.2024.124142.

[89] L. Ding, Q.-L. Han, and X.-M. Zhang, "Distributed Secondary Control for Active Power Sharing and Frequency Regulation in Islanded Microgrids Using an Event-Triggered Communication Mechanism," *Ieee Transactions on Industrial Informatics,* Article vol. 15, no. 7, pp. 3910-3922, Jul 2019, doi: 10.1109/tii.2018.2884494.

[90] Z. Luo, H. Liu, N. Wang, T. Zhao, and J. Tian, "Optimal adaptive decentralized under-frequency load shedding for islanded smart distribution network considering wind power uncertainty," *Applied Energy,* Article vol. 365, Jul 1 2024, Art no. 123162, doi: 10.1016/j.apenergy.2024.123162.

[91] M. Mousavi and M. Wu, "A DSO Framework for Market Participation of DER Aggregators in Unbalanced Distribution Networks," *Ieee Transactions on Power Systems,* Article vol. 37, no. 3, pp. 2247-2258, May 2022, doi: 10.1109/tpwrs.2021.3117571.

[92] H. Zhang, S. Ma, T. Ding, Y. Lin, and M. Shahidehpour, "Multi-Stage Multi-Zone Defender-Attacker-Defender Model for Optimal Resilience Strategy With Distribution Line Hardening and Energy Storage System Deployment," *Ieee Transactions on Smart Grid,* Article vol. 12, no. 2, pp. 1194-1205, Mar 2021, doi: 10.1109/tsg.2020.3027767.

[93] X. Wu and A. J. Conejo, "Security-Constrained ACOPF: Incorporating Worst Contingencies and Discrete Controllers," *Ieee Transactions on Power Systems,* Article vol. 35, no. 3, pp. 1936-1945, May 2020, doi: 10.1109/tpwrs.2019.2937105.

[94] L. Zeng, M. Sun, X. Wan, Z. Zhang, R. Deng, and Y. Xu, "Physics-Constrained Vulnerability Assessment of Deep Reinforcement Learning-Based SCOPF," *Ieee Transactions on Power Systems,* Article vol. 38, no. 3, pp. 2690-2704, May 2023, doi: 10.1109/tpwrs.2022.3192558.

[95] X. Kong, Y. Sun, M. A. Khan, L. Zheng, J. Qin, and X. Ji, "Cyber-physical system planning for VPPs supporting frequency regulation considering hierarchical control

and multidimensional uncertainties," *Applied Energy,* Article vol. 353, Jan 1 2024, Art no. 122104, doi: 10.1016/j.apenergy.2023.122104.

[96] S. Wang *et al.*, "A Data-Driven Multi-Agent Autonomous Voltage Control Framework Using Deep Reinforcement Learning," *Ieee Transactions on Power Systems,* Article vol. 35, no. 6, pp. 4644-4654, Nov 2020, doi: 10.1109/tpwrs.2020.2990179.

[97] D. Cao, J. Zhao, W. Hu, F. Ding, Q. Huang, and Z. Chen, "Attention Enabled Multi-Agent DRL for Decentralized Volt-VAR Control of Active Distribution System Using PV Inverters and SVCs," *Ieee Transactions on Sustainable Energy,* Article vol. 12, no. 3, pp. 1582-1592, Jul 2021, doi: 10.1109/tste.2021.3057090.

[98] J. Zhao, F. Li, S. Mukherjee, and C. Sticht, "Deep Reinforcement Learning-Based Model-Free On-Line Dynamic Multi-Microgrid Formation to Enhance Resilience," *Ieee Transactions on Smart Grid,* Article vol. 13, no. 4, pp. 2557-2567, Jul 2022, doi: 10.1109/tsg.2022.3160387.

[99] A. R. Sayed, C. Wang, H. I. Anis, and T. Bi, "Feasibility Constrained Online Calculation for Real-Time Optimal Power Flow: A Convex Constrained Deep Reinforcement Learning Approach," *Ieee Transactions on Power Systems,* Article vol. 38, no. 6, pp. 5215-5227, Nov 2023, doi: 10.1109/tpwrs.2022.3220799.

[100] S. Jung, C. Park, M. Levorato, J.-H. Kim, and J. Kim, "Two-Stage Self-Adaptive Task Outsourcing Decision Making for Edge-Assisted Multi-UAV Networks," *Ieee Transactions on Vehicular Technology,* Article vol. 72, no. 11, pp. 14889-14905, Nov 2023, doi: 10.1109/tvt.2023.3283404.

[101] Y. Li, R. Wang, Y. Li, M. Zhang, and C. Long, "Wind power forecasting considering data privacy protection: A federated deep reinforcement learning approach," *Applied Energy,* Article vol. 329, Jan 1 2023, doi: 10.1016/j.apenergy.2022.120291.

[102] P. Huang, M. Xu, F. Fang, and D. Zhao, "Robust Reinforcement Learning as a Stackelberg Game via Adaptively-Regularized Adversarial Training," in *31st International Joint Conference on Artificial Intelligence (IJCAI)*, Vienna, AUSTRIA, 2022, Jul 23-29 2022, 2022, pp. 3099-3106. [Online]. Available: <Go to ISI>://WOS:001202342303032. [Online]. Available: <Go to ISI>://WOS:001202342303032

[103] J. Yan *et al.*, "Scheduling Post-Disaster Power System Repair With Incomplete Failure Information: A Learning-to-Rank Approach," *Ieee Transactions on Power Systems,* Article vol. 37, no. 6, pp. 4630-4641, Nov 2022, doi: 10.1109/tpwrs.2022.3149983.

[104] Z. Liu and L. Wang, "A Distributionally Robust Defender-Attacker-Defender Model for Resilience Enhancement of Power Systems Against Malicious Cyberattacks," *Ieee Transactions on Power Systems,* Article vol. 38, no. 6, pp. 4986-4997, Nov 2023, doi: 10.1109/tpwrs.2022.3222309.

[105] T. Haarnoja *et al.*, "Soft Actor-Critic Algorithms and Applications," *Arxiv,* preprint Jan 29 2019, doi: arXiv:1812.05905.

[106] P. Christodoulou, "Soft Actor-Critic for Discrete Action Settings," *Arxiv,* preprint Oct 18 2019, doi: arXiv:1910.07207.

[107] A. Abusorrah, A. Alabdulwahab, Z. Li, and M. Shahidehpour, "Minimax-Regret Robust Defensive Strategy Against False Data Injection Attacks," *Ieee Transactions on Smart Grid,* Article vol. 10, no. 2, pp. 2068-2079, Mar 2019, doi: 10.1109/tsg.2017.2788040.

[108] R. Kour, R. Karim, and P. Dersin, *Game Theory and Cyber Kill Chain: A Strategic Approach to Cybersecurity* (International Congress and Workshop on Industrial AI and eMaintenance 2023. Lecture Notes in Mechanical Engineering). 2024, pp. 451-63.

[109] J. Heinrich and D. Silver, "Deep Reinforcement Learning from Self-Play in Imperfect-Information Games," *Arxiv,* preprint Jun 28 2016, doi: arXiv:1603.01121.

[110] S. Bohez, A. Abdolmaleki, M. Neunert, J. Buchli, N. Heess, and R. Hadsell, "Value constrained model-free continuous control," *Arxiv,* preprint Feb 12 2019, doi: arXiv:1902.04623.

[111] G. Gutierrez-Alcaraz, B. Diaz-Lopez, J. M. Arroyo, and V. H. Hinojosa, "Large-Scale Preventive Security-Constrained Unit Commitment Considering <i>N-k</i> Line Outages and Transmission Losses," *Ieee Transactions on Power Systems,* Article vol. 37, no. 3, pp. 2032-2041, May 2022, doi: 10.1109/tpwrs.2021.3116462.

[112] L. Huang, C. S. Lai, Z. Zhao, G. Yang, B. Zhong, and L. L. Lai, "Robust N - k Security-constrained Optimal Power Flow Incorporating Preventive and Corrective Generation Dispatch to Improve Power System Reliability," *Csee Journal of Power and Energy Systems,* Article vol. 9, no. 1, pp. 351-364, Jan 2023, doi: 10.17775/cseejpes.2021.06560.

[113] T. Wu, S. Bu, X. Wei, G. Wang, and B. Zhou, "Multitasking multi-objective operation optimization of integrated energy system considering biogas-solar-wind renewables," *Energy Conversion and Management,* Article vol. 229, Feb 1 2021, Art no. 113736, doi: 10.1016/j.enconman.2020.113736.

[114] H. Li and H. He, "Learning to Operate Distribution Networks With Safe Deep Reinforcement Learning," *Ieee Transactions on Smart Grid,* Article vol. 13, no. 3, pp. 1860-1872, May 2022, doi: 10.1109/tsg.2022.3142961.

[115] B. Park, J. Holzer, and C. L. DeMarco, "A Sparse Tableau Formulation for Node-Breaker Representations in Security-Constrained Optimal Power Flow," *Ieee Transactions on Power Systems,* Article vol. 34, no. 1, pp. 637-647, Jan 2019, doi: 10.1109/tpwrs.2018.2869705.

[116] Y. Zhao, G. Zhang, W. Hu, Q. Huang, Z. Chen, and F. Blaabjerg, "Meta-Learning Based Voltage Control for Renewable Energy Integrated Active Distribution Network Against Topology Change," *Ieee Transactions on Power Systems,* Article vol. 38, no. 6, pp. 5937-5940, Nov 2023, doi: 10.1109/tpwrs.2023.3309536.

[117] M. G. Kashani, M. Mobarrez, S. Bhattacharya, and Ieee, "Smart Inverter Volt-Watt Control Design in High PV Penetrated Distribution Systems," in *9th Annual IEEE Energy Conversion Congress and Exposition (ECCE)*, Cincinnati, OH, 2017, Oct 01-10 2017, in IEEE Energy Conversion Congress and Exposition, 2017, pp. 4447-4452. [Online]. Available: <Go to ISI>://WOS:000426847404108. [Online]. Available: <Go to ISI>://WOS:000426847404108

[118] H. Liu, W. Wu, and Y. Wang, "Bi-Level Off-Policy Reinforcement Learning for Two-Timescale Volt/VAR Control in Active Distribution Networks," *Ieee Transactions on Power Systems,* Article vol. 38, no. 1, pp. 385-395, Jan 2023, doi: 10.1109/tpwrs.2022.3168700.

[119] T. Zhang, L. Yu, D. Yue, C. Dou, X. Xie, and G. P. Hancke, "Two-Timescale Coordinated Voltage Regulation for High Renewable-Penetrated Active Distribution Networks Considering Hybrid Devices," *Ieee Transactions on*

Industrial Informatics, Article vol. 20, no. 3, pp. 3456-3467, Mar 2024, doi: 10.1109/tii.2023.3308348.

[120] K. P. Schneider *et al.*, "Analytic Considerations and Design Basis for the IEEE Distribution Test Feeders," *Ieee Transactions on Power Systems,* Article vol. 33, no. 3, pp. 3181-3188, May 2018, doi: 10.1109/tpwrs.2017.2760011.

[121] M. Menghwar, J. Yan, Y. Chi, M. A. Amin, and Y. Liu, "A market-based real-time algorithm for congestion alleviation incorporating EV demand in active distribution networks," *Applied Energy,* Article vol. 356, Feb 15 2024, Art no. 122426, doi: 10.1016/j.apenergy.2023.122426.

[122] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *35th International Conference on Machine Learning (ICML)*, Stockholm, SWEDEN, 2018, Jul 10-15 2018, vol. 80, in Proceedings of Machine Learning Research, 2018. [Online]. Available: <Go to ISI>://WOS:000683379201099. [Online]. Available: <Go to ISI>://WOS:000683379201099

[123] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei, "MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers," *Arxiv,* preprint Jun 27 2021, doi: arXiv:2012.15828.

[124] Z. Zhang *et al.*, "Attention Entropy is a Key Factor: An Analysis of Parallel Context Encoding with Full-attention-based Pre-trained Language Models," *Arxiv,* preprint Dec 21 2024, doi: arXiv:2412.16545.

[125] X. Wu, X. Li, J. Li, P. C. Ching, V. C. M. Leung, and H. Vincent Poor, "Caching Transient Content for IoT Sensing: Multi-Agent Soft Actor-Critic," *Arxiv,* preprint Aug 30 2020, doi: arXiv:2008.13191.

[126] C. Samende, J. Cao, and Z. Fan, "Multi-agent deep deterministic policy gradient algorithm for peer-to-peer energy trading considering distribution network constraints," *Applied Energy,* Article vol. 317, Jul 1 2022, Art no. 119123, doi: 10.1016/j.apenergy.2022.119123.

[127] C. Yu *et al.*, "The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games," *Arxiv,* preprint Jul 21 2022, doi: arXiv:2103.01955.

[128] Q. Lin and H. Ma, "SACHA: Soft Actor-Critic with Heuristic-Based Attention for Partially Observable Multi-Agent Path Finding," *Arxiv,* preprint Jul 05 2023, doi: arXiv:2307.02691.

[129] S. Thai *et al.*, "Stabilizing Transformer Training by Preventing Attention Entropy Collapse," in *40th International Conference on Machine Learning*, Honolulu, HI, 2023, Jul 23-29 2023, vol. 202, in Proceedings of Machine Learning Research, 2023. [Online]. Available: <Go to ISI>://WOS:001372555102039. [Online]. Available: <Go to ISI>://WOS:001372555102039