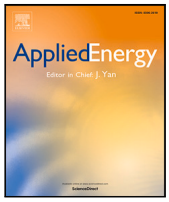
Applied Energy 335 (2023) 120759

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach

Yi Wang [a], Dawei Qiu [a,*], Mingyang Sun [b], Goran Strbac [a], Zhiwei Gao [c]

[a] Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK
[b] Department of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China
[c] Faculty of Engineering and Environment, University of Northumbria, Newcastle upon Tyne, NE1 8ST, UK

ARTICLE INFO

Keywords:
Multi-energy microgrid
Energy management
Dynamic security assessment
Physical-informed safety layer
Reinforcement learning

ABSTRACT

The large-scale integration of distributed energy resources into the energy industry enables the fast transition to a decarbonized future but raises some potential challenges of insecure and unreliable operations. Multi-energy Microgrids (MEMGs), as localized small multi-energy systems, can effectively integrate a variety of energy components with multiple energy sectors, which have been recently recognized as a valid solution to improve the operational security and reliability. As a result, a massive amount of research has been conducted to investigate MEMG energy management problems, including both model-based optimization and model-free learning approaches. Compared to optimization approaches, reinforcement learning is being widely deployed in MEMG energy management problems owing to its ability to handle highly dynamic and stochastic processes without knowing any system knowledge. However, it is still difficult for conventional model-free reinforcement learning methods to capture the physical constraints of the MEMG model, which may therefore destroy its secure operation. To address this research challenge, this paper proposes a novel safe reinforcement learning method by learning a dynamic security assessment rule to abstract a physical-informed safety layer on top of the conventional model-free reinforcement learning energy management policy, which can respect all the physical constraints through mathematically solving an action correction formulation. In this setting, the secure energy management of the MEMG can be guaranteed for both training and test procedures. Extensive case studies based on two integrated systems (i.e., a small 6-bus power and 7-node gas network, and a large 33-bus power and 20-node gas network) are carried out to verify the superior performance of the proposed physical-informed reinforcement learning method in achieving a cost-effective MEMG energy management performance while respecting all the physical constraints, compared to conventional reinforcement learning and optimization approaches.

1. Introduction

Over the last decades, the energy industry has undergone major changes due to various technical, economic, and environmental factors. One of the most remarkable things is related to deregulation and decarbonization, which promise a global energy transition and open up new challenges on both the generation and distribution sides [1]. Distributed energy resources (DERs) (e.g., conventional diesel generators (DGs), renewable energy resources (RESs), and energy storage systems (ESs)) are rapidly becoming attractive due to their high efficiency, increased reliability, and less environmental impact [2]. However, the large-scale integration of DERs into distribution networks also imposes significant operational issues, e.g., demand–supply imbalance, power quality, voltage instability, etc. [3]. Furthermore, the increasing integration of variable and uncertain RESs can influence the secure operation of multi-energy systems (MESs) due to the close interconnection between different energy vectors [4]. In this context, multi-energy microgrids (MEMGs), as an effective and secure coordinated management solution integrating multiple energy vectors, have recently attracted great interest from both the academy and the industry in various aspects, e.g., supporting system demand–supply balances, reducing energy costs, deferring or avoiding generation and transmission reinforcements, etc. [5].

Driven by this desire, there have been substantial efforts focused on the areas of MEMG operation control and energy management at the distribution system level, including both model-based optimization and model-free learning approaches. On one hand, the optimization approaches acquire the complete knowledge (e.g., operation models of DERs and distribution networks) of the system and formulate it as

* Corresponding author.
E-mail address: d.qiu15@imperial.ac.uk (D. Qiu).

https://doi.org/10.1016/j.apenergy.2023.120759
Received 18 August 2022; Received in revised form 19 October 2022; Accepted 22 January 2023
Available online 30 January 2023

Nomenclature

Indices and Sets

$t \in T$	Index and set of time steps (hours)
$b \in EB$	Index and set of electric buses (EBs) in power network
$b \in GB$	Index and set of gas nodes (GBs) in gas network
$l \in PL$	Index and set of branches (PLs) in power network
$l \in GL$	Index and set of pipelines (GLs) in gas network
$g \in DG$	Index and set of diesel generators (DGs)
$g \in GG$	Index and set of gas-fired generators (GGs)
$g \in GW$	Index and set of gas wells (GWs)
$k \in ES$	Index and set of energy storage systems (ESs)
$k \in GS$	Index and set of gas storage systems (GSs)
B_{pgd}	Bus set of upstream power grid
B_{ggd}	Bus set of upstream gas grid
B_{ed}	Bus set of electric demand (ED) d
B_{gd}	Node set of gas demand (GD) d
B_{dg}	Bus set of DG g
B_{gg}	Bus set of GG g
B_{gw}	Node set of GW g
B_{es}	Bus set of ES k
B_{gs}	Node set of GS k
B_{res}	Bus set of renewable energy sources (RES) g

Parameters

Δt	Time resolution (1 h)
λ_t^{p+}	Grid active power buying price at time t (£/kWh)
λ_t^{p-}	Grid active power selling price at time t (£/kWh)
λ_t^{q+}	Grid reactive power buying price at time t (£/kVARh)
λ_t^{q-}	Grid reactive power selling price at time t (£/kVARh)
$c_g^{dg,p}$	Production cost of active power of DG g (£/kWh)
$c_g^{dg,q}$	Production cost of reactive power of DG g (£/kVARh)
$c_g^{gg,p}$	Production cost of active power of GG g (£/kWh)
$c_g^{gg,q}$	Production cost of reactive power of GG g (£/kVARh)
$P_{d,t}^{ed}$	Active demand d at time t (kW)
$Q_{d,t}^{ed}$	Reactive demand d at time t (kVAR)
$\overline{P}_g^{dg}$	Maximum active power of DG g (kW)
$\underline{P}_g^{dg}$	Minimum active power of DG g (kW)
$\overline{Q}_g^{dg}$	Maximum reactive power of DG g (kVAR)
$\underline{Q}_g^{dg}$	Minimum reactive power of DG g (kVAR)
δ_g^{dg}	Rated power factor of DG g
$\overline{P}_g^{gg}$	Maximum active power of GG g (kW)
$\underline{P}_g^{gg}$	Minimum active power of GG g (kW)
$\overline{Q}_g^{gg}$	Maximum reactive power of GG g (kVAR)
$\underline{Q}_g^{gg}$	Minimum reactive power of GG g (kVAR)
δ_g^{gg}	Rated power factor of GG g
b_g^{gg}	Coefficient for gas consumption of GG g (Sm^3/kWh)
$\overline{P}_k^{es}$	Active power capacity of ES k (kW)
$\overline{E}_k^{es}$	Energy capacity of ES k (kWh)
$\underline{E}_k^{es}$	Maximum depth of ES discharge k (kWh)
η_k^{esc}	Charging efficiency of ES k (%)
η_k^{esd}	Discharging efficiency of ES k (%)
$\overline{G}_g^{gw}$	Maximum gas output of GW g (Sm^3/h)
$\underline{G}_g^{gw}$	Minimum gas output of GW g (Sm^3/h)
$\overline{F}_k^{gs}$	Inflation/deflation capacity of GS k (Sm^3/h)
$\overline{G}_k^{gs}$	Gas capacity of GS k (Sm^3)
$\overline{V}$	Maximum permissible voltage (p.u.)
$\underline{V}$	Minimum permissible voltage (p.u.)
B_l	Susceptance of line l (p.u.)
G_l	Conductance of line l (p.u.)
$\overline{S}_l$	Capacity limit of line l (kVA)
$\overline{P}_g^{pgd}$	Active power import limit of upstream power grid g (kW)
$\underline{P}_g^{pgd}$	Active power export limit of upstream power grid g (kW)
$\overline{Q}_g^{pgd}$	Reactive power import limit of upstream power grid g (kVAR)
$\underline{Q}_g^{pgd}$	Reactive power export limit of upstream power grid g (kVAR)
$\overline{G}_l$	Capacity limit of gas pipeline l (Sm^3/h)
λ_l	Compression factor of compressor on gas pipeline l
η_l	Parameter for gas flow and pressure relationship on gas pipeline l ($Sm^3/(h{\cdot}bar^2)$)
$\overline{G}_g^{ggd}$	Gas supply limit of upstream gas grid g (Sm^3/h)

Variables

$P_{g,t}^{dg}$	Active power generation of DG g at time t (kW)
$F_{k,t}^{gsc}$	Inflating gas of GS k at time t (Sm^3/h)
$F_{k,t}^{gsd}$	Deflating gas of GS k at time t (Sm^3/h)
$G_{k,t}^{gs}$	Energy content of GS k at time t (Sm^3)
$P_{g,t}^{pgd}$	Active power exchange with the upstream power grid g at time t (kW)
$Q_{g,t}^{pgd}$	Reactive power exchange with the upstream power grid g at time t (kVAR)
$G_{g,t}^{ggd}$	Gas supply from the upstream gas grid g at time t (Sm^3/h)
$V_{b,t}$	Voltage magnitude at EB b at time t (p.u.)
$\delta_{bp,t}$	Voltage angle difference between EB b and bus p at time t (°)
$P_{b,t}^{ex}$	Active power exchange between EB b and other buses at time t (kW)

a constrained optimization problem. In this setting, all the physical constraints can be satisfied. However, it is normally impractical to obtain accurate system knowledge due to aging and privacy issues [6]. Furthermore, solving an optimization problem for each state condition is time-consuming, especially when taking various system uncertainties and dynamics into account [7]. On the other hand, the model-free

$Q^{ex}_{b,t}$	Reactive power exchange between EB b and other buses at time t (kVAR)
$Q^{dg}_{g,t}$	Reactive power generation of DG g at time t (kVAR)
$P_{l,t}$	Active power of PL l at t (kW)
$Q_{l,t}$	Reactive power of PL l at t (kVAR)
$P^{res}_{g,t}$	Active power output of RES g at time t (kW)
$G_{l,t}$	Gas flow of GL l at time step t (Sm^3/h)
$\rho_{b,t}$	Gas pressure of GB b at time step t (bar)
$P^{gg}_{g,t}$	Active power generation of GG g at time t (kW)
$Q^{gg}_{g,t}$	Reactive power generation of GG g at time t (kVAR)
$P^{esc}_{k,t}$	Charging power of ES k at time t (kW)
$P^{esd}_{k,t}$	Discharging power of ES k at time t (kW)
$E^{es}_{k,t}$	Energy content of ES k at time t (kWh)
$u^{es}_{k,t}$	Binary indicating the battery status of charging ($u^{es}_{k,t} = 1$) or discharging/idle ($u^{es}_{k,t} = 0$)
$G^{gw}_{g,t}$	Gas output of GW g at time step t (Sm^3/h)

learning approaches can be deployed in real-time control and do not depend on any prior knowledge of the system. Furthermore, the well-learned control policies can adapt to various state conditions of the system uncertainties and dynamics. However, it is difficult for the model-free learning approaches to capture the physical constraints of the studied system, which can be prone to insecure operations. This is because the system model is assumed to be a black-box, and the microgrid central controller (MGCC) has no idea how to safely operate the studied system without its complete knowledge. As such, this paper proposes a novel physical-informed reinforcement learning (RL) approach for the secure operation of MEMGs that can satisfy all the physical constraints appropriately while also ensuring their effective energy management.

1.1. Literature review on model-based approaches

So far, model-based optimization approaches have contributed to most of the existing literature on microgrid (MG) energy management problems [8]. In [9], a game-theoretic modeling approach is proposed to integrate the supply-side and demand-side responses for the effective energy management of an isolated MG. However, this paper assumes perfect forecasts of demand profiles and renewable energy. The obtained solutions, thus, may not be consistent with reality and practicality. In order to capture system uncertainties of RESs, demand, and electricity prices, a risk-constrained stochastic framework has been developed in [10] for the joint energy and reserve scheduling of autonomous MGs. In [11], a two-stage robust bi-level energy sharing framework is proposed to overcome the impact of uncertainties associated with market prices and RESs for a prosumer MG. Furthermore, in a shorter amount of time, an online optimization approach based on model predictive control (MPC) has been developed in [12] for the real-time energy management of MGs. However, it is worth noting that MPC-based approaches need to consider the impact of future scenarios for MG operations and solve a time-coupled physical model at each time step, which can be time-consuming [13]. Additionally, the length of the rolling horizon is empirical and normally difficult to select.

Apart from MGs involving only power sector, integrated MEMGs have recently attracted much interest in terms of energy management owing to their various advantages, such as increased system reliability and efficiency, reduced fuel consumption, energy cost, and carbon emissions [4]. For instance, in [14], a two-stage stochastic optimization approach based on scenario analysis is proposed for the energy management of MEMGs, considering the stochastic processes of wind power generation and demand profiles. In [15], an adjustable and robust formulation is developed for the optimal energy management of MEMGs, capturing uncertainties associated with energy demand and RESs. In [16], a multi-timescale coordinated adaptive robust optimization approach is suggested for the energy management problem of industrial MEMGs to handle the uncertain renewable generation. However, all the above papers entirely ignore the network modeling of integrated MESs, which cannot capture technical constraints related to system stability properties, e.g., nodal voltage and power flow limits in power networks as well as nodal pressure and gas flow limits in gas networks. In [17], a two-stage robust framework involving internal energy network constraints is proposed for the multi-temporal time-ahead energy management of smart MEMGs. In [18], the operational constraints of both gas and power networks are included in an edge-based MEMG modeling framework. In [19], a distributed multi-period operational model is proposed for the energy management of MEMGs, characterized by the detailed modeling of coupled power, heating, and gas energy networks.

Overall, extensive efforts have been made to develop model-based optimization approaches to study MG or MEMG energy management problems for different purposes. However, the limitations of the above research cannot be erased and are summarized hereafter. First, all the above papers rely on the accuracy of uncertainty forecasting, which is normally impractical [20]. Second, uncertainties are handled via stochastic programming or robust optimization approaches in most of the above research, which may only be able to capture a small number of representative scenarios or lead to very conservative optimization results. Meanwhile, stochastic programming approaches can be time-consuming, especially when a large number of scenarios are involved. Hence, they are not capable of providing timely services for real-time MEMG energy management. Third, considering the highly stochastic and dynamic real-world environment, it is difficult to explicitly acquire the mathematical models and technical parameters of all resources and networks inside the investigated MEMG, which are crucial for model-based MEMG energy management to obtain secure solutions.

1.2. Literature review on model-free RL approaches

In view of the aforementioned shortcomings in model-based optimization approaches, *reinforcement learning* (RL) [21] is a model-free control method for studying the sequential decision-making process of an agent who can gradually learn the optimal control strategies based on the experiences gained from repeated interactions with the environment, without a *prior* knowledge. Furthermore, being an online learning method, RL can make efficient use of growing amounts of data from the environment, thereby capturing system uncertainties and adapting to various state dynamics. Finally, once the RL method is well trained, its policy can be delivered to the online test set on timescales of milliseconds without requiring any identification. Therefore, RL is claimed as an effective tool for real-time automatic energy management problems.

As reviewed in [8], previous work has successfully utilized various RL methods to solve MG or MEMG energy management problems. For example, a conventional Q-learning (QL) method is used to learn the optimal strategies for energy management and demand scheduling of an MG [22]. However, QL depends on a look-up Q-table that discretizes both the state and action domains, thereby suffering from the curse of dimensionality [21]. To address this issue, a deep Q-network (DQN) [23] method is proposed to learn the comprehensive state features through a deep neural network (DNN) to approximate the Q-value function. In the literature, DQN has been applied to enhance the energy management system of an MG that coordinates different flexible sources [24]. To further represent the continuous action space,

the policy-based RL methods such as deep deterministic policy gradient (DDPG) [25] and proximal policy optimization (PPO) [26] have been successfully applied to the MG energy management problems. Regarding the application of RL methods in MEMGs, a real-time autonomous energy management strategy combining the DDPG method with prioritized experience replay is proposed in [7] for the energy management of a residential MEMG. In [27], the scheduling policies of different types of loads in MEMGs are optimized via a PPO method. In [28], a model-free safe RL method is proposed to solve the optimal control problem of a renewable-based MEMG while satisfying the operating constraints of all its controllable devices. However, it is worth noting that the RL methods developed in the above papers drive MGs and MEMGs to make energy management decisions without considering any physical constraints of the power and/or gas networks inside the operation model, e.g., nodal voltage limit, power line capacity, nodal gas pressure limit, and pipeline capacity, which can lead to insecure operations. This is because the DNN training process for conventional RL methods is an unconstrained learning problem that ignores system physical constraints.

To address this practical issue, [25,26] have formulated the physical constraint violations as penalty terms added to the reward function, which aims to satisfy the operation constraints of battery energy capacity and main grid power exchange limit. However, designing a reward function becomes a complicated and challenging task, especially as the number of constraints increases. As a result, the constrained Markov decision process (CMDP) based on the Lagrange multipliers is proposed in [29] for MGs, which formulates the power flow constraints into a gradient descent algorithm during the training process. In [30], a constrained soft actor–critic (SAC) method based on a CMDP framework is proposed to solve the voltage control problem of active distribution networks (ADNs). In [31], a safe RL method based on CMDP has been developed to solve the optimal operation problem of distribution networks with the objectives of voltage regulation and energy cost minimization. However, running gradient descent with every policy query (i.e., forward propagation) requires sophisticated in-graph implementation and is computationally intensive, while the gradient calculation can cause numerical instabilities and long convergence times, and requires careful step-size selection [32]. Additionally, determining gradient factors (i.e., Lagrange multipliers) in a CMDP requires complete knowledge of distribution network models and parameters. Furthermore, it is noted that reliability and security issues are crucial in energy systems, so the safety of the system operation model has to be guaranteed even during the initial exploration of RL training process, which is unattainable in the above papers.

To ensure the safe operation of the entire RL training process, a CMDP is proposed in [33] to solve the optimal voltage control problem of an ADN, while a safety layer is introduced to correct the voltage control actions to maintain all bus voltages within the acceptable range. The rationale behind the safety layer is to approximate voltage constraints with a first-order linear programming. After that, this safety layer can be added on top of the RL policy to solve a constrained optimization problem with a closed-form analytical solution of the voltage action corrections. However, the proposed safety layer is pre-trained and does not continue updating during the RL training process, which may lead to an inaccurate approximation of the voltage constraints when a growing number of new system states are generated. In [34], a model-augmented RL method featuring a safety layer is proposed for the same voltage control problem of an ADN, where a mutual information regularizer (i.e., safety layer) is developed to improve the approximation quality of the voltage constraints. However, it is worth noting that the two above papers only focus on avoiding the violations of voltage constraints, while ignoring other technical constraints, such as the power flow constraints and the DER operating constraints, which cannot completely ensure the operation security of the studied system. Additionally, the approximation of an ADN focuses on the power network sector rather than an integrated MEMG setup that is capable of capturing both the power and gas networks. Finally, the above two papers assume a virtual ADN environment to represent the RL state transition. This is, however, inaccurate since the virtual version cannot exactly simulate the real ADN, which may raise potential safety issues when applied to real-world test models.

1.3. Paper contributions

To highlight the contributions of this paper, existing literature associated with the energy management problems of both MG and MEMG has been systematically organized in Table 1. On one hand, compared to the model-based optimization methods [9–12,14–19], this paper employs a model-free safe RL method that can learn the energy management control policy of an MEMG, while ensuring the secure network operation. On the other hand, compared to the existing RL methods [7,22,24–31,33,34], a significant research gap has been identified, which drives the motivation behind this paper: no previous work has developed a safe and automatic control method for the real-time energy management of MEMGs that can satisfy all the physical constraints pertaining to the MEMG model. To fill this research gap, this paper proposes for the first time a novel physical-informed RL method that integrates the benefit of physical models for secure MEMG energy management, inspired by recent advances in human intervention [35] and shielding system [36] towards safe RL concept. More specifically, this paper employs a safety layer that can correct unsafe actions to satisfy the network constraints of nodal voltage limits, bus pressure limits, as well as power and gas flow limits. It should be noted that these constraints can even be satisfied during the RL training process. In more detail, the contributions of this paper have been summarized as below:

(1) *Application*: In contrast to previous work [25,26,29–31,33,34] that only models the power operation, this paper employs an MEMG that integrates both electricity and gas sectors. As a result, the flexibility and synergy among the multi-energy sectors of the MEMG can be obtained. The MEMG energy management problem is then formulated as a finite CMDP [21] subject to all the physical constraints. In this context, the mathematical models and technical parameters related to the MEMG are unnecessary. Meanwhile, the system uncertainties and dynamics can be captured without requiring their statistical knowledge.
(2) *Security assessment*: In contrast to previous work that uses penalization method added to the reward function [25,26], constrained policy gradient method [29–31], and conventional safety layer capturing voltage constraint only [33,34], this paper proposes a more robust dynamic security assessment model. Specifically, the security assessment rule is trained by supervised learning techniques to classify whether an MEMG operating point (operation under voltage, pressure, power, and gas flow constraints) is secure or not via a binary classification (1 if secure; 0 otherwise). Once the security assessment rule is well trained, it can be extracted to formulate a safety layer for action corrections when an unsafe operating point does exist. Furthermore, the pre-trained security assessment rule can be further updated through new operating points generated from the RL training procedure.
(3) *Safe policy*: A novel physical-informed (PI) RL method called PI-SPPO is proposed to efficiently solve the CMDP by casting the security assessment rule of the MEMG into a safety layer on top of a PPO policy [37]. The safety layer corrects the unsafe action in a mathematical manner by solving an optimization problem to discover the minimal change to the original PPO-based action that satisfies all the physical constraints.

Table 1
Summary of existing literature associated with the MG/MEMG energy management problem.

Paper	Method	MG model	Uncertainties	Algorithm	Physical constraints (solution)
[9]	Model-based	MG	Deterministic	Bi-level optimization	No
[10]	Model-based	MG	RES, demand, price	Stochastic programming	Voltage, power flow (optimization)
[11]	Model-based	MG	RES, price	Robust optimization	No
[12]	Model-based	MG	RES, demand	Model predictive control	No
[14]	Model-based	MEMG	RES, demand	Stochastic programming	No
[15]	Model-based	MEMG	RES, demand	Robust optimization	No
[16]	Model-based	MEMG	RES	Robust optimization	No
[17]	Model-based	MEMG	RES, demand	Robust optimization	Power, gas, heat flow (optimization)
[18]	Model-based	MEMG	RES	Dynamic programming	Power and gas flow (state transition)
[19]	Model-based	MEMG	RES, demand, price	Consensus-based ADMM	Power, gas, heat flow (optimization)
[22]	Model-free	MG	RES, demand	Q-learning	No
[24]	Model-free	MG	RES, demand, price	DQN	No
[7]	Model-free	MEMG	RES, demand	DDPG	No
[27]	Model-free	MEMG	RES, demand, price	PPO	No
[28]	Model-free	MEMG	RES, demand	DDPG	No
[25]	Model-free	MG	RES, demand	DDPG	Power balance (penalty)
[26]	Model-free	MG	RES, demand	PPO	Power exchange (penalty)
[29]	Model-free	MG	RES, demand	PG	Voltage, power flow (CMDP)
[30]	Model-free	ADN	RES, demand	SAC	Voltage (CMDP)
[31]	Model-free	ADN	RES, demand, price	PG	Voltage, power flow (CMDP)
[33]	Model-free	ADN	RES, demand	DDPG	Voltage (safety layer)
[34]	Model-free	ADN	RES, demand	AC	Voltage (safety layer)
This	**Model-free**	**MEMG**	**RES, demand, price**	**PPO**	**Voltage, pressure, gas and power flow (safety layer)**

To the best of the authors' knowledge, this is the first work to adopt a safe RL method to study the MEMG energy management problem. It is believed that this work contributes to the field of MEMG energy management problems. Moreover, the proposed PI-SPPO method contributes to the methodology for both the dynamic security assessment rule of a coupled power-gas network as well as the development of an advanced safe RL method.

Given the above list of contributions, extensive case studies have been carried out to verify the effectiveness of the proposed safe RL method. In detail, the proposed PI-SPPO firstly shows its superior performance in handling the physical constraints of the MEMG model (e.g., nodal voltage, line capacity, nodal gas pressure, pipeline capacity, etc.) for both the training and test processes compared to the conventional RL methods. Secondly, the proposed PI-SPPO method produces a real-time and automatic energy management policy that can adapt to various system state conditions in a 6-bus power and 7-node gas network. Thirdly, the scalability of the proposed PI-SPPO method has been demonstrated via three interconnected MEMGs in a larger 33-bus power and 20-node gas network.

1.4. Paper organization

The rest of this paper is organized as follows. Section 2 describes the studied problem and presents the mathematical formulations of various MES components and the integrated power-gas network. Section 3 provides the detailed CMDP formulation of the proposed secure MEMG energy management problem. Section 4 demonstrates the detailed algorithm of the proposed PI-SPPO that can efficiently solve the studied problem. The input data and experiment setup are presented in Section 5, followed by case studies and discussion developed in Section 6 and Section 7, respectively. Finally, conclusions and future extensions are drawn in Section 8.

2. Problem formulation

As illustrated in Fig. 1, we focus on the energy management problem of a multi-energy microgrid (MEMG), which includes an integrated power and gas network. In the power network, a group of DERs are appropriately installed, categorized into conventional generation (e.g., diesel generators (DGs)), gas-fired generators (GGs), renewable-based generation (e.g., photovoltaics (PVs) and wind turbines (WTs)), and energy storage systems (ESs). The electric demands (EDs) capturing normal energy consumption are also located in the power network. In the gas network, gas wells (GWs), gas storage systems (GSs), gas demands (GDs), and GGs are deployed on certain nodes suitably. It is worth noting that the power network and gas network are coupled through GGs.

In order to effectively manage the above utilized energy components (i.e., DGs, GGs, ESs, GWs, and GSs) in both power and gas networks, the MEMG equips a microgrid central controller (MGCC) that can regulate the power and gas dispatches based on (1) the grid information of power and gas price signals; (2) the local information of demand and renewable generation; and (3) the battery information of ES and GS energy content. It is, however, noted that the integrated power and gas network is regarded as a black-box for MGCC, which means unknown network topology, line parameters, uncertain renewable generation and demand fluctuation as well as uncertain price signals associated with the grid electricity prices, including the prices for buying electricity from the upstream grid and selling electricity to the upstream grid. Therefore, the MGCC has to use the limited and observable system information to learn the optimal energy management scheme and make scheduling decisions for secure operation. To better understand the operation model of the examined MEMG, the following subsections aim to express the operational characteristics of its controllable energy components and the constraints of its power and gas networks.

2.1. Operational characteristics of controllable components

This subsection aims at providing the detailed mathematical models of three types of energy generators (DGs, GGs, GWs) and two types of storage units (ESs, GSs).

2.1.1. Dispatchable diesel generators

Small-scale DGs are rapidly becoming attractive in MEMG due to their advantages of easy installation and high reliability. In general, the operational characteristics of a DG unit g can be formulated as:

$$\underline{P}_g^{dg} \le P_{g,t}^{dg} \le \overline{P}_g^{dg}, \forall g \in DG, \forall t \in T, \tag{1}$$

$$\underline{Q}_g^{dg} \le Q_{g,t}^{dg} \le \overline{Q}_g^{dg}, \forall g \in DG, \forall t \in T, \tag{2}$$

$$|Q_{g,t}^{dg}| \le P_{g,t}^{dg} \tan(\cos^{-1} \delta_g^{dg}), \forall g \in DG, \forall t \in T, \tag{3}$$

where $P_{g,t}^{dg}$ and $Q_{g,t}^{dg}$ correspond to the active and reactive power outputs of the DG unit g, which are restricted by their power limits $\overline{P}_g^{dg}/\underline{P}_g^{dg}$

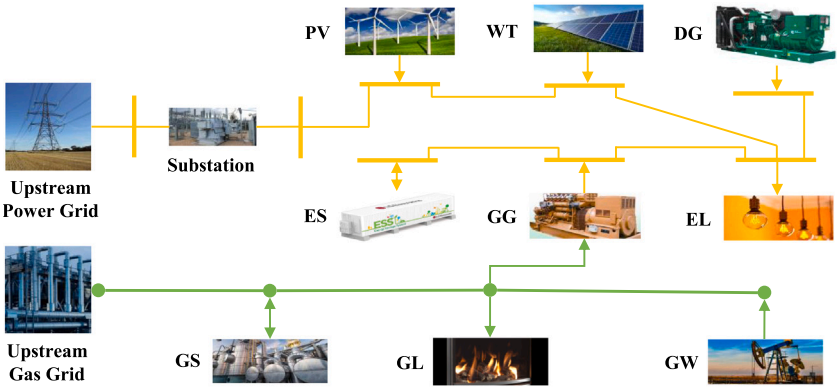


Fig. 1. The utilized MEMG integrating power and gas networks.

in (1) and $\overline{Q}_g^{dg}/\underline{Q}_g^{dg}$ in (2), respectively. Constraint (3) refers to the influence of its rated power factor δ_g^{dg} on active and reactive power generation of the DG unit g [38].

2.1.2. Energy storage systems

As a flexible option, ESs can assist the MEMG to deal with various uncertainties and dynamics via reasonable charging and discharging behaviors, in which the operational characteristics of an ES unit k can be formulated as:

$$E_{k,t+1}^{es} = E_{k,t}^{es} + P_{k,t}^{esc}\eta_k^{esc}\Delta t + \frac{P_{k,t}^{esd}\Delta t}{\eta_k^{esd}}, \forall k \in ES, \forall t \in T, \tag{4}$$

$$\underline{E}_k^{es} \le E_{k,t}^{es} \le \overline{E}_k^{es}, \forall k \in ES, \forall t \in T, \tag{5}$$

$$0 \le P_{k,t}^{esc} \le \overline{P}_k^{es} u_{k,t}^{es}, \forall k \in ES, \forall t \in T, \tag{6}$$

$$\overline{P}_k^{es}(u_{k,t}^{es} - 1) \le P_{k,t}^{esd} \le 0, \forall k \in ES, \forall t \in T, \tag{7}$$

where Eq. (4) refers to the battery's energy dynamics that take the energy losses caused by charging and discharging efficiencies $\eta_k^{esc}, \eta_k^{esd}$ into account. The battery energy content $E_{k,t}^{es}$ as well as the charging and discharging power $P_{k,t}^{esc}, P_{k,t}^{esd}$ are constrained in (5)–(7), where the binary variable $u_{k,t}^{es}$ is introduced to indicate the battery status of charging ($u_{k,t}^{es} = 1$) or discharging ($u_{k,t}^{es} = 0$), since the charging and discharging statuses cannot occur simultaneously.

2.1.3. Gas-fired generators

The GGs realize the coupling between the power and gas networks. In general, GGs consume natural gas as a demand in the gas network and supply electricity as a source to the power network. The operational characteristics of a GG unit g can be formulated as:

$$\underline{P}_g^{gg} \le P_{g,t}^{gg} \le \overline{P}_g^{gg}, \forall g \in GG, \forall t \in T, \tag{8}$$

$$\underline{Q}_g^{gg} \le Q_{g,t}^{gg} \le \overline{Q}_g^{gg}, \forall g \in GG, \forall t \in T, \tag{9}$$

$$P_{g,t}^{gg} = G_{g,t}^{gg}/\eta_g^{gg}, \forall g \in GG, \forall t \in T, \tag{10}$$

$$|Q_{g,t}^{gg}| \le P_{g,t}^{gg}\tan(\cos^{-1}\delta_g^{gg}), \forall g \in GG, \forall t \in T, \tag{11}$$

where $P_{g,t}^{gg}$ and $Q_{g,t}^{gg}$ correspond to the active and reactive power outputs of the GG unit g, which are restricted by the power limits $\overline{P}_g^{gg}/\underline{P}_g^{gg}$ in (8) and $\overline{Q}_g^{gg}/\underline{Q}_g^{gg}$ in (9), respectively. The energy conversion between the power generation $P_{g,t}^{gg}$ and the gas consumption $G_{g,t}^{gg}$ of the GG unit g at time step t is expressed by Eq. (10), where η_g^{gg} represents the energy conversion coefficient in Sm^3/kWh. Constraint (11) refers to the influence of its rated power factor δ_g^{gg} on active and reactive power generation of the GG unit g.

2.1.4. Gas wells

GWs are the main source of the gas network. They are wells drilled specifically for natural gas and contain little or no oil [39]. In general, the operation of a traditional GW unit g is limited by:

$$\underline{G}_g^{gw} \le G_{g,t}^{gw} \le \overline{G}_g^{gw}, \ \forall g \in GW, \ \forall t \in T, \tag{12}$$

where $\underline{G}_g^{gw}$ and $\overline{G}_g^{gw}$ correspond to the lower and upper limits of GW output, respectively.

2.1.5. Gas storage systems

As a flexible option in the gas network, GSs are deployed to store unusable natural gas and then release the stored natural gas to the gas network during the peak period of natural gas demand [39], which can be formulated as:

$$G_{k,t+1}^{gs} = G_{k,t}^{gs} + F_{k,t}^{gsc}\eta_k^{gsc}\Delta t + \frac{F_{k,t}^{gsd}\Delta t}{\eta_k^{gsd}}, \forall k \in GS, \forall t \in T, \tag{13}$$

$$\underline{G}_k^{gs} \le G_{k,t}^{gs} \le \overline{G}_k^{gs}, \forall k \in GS, \forall t \in T, \tag{14}$$

$$0 \le F_{k,t}^{gsc} \le \overline{F}_k^{gs} u_{k,t}^{gs}, \forall k \in GS, \forall t \in T, \tag{15}$$

$$\overline{F}_k^{gs}(u_{k,t}^{gs} - 1) \le F_{k,t}^{gsd} \le 0, \forall k \in GS, \forall t \in T, \tag{16}$$

where Eq. (13) refers to the GS energy dynamics that take inflating and deflating efficiencies $\eta_k^{gsc}, \eta_k^{gsd}$ into account. The GS energy content $G_{k,t}^{gs}$ as well as the gas inflation and deflation quantities $F_{k,t}^{gsc}, F_{k,t}^{gsd}$ are constrained in (14)–(16), where the binary variable $u_{k,t}^{gs}$ is introduced to ensure that gas inflation and deflation cannot happen simultaneously.

2.2. Constraints and limits of power and gas networks

This subsection aims at providing the detailed limits and constraints of the power and gas networks that shall maintain the MEMG secure operation.

2.2.1. Power network

For each time step t, the secure operation of the power network should be always guaranteed, which corresponds to the following limits and constraints:

$$\begin{aligned}&\sum_{g\in B_{pgd}} P_{g,t}^{pgd} + \sum_{g\in B_{dg}} P_{g,t}^{dg} + \sum_{g\in B_{gg}} P_{g,t}^{gg} + \sum_{g\in B_{res}} P_{g,t}^{res} \\ &= \sum_{d\in B_{ed}} P_{d,t}^{ed} + \sum_{k\in B_{es}} (P_{k,t}^{esc} + P_{k,t}^{esd}) + P_{b,t}^{ex}, \forall b \in EB, \forall t \in T,\end{aligned} \tag{17}$$

$$\sum_{g\in B_{pgd}} Q_{g,t}^{pgd} + \sum_{g\in B_{dg}} Q_{g,t}^{dg} + \sum_{g\in B_{gg}} Q_{g,t}^{gg} = \sum_{d\in B_{ed}} Q_{d,t}^{ed} + Q_{b,t}^{ex}, \forall b \in EB, \forall t \in T, \tag{18}$$

$$P_{b,t}^{ex} = \sum_{p\in B} V_{b,t}V_{p,t}(G_{bp}\cos\delta_{bp,t} + B_{bp}\sin\delta_{bp,t}), \forall b \in EB, \forall t, \tag{19}$$

$$Q_{b,t}^{ex} = \sum_{p\in B} V_{b,t}V_{p,t}(G_{bp}\sin\delta_{bp,t} - B_{bp}\cos\delta_{bp,t}), \forall b \in EB, \forall t \in T, \tag{20}$$

$$P_{l,t}^2 + Q_{l,t}^2 \le \overline{S}_l, \forall l \in PL, \forall t \in T, \tag{21}$$

$$\underline{V} \le V_{b,t} \le \overline{V}, \forall b \in EB, \forall t \in T, \tag{22}$$

$$\underline{P}_g^{pgd} \le P_{g,t}^{pgd} \le \overline{P}_g^{pgd}, \forall g \in PGD, \forall t \in T, \tag{23}$$

$$\underline{Q}_g^{pgd} \le Q_{g,t}^{pgd} \le \overline{Q}_g^{pgd}, \forall g \in PGD, \forall t \in T, \tag{24}$$

where the nodal active/reactive power balance at a certain bus b and the AC power flow equations capturing the power network topology are presented in (17)–(18) and (19)–(20), respectively [40]. The sets B_{pgd}, B_{ed}, B_{res}, B_{dg}, B_{gg}, and B_{es} correspond to the bus sets connected with the upstream power grid, EDs, RESs, DGs, GGs, and ESs located at bus b, respectively. Constraints (21) and (22) represent the operational constraints of line power flow and nodal voltage magnitudes, restricted by line capacity $\overline{S}_{bp}$ and voltage limits $\underline{V}, \overline{V}$, respectively. Constraints (23)–(24) restrict the power exchange at the substation between the MEMG and the upstream power grid, where $P_{g,t}^{pgd}$ and $Q_{g,t}^{pgd}$ denote the active and reactive power exchange, respectively.

2.2.2. Gas network

Regarding the natural gas network, a steady state natural gas operation is proposed [41], in which the secure operation of the gas network should be always guaranteed, corresponding to the following limits and constraints:

$$\sum_{g\in B_{ggd}} G_{g,t}^{ggd} + \sum_{g\in B_{gw}} G_{g,t}^{gw} = \sum_{d\in B_{gd}} G_{d,t}^{gd} + \sum_{k\in B_{gs}} (F_{k,t}^{gsc} + F_{k,t}^{gsd}) + \sum_{d\in B_{gg}} G_{g,t}^{gg} + \sum_{pb\in GL} G_{pb,t} - \sum_{bp\in GL} G_{bp,t}, \forall b \in GB, \forall t \in T, \tag{25}$$

$$\rho_{b,t} \le \rho_{p,t} \le \lambda_l \rho_{b,t}, \ \forall l \in GL^{act}, \ \forall t \in T, \tag{26}$$

$$\underline{\rho}_b \le \rho_{b,t} \le \overline{\rho}_b, \ \forall b \in GB, \ \forall t \in T, \tag{27}$$

$$G_{l,t}^2 - \eta_l(\rho_{b,t}^2 - \rho_{p,t}^2) = 0, \ \forall l \in GL^{ina}, \ \forall t \in T, \tag{28}$$

$$\underline{G}_l \le G_{l,t} \le \overline{G}_l, \ \forall l \in GL, \ \forall t \in T, \tag{29}$$

$$\underline{G}_g^{ggd} \le G_{g,t}^{ggd} \le \overline{G}_g^{ggd}, \forall g \in GGD, \forall t \in T, \tag{30}$$

where the nodal gas balance at a certain node b is presented in (25). The sets B_{ggd}, B_{gd}, B_{gw}, B_{gs}, and B_{gg} correspond to the upstream gas grid, GDs, GWs, GSs, and GGs located at node b, respectively. Pipelines without compressors are denoted as inactive pipelines belonging to GL^{ina}, while those with compressors are active pipelines belonging to GL^{act}. The nodal gas pressure $\rho_{b,t}$ for a compressor with the gas flow from node b to node p in GL^{act} is constrained by (26), where λ_l indicates the compressor's compression factor at pipeline l. Constraint (27) ensures that gas pressure at each node stays within a preset range. For an inactive gas pipeline l with gas flow from node b to node p in GL^{ina}, the relationship between gas flow and nodal gas pressure is represented by (28), where η_l represents the relationship between gas flow and pressure based on Weymouth equation [41]. Furthermore, as expressed in (29), the gas flow is limited by the pipeline capacity, while constraint (30) restricts the gas supply from the upstream gas grid.

2.3. Objective function of the MEMG

The objective function of the MEMG is the expectation of cost minimization, taking over the randomness of system uncertainties and stochastic control variables, which can be expressed as:

$$\min \sum_{t=1}^{T} \mathbb{E}\Big\{ \lambda_t^{p+} \sum_{g\in PGD} [P_{g,t}^{pgd}]^+ + \lambda_t^{p-} \sum_{g\in PGD} [P_{g,t}^{pgd}]^- + \lambda_t^{q+} \sum_{g\in PGD} [Q_{g,t}^{pgd}]^+ + \lambda_t^{q-} \sum_{g\in PGD} [Q_{g,t}^{pgd}]^- + \sum_{g\in DG} c_g^{dg,p} P_{g,t}^{dg} + \sum_{g\in DG} c_g^{dg,q} |Q_{g,t}^{dg}| + \sum_{g\in GG} c_g^{gg,p} P_{g,t}^{gg} + \sum_{g\in GG} c_g^{gg,q} |Q_{g,t}^{gg}| + \sum_{g\in GGD} \lambda_t^g G_{g,t}^{ggd} + \sum_{g\in GW} c_g^{gw} G_{g,t}^{gw} \Big\} \tag{31}$$

where the max/min operator $[\cdot]^{+/-} = \max / \min\{\cdot, 0\}$ indicates taking the maximum/minimum value between $\cdot$ and 0. In detail, the operation cost includes: (1) the electricity net cost with the upstream power grid; (2) the DG generation cost; (3) the gas supply cost from the upstream gas grid; and (4) the GW generation cost. Furthermore, in (31), λ_t^{p+} and λ_t^{p-} indicate the buy and sell prices of grid active power $P_{g,t}^{pgd}$; λ_t^{q+} and λ_t^{q-} indicate the buy and sell prices of grid reactive power $Q_{g,t}^{pgd}$; $c_g^{dg,p}$ and $c_g^{dg,q}$ indicate the generation costs of DG active power $P_{g,t}^{dg}$ and reactive power $Q_{g,t}^{dg}$; $c_g^{gg,p}$ and $c_g^{gg,q}$ indicate the generation costs of GG active power $P_{g,t}^{gg}$ and reactive power $Q_{g,t}^{gg}$; λ_t^g indicates the price of gas supply $G_{g,t}^{ggd}$ from the upstream gas grid; and c_g^{gw} indicates the generation cost of GW gas output $G_{g,t}^{gw}$.

2.4. Challenges

Solving the above constrained optimization for an MEMG energy management problem is very challenging. First, the MGCC becomes clueless if the mathematical models and technical parameters of the utilized energy components and networks are unknown, since the optimization problem (1)–(31) cannot be even formulated [6]. Second, the MEMG is characterized by various uncertainties (e.g., demand, renewable generation, and price signals); nevertheless, it is difficult to obtain accurate probability distributions of uncertainties so as to not capture the model representation. Third, solving a time-coupled optimization problem may take a long time, especially when a vast number of high-dimensional stochastic variables are required to optimize [7]. Fourth, it is hard to develop a generalized control scheme that can be applied to any state condition of the MEMG environment, since an independent optimization needs to be resolved for a new state condition.

To this end, an alternative data-driven model-free RL-based method could be proposed to solve the above MEMG energy management problem. In this setting, the MGCC does not require the system knowledge but learns the optimal scheduling decisions by interacting with the unknown environment. In addition, extensive interactions with the environment throughout the learning process can effectively capture system uncertainties. Once the RL method has been trained thoroughly, the control policies can be deployed in milliseconds for realistic energy management decisions in response to any new state condition. Finally, to ensure secure operations, the RL method shall also be capable of taking all the physical constraints of the MEMG setup into account, which leads to the requirement for a physical-informed safe RL method.

3. Constrained Markov decision process

Since the MGCC needs to manage the energy schedules of MEMG's controllable components in a dynamic process with Markovian decisions respecting all the physical constraints of the power and gas networks, it is reasonable to formulate the above MEMG energy management problem (1)–(31) as a *Constrained Markov decision process* (CMDP) [42], as depicted in Fig. 2. The CMDP is defined by $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, C, \gamma \rangle$, including:

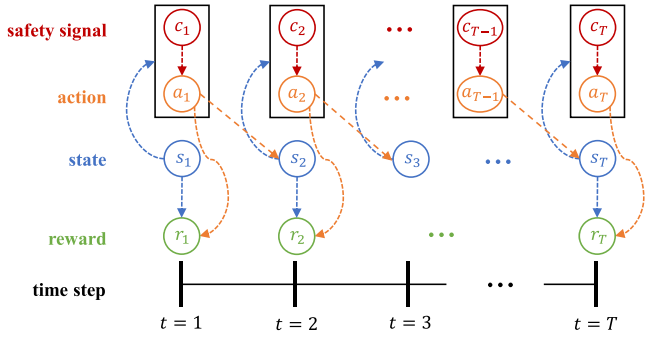


Fig. 2. Architecture of the proposed CMDP.

- a state space $s \in \mathcal{S}$;
- an action space $a \in \mathcal{A}$;
- an immediate scalar reward $r(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$;
- a state transition $\mathcal{T}(s,a,\omega) : \mathcal{S}\times\mathcal{A}\times\mathcal{W} \to \mathcal{S}$ following a conditional probability function $P(s'|s,a,\omega) : \mathcal{S}\times\mathcal{A}\times\mathcal{W}\times\mathcal{S} \to \mathbb{R}$, where $\omega \in \mathcal{W}$ indicates the stochasticity (e.g., renewable, demand, price) in the environment $\mathcal{E}$;
- a set of immediate constraint functions $C = \{c(s,a) : \mathcal{S}\times\mathcal{A} \to \mathbb{R}\}$, where $c(s,a)$ is defined as the safety signal influenced by the state s and action a;
- a discount factor $\gamma \in [0,1)$ used to expect the long-term return of the agent's objective, i.e., cumulative discounted reward $R = \sum_{t=0}^{T} \gamma^t r_t$.

In this case, the MGCC is defined as the agent who employs a policy π to interact with the CMDP and emits a trajectory of states, actions, safety signals, and rewards: $s_1, a_1, c_1, r_1, s_2, a_2, c_2, \ldots, r_T$ over $\mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$. In detail, at each time step t, the agent chooses an action a_t according to the policy $\pi(a|s) : \mathcal{S} \to P(\mathcal{A})$ based upon the current state s_t observed from the environment. The MEMG environment then moves into the next state according to the state transition function $\mathcal{T}(s,a,\omega)$ conditioned on the current state s_t, the executed action a_t, and the stochastic parameters ω_t. The agent then obtains a reward r_t and a new state s_{t+1}. At the same time, the environment also generates the safety signals c_t upon the observed state s_t and the associated action a_t. To summarize, we study a constrained policy optimization of the CMDP to maximize the cumulative discounted reward:

$$\max_{a_t \sim \pi(s_t)} \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)] \tag{32}$$

$$\text{s.t.} \quad c_t(s_t, a_t) \le \overline{c}, \forall t \in T, \tag{33}$$

where the executed action a is sampling from the constrained policy $a \sim \pi(s)$. The safety signals $c(s,a)$ at each state–action pair (s,a) are upper bounded by the corresponding physical limits $\overline{c}$. In this CMDP, the time interval between two consecutive time steps $\Delta t = 1$ h, while $T = 24$ hours is the time horizon of the operation problem. In this context, the main components of the proposed CMDP formulation, including state, action, constraints, state transition, and reward, can be expressed in the following subsections.

3.1. State

The state s_t observed by the MGCC agent at time step t can be defined as:

$$s_t = [\lambda_t^{p+}, \lambda_t^{p-}, \lambda_t^{g}, P_{d,t}^{ed}, G_{d,t}^{gd}, P_{g,t}^{res}, E_{k,t}^{es}, G_{k,t}^{gs}], \forall t \in T, \tag{34}$$

which consists of two parts: (1) the exogenous state representing the local information not affected by the action, including the electricity grid buy and sell prices of active power $\lambda_t^{p+}, \lambda_t^{p-}$ (the electricity grid buy and sell prices of reactive power $\lambda_t^{q+}, \lambda_t^{q-}$ are not considered since they follow the same pattern as the active prices [43]), the price of gas supply from the upstream gas grid λ_t^{g}, the ED $P_{d,t}^{ed}$, the GD $G_{d,t}^{gd}$, and the RES $P_{g,t}^{res}$; and (2) the endogenous state serving as the feedback signals of its executed action, including the current energy storage content $E_{k,t}^{es}$ of the ES unit k and the current gas storage content $G_{k,t}^{gs}$ of the GS unit k.

3.2. Action

The action a_t at time step t indicates the energy schedules of all controllable components in the MEMG, which can be defined as:

$$a_t = [a_{g,t}^{dg,p}, a_{g,t}^{dg,q}, a_{g,t}^{gg,p}, a_{g,t}^{gg,q}, a_{k,t}^{es}, a_{g,t}^{gw}, a_{k,t}^{gs}], \forall t \in T, \tag{35}$$

where actions $a_{g,t}^{dg,p}, a_{g,t}^{dg,q} \in [0,1]$ correspond to the active and reactive power generation magnitudes of the DG unit g as a percentage of its power capacities $[\underline{P}_g^{dg}, \overline{P}_g^{dg}], [\underline{Q}_g^{dg}, \overline{Q}_g^{dg}]$; actions $a_{g,t}^{gg,p}, a_{g,t}^{gg,q} \in [0,1]$ correspond to the active and reactive power generation magnitudes of the GG unit g as a percentage of its power capacities $[\underline{P}_g^{gg}, \overline{P}_g^{gg}], [\underline{Q}_g^{gg}, \overline{Q}_g^{gg}]$; action $a_{k,t}^{es} \in [-1,1]$ represents the magnitude of charging (positive) and discharging (negative) power of the ES unit k as a percentage of its power capacity $[-\overline{P}_k^{es}, \overline{P}_k^{es}]$; actions $a_{g,t}^{gw} \in [0,1]$ correspond to the gas output magnitude of the GW unit g as a percentage of its capacity limit $[\underline{G}_g^{gw}, \overline{G}_g^{gw}]$; and action $a_{k,t}^{gs} \in [-1,1]$ represents the magnitude of gas inflation (positive) and deflation (negative) of the GS unit k as a percentage of its power capacity $[-\overline{F}_k^{gs}, \overline{F}_k^{gs}]$.

3.3. Constraints

The physical constraints considered in this MEMG have been discussed in Sections 2.1 and 2.2, which can be generally categorized into two parts: (1) the operational constraints of all controllable components, i.e., DG (1)–(3), ES (4)–(7), GG (8)–(11), GW (12), and GS (13)–(16); and (2) the operational constraints of power network (17)–(24) and gas network (25)–(30).

In the first category, constraints (1)–(2), (6)–(10), (12), and (15)–(16) associated with power and gas capacities are time-independent and represented by their corresponding lower and upper bounds, which can be directly handled by the following expressions:

$$P_{g,t}^{dg} = a_{g,t}^{dg,p}(\overline{P}_g^{dg} - \underline{P}_g^{dg}) + \underline{P}_g^{dg}, \forall g \in DG, \forall t \in T, \tag{36}$$

$$Q_{g,t}^{dg} = a_{g,t}^{dg,q}(\overline{Q}_g^{dg} - \underline{Q}_g^{dg}) + \underline{Q}_g^{dg}, \forall g \in DG, \forall t \in T, \tag{37}$$

$$P_{g,t}^{gg} = a_{g,t}^{gg,p}(\overline{P}_g^{gg} - \underline{P}_g^{gg}) + \underline{P}_g^{gg}, \forall g \in GG, \forall t \in T, \tag{38}$$

$$Q_{g,t}^{gg} = a_{g,t}^{gg,q}(\overline{Q}_g^{gg} - \underline{Q}_g^{gg}) + \underline{Q}_g^{gg}, \forall g \in GG, \forall t \in T, \tag{39}$$

$$G_{g,t}^{gw} = a_{g,t}^{gw}(\overline{G}_g^{gw} - \underline{G}_g^{gw}) + \underline{G}_g^{gw}, \forall g \in GW, \forall t \in T, \tag{40}$$

$$P_{k,t}^{esc} = [a_{k,t}^{es} \overline{P}_k^{es}]^+, \forall k \in ES, \forall t \in T, \tag{41}$$

$$P_{k,t}^{esd} = [a_{k,t}^{es} \overline{P}_k^{es}]^-, \forall k \in ES, \forall t \in T, \tag{42}$$

$$F_{k,t}^{gsc} = [a_{k,t}^{gs} \overline{G}_k^{gs}]^+, \forall k \in GS, \forall t \in T, \tag{43}$$

$$F_{k,t}^{gsd} = [a_{k,t}^{gs} \overline{G}_k^{gs}]^-, \forall k \in GS, \forall t \in T. \tag{44}$$

However, the power network constraints (17)–(24), the gas network constraints (25)–(30), the power factor constraints (3) and (11), as well as the energy content constraints (5) and (14) cannot be handled via the above straightforward manner. This is because the above mentioned constraints are not directly determined by the action a with a certain level of percentage ratio, but are affected by many factors. On one

hand, the power and gas flow constraints (17)–(24) and (25)–(30) as well as power factor constraints (3) and (11) are highly complex and coupled, while these constraints cannot be handled individually. On the other hand, the storage energy contents (5) and (14) are time-coupling constraints, which are not only affected by the charging/discharging quantities and energy loss efficiencies at the current time step but also affected by the value of energy content in the previous time step. To this end, we develop a set of immediate constraint functions C to describe these unmanageable constraints.

$$\{(17)-(24),(25)-(30),(5),(14),(3),(11)\}=C,\forall t\in T. \tag{45}$$

3.4. State transition

The state transition process from time step t to $t+1$ is governed by $s_{t+1}=\mathcal{T}(s_t,a_t,\omega_t)$ with a probability function $P(s_{t+1}|s_t,a_t,\omega_t)$, which is influenced by the combination of environment current state s_t, agent's action a_t, and environment stochasticity ω_t. In this problem, $\omega_t=[\lambda_t^{p+},\lambda_t^{p-},P_{d,t}^{ed},P_{g,t}^{res},G_{d,t}^{gd}]$ corresponds to the exogenous state that is independent of the agent's action and has intrinsic variability. Since ω_t is influenced by numerous exogenous factors, such as market pricing schemes, energy usage behaviors, solar radiation, wind speed, etc., it poses substantial challenges to discover appropriate probabilistic models that can fully represent such unpredictability. In the machine learning area, RL can overcome this issue by introducing a data-driven solution that does not depend on precise probability distributions for various uncertainties but instead learns state features from the dataset itself [21].

By contrast, the state transition for the endogenous state features $E_{k,t}^{es}$ and $G_{k,t}^{gs}$ are determined by the actions $a_{k,t}^{es}$ and $a_{k,t}^{gs}$, respectively. As discussed in Section 3.3, the ES's charging and discharging power quantities $P_{k,t}^{esc}$, $P_{k,t}^{esd}$ as well as the GS's inflation and deflation quantities $F_{k,t}^{gsc}$, $F_{k,t}^{gsd}$ have been expressed in (41)–(42) and (43)–(44), respectively. The state transition $E_{k,t+1}^{es}$ of the ES unit k and the state transition $G_{k,t+1}^{gs}$ of the GS unit k from time step t to $t+1$ can be automatically calculated as (4) and (13), respectively.

3.5. Reward

The reward function for the MGCC agent at time step t is designed as the negative operation costs of the MEMG (31), which can be expressed as:

$$\begin{aligned}r_t=&-\lambda_t^{p+}\sum_{g\in PUG}[P_{g,t}^{pug}]^+-\lambda_t^{p-}\sum_{g\in PUG}[P_{g,t}^{pug}]^--\lambda_t^{q+}\sum_{g\in PUG}[Q_{g,t}^{pug}]^+\\&-\lambda_t^{q-}\sum_{g\in PUG}[Q_{g,t}^{pug}]^-\\&-\sum_{g\in DG}c_g^{dg,p}P_{g,t}^{dg}-\sum_{g\in DG}c_g^{dg,q}|Q_{g,t}^{dg}|-\sum_{g\in GG}c_g^{gg,p}P_{g,t}^{gg}-\sum_{g\in GG}c_g^{gg,q}|Q_{g,t}^{gg}|\\&-\sum_{g\in GGD}\lambda_t^{g}G_{g,t}^{ggd}-\sum_{g\in GW}c_g^{gw}G_{g,t}^{gw},\ \forall t\in T.\end{aligned} \tag{46}$$

4. Proposed physical-informed reinforcement learning

In order to properly solve the above CMDP, we propose a novel safe RL method called PI-SPPO, with its general architecture being depicted in Fig. 3. Specifically, PI-SPPO generates three practical implementation details for the studied MEMG energy management problem respecting all the physical constraints:

(1) *Security assessment rule*: approximate a safe operation region of the safety constraint set C in (45) of the examined MEMG, and then embed the approximated security assessment rule in a physical-informed safety layer on top of the model-free RL-based control policy.

(2) *Model-free PPO control policy*: utilize the actor–critic architecture of the proximal policy optimization (PPO) algorithm [37] that is capable of handling the high-dimensional continuous state and action spaces of the MEMG energy management problem, with a stable learning performance, high sampling efficiency, and little hyperparameter tuning.

(3) *Physical-informed safety layer*: construct a safety layer that can auto-correct the action computed from the model-free PPO control policy to maintain the secure operation of the safety constraint set C by mathematically solving an analytical optimization problem subject to the security assessment rule.

In this context, the proposed PI-SPPO is realized as a completely model-free method, which means the MGCC agent has no knowledge about the investigated MEMG environment. However, the MGCC agent can learn the system characteristics and the control policy through its interactions with the MEMG environment. It is worth noting that the three implementation details listed above are all coupled and highly interacted with each other. On one hand, the security assessment rule can be pre-trained offline via a supervised learning (SL) technique and then be embedded into the physical-informed safety layer on top of the model-free PPO control policy, which can assist the MGCC agent to generate safe MEMG operating actions that respect all the physical constraints. On the other hand, the model-free PPO control policy can be combined with the safety layer to continuously generate new operating points for the online learning update of the approximated security assessment rule during the RL training procedure, which enhances the ability of the safety layer to adapt to new MEMG operating points for more accurate classification. In detail, the security assessment rule, the model-free PPO control policy, and the physical-informed safety layer are described in the following subsections.

4.1. Supervised learning for security assessment

We consider a supervised learning (SL) classification method that can predict the security of an MEMG operating point involving all the components. For such a task, the common approach is to use a binary class label (i.e., safe or unsafe) corresponding to the state of the MEMG system subject to a set of user-specified binary criterion (e.g., line overloads, over-voltages, over-pressure, etc.) [44]. In this case, let $x_t=[x_{g,t}^{dg,p},x_{g,t}^{dg,q},x_{g,t}^{gg,p},x_{g,t}^{gg,q},x_{k,t}^{es},x_{g,t}^{gw},x_{k,t}^{gs}]$ be a normalized MEMG operating point (equivalent to RL action a_t) at time step t that includes DGs, GGs, ESs, GWs, and GSs, respectively. The system's security can be then expressed as a function:

$$f^s(x_t)\rightarrow y_t=\begin{cases}1, & \text{safe}\\ 0, & \text{unsafe},\end{cases}\ \forall t\in T, \tag{47}$$

where $y_t\in\{0,1\}$ corresponds to the class (binary) label, i.e., $y_t=1$ and $y_t=0$ signifying safe and unsafe MEMG operations, respectively. Given the operating point x_t, the probability $f_\beta^s(x_t)$ of $y_t=1$ can be estimated as:

$$f_\beta^s(x_t)=\Pr(y_t=1|x_t)=\frac{1}{1+\exp(-\beta\cdot x_t)},\forall t\in T, \tag{48}$$

where the vector of parameter β can be optimized via gradient descent algorithm that aims to search the optimal parameters that a hyper plane can partition the data points into its respective classes with maximum accuracy [45]. In the training procedure, the performance of the classifier with an operating point can be measured based on the cross-entropy:

$$L^{safe}(\beta)=-[p_t\log p_t+(1-p_t)\log(1-p_t)],\forall t\in T, \tag{49}$$

where p_t represents the true probability of $y_t=1$. Then backpropagation is employed to fine-tune the weights and bias of the DNN for minimizing L^{safe}. Finally, it is noted that the pairs of operating points and labels $[x,y]$ used for training the security assessment rule can be obtained via the offline simulation of the power and gas networks (17)–(30).

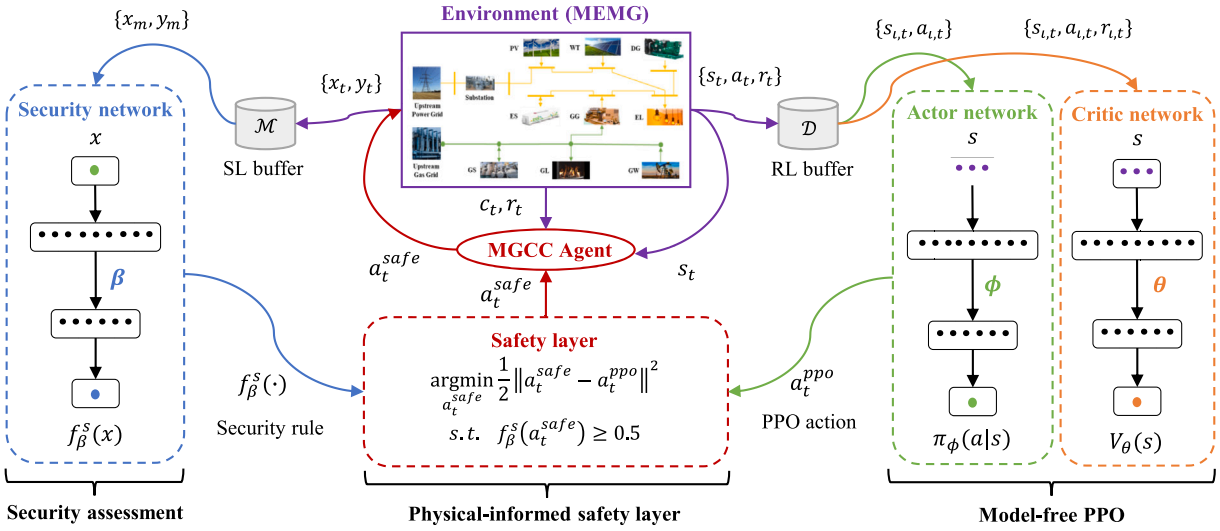


Fig. 3. Architecture of the proposed PI-SPPO method.

4.2. Reinforcement learning for energy management

PPO is an advanced policy gradient method that can achieve a balance between the ease of implementation, sampling efficiency, and ease of tuning [37]. In other words, training a relatively good performance in the vanilla policy gradient method is very challenging, since it is very sensitive to the learning rate, i.e., a small learning rate takes a long time to make the training converge, while a large learning rate easily falls into the local optimum. In addition, the vanilla policy gradient method updates the policy network based on the transitions generated by the current policy network, thereby suffering from the poor sampling efficiency, since the prior transitions cannot be utilized frequently to update the policy network. However, PPO can improve the sample efficiency by making use of the importance sampling technique [46] to obtain the data for training. The idea of importance sampling is sampling the training data from a proposal distribution to approximate the expectation on average. In this context, PPO proposes two policy networks: a new policy $\pi_{\phi}^{ppo}(a|s)$ and an old policy $\pi_{\phi\text{old}}^{ppo}(a|s)$. Specifically, the new policy $\pi_{\phi}^{ppo}(a|s)$ is evaluated with samples collected from the old policy $\pi_{\phi\text{old}}^{ppo}(a|s)$. To further reduce the variance of the estimate between these two policies, PPO constructs a probability ratio between the new and old policies, and then clips them within a stable interval. In this case, the policy of PPO can be updated within a trust region. Similar to many policy gradient methods, PPO is also characterized by an actor–critic architecture and is applicable to high-dimensional continuous state and action spaces.

To model the action characteristics in (35), a set of Gaussian distributions are generated for the actor network parameterized by ϕ to output the corresponding mean and standard deviation for all energy schedules, and then sample the optimal action a_t in environment state s_t using the stochastic policy $\pi_{\phi}^{ppo}(a|s)$. Specifically, the stochastic policy $\pi_{\phi}^{ppo}(a|s)$ is updated using the PPO algorithm [37], which maximizes its clipped surrogate objective that considers the restriction of policy update:

$$L^{clip}(\phi) = \hat{\mathbb{E}}_t\left[\min(\zeta_t \hat{A}_t, \text{clip}(\zeta_t, 1-\epsilon, 1+\epsilon)\hat{A}_t)\right], \forall t \in T, \tag{50}$$

where the first term $\zeta_t \hat{A}_t$ within the operator $\min\{\cdot\}$ indicates the normal policy gradient, while the second term $\text{clip}(\zeta_t, 1-\epsilon, 1+\epsilon)\hat{A}_t$ within the operator $\min\{\cdot\}$ trims the policy gradient by clipping the probability ratio ζ_t^d between $[1-\epsilon, 1+\epsilon]$. The hyperparameter $\epsilon \in [0, 1]$ is used to truncate the gradient update of the new policy from the old version. In other words, the advantage function $\hat{A}_t$ will be clipped if the probability ratio goes beyond the range $[1-\epsilon, 1+\epsilon]$. In the PPO policy, the probability ratio ζ_t can be expressed as:

$$\zeta_t = \frac{\pi_{\phi}^{ppo}(a_t|s_t)}{\pi_{\phi\text{old}}^{ppo}(a_t|s_t)}, \forall t \in T, \tag{51}$$

In addition, the generalized advantage function $\hat{A}_t$ in (50) can be expressed as:

$$\hat{A}_t = \delta_t + \gamma\delta_{t+1} + \cdots + \gamma^{T-t+1}\delta_{T-1}, \forall t \in T, \tag{52}$$

where

$$\delta_t = r_t + \gamma V_{\theta}(s_{t+1}) - V_{\theta}(s_t), \forall t \in T, \tag{53}$$

here $V_{\theta}(s)$ is the state-value function, which is approximated by the critic network parameterized by θ.

4.3. Physical-informed safety layer

As previously stated, deploying the action a_t^{ppo} computed by the PPO policy directly to the environment may result in physical constraint violations of power and gas networks. In order to address this issue properly, the pre-trained security assessment rule is embedded into a safety layer on top of the PPO policy to ensure the MEMG secure operation with the minimum interference, as depicted in Fig. 3. In other words, the original action a_t^{ppo} resulting from the PPO policy $\pi_{\phi}^{ppo}(a_t|s_t)$ will be corrected as little as possible (only if it endangers the safety) to the safe action a_t^{safe}, following the security rule $f_{\beta}^{s}(a_t^{safe}) \to 1$. Mathematically, the safety layer added on top of PPO policy aims to solve:

$$\arg\min_{a_t^{safe}} \frac{1}{2}\|a_t^{safe} - a_t^{ppo}\|^2 \tag{54}$$

$$\text{s.t.} \quad f_{\beta}^{s}(a_t^{safe}) \geq 0.5, \forall t \in T, \tag{55}$$

where the objective (54) is finding the safe action a_t^{safe} that perturbs the original PPO action a_t^{ppo} as little as possible in the Euclidean norm in order to ensure secure MEMG operations, i.e., satisfying the physical safety constraints of the integrated power and gas network C in (45). In this technique, the correction optimization program of actions can be solved in a mathematical manner and further ensure the secure operation of the power and gas networks per time step and environment state, even during the training process. It can be found that if there is no constraint violation, the PPO action a_t^{ppo} can be converted into real energy schedules according to (36)–(44) and then directly executed

to the environment, meanwhile the corresponding reward (46) can be calculated and obtained by the MGCC agent. However, if the power and gas network constraints C are not satisfied (e.g., over the limits of line capacity, bus voltage, and nodal pressure), the PPO action a_t^{ppo} are corrected to a_t^{safe} with the minimum interference to ensure safe MEMG operations. Afterwards, a_t^{safe} can be transferred back to the component energy schedules, and then the reward r_t and the next state s_{t+1} can be obtained accordingly.

4.4. Training process

During the training process, PI-SPPO runs for MGCC agent by its PPO old policy $\pi_{\phi\text{old}}^{ppo}(a|s)$ together with the safety layer $f_\beta^s(a)$ to generate safe action for each training episode (T time steps), and then collects the trajectory $\tau = [s_1, a_1, c_1, r_1, s_2, \ldots, r_T]$ via the interactions with the environment. After a batch of trajectories are gathered from the RL buffer $D = \{\tau_\iota\}_{\iota=1}^{J}$, the MGCC agent can then utilize them to calculate the discounted reward-to-go $\hat{R}_{\iota,t} = \sum_{h=t}^{T} \gamma^{h-t} r_{\iota,h}$ and the advantage function $\hat{A}_{\iota,t}$ for each trajectory ι and time step t. Then, the actor network is trained by maximizing its objective as below:

$$\mathcal{L}^{clip}(\phi) = \frac{1}{J \times T} \sum_{\iota=1}^{J} \sum_{t=1}^{T} \min\left(\zeta_{\iota,t}\hat{A}_{\iota,t}, \text{clip}(\zeta_{\iota,t}, 1-\epsilon, 1+\epsilon)\hat{A}_{\iota,t}\right), \tag{56}$$

where J indicates the training batch size. The critic network is trained by minimizing the loss function of mean-squared error:

$$\mathcal{L}^{loss}(\theta) = \frac{1}{J \times T} \sum_{\iota=1}^{J} \sum_{t=1}^{T} \left(\hat{R}_{\iota,t} - V_\theta(s_{\iota,t})\right)^2. \tag{57}$$

Given the above optimizations, the network weights of actor and critic can be respectively updated as below:

$$\phi \leftarrow \phi + \alpha^\phi \nabla_\phi \mathcal{L}^{clip}(\phi), \tag{58}$$

$$\theta \leftarrow \theta + \alpha^\theta \nabla_\theta \mathcal{L}^{loss}(\theta), \tag{59}$$

where α^ϕ, and α^θ indicate the learning rates of the gradient ascent and descent algorithms for actor and critic networks, respectively.

It is worth noting that the corrected action a_t^{safe} in optimization (54)–(55) might be still possible to cause the violations of physical constraints C, since the approximated security assessment rule cannot be 100% accurate theoretically [31,34]. In addition, the dynamic interactions with the environment during the RL training process may introduce new operating points that are unseen in the pre-training procedure. To this end, we continue the online training procedure of the embedded security assessment rule to further improve its accuracy and adaptability to new operating points. Specifically, during the RL training process, the action corrected by the safety layer will be sent to a real MEMG environment for the final verification of safety. If the action is safe, it will be labeled with 1; otherwise, 0. In this context, a new pair of MEMG operating point and label $[x, y]$ is generated, which can be added into the SL data buffer $\mathcal{M}$ for online training:

$$\mathcal{L}^{safe}(\beta) = -\frac{1}{M} \sum_{m=1}^{M} [p_m \log p_m + (1-p_m)\log(1-p_m)] \tag{60}$$

then, the network weights of the approximated security assessment rule can be updated as below:

$$\beta \leftarrow \beta + \alpha^\beta \nabla_\beta \mathcal{L}^{safe}(\beta), \tag{61}$$

where α^β indicates the learning rate of the gradient descent algorithm for training the security assessment rule network. In this case, the security assessment rule can be updated with new MEMG operating data during the RL training process, which leads to enhanced classification ability, especially for the system state features generated by the RL algorithm. Furthermore, even though there are occasional unsafe RL actions due to stochasticity, poor observations, etc., bad data detection mechanisms are normally deployed in modern energy systems, which can effectively detect these actions that do not meet the physical constraints and thus prevent the system from potential damage [47].

Finally, the pseudo-code of PI-SPPO for training process is shown as Algorithm 1:

Algorithm 1 PI-SPPO for training process

Initialize weights ϕ, θ, and β for actor, critic, and security networks, respectively
Set learning rates α^ϕ, α^θ, and α^β for actor, critic, and security network, respectively. Set clip factor ϵ
Set RL buffer D and SL buffer $\mathcal{M}$
for episode (i.e., day) $e = 1$ to E **do**
  Initialize the environment state s_0
  Set an empty trajectory $\tau = []$
  for time step (i.e., 1 hour) $t = 1$ to T **do**
    Select action a_t^{ppo} according to PPO policy $\pi_\phi^{ppo}(a|s)$
    Correct PPO action a_t^{ppo} to safe action a_t^{safe} using (54))-(55) according to the security rule $f_\beta^s(a) \rightarrow 1$
    Execute the safe action a_t^{safe} to the real MEMG environment
    Observe reward r_t and next state s_{t+1}
    Collect the safety signal c_t
    Store one experience to trajectory τ += $[s_t, a_t, c_t, r_t, s_{t+1}]$
    Store the new pair of MEMG operating point and label to the SL buffer $\mathcal{M}$
    Update state $s_t \leftarrow s_{t+1}$
  end for
  Collect the batch of trajectories from RL buffer $\{\tau_\iota\}_{\iota=1}^{J} \sim D$, then compute advantage function $\hat{A}_{\tau,t}$ and discounted reward-to-go $\hat{R}_{\tau,t}$ for each trajectory τ and time step t
  Update actor and critic network weights ψ, θ in (58)-(59)
  Update security network weight β in (61) using the training data from SL buffer $\mathcal{M}$
end for

5. Experiment setup and input data

5.1. Experiment setup

The examined MEMG includes a 6-bus power and 7-node gas network modified from [48], as illustrated in Fig. 4. Components, including 1 DG, 1 GG, 1 PV, 1 WT, 1 ES, 1 GS, and 2 GWs, are appropriately deployed in the integrated power and gas network. The operation data of ED and RES generation in the power network are obtained from a real-world open-source dataset Ausgrid [49]. The operation data of GD in the gas network is obtained from [48]. The grid active power buying prices of electricity are collected from Nord-Pool group [50], while the grid reactive power buying prices of electricity are 10% of the grid active power buying prices [43]. Furthermore, the grid active and reactive power selling prices are both assumed to be 50% of their buying prices, respectively. In order to capture uncertainties associated with demand, RES generation, and electricity price signals, a yearly dataset capturing various data characteristics is utilized. For illustration, the daily mean and standard deviations of these time-series data over the year are plotted in Fig. 5. Afterwards, we split it into two pieces, with the first 11 months being the training data and the last month being the test data, for the purpose of RL method evaluation. Finally, the gas price is provided by the British Gas tariff plan supplier fixed at 0.0325 £/kWh [51].

The technical parameters of 1 DG, 1 GG, 1 ES, 2 GWs, and 1 GS are presented in Table 2. The amplitudes of all bus voltages are bounded between 0.95 p.u. and 1.05 p.u. [52]. In addition, the reactive power cost is assumed equal to 10% the value of active cost for DG and GG generation costs [43].

Table 2
Technical parameters of DG, GG, and GW.

Component	Parameters	Values
DG	$\underline{P}^{dg}$, $\overline{P}^{dg}$, $\underline{Q}^{dg}$, $\overline{Q}^{dg}$	0 kW, 75 kW, −40 kVAR, 50 kVAR
GG	$\underline{P}^{gg}$, $\overline{P}^{gg}$, $\underline{Q}^{gg}$, $\overline{Q}^{gg}$, b^{gg}	0 kW, 50 kW, −30 kVAR, 35 kVAR, 0.5 Sm^3/kWh
GW1	$\underline{G}^{gw}$, $\overline{G}^{gw}$	0 Sm^3/h, 150 Sm^3/h
GW2	$\underline{G}^{gw}$, $\overline{G}^{gw}$	0 Sm^3/h, 200 Sm^3/h
ES	$\underline{E}^{es}$, $\overline{E}^{es}$, $\overline{P}^{es}$, η^{esc}/η^{esd}	0 kWh, 200 kWh, 50 kW, 0.9
GS	$\underline{G}^{gs}$, $\overline{G}^{gs}$, $\overline{F}^{gs}$, η^{gsc}/η^{gsd}	0 Sm^3, 240 Sm^3, 60 Sm^3/h, 0.95

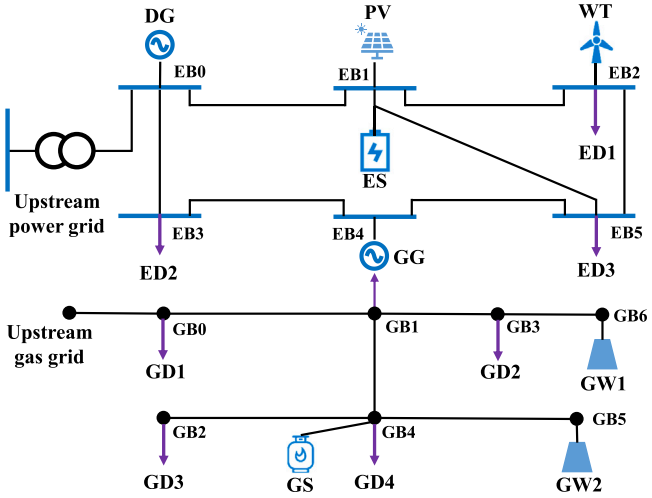


Fig. 4. MEMG of a 6-bus power and 7-node gas network.

5.2. Benchmarks

In order to validate the superior performance of our proposed PI-SPPO in the examined MEMG energy management problem, we compare it against two model-based optimization approaches (Perfect-MILP and Stochastic-MPC) and one state-of-the-art RL method (Penalty-PPO) described as below:

(1) Perfect-MILP [9]: the MGCC agent solves a deterministic MILP for the daily optimization problem. To apply MILP to the examined MEMG energy management problem, a model-based optimization is constructed with the objective function (31) and constraints (1)–(30), which assumes the perfect information of the investigated MEMG's mathematical models, technical parameters, and system uncertainties.
(2) Stochastic-MPC [12]: the MGCC agent solves an hourly-specific stochastic MPC optimization problem that allows the current time step to be optimized while taking future time steps into account and satisfying a set of constraints. To apply MPC to the examined MEMG energy management problem, model-based optimization is constructed with the objective function (31) and constraints (1)–(30), which assumes the perfect information of the investigated MEMG's mathematical models and technical parameters, but captures the stochasticity of system uncertainties via scenario generation and reduction techniques.
(3) Penalty-PPO [25]: the MGCC agent adopts a model-free PPO method with penalty terms added to the reward function (46) to penalize the violations of physical safety constraints C in (45).

5.3. Implementations

5.3.1. Supervised learning for security assessment

The network structure of the proposed security assessment model is presented in Table 3 and explained as follows: the input is the operating point in X_DIM = 7 dimensions and the output is the probability in Y_DIM = 1 dimension with a `sigmoid` activation function. The hidden layer is constructed with 64 units using a `ReLU` activation function. To train this security assessment model, we use the Adam optimizer [53] with a learning rate $\alpha^{\beta} = 10^{-3}$ and a binary cross-entropy loss function. The total number of batch sizes and training epochs is 32 and 500, respectively. The size of the SL data buffer is set at 8,000. More specifically, we split the data set of the SL buffer into 7,200 training data and 1,800 test data.

Table 3
The general specifications of supervised learning and reinforcement learning models.

Model	Network structure
Security	linear(x_dim, 16) → ReLU() → sigmoid(16, y_dim)
Actor	linear(s_dim, 64) → ReLU() → sigmoid(64, a_dim) + softplus(64, a_dim)
Critic	linear(s_dim, 64) → ReLU() → linear(64, 1)

5.3.2. Reinforcement learning for energy management

The detailed specification of actor and critic networks for the proposed PPO model is presented in Table 3. The input of the actor network is the observed state in S_DIM = 8 dimensions, while the output is the executed action in A_DIM = 7 dimensions, constructed by a Gaussian policy with `sigmoid` and `softplus` activation functions corresponding to its mean and standard deviation, respectively. For the critic network, a linear activation function is used for the output layer, while its input includes S_DIM = 8 dimensions. Finally, we construct one hidden layer with 64 units using `ReLU` as the activation function for both actor and critic networks.

To make the experiments comparable, we run 3,000 episodes with $T = 24$ time steps for the proposed RL algorithm to evaluate their training performance with the same 10 random seeds for the environment and weights initialization. During the training process, we use the Adam optimizer [53] for actor and critic networks with a learning rate $\alpha^{\psi} = 10^{-4}$ and $\alpha^{\theta} = 10^{-3}$, respectively. The batch size $J = 24$ refers to the number of collected trajectories per episode for updating networks. We employ a clip rate $\epsilon = 0.2$ and a discount rate $\gamma = 0.9$ used to expect a long-term return within an operation day of 24 time steps.

6. Case studies

6.1. Performance evaluation of safe RL

This section evaluates the training and test performance of the proposed PI-SPPO method for the examined MEMG energy management problem. Since the learning procedure of the proposed PI-SPPO consists of two parts: (1) supervised learning for the security assessment model and (2) reinforcement learning for the energy management model, we would like to evaluate their learning performance separately for each of the parts.

6.1.1. Security assessment model

Fig. 6(a) and (b) respectively illustrate the accuracy score and loss of both training and validation parts against the number of epochs during the training process, while Fig. 6(c) evaluates the well-trained security assessment model on the test data by using the confusion metric.

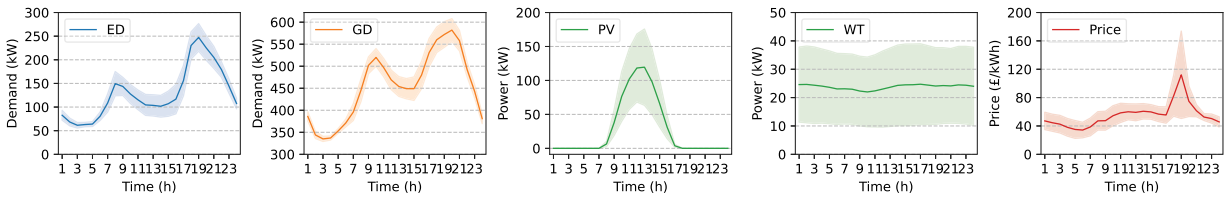
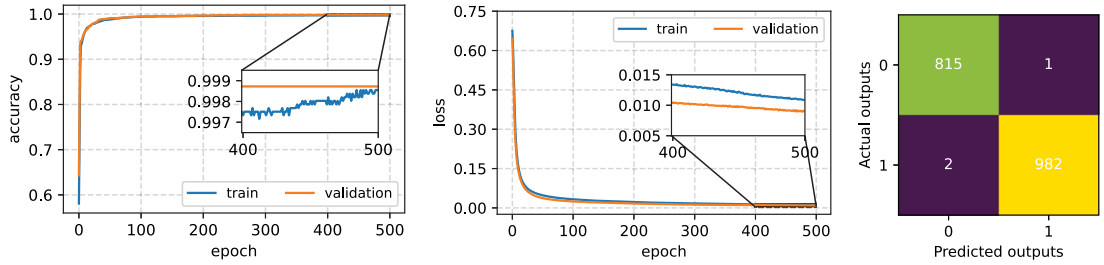
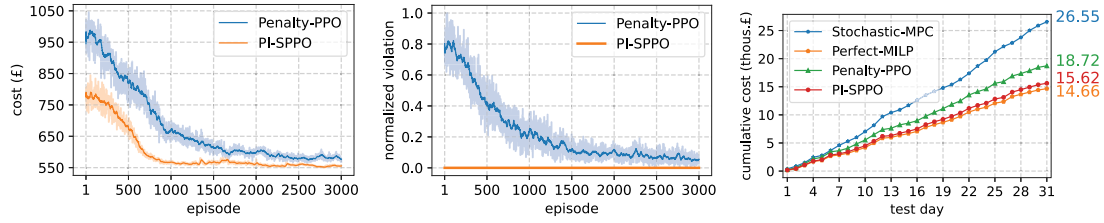


Fig. 5. Daily profiles of electric demand, gas demand, PV and wind power generation, and grid (active power buying) prices of electricity. Lines and areas respectively indicate the mean and standard deviations over the year.



(a) training accuracy (b) training loss (c) test metric

Fig. 6. The performance of security assessment model in (a) training accuracy, (b) training loss, and (c) test metric.



(a) training cost (b) normalized constraint violation (c) test cost

Fig. 7. The performance of energy management model in (a) training cost, (b) normalized constraint violation, and (c) test cost.

First of all, it can be observed from Fig. 6(a) and (b) that the training accuracy and loss of the security assessment model show continued improvement for the duration of 500 epochs and reach convergence around 400 epochs. Specifically, the train accuracy goes up to 99.85% and the train loss drops down to 0.0109. Except for the train set, the validation accuracy (99.87%) and loss (0.0091) also show good performance that is close to the train results. Once the security assessment model is well trained, we can evaluate the model using the `evaluate()` method, which returns a test accuracy of 99.83% and a test loss of 0.0137 upon the test set. Finally, the confusion metric in Fig. 6(c) shows that there are only 3 samples with incorrect predictions among the 1800 test samples.

6.1.2. Energy management model

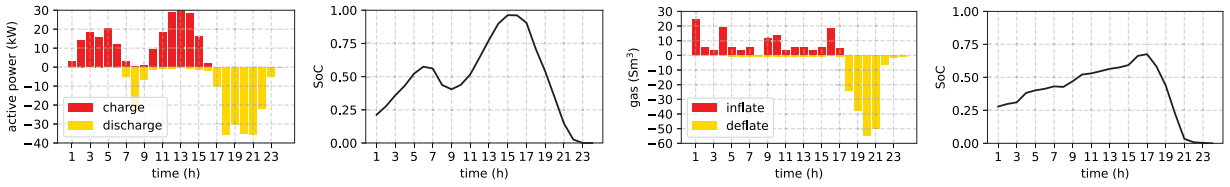
Fig. 7(a) illustrates the convergence curve of the episodic cost for Penalty-PPO (sum of energy cost and penalty cost) and PI-SPPO (energy cost only) methods, where the solid lines and the shaded areas respectively depict the moving average over 100 episodes and the oscillations of the cost during the training process. Fig. 7(b) shows the normalized constraint violations of the constraint function C for Penalty-PPO and PI-SPPO methods. Finally, Fig. 7(c) shows the cumulative daily cost over the 31 test days for Penalty-PPO and PI-SPPO methods, as well as the model-based Perfect-MILP and Stochastic-MPC methods, where their corresponding values are illustrated on the right vertical axis of the figure.

Our first observation from Fig. 7(a) is that Penalty-PPO (blue) and PI-SPPO (orange) both exhibit a continuous downward trend during the training process, and finally reach convergence within 3,000 episodes. However, the convergence speed of PI-SPPO (around 1,000 episodes) is much faster than Penalty-PPO (around 2,500 episodes), this is because learning PI-SPPO with cost minimization only is easier than Penalty-PPO that additionally considers the penalty of C. Moreover, PI-SPPO shows superior performance over Penalty-PPO in terms of lower operation cost and higher stability (i.e., lower standard deviation). Go further, we can observe from Fig. 7(b) that the normalized constraint violation of C in Penalty-PPO is reduced with the increase of episode number. Nevertheless, violation value cannot reduce to zero, thus still destroying the secure operation of the MEMG. On the other hand, the empirical results shown in Fig. 7(b) demonstrate the effectiveness of the security assessment model in handling the physical constraints of MEMG and can always ensure its safety.

Once two RL policies are well-trained, they can be directly deployed to the test performance. It can be observed from Fig. 7(c) that PI-SPPO reaches close to the theoretical optimal Perfect-MILP (6.57% difference), while 16.56% and 41.17% lower than Penalty-PPO (with penalty cost) and Stochastic-MPC, respectively. Finally, it is noted that there is completely no constraint violation of MEMG in PI-SPPO, but 56.74 kW line power flow limit violation; 0.26 p.u. nodal voltage limit violation; 105.16 Sm^3/h pipeline gas flow limit violation; 183.85 bar nodal gas pressure limit violation; 11.05 kWh ES energy content limit violation; and 14.32 Sm^3 GS gas content limit violation in Penalty-PPO averaged over the 31 test days.

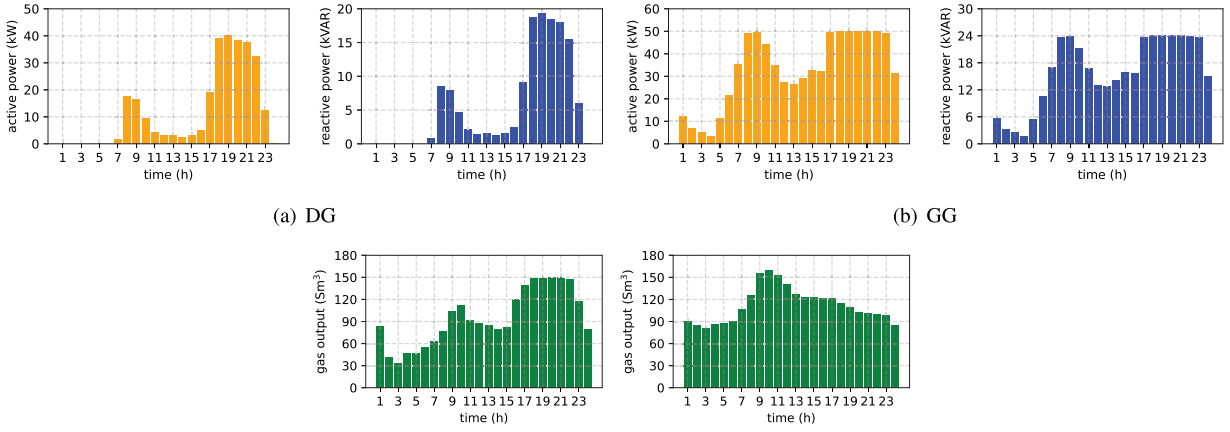
6.2. MEMG energy management analysis

Having evaluated the performance of the proposed security rule and demonstrated the superiority of the proposed PI-SPPO over the conventional Penalty-PPO and two optimization (Perfect-MILP and Stochastic-MPC) methods in both training and test processes, this section aims to



(a) ES (b) GS

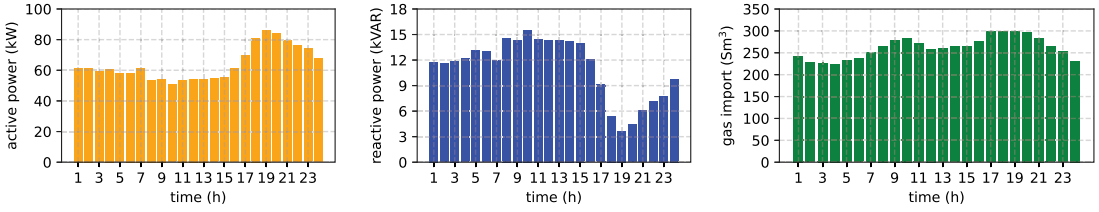
Fig. 8. Charging and discharging behaviors and SoC of (a) ES and (b) GS.



(a) DG (b) GG

(c) GW1 and GW2

Fig. 9. Active and reactive power generation of (a) DG and (b) GG, and gas generation of (c) two GWs.



(a) Power Grid (b) Gas Grid

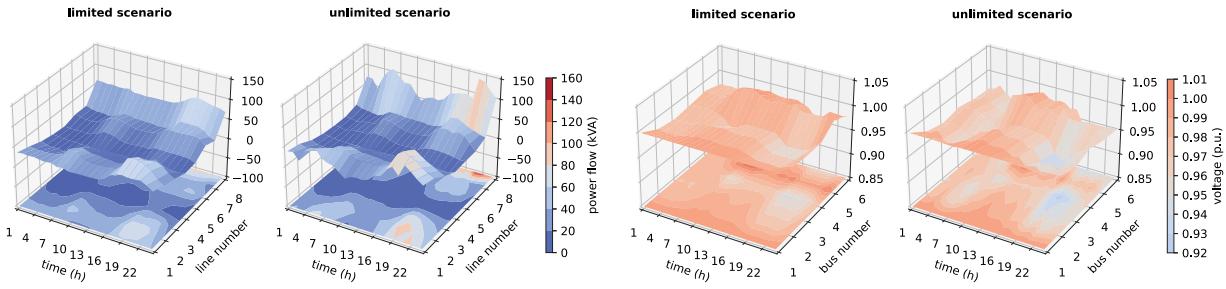
Fig. 10. Active and reactive power supply from upstream (a) power grid and gas supply from upstream (b) gas grid.

analyze the detailed energy management of the examined MEMG. To this end, we plot the averaged power charge and discharge of ES as well as the averaged gas inflation and deflation of GS over the 31 test days in Fig. 8; the averaged active and reactive power generation of DG and GG as well as the averaged gas generation of two GWs over the 31 test days in Fig. 9; additionally, the averaged active and reactive power supply from the upstream power grid as well as the gas supply from the upstream gas grid over the 31 test days in Fig. 10.

It can be observed from Figs. 8 and 9 that the power and gas dispatches of all the components in the studied MEMG can be managed within their operation limits. Firstly, the ES charges power in the early morning and midday, as shown in Fig. 8(a), when the electricity price is at the lowest level and PV generation is extremely high, respectively. On the other hand, the ES mainly discharges power in the evening, as shown in Fig. 8(a), when electric demand is at its peak. It is noted that the combination of ES charging and discharging behaviors can effectively reduce the MEMG operation cost via energy arbitrage and mitigate the PV curtailment via RES absorption. Similarly, the GS inflates gas in the early morning and midday when gas demand is relatively low, while deflating gas into the gas network in the evening when gas demand significantly increases, as illustrated in Figs. 8 (b). The SoC dynamics of the ES and GS are also presented in Fig. 8(a) and (b), respectively. The energy contents of both ES and GS are fully utilized given the SoC of zero by the end of the day. Secondly, the DG and GG units provide both active and reactive power support for the MEMG, especially in the evening, as shown in Fig. 9(a) and (b) respectively, since both active and reactive demands reach their peak in this period of most test days. Meanwhile, two GW units provide gas generation for the gas network as shown in Fig. 9(c), especially in the evening and the morning, respectively. This is because GW1 is mainly used to supply the gas demand at nodes 4 and 2, which reach the peak level in the evening, while GW2 is close to the gas demands at node 1 and node 3, which reach the peak level in the morning. Finally, as shown in Fig. 10, the MEMG needs to import a certain level of electricity (active and reactive power) and gas from the upstream power and gas grids, respectively, to satisfy the energy deficits of power and gas systems apart from the energy supplied from various components.

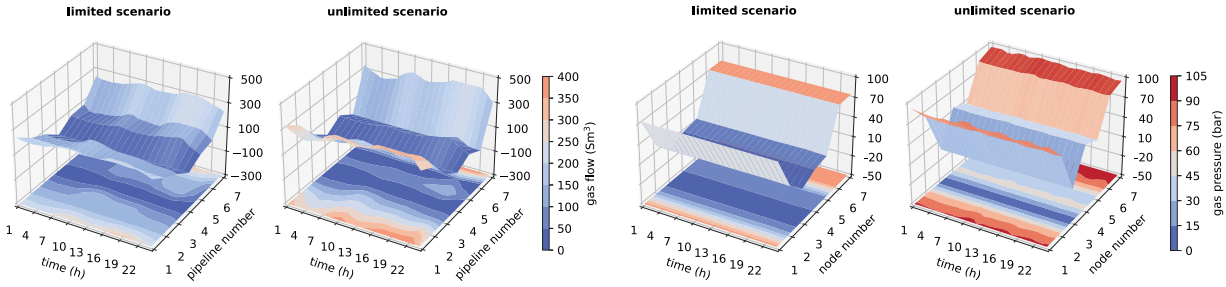
6.3. Power and gas network operation analysis

In order to investigate the impact of the proposed physic-informed safety layer on the secure operation of the studied MEMG, this section also analyzes the status of line power flows and bus voltage magnitudes pertaining to the power network as well as the status of pipeline gas



(a) power flows (b) nodal voltages

Fig. 11. Line (a) power flows and (b) nodal voltages with limited and unlimited power and gas networks.



(a) gas flows (b) nodal pressures

Fig. 12. Pipeline (a) gas flows and (b) nodal pressures with limited and unlimited power and gas networks.

flows and node pressure magnitudes pertaining to the gas network. Specifically, two scenarios with limited and unlimited MEMG energy flows are conducted for comparison (i.e., in the unlimited scenario, there is no safety layer), of which their averaged power line flows and nodal voltages, as well as the pipeline gas flows and nodal pressures over the 31 test days are illustrated in Figs. 11 and 12, respectively.

As far as the line power flows and nodal voltages are concerned, there are apparent constraint violations of line thermal capacity at line ID 1, 2, and 8 under the unlimited scenario of Fig. 11(a). Under the unlimited scenario, the maximum power flows of these three lines reach 90 kVA, 90 kVA, and 140 kVA, respectively, far exceeding their line capacities of 60 kVA, 80 kVA, and 120 kVA. However, when the proposed safety layer is employed, all the line power flows can be restricted within their line capacity limits, leading to zero constraint violations, as illustrated in the limited scenario of Fig. 11(a), i.e., there is no red color in the surface area. Regarding nodal voltages, constraint violations (below 0.95 p.u.) do exist at bus ID 3, 4, and 6 under the unlimited scenario of Fig. 11(b). In particular, the averaged voltage levels of bus 4 even drop to 0.92 p.u. between hours 19–22, which can cause severe safety issues in practice, e.g., line outages. However, the nodal voltages can always be restricted between 0.95 and 1.05 p.u. under the limited scenario of Fig. 11(b). Similarly, in the gas network, apparent constraint violations (over 300 Sm^3/h) of pipeline capacity occur at pipeline ID 1, 2, and 7 under the unlimited scenario of Fig. 12(a), while severe constraint violations (over 75 bar) occur at node ID 1, 6, and 7 under the unlimited scenario of Fig. 12(b). On the contrary, when the proposed safety layer is employed, there are no constraint violations in both gas flow and nodal pressure, as depicted in the limited scenario of Fig. 12(a) and (b). It can be concluded from the above comparison that the proposed safety layer in PI-SPPO shows its effectiveness in ensuring secure MEMG network operation.

6.4. Multi-agent setup in a 33-bus power and 20-node gas network

To further investigate the scalability of the proposed PI-SPPO for the MEMG energy management problem, a larger operation system (an integrated 33-bus power and 20-node gas network) is utilized in this subsection, including 1 DG, 5 PVs, 3 WTs, 3 GGs, 3 ESs, 4 GWs, and 3 GSs, of which its network structure is shown in Fig. 13. It can be found that the large integrated power-gas network is separated into three regions corresponding to three MEMGs, since using a single agent for the entire network operation may cause the curse of dimensionality, thereby exploding the learning performance. The implementation steps of this multi-agent setup are similar to those of the single-agent setup in the previous 6-bus power and 7-node gas network. Specifically, each MGCC (1) approximates the safety constraint set of its own region and then embeds the approximated security assessment rule into a safety layer on top of the RL-based control policy; (2) trains an RL-based control policy based on the PPO algorithm, with the action dimensions represented as all the DER power dispatches of the considered region; and (3) manages its own DERs via the trained PI-SPPO policy respecting all its regional network constraints.

It can be observed from Figs. 14 and 15 that the power and gas dispatches of all the components in these three MEMGs are still restricted by their operating limits. Specifically, ESs in Fig. 14(a) present reasonable charging and discharging behaviors, e.g., charging in the morning and midday when electricity price is low and PV generation is high while discharging in the evening when electric demand is high. Similarly, GSs in Fig. 14(b) inflate in the morning and afternoon and deflate in the evening. DGs and GGs in Fig. 15(a) mainly choose to generate active and reactive power in the evening when the electric demand is high, while GG2 in MEMG 2 provides a large amount of active and reactive power for the power network over the day due to its relatively low generation cost. Regarding GWs in Fig. 15(b), GW 1 in MEMG 1, and GWs 2 and 3 in MEMG 2 contribute almost entirely to the gas demand and reach their capacity during the high demand periods. The gas output of GW 4 in MEMG 3 is relatively low due to the high generation cost. Finally, as shown in Fig. 16(a) and (b), these MEMGs also imports a certain level of active and reactive power and gas from the upstream grids to meet the remaining demand requirements of the whole multi-energy system. Regarding the safe operation of these MEMGs, it can be observed from Figs. 17 (a)–(d) that line power flows,

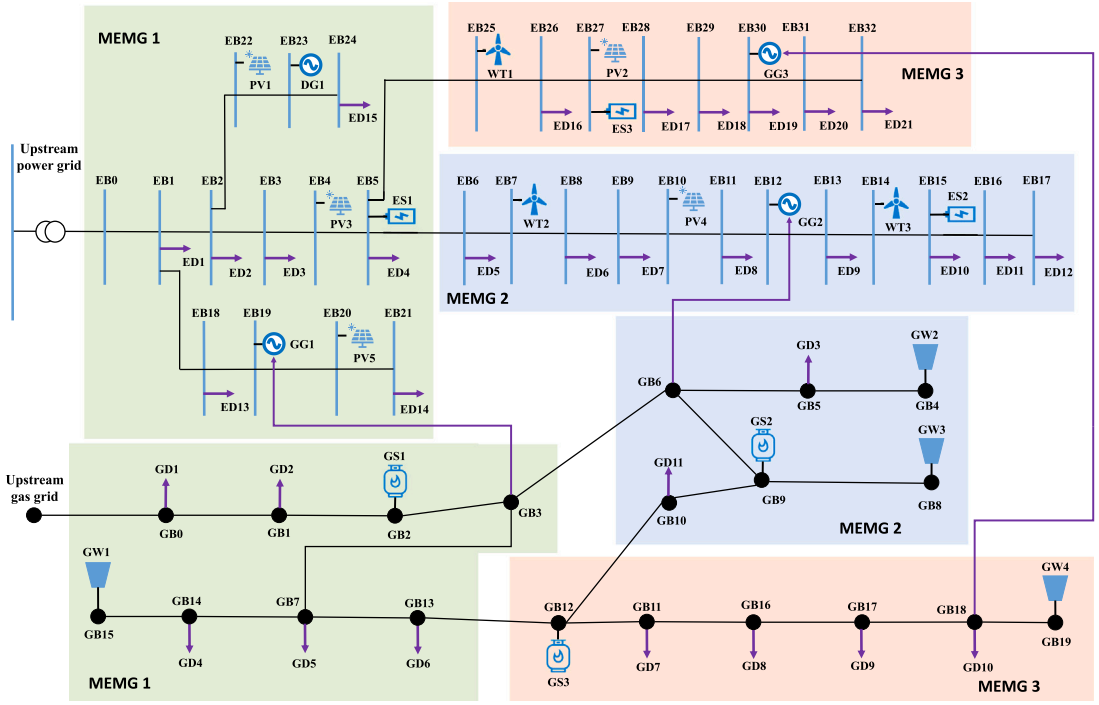
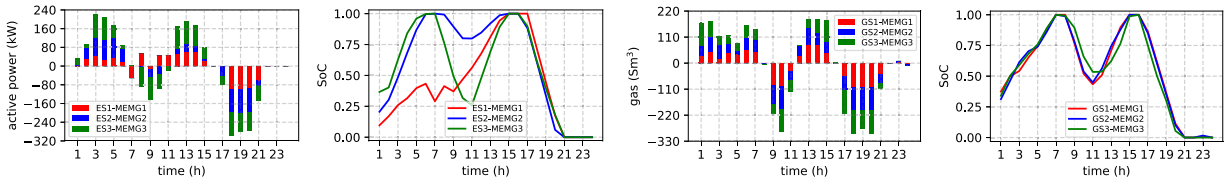
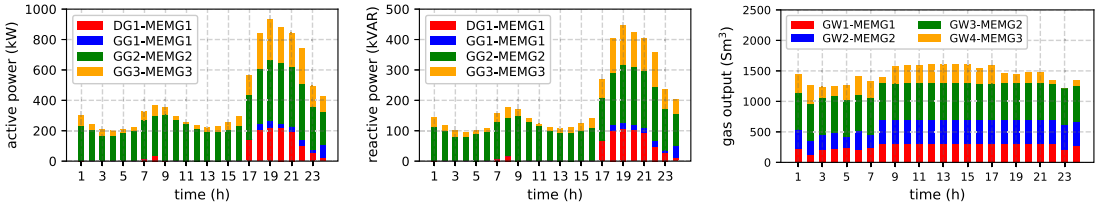


Fig. 13. The 33-bus power and 20-node gas network with three MEMGs.



(a) ESs (b) GSs

Fig. 14. Charging and discharging behaviors and SoC of (a) three ESs and (b) three GSs.



(a) DGs and GGs (b) GWs

Fig. 15. Active and reactive power generations of (a) two DGs and (b) two GGs, and gas generations of (c) four GWs.

nodal voltages, pipeline gas flows, and nodal pressures are all restricted within their acceptable limits without any constraint violation, which further verifies the effectiveness of the proposed PI-SPPO in ensuring secure and reliable MEMG energy management and network operation.

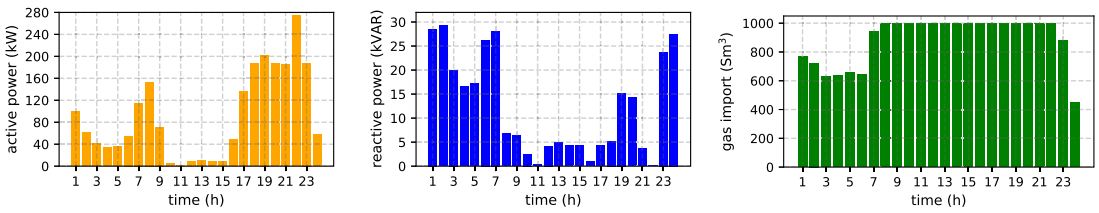
7. Discussion

7.1. Key findings from empirical results

As presented in Section 6, the case studies include the performance evaluation of both dynamic security assessment rule and PPO control policy, the energy management analysis as well as the power and gas network operation of a small 6-bus power and 7-node gas network, and a multi-agent setup in a large 33-bus power and 20-node gas network. Overall, the key findings are summarized as below:

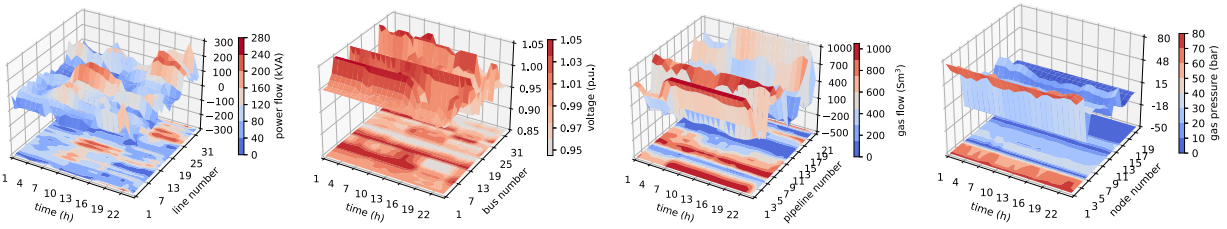
(1) The training and test performance of the SL-based security assessment model and the RL-based energy management model have been evaluated in Section 6.1.1 and Section 6.1.2, respectively. On one hand, numerical results show that the security assessment model achieves good performance with 99.85% training accuracy, 99.87% validation accuracy, and 99.83% test accuracy for the 6-bus power and 7-node gas network. On the other hand, the proposed PI-SPPO achieves better performance than the benchmark Penalty-PPO in both the operation cost and constraint violation. Furthermore, the cumulative operation cost of the proposed PI-SPPO is only 6.57% higher than the Perfect-MILP with theoretic solutions, over the 31 test days.

(2) The scheduling behaviors of all components are well learned in PI-SPPO to fully supply power and gas demands in the examined



(a) Power Grid (b) Gas Grid

Fig. 16. Active and reactive power supply from upstream (a) power grid and gas supply from upstream (b) gas grid.



(a) power flows (b) nodal voltages (c) gas flows (d) nodal pressures

Fig. 17. (a) power flows, (b) nodal voltages, (c) gas flows, and (d) nodal pressures in the integrated 33-bus power and 20-node gas network.

MEMG, while GGs are efficiently utilized through the energy transitions from gas to power sectors. Meanwhile, the impact of the proposed safety layer is investigated in Section 6.3 by analyzing the status of line power flows and bus voltage magnitudes pertaining to the power network as well as the status of pipeline gas flows and node pressure magnitudes pertaining to the gas network between two scenarios with and without a safety layer.

(3) The scalability of the proposed PI-SPPO has been demonstrated in the context of a large 33-bus power and 20-node gas network. In order to avoid exploding the learning performance, the large network has been divided into three regions, where each region is represented by an MEMG managed and operated by its individual MGCC. As a result, the 33-bus power and 20-node gas network can be reformulated into three networked MEMGs. The numerical results show that the proposed PI-SPPO can be effectively deployed to a larger network with reasonable DERs' dispatches while also satisfying all the physical constraints of both power and gas networks.

7.2. Real-world applications

Extensive case studies have been carried out to show that the proposed PI-SPPO method has the ability to ensure secure MEMG energy management without constraint violations. It is worth noting that this feature is extremely important for advancing the real-world applications of RL methods in integrated energy systems that are normally regarded as critical infrastructures in modern societies. As such, it is anticipated that such a safe RL method has a better chance to be deployed in real-world scenarios than conventional RL methods.

In fact, there is no research deploying and testing their trained RL policies in real-world energy system applications [54]. To achieve such a practical implementation, as a start, the well-trained PI-SPPO method may be extended to cover a broader range of parameter settings (e.g., use of finer decision time-slots), and then be validated with hardware circuit experiments or semi-physical simulation experiments, which can further improve the safety and interoperability of RL methods. After the comprehensive validation, the proposed safe RL method may be able to be deployed in industrial applications and conduct real-world operational tests.

To train a good RL policy with high efficiency and accuracy, numerous real-world datasets are normally a necessity, which leads to the challenges associated with data quantity, data quality, and data availability [54]. On one hand, virtual sample generation techniques may be a potential option to construct larger-scale training samples from existing operational datasets; on the other hand, advanced sensors, smart meters, and other communication technologies can be deployed to improve data quality and availability. Additionally, data privacy should be preserved during the RL training process, which leads to the requirements for blockchain and cyber security technologies.

Decentralization and digitalization are rapidly transforming the energy industry, leading to the requirements for the decentralized setup of multi-interconnected MEMGs. In this context, each MEMG may actively seek energy trading opportunities with neighboring MEMGs in order to reduce its own energy costs, while handling secure energy management in its own region. To apply the proposed PI-SPPO method to this scenario, a multi-agent extension is necessary as presented in Section 6.4. Furthermore, integrating reasonable reputation or credit-based marketing mechanisms [55] into RL methods can also be of high importance for effective energy trading among these MEMGs. For instance, when there are a group of interconnected MEMGs in the network that can trade energy with each other for profits, an MGCC agent may prefer to first interact with MGCCs with relatively high safety performance. In other words, the security of energy management in an MEMG may be linked to its reputation or credit, and this may influence its priority in energy trading markets, incentivizing the MEMG to pursue a higher level of safety by improving the accuracy of its safe RL method.

8. Conclusions and future work

This paper has proposed a novel physical-informed safe reinforcement learning algorithm named PI-SPPO to solve an MEMG energy management problem involving various energy resources (e.g., diesel generators, gas-fired generators, energy storage systems, gas wells, and gas storage systems) in an integrated power-gas network environment. The proposed PI-SPPO algorithm (1) takes advantage of the conventional *Proximal Policy Optimization* algorithm on sampling efficiency and hyperparameter tuning, thereby being able to address the high-dimensional continuous state and action space; (2) uses supervised learning techniques to train a security assessment rule for the MEMG, which is formulated as a safety layer on top of the *Proximal Policy*

Optimization policy to mathematically solve an action correction formulation for MEMG secure operations; (3) captures uncertainties associated with grid price signals, renewable energy resources, and demand profiles through the learning procedure. Extensive case studies based on two MEMGs (i.e., a small 6-bus power and 7-node gas network, a large 33-bus power and 20-node gas network) have demonstrated the effectiveness of the proposed PI-SPPO algorithm in generating realistic energy scheduling decisions, reducing energy management costs, and maintaining the secure operation of the investigated MEMGs.

Future extensions of this work can move in the following three directions. First, this paper only focuses the energy management problem of one MEMG. Future work will include the coordinated energy management of multiple networked MEMGs and solve it using multi-agent reinforcement learning algorithm with reputation-based marketing mechanisms. Second, this paper only considers the energy integration between power and gas sectors. Future work will capture the additional heating and cooling sectors. Third, the exogenous state features (e.g., system demand, PV generation, price signals) unaffected by actions can slow down the training process by injecting uncontrolled variation into the reward signal. Future work will develop a fast and robust learning algorithm for the exogenous state (noise) of the experiment environment.

CRediT authorship contribution statement

Yi Wang: Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Dawei Qiu:** Methodology, Software, Data curation, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Mingyang Sun:** Methodology, Writing – original draft, Writing – review & editing. **Goran Strbac:** Conceptualization, Project administration, Methodology, Supervision, Funding acquisition. **Zhiwei Gao:** Writing – original draft, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported by two UK EPSRC projects: ‘Integrated Development of Low-Carbon Energy Systems (IDLES): A Whole-System Paradigm for Creating a National Strategy’ (project code: EP/R045518/1) and UK-China project - ‘Technology Transformation to Support Flexible and Resilient Local Energy Systems’ (project code: EP/T021780/1), and one Horizon Europe project: ‘Reliability, Resilience and Defense technology for the griD’ (Grant agreement ID: 101075714) as well as the National Natural Science Foundation of China under Grants 62103371, 52161135201, U20A20159, 62061130220.

References

[1] Chinmoy L, Iniyan S, Goic R. Modeling wind power investments, policies and social benefits for deregulated electricity market–A review. Appl Energy 2019;242:364–77.
[2] Wang Q, Zhang C, Ding Y, Xydis G, Wang J, Østergaard J. Review of real-time electricity markets for integrating distributed energy resources and demand response. Appl Energy 2015;138:695–706.
[3] Quadri IA, Bhowmick S, Joshi D. A comprehensive technique for optimal allocation of distributed energy resources in radial distribution systems. Appl Energy 2018;211:1245–60.
[4] Wang J, Zhong H, Ma Z, Xia Q, Kang C. Review and prospect of integrated demand response in the multi-energy system. Appl Energy 2017;202:772–82.
[5] Alam MN, Chakrabarti S, Ghosh A. Networked microgrids: State-of-the-art and future perspectives. IEEE Trans Ind Inf 2019;15(3):1238–50.
[6] Christakou K, Paolone M, Abur A. Voltage control in active distribution networks under uncertainty in the system model: A robust optimization approach. IEEE Trans Smart Grid 2017;9(6):5631–42.
[7] Ye Y, Qiu D, Wu X, Strbac G, Ward J. Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning. IEEE Trans Smart Grid 2020;11(4):3068–82.
[8] Zia MF, Elbouchikhi E, Benbouzid M. Microgrids energy management systems: A critical review on methods, solutions, and prospects. Appl Energy 2018;222:1033–55.
[9] Azzam SM, Elshabrawy T, Ashour M. A bi-level framework for supply and demand side energy management in an islanded microgrid. IEEE Trans Ind Inf 2022.
[10] Vahedipour-Dahraie M, Rashidizadeh-Kermani H, Anvari-Moghaddam A, Guerrero JM. Stochastic risk-constrained scheduling of renewable-powered autonomous microgrids with demand response actions: Reliability and economic implications. IEEE Trans Ind Appl 2019;56(2):1882–95.
[11] Cui S, Wang Y-W, Xiao J-W, Liu N. A two-stage robust energy sharing management for prosumer microgrid. IEEE Trans Ind Inf 2018;15(5):2741–52.
[12] Tobajas J, Garcia-Torres F, Roncero-Sánchez P, Vázquez J, Bellatreche L, Nieto E. Resilience-oriented schedule of microgrids with hybrid energy storage system using model predictive control. Appl Energy 2022;306:118092.
[13] Yang S, Wan MP, Chen W, Ng BF, Dubey S. Model predictive control with adaptive machine-learning-based model for building energy efficiency and comfort optimization. Appl Energy 2020;271:115147.
[14] Li K, Yang F, Wang L, Yan Y, Wang H, Zhang C. A scenario-based two-stage stochastic optimization approach for multi-energy microgrids. Appl Energy 2022;322:119388.
[15] Moretti L, Martelli E, Manzolini G. An efficient robust optimization model for the unit commitment and dispatch of multi-energy systems and microgrids. Appl Energy 2020;261:113859.
[16] Zhang C, Xu Y, Dong ZY, Yang LF. Multitimescale coordinated adaptive robust operation for industrial multienergy microgrids with load allocation. IEEE Trans Ind Inf 2019;16(5):3051–63.
[17] Ceseña EAM, Mancarella P. Energy systems integration in smart districts: robust optimisation of multi-energy flows in integrated electricity, heat and gas networks. IEEE Trans Smart Grid 2018;10(1):1122–31.
[18] Xu Y, Ding T, Qu M, Du P. Adaptive dynamic programming for gas-power network constrained unit commitment to accommodate renewable energy with combined-cycle units. IEEE Trans Sustain Energy 2019;11(3):2028–39.
[19] Xu D, Wu Q, Zhou B, Li C, Bai L, Huang S. Distributed multi-energy operation of coupled electricity, heating, and natural gas networks. IEEE Trans Sustain Energy 2019;11(4):2457–69.
[20] Birge JR, Louveaux F. Introduction to stochastic programming. 2nd ed.. New York, NY, USA: Springer; 2011.
[21] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT Press; 2018.
[22] Foruzan E, Soh L-K, Asgarpoor S. Reinforcement learning approach for optimal distributed energy management in a microgrid. IEEE Trans Power Syst 2018;33(5):5749–58.
[23] Mnih V, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[24] Ji Y, Wang J, Xu J, Fang X, Zhang H. Real-time energy management of a microgrid using deep reinforcement learning. Energies 2019;12(12):2291.
[25] Lei L, Tan Y, Dahlenburg G, Xiang W, Zheng K. Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids. IEEE Internet Things J 2021;8(10):7938–53.
[26] Guo C, Wang X, Zheng Y, Zhang F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. Energy 2022;238:121873.
[27] Zhao L, Yang T, Li W, Zomaya AY. Deep reinforcement learning-based joint load scheduling for household multi-energy system. Appl Energy 2022;119346.
[28] Qiu D, Dong Z, Zhang X, Wang Y, Strbac G. Safe reinforcement learning for real-time automatic control in a smart energy-hub. Appl Energy 2022;309:118403.
[29] Zhang Q, Dehghanpour K, Wang Z, Qiu F, Zhao D. Multi-agent safe policy learning for power management of networked microgrids. IEEE Trans Smart Grid 2021;12(2):1048–62.
[30] Wang W, Yu N, Gao Y, Shi J. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. IEEE Trans Smart Grid 2019;11(4):3008–18.
[31] Li H, He H. Learning to operate distribution networks with safe deep reinforcement learning. IEEE Trans Smart Grid 2022.
[32] Dalal G, Dvijotham K, Vecerik M, Hester T, Paduraru C, Tassa Y. Safe exploration in continuous action spaces. 2018, arXiv preprint arXiv:1801.08757.
[33] Kou P, Liang D, Wang C, Wu Z, Gao L. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. Appl Energy 2020;264:114772.
[34] Gao Y, Yu N. Model-augmented safe reinforcement learning for Volt-VAR control in power distribution networks. Appl Energy 2022;313:118762.
[35] Saunders W, Sastry G, Stuhlmueller A, Evans O. Trial without error: Towards safe reinforcement learning via human intervention. 2017, arXiv preprint arXiv:1707.05173.
[36] Alshiekh M, Bloem R, Ehlers R, Könighofer B, Niekum S, Topcu U. Safe reinforcement learning via shielding. In: Proceedings of the AAAI conference on artificial intelligence. 32, (1). 2018.

[37] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017, arXiv preprint arXiv:1707.06347.
[38] Ding T, Lin Y, Bie Z, Chen C. A resilient microgrid formation strategy for load restoration considering master-slave distributed generators and topology reconfiguration. Appl Energy 2017;199:205–16.
[39] Duan J, Yang Y, Liu F. Distributed optimization of integrated electricity-natural gas distribution networks considering wind power uncertainties. Int J Electr Power Energy Syst 2022;135:107460.
[40] Franco JF, Ochoa LF, Romero R. AC OPF for smart distribution networks: An efficient and robust quadratic approach. IEEE Trans Smart Grid 2017;9(5):4613–23.
[41] Lin Y, Chen B, Wang J, Bie Z. A combined repair crew dispatch problem for resilient electric and natural gas system considering reconfiguration and DG islanding. IEEE Trans Power Syst 2019;34(4):2755–67.
[42] Altman E. Constrained Markov decision processes: Stochastic modeling. Routledge; 1999.
[43] Andrianesis P, Caramanis M, Li N. Optimal distributed energy resource coordination: A decomposition method based on distribution locational marginal costs. IEEE Trans Smart Grid 2021;13(2):1200–12.
[44] Cremer JL, Konstantelos I, Tindemans SH, Strbac G. Data-driven power system operation: Exploring the balance between cost and risk. IEEE Trans Power Syst 2018;34(1):791–801.
[45] Prasad PD, Halahalli HN, John JP, Majumdar KK. Single-trial EEG classification using logistic regression based on ensemble synchronization. IEEE J Biomed Health Inform 2013;18(3):1074–80.
[46] Metelli AM, Papini M, Faccio F, Restelli M. Policy optimization via importance sampling. Adv Neural Inf Process Syst 2018;31.
[47] Zeng L, Sun M, Wan X, Zhang Z, Deng R, Xu Y. Physics-constrained vulnerability assessment of deep reinforcement learning-based SCOPF. IEEE Trans Power Syst 2022.
[48] Wang C, Wei W, Wang J, Liu F, Qiu F, Correa-Posada CM, et al. Robust defense strategy for gas–electric systems against malicious attacks. IEEE Trans Power Syst 2017;32(4):2953–65.
[49] Ratnam EL, Weller SR, Kellett CM, Murray AT. Residential load and rooftop PV generation: an Australian distribution network dataset. Int J Sustain Energy 2017;36(8):787–806.
[50] Nord Pool. Historical market data. 2021, URL https://www.nordpoolgroup.com/historical-market-data/.
[51] British gas tariffs 2021: Compare energy prices. 2021, URL https://www.switchplan.co.uk/suppliers/british-gas/.
[52] Ma W, Wang W, Chen Z, Wu X, Hu R, Tang F, et al. Voltage regulation methods for active distribution networks considering the reactive power optimization of substations. Appl Energy 2021;284:116347.
[53] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. 3rd int. conf. learn. represent.. ICLR, San Diego, USA; 2015, p. 1–15.
[54] Chen X, Qu G, Tang Y, Low S, Li N. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. IEEE Trans Smart Grid 2022;13(4):2935–58.
[55] Janko S, Johnson NG. Reputation-based competitive pricing negotiation and power trading for grid-connected microgrid networks. Appl Energy 2020;277:115598.