Article

Integrated Demand Response in Multi-Energy Microgrids: A Deep Reinforcement Learning-Based Approach

Chenhui Xu * and Yunkai Huang

School of Electrical Engineering, Southeast University, Nanjing 210096, China; huangyk@seu.edu.cn
* Correspondence: 230199099@seu.edu.cn

Abstract: The increasing complexity of multi-energy coordinated microgrids presents a challenge for traditional demand response providers to adapt to end users’ multi-energy interactions. The primary aim of demand response providers is to maximize their total profits via designing a pricing strategy for end users. The main challenge lies in the fact that DRPs have no access to the end users’ private preferences. To address this challenge, we propose a deep reinforcement learning-based approach to devise a coordinated scheduling and pricing strategy without requiring any private information. First, we develop an integrated scheduling model that combines power and gas demand response by converting multiple energy sources with different types of residential end users. Then, we formulate the pricing strategy as a Markov Decision Process with an unknown transition. The novel soft actor-critic algorithm is utilized to efficiently train neural networks with the entropy function and to learn the pricing strategies to maximize demand response providers’ profits under various sources of uncertainties. Case studies are conducted to demonstrate the effectiveness of the proposed approach in both deterministic and stochastic environment settings. Our proposed approach is also shown to be effective in handling different levels of uncertainties and achieving the near-optimal pricing strategy.

Keywords: demand response; multi-energy microgrids; deep reinforcement learning; uncertainties

1. Introduction

With the increasing penetration of renewable energy sources and distributed energy resources (DERs), microgrids have emerged as a promising solution for the sustainable and reliable supply of energy [1–3]. In a multi-energy microgrid, various energy systems—such as electricity, heat, and gas—are interconnected, which provides flexibility in energy management and enhances energy efficiency [4,5]. However, managing energy systems in a multi-energy microgrid is challenging due to the high variability of energy supply and demand. The integration of these different energy systems presents new challenges, such as multi-energy load balancing and optimal energy management.

Demand response (DR) has emerged as a promising solution for effectively managing energy systems in multi-energy microgrids [6–9]. DR programs can be broadly classified into two categories, namely price-based DR [10,11] and incentive-based DR [12–14]. The former adjusts dynamic price signals to reshape the load profiles of the end participants, while the latter provides rewards or compensation to loads for reducing peak demand. Although both types of DR have facilitated the development of demand-side management, price-based DR is more prevalent and convenient for demand response providers in practice. This is because the dynamic pricing signals, which are integral to price-based DR, offer a more responsive and flexible approach for optimizing energy consumption patterns compared to incentive-based DR [15]. Therefore, this paper primarily focuses on price-based DR from the DRPs’ perspective.

Accordingly, DRPs enable end users to adjust their energy consumption patterns in response to the pricing signals, which can reduce the peak demand, balance the load,
and improve the stability and reliability of the grid. Numerous studies have explored the potential of DR in multi-energy systems. In [16], a multi-objective optimization and multi-energy DR model is devised via considering the uncertainties of the system for the economic-emission dispatch problem. In [17], an ensemble control scheme is developed to unlock the potential to regulate the energy consumption of an aggregation of many residential, commercial, and industrial consumers within a less centralized operating paradigm. In [18], an multi-energy microgrid consisting of power generation, combined heat and power, and gas units is formulated as day-ahead energy management with demand response. Price-based load management is taken into account to lower the costs due to the transfer of information between the consumer and the generator. In addition, stochastic energy-reserve mixed integer linear programming [19] and a stochastic decision-making model with different types of end users [20] are utilized to meet multi-energy demands. A distributed multi-energy demand DR method is discussed in [21] for the optimal coordinated operation of smart building clusters. Finally, in [22], a novel Internet of Things-enabled approach is introduced for optimizing the multi-energy operation costs and enhancing the network reliability with demand response.

Although traditional DR methods shed light on multi-energy demand management, they mostly rely on pre-defined rules, schedules, and incentives to impose impacts on the consumption behavior of end users. However, these methods are limited in their ability to respond to dynamic changes in the multi-energy microgrids due to the lack of access to the end users’ private preferences of energy consumptions in DR programs. Deep reinforcement learning (DRL) has shown great potential for developing intelligent DR systems that can adapt to the dynamic changes in the behavior of DR end users [23,24]. DRL is a subfield of artificial intelligence that combines deep neural networks and reinforcement learning to enable agents to learn the optimal actions in an unknown environment [25–27]. DRL-based methods are able to learn from historical data and real-time feedback to optimize energy consumption and improve the efficiency and reliability of the power systems [28–31].

The literature presents several studies that have utilized DRL algorithms to optimize demand response (DR) programs. In one such study, an algorithm based on DRL is described in [32] that enables the real-time DR of home appliances. The algorithm jointly decides both discrete and continuous actions to optimize the schedules of different types of appliances. In [33], a DRL-based algorithm is implemented to optimize the decision-making process of DR providers (DRPs) for maximizing their profits through DR scheduling to improve the system reliability. Similarly, in [34], an incentive-based DR program is proposed for virtual power plants to minimize the deviation penalty from participating in the power market. The model-free DRL-based approach deals with the randomness existing in the model and adaptively determines the optimal DR pricing strategy. Additionally, [35] proposes a model-free DRL paradigm for DRPs to achieve the maximum long-term profit by handling uncertainties of electricity prices and participants’ behavior patterns. Finally, ref. [36] formulates a DR problem as a stochastic Stackelberg game and designs a two-timescale reinforcement learning algorithm to determine the DR scheduling without sharing participants’ private information.

The above-mentioned studies demonstrate the efficacy of DRL-based approaches in solving DR optimization problems while considering the uncertainties and complexities involved in the DR process. However, the application of DRL algorithms to optimizing demand response programs has primarily focused on single-energy systems, neglecting the unique challenges posed by multi-energy demand response. Multi-energy demand response programs require the simultaneous optimization of energy usage from different sources—such as electricity, gas, and heat—to achieve optimal energy management. The integration of these diverse energy sources, coupled with the unknown preferences of end users, presents a significant research gap in the successful application of DRL for multi-energy demand response. The complexities of coupling multiple energy sources and incorporating end users’ preferences into the optimization process require the development of novel DRL algorithms specifically tailored to address these challenges. Thus, there is a
pressing need to bridge this gap and develop effective DRL-based approaches capable of handling the complex multi-energy characteristics of demand response programs.

In this study, we address the problem of optimizing energy consumption in a multi-energy microgrid through a novel DR framework based on DRL. The objective of our framework is to maximize the total profits of DRPs by effectively managing the energy usage across multiple energy systems. We recognize the interdependencies between different energy systems and the need to adapt to dynamic and uncertain changes in end users’ behavior, without any prior information about their preferences.

To tackle this problem, we first develop an integrated scheduling model that combines power and gas demand response, accounting for the diverse energy sources and residential end users. We then formulate the pricing strategy of MDP with an unknown transition. To effectively train neural networks and learn pricing strategies, we employ the novel SAC algorithm with the entropy function as regularization terms. This approach allows us to maximize DRPs’ profits in the presence of uncertain electricity prices and PV output.

To evaluate the effectiveness of our proposed approach, we conduct case studies in both deterministic and stochastic environment settings. The results demonstrate the efficacy of our framework in optimizing energy consumption, showcasing its potential for practical applications in the design and optimization of multi-energy microgrids. By promoting the integration of renewable energy sources and improving the sustainability of the grid, our proposed framework contributes to advancing the state-of-the-art in multi-energy microgrid design and optimization.

Our main contributions are threefold:

1. Development of an integrated DR scheduling and pricing model in multi-energy microgrids: The proposed approach presents an integrated scheduling model that combines power and gas demand response for different types of residential end users. This model can effectively address the increasing complexity of multi-energy coordinated microgrids and provide a coordinated scheduling and pricing strategy for DRPs.


3. Effective handling of multiple sources of uncertainties: The proposed approach is shown to be effective in handling different levels of uncertainties and achieving a near-optimal pricing strategy. Case studies conducted in both deterministic and stochastic environment settings demonstrate the effectiveness of the proposed approach in optimizing energy consumption and maximizing DRPs’ total profits without requiring end users’ private information.

The rest of the paper is organized as follows. Section 2 presents the formulation of an integrated DR scheduling and pricing model in multi-energy microgrids. Section 3 describes the proposed DRL-based DR framework with the novel SAC algorithm. Section 4 presents the simulation experiments and results. Finally, Section 5 concludes the paper and discusses future research directions.

2. Formulation of Integrated DR Scheduling and Pricing Model

In this section, the integrated DR scheduling and pricing model is formulated using bilevel programming, where the lower-level is the optimal response by an end user with multi-energy demands and the upper-level is the optimal scheduling of multi-energy microgrids and the dynamic prices setting from the DRP’s perspective.

2.1. The Overall Scheme

Figure 1 illustrates the proposed integrated DR scheduling and pricing model designed to optimize the scheduling of multi-energy microgrids and maximize the total profits of
DRPs. The multi-energy microgrids include three energy flows: power flow, heat flow, and gas flow. The power flow consists of electricity purchased from or sold to the upper-level grid, uncertain output of PVs, output of a gas turbine, and electricity used by electric heat pump and end users. The heat flow consists of heat generated by the gas turbine and electric heat pump, as well as heat consumed by end users. The gas flow describes the balance between the gas supply and the gas consumption by the gas turbine and end user. End users are subject to energy prices set by DRPs and adjust their consumption behavior of different energy sources based on their private preferences, which remain unknown to DRPs. Therefore, DRPs face two main challenges in solving the proposed model: (1) how to address uncertainties of electricity prices and PV output; (2) how to set prices of different energy sources for end users with unknown preferences. The proposed model is formulated using bilevel programming in the following sections.

\[ \text{Figure 1. The overall scheme of integrated DR scheduling and pricing model.} \]

### 2.2. The Overall Scheme

The upper level describes the optimal scheduling of multi-energy microgrids and the dynamic prices setting from the DRP’s perspective. In particular, the objective function is to maximize the total profits, as follows:

\[ \max_{t \in T} \sum_{k \in K} \left( \sum_{i \in I} \left( \lambda_i^{e_k} p_i^k + \lambda_i^{k,g} g_i^k + \lambda_i^{k,h} h_i^k - \lambda_i^{e_{\text{in}}} p_{i,t}^{\text{in}} - \lambda_i^{g_{\text{in}}} g_{i,t}^{\text{in}} + \lambda_i^{e_{\text{out}}} p_{i,t}^{\text{out}} \right) \right) \]  

(1)

where \( p_i^k, g_i^k, \) and \( h_i^k \) denote the power, gas, and heat demands of end user \( k \) at period \( t \), which are determined in the lower-level problem based on their reaction to the electricity prices \( \lambda_i^{e_k} \), gas prices \( \lambda_i^{k,g} \), and heat prices \( \lambda_i^{k,h} \) set by DRPs. In addition, \( \lambda_i^{e_{\text{in}}} \), \( \lambda_i^{e_{\text{out}}} \), and \( \lambda_i^{g_{\text{in}}} \) represent the purchase, sell electricity prices, and purchase gas price at period \( t \). The relationship among the above-mentioned prices can be presented as follows:

\[ \lambda_i^{e_{\text{in}}} \leq \lambda_i^{e_k} \leq \lambda_i^{e_{\text{max}}}, \forall t, k \]  

(2)

\[ \lambda_i^{g_{\text{in}}} \leq \lambda_i^{k,g} \leq \lambda_i^{g_{\text{max}}}, \forall t, k \]  

(3)

\[ \lambda_i^{h_{\text{in}}} \leq \lambda_i^{k,h} \leq \lambda_i^{h_{\text{max}}}, \forall t, k \]  

(4)

where constraints (2)–(4) restrict the bound of the prices of different demands set by DRPs. Accordingly, end users will react to these prices based on their preferences and adjust their consumption behavior. Hence, the multi-energy balance constraints can be formulated as follows:

\[ p_{i,t}^{\text{in}} - p_{i,t}^{\text{out}} + p_{i,t}^{e} + p_{i,t}^{PV} = \sum_k p_i^k + p_i^h, \forall t \]  

(5)
where constraint (15) denotes the concave welfare function of the $k$th end user, and $\lambda^{k,e}_t$, $\lambda^{k,g}_t$, $\lambda^{k,h}_t$, and $\lambda^{k,w}_t$ represent the preferences parameters that have an influence on the reactions of the end user to the prices of multi-energy demands. In addition,

$$h^p_t + h^g_t = \sum_k h^k_t, \forall t$$  \hspace{1cm} (6)$$

$$s^{in}_t = s^g_t + \sum_k s^k_t, \forall t$$ \hspace{1cm} (7)

where constraints (5)–(7) represent the power, heat, and gas balance, respectively. In particular, $p^{in}_t$ and $p^{out}_t$ denote the purchase and sell electricity from the upper-level grid. $p^{PV}_t$ and $p^{PV}_t$ are the output of the gas turbine and PV with uncertainties. In addition, $p^{h}_t$ is the power of the electric heat pump. As for the heat and gas balance, $h^p_t$ and $h^g_t$ indicate the heat generated by the electric heat pump and the gas turbine, while $s^{in}_t$ and $s^{g}_t$ are the purchased and consumed gas. These decision variables must obey the following restrictions:

$$p^{k,\text{min}}_t \leq p^k_t \leq p^{k,\text{max}}_t, \forall t$$ \hspace{1cm} (8)$$

$$0 \leq p^{PV}_t \leq \overline{p}^{PV}_t, \forall t$$ \hspace{1cm} (9)$$

$$0 \leq p^h_t \leq p^{h,\text{max}}_t, \forall t$$ \hspace{1cm} (10)$$

$$\eta^{h} p^h_t = h^p_t, \forall t$$ \hspace{1cm} (11)$$

$$\eta^{g} s^g_t = h^g_t, \forall t$$ \hspace{1cm} (12)$$

$$\eta^{w} s^{w}_t = p^w_t, \forall t$$ \hspace{1cm} (13)$$

where constraint (8) describes the bound of the gas turbine output $p^{k,\text{min}}_t$ and $p^{k,\text{max}}_t$. Constraint (9) ensures that the PV output is not greater than the actual output $\overline{p}^{PV}_t$. Constraint (10) limits the maximum power of the electric heat pump. Constraints (11)–(13) describe the process of converting one energy into another, where $\eta^{h}$, $\eta^{g}$, and $\eta^{w}$ denote the corresponding conversion efficiency. Specifically, $\eta^{h}$ is the efficiency of the electric heat pump converting electricity into heat, while $\eta^{g}$ and $\eta^{w}$ represent the heat and power conversion efficiency of the gas turbine.

### 2.3. The Formulation of Lower-Level Problem

Once the DRPs set the prices of multi-energy demands, the end users will react to the price signals and start adjusting their consumption behavior according to their preferences to maximize their welfares. Based on that, DRPs will observe the multi-energy demands and perform the corresponding optimal scheduling. The details of the $k$th end user’s reaction in the lower-level problem can be formulated as follows:

$$\max_{p^k_t, s^g_t, h^k_t} \sum_{t \in T} \left( U \left(p^k_t, s^g_t, h^k_t \right) - \lambda^{k,e}_t p^k_t - \lambda^{k,g}_t s^g_t - \lambda^{k,h}_t h^k_t \right)$$ \hspace{1cm} (14)$$

$$U \left(p^k_t, s^g_t, h^k_t \right) = a^{k,p}_t \left(p^k_t \right)^2 + b^{k,p}_t p^k_t + c^{k,\text{in}}_t s^g_t + c^{k,\text{out}}_t + d^{k,\text{h}}_t h^k_t + e^{k,\text{w}}_t, \forall t, k$$ \hspace{1cm} (15)$$

$$0 \leq p^k_t \leq p^{k,\text{max}}_t, \forall t, k$$ \hspace{1cm} (16)$$

$$0 \leq s^g_t \leq s^{g,\text{max}}_t, \forall k, t$$ \hspace{1cm} (17)$$

$$0 \leq h^k_t \leq h^{k,\text{max}}_t, \forall k, t$$ \hspace{1cm} (18)$$

where constraint (15) denotes the concave welfare function of the $k$th end user, and $a^{k,p}_t$, $b^{k,p}_t$, $c^{k,\text{in}}_t$, $c^{k,\text{out}}_t$, $d^{k,\text{h}}_t$, and $e^{k,\text{w}}_t$ represent the preferences parameters that have an influence on the reactions of the end user to the prices of multi-energy demands. In addition,
constraints (16)–(18) describe the upper limits on the energy consumption of different energies. The aforementioned parameters, such as the private preferences of the end users and energy prices, are not disclosed to the DRPs. This limitation motivates the development of a Markov Decision Process (MDP)-based formulation in the following sections.

3. MDP Formulation for DRP’s Pricing Strategy

To determine the DRP’s pricing strategy in a model-free manner, the previously mentioned bi-level programming is reformulated into an MDP problem. This allows for the development of a DRL-based framework to learn the near-optimal pricing strategy, given the uncertainties of electricity prices and the unknown preferences parameters of the end users. The proposed framework operates on a finite MDP with an unknown transition probability, enabling the DRP to devise a pricing strategy without requiring any private information from the end users.

At each period, the DRP receives information regarding electricity and gas prices \( \lambda_{t}^{e,\text{in}}, \lambda_{t}^{e,\text{out}}, \) and \( \lambda_{t}^{g,\text{in}} \) from the upper-level grid and gas supply, respectively. Using this information and the state information from previous periods, the DRP forms a state \( s_{t} \). Based on this state, the policy network of the DRP determines its action \( a_{t} \), which includes the electricity prices \( \lambda_{t}^{k,e} \), gas prices \( \lambda_{t}^{k,g} \), and heat prices \( \lambda_{t}^{k,h} \). These prices are then communicated to each end user at period \( t \). Upon receiving the dynamic price, each end user solves its own welfare-maximizing problem and adjusts its consumption behavior accordingly. The DRP is then able to calculate its profits based on these adjustments. Next, the DRP observes the new state \( s_{t+1} \) with unknown transition probability and generates a new action \( a_{t+1} \). This process is repeated in an online manner until the DRP’s pricing strategy converges. The details of the MDP formulation are further illustrated below.

3.1. States

The upper level describes the optimal scheduling of multi-energy microgrids and the dynamic prices setting from the DRP’s perspective. In particular, the objective function is to maximize the total profits, as follows:

The states of a DRP include the electricity and gas prices \( \lambda_{t}^{e,\text{in}}, \lambda_{t}^{e,\text{out}}, \) and \( \lambda_{t}^{g,\text{in}} \) electricity prices from the upper-level grid, and the gas supply and reaction of the end users at the process is repeated in an online manner until the DRP’s pricing strategy converges. The \( \lambda \) probability, enabling the DRP to devise a pricing strategy without requiring any private information from the end users.

At each period, the DRP receives information regarding electricity and gas prices \( \lambda_{t}^{e,\text{in}}, \lambda_{t}^{e,\text{out}}, \) and \( \lambda_{t}^{g,\text{in}} \) from the upper-level grid and gas supply, respectively. Using this information and the state information from previous periods, the DRP forms a state \( s_{t} \). Based on this state, the policy network of the DRP determines its action \( a_{t} \), which includes the electricity prices \( \lambda_{t}^{k,e} \), gas prices \( \lambda_{t}^{k,g} \), and heat prices \( \lambda_{t}^{k,h} \). These prices are then communicated to each end user at period \( t \). Upon receiving the dynamic price, each end user solves its own welfare-maximizing problem and adjusts its consumption behavior accordingly. The DRP is then able to calculate its profits based on these adjustments. Next, the DRP observes the new state \( s_{t+1} \) with unknown transition probability and generates a new action \( a_{t+1} \). This process is repeated in an online manner until the DRP’s pricing strategy converges. The details of the MDP formulation are further illustrated below.

3.1. States

The upper level describes the optimal scheduling of multi-energy microgrids and the dynamic prices setting from the DRP’s perspective. In particular, the objective function is to maximize the total profits, as follows:

The states of a DRP include the electricity and gas prices \( \lambda_{t}^{e,\text{in}}, \lambda_{t}^{e,\text{out}}, \) and \( \lambda_{t}^{g,\text{in}} \) electricity prices from the upper-level grid, and the gas supply and reaction of the end users at the previous \( N \) and \( M \) periods, concatenated as follows:

\[
s_{t} = \left( \lambda_{t}^{e,\text{in}} \right)_{k=1}^{K} \lambda_{t}^{e,\text{out}} \lambda_{t}^{g,\text{in}} \lambda_{t}^{k,e} \lambda_{t}^{k,g} \lambda_{t}^{k,h} \right)_{k=1}^{K}, \quad t \in T
\]  

(19)

3.2. Actions

Observing the states \( s_{t} \), the DRPs take an action, representing the dynamic price of multi-energy demands sent to each end user \( k \) at period:

\[
a_{t} = \left[ \lambda_{t}^{k,e}, \lambda_{t}^{k,g}, \lambda_{t}^{k,h} \right]_{k=1}^{K}, \quad \forall t
\]  

(20)

3.3. Reward

In response to the prices set by the DRPs, the end users adjust their loads, leading to a resulting reward for the DRPs at time period \( t \). The reward function based on the objective function (1) is expressed as follows:

\[
r_{t} = \sum_{k \in K} \left( \lambda_{t}^{k,e} p_{t}^{k} + \lambda_{t}^{k,g} q_{t}^{k} + \lambda_{t}^{k,h} h_{t}^{k} - \lambda_{t}^{e,\text{in}} p_{t}^{\text{in}} - \lambda_{t}^{g,\text{in}} q_{t}^{\text{in}} + \lambda_{t}^{e,\text{out}} p_{t}^{\text{out}} \right), \quad \forall t
\]  

(21)

3.4. Transition Function

The states transition from \( s_{t} \) to \( s_{t+1} \) is influenced jointly by the action \( a_{t} \) taken by DRPs and the unknown parameter of the end user \( k \) as follows:

\[
s_{t+1} = f(a_{t}, s_{t}, \omega)
\]  

(22)
where $\omega$ represents the exogenous randomness variable in the environment.

4. The DRL-Based Algorithm for Solving MDP Formulation

In this section, we propose the soft actor critic (SAC) algorithm to approximate the solution to the MDP formulation of the DRPs’ pricing problem and learn the near-optimal pricing strategy in an adaptive manner.

The Soft Actor-Critic (SAC) algorithm is a key component of our research, designed to optimize the demand response in multi-energy systems. SAC is an off-policy reinforcement learning algorithm that combines the actor-critic framework with maximum entropy reinforcement learning. It is particularly well-suited for problems with continuous action spaces and stochastic environments.

At its core, SAC employs an actor-critic architecture, where the actor network learns the policy, mapping observations to actions, while the critic network estimates the value function, providing a measure of the expected return. The policy is updated based on the advantage function, which quantifies the advantage of taking a particular action given the current state and value estimates.

Exploration-exploitation trade-off is a crucial aspect of SAC. To maintain a balance between exploration and exploitation, SAC incorporates entropy regularization. By maximizing the entropy of the policy, SAC encourages exploration, preventing the agent from getting stuck in suboptimal solutions and promoting the discovery of diverse and potentially better strategies. This property is particularly valuable in dynamic multi-energy systems, where the optimal solutions can change over time.

In the context of our research, we have made specific adaptations to the SAC algorithm to address the complexities of the multi-energy demand response problem. We have integrated additional components to handle uncertainties arising from the PV output and electricity prices. By incorporating these uncertainties into the algorithm, SAC can effectively adapt to dynamic changes in the system and provide robust policies that maximize the DRPs’ profits.

Overall, the SAC algorithm provides a powerful framework for learning optimal policies in multi-energy demand response. It combines the advantages of actor-critic architecture, exploration through entropy regularization, and adaptability to uncertain environments. By elaborating on the SAC algorithm, we aim to provide readers with a comprehensive understanding of its functioning and its significance in achieving the objectives of our study. In this paper, the neural network architecture for the SAC algorithm is designed to consist of three components: an actor network, a critic network, and a temperature parameter network. The actor network learns to approximate the optimal policy, which maps the states to the action space. The critic network estimates the state-value function, which evaluates the goodness of the states. The temperature parameter network is responsible for scaling the exploration noise added to the policy during the training process.

The training process of the SAC algorithm includes two main stages: the actor-critic learning stage and the temperature learning stage. In the actor-critic learning stage, the actor and critic networks are updated based on the temporal difference (TD) error between the predicted and actual state-value functions. In the temperature learning stage, the temperature parameter network is updated to control the entropy in the Q-function. The training process is repeated until convergence or a predefined stopping criterion is reached.

The detailed structure of the neural network and training process is discussed in the following sections.

4.1. Preliminaries

The upper level describes the optimal scheduling of multi-energy microgrids and the dynamic prices setting from the DRP’s perspective. In particular, the objective function is to maximize the total profits, as follows:
The soft actor critic (SAC) algorithm is a maximum-entropy off-policy DRL method that offers improved sample efficiency and robust training results. In contrast to traditional deep Q networks (DQN), SAC employs soft updates to the Q-network, which encourages exploration over the action space during the training process. In addition to the reward obtained from the environment, the agent also receives an entropy bonus that captures the randomness of the current policy to prevent premature convergence to suboptimal policies. Therefore, the objective of training is to maximize the total reward (or minimize the total cost) given by:

$$\pi_\theta^* = \arg\max_\theta \sum_{i=1}^T \gamma^i \mathbb{E}[r_i(s_i, a_i) + \alpha H(\pi_\theta(s_i))]$$

where $\pi_\theta$ represents the policy with parameters $\theta$ and $\gamma$ indicates the discount factor. $H(\cdot)$ is the entropy function and $\alpha$ is the temperature factor that denotes the importance of the entropy term.

### 4.2. Training Process

The upper level describes the optimal scheduling of multi-energy microgrids and the dynamic prices setting from the DRP’s perspective. In particular, the objective function is to maximize the total profits as follows:

Figure 2 depicts the architecture of the SAC algorithm, which is comprised of three fully connected neural networks: the Actor, Critic, and temperature parameter networks. The Critic network is parameterized by $\varphi$ and is trained to approximate the soft Bellman optimal equation. To address the overestimation of Q values, the clipped double DQN is adopted with target networks parameterized by $\overline{\varphi}$.

$$y = \min_{i=1,2} r_i + \gamma \left( Q^\overline{\varphi}(s_{t+1}, a_{t+1}) - \alpha \log \pi_\theta(a_{t+1}|s_{t+1}) \right)$$

$$J(\varphi_i) = \mathbb{E}_{(s_i, a_i, r_i, s_{i+1}) \sim B} \left[ (Q_i(s_i, a_i) - y)^2 \right], i = 1, 2$$

where $y$ is the update target of two soft $q$ functions and $\log \pi_\theta(\cdot)$ indicates the policy entropy. The policy $\pi_\theta$ of Actor aims to maximize the sum of the minimum of the two $q$ functions and the entropy bonus:

$$a_\theta = \tanh(\mu + \varepsilon \sigma), \varepsilon \sim N$$
\[
J(\theta) = \max_{\theta} \mathbb{E}_{\epsilon \sim N} \left[ \min_{i=1,2} Q^{\theta_i}(s_t, a_\theta) - \alpha \log \pi_\theta(a_\theta | s_t) \right]
\] (27)

where \(a_\theta\) is the action sample calculated by states, the output of Actor, and independent noise, which allows gradients being backpropagated through the Actor network.

The temperature factor \(\alpha\) plays a crucial role in balancing the exploration and exploitation. Hence, the entropy automated adjusting method is utilized to fine-tune the \(\alpha\) via minimizing the objective function, as follows:

\[
J(\alpha) = \mathbb{E}_{\epsilon \sim N} \left[ -\alpha \log \pi_\theta(a_\theta | s_t) - \alpha \mathcal{H} \right]
\] (28)

where the target entropy \(\mathcal{H}\) is set as the negative of the action dimension. In addition, target networks are applied to smooth the approximation of the q functions.

Accordingly, the specific training process with the SAC algorithm is presented in Algorithm 1. In Steps 1–2, the parameters of the q-function network and the policy network are initialized, and an empty replay buffer is created. In Steps 5–8, we utilize the policy network to construct the prices of multi-energy demands from the DRP’s perspective and observe the reaction of different end users based on the lower-level problem, where the private preferences remain unknown. Then, the scheduling model of the multi-energy microgrid is solved to calculate the total operation cost as a reward signal. The corresponding state transition will be recorded in the replay buffer. Finally, in Steps 11–14 of each gradient step, the parameters of the Actor, Critic, target networks and temperature factor are updated, respectively.

---

**Algorithm 1** The Proposed DRL-based Integrated DR with SAC Algorithm

1. Initialize replay buffer \(B\)
2. Initialize Actor \(\theta\), Critic \(\phi_i, \alpha\) and target network \(\overline{\phi}_i\)
3. for each epoch do
4. for each state transition step do
5. Given \(s_t\), take actions \(a_t\) based on (26)
6. Observe the multi-energy demands \((14)\) with \(a_t\) as prices
7. Solve the scheduling model and obtain operation costs
8. Receive \(r_t, s_{t+1}\) and record them in buffer \(B\)
9. end for
10. for each gradient step do
11. \(\theta \leftarrow \theta - \lambda_\theta \nabla J(\theta)\)
12. \(\phi_i \leftarrow \theta - \lambda_{\phi_i} \nabla J(\phi_i)\)
13. \(\alpha \leftarrow \alpha - \lambda_\alpha \nabla J(\alpha)\)
14. \(\overline{\phi}_i \leftarrow \tau \overline{\phi}_i - (1 - \tau) \phi_i\)
15. end for
16. end for

The SAC algorithm, as adopted in this study, introduces a trade-off between bias and variance that warrants discussion. This trade-off arises from potential inconsistencies between the critic’s value function and the actor’s policy, which can have implications for the quality of the policy gradient.

One aspect to consider is the potential bias introduced by an inaccurate or inconsistent value function. If the critic’s estimate of the value function is biased, it can lead to suboptimal policy updates. In such cases, the actor’s policy may not be guided effectively towards the global optimum, resulting in reduced performance and suboptimal solutions. It is crucial to acknowledge this bias and its potential impact on the overall performance of the SAC algorithm. On the other hand, the variance of the policy gradient estimates can also affect the performance of SAC. High variance may lead to unstable training and hinder the convergence of the algorithm. Therefore, achieving an appropriate balance between bias and variance is crucial to ensure stable and effective learning.
To address this trade-off, future improvements to the adopted SAC algorithm can be explored. One potential avenue is the utilization of advanced value function approximation techniques. These techniques aim to reduce bias by improving the accuracy and consistency of the critic’s value function estimate. Examples include bootstrapping methods, ensemble methods, or incorporating other value function approximation architectures. Additionally, the integration of regularization methods can help manage the bias-variance trade-off. Regularization techniques, such as entropy regularization, can help to balance exploration and exploitation, leading to more stable and robust policy learning. By carefully adjusting the regularization terms, it is possible to mitigate biases and stabilize the training process while maintaining the exploration capabilities. Furthermore, ensemble methods can be employed to address both bias and variance concerns. By training multiple critics or actors concurrently and aggregating their outputs, ensemble methods can provide more accurate value function estimates and policy updates. This approach can help reduce bias and variance, leading to improved performance and convergence properties.

5. Case Studies and Discussion

In the section, we present multiple case studies to verify the effectiveness of the proposed DRL-based framework for integrated scheduling and DR in multi-energy microgrids. First, the experimental settings are presented, and the training process that learns the optimal DRP pricing strategy is shown. Then, we illustrate the performance of the proposed framework under uncertainties in loads and electricity prices. Finally, to manifest its huge potential in the real-world DR application, we further test the ability of the generalization of the proposed framework under different levels of uncertainties.

5.1. Training Process

Three classes of end users are considered with different time-varying preferences of multi-energy demands. In particular, each type of end user has different parameters \(a_k^p\), \(b_k^p\), \(a_k^g\), \(b_k^g\) for electric and gas demands, while the heat demands are fixed. Firstly, the neural networks are trained with deterministic data to illustrate their ability to approximate the near-optimal pricing strategies. Secondly, the neural networks are trained in an uncertain environment, where the electricity prices of the upper-level grid and PV output follow truncated normal distributions. In the training under uncertainties, the variances of the uncertain prices and PV output are taken as 5% to their means. To further illustrate its generalization ability, in Section 5.2, we change the levels of uncertainties from 0% to 25%.

This part presents the results of the training process of neural networks based on deterministic data, which has been conducted over 10,000 epochs, as depicted in Figure 3. The effectiveness of the deep reinforcement learning (DRL)-based framework has been evaluated by calculating the ratio of the optimal return obtained by the bi-level model and that achieved by the neural network. It should be noted that the bi-level model was provided with the full private preferences of the end users, whereas the proposed DRL-based approach does not require any prior information. The graph shows that the ratio of the neural network varies significantly between 0.90 and 0.67 in the first 2000 training epochs due to the SAC algorithm, which encourages exploration and makes the neural network more inclined to choose actions with higher entropy to explore the action space. These actions are seldom selected, thereby enhancing the accuracy of the neural network for global value estimation. As the number of training epochs increases, the temperature neural network continually updates its own parameters, and the reward for action space exploration is gradually reduced. As a result, the policy neural network begins to choose the optimal strategy based on the current Q-function fitting. Therefore, the ratio increases rapidly to 0.95 from the 2000th to the 4000th training epoch, indicating that the neural network has an excellent ability to learn the optimal strategy. After the 4000th training epoch, the performance of the neural network becomes stable, and it converges around 0.97.
In this part, we compare the effectiveness of the bi-level model and policy neural network in regulating the end users’ demands in the demand response. The bi-level model is employed to obtain the optimal multi-energy demand response curves of end users under the assumption that all end users’ preferences are known. On the other hand, the policy neural network is trained to output the pricing of DRP, which is sent to different users, and their energy consumption feedback is observed. The results are presented in Figures 4 and 5, which demonstrate that different end users have distinct energy preferences. Specifically, end user A tends to use more electricity loads in the early hours of the morning, while user B prefers electricity loads between 18:00 and 24:00 to maximize their utilities. End user B is more likely to use the gas load between 1 and 5 and between 18:00 and 24:00. Our proposed deep reinforcement learning (DRL)-based method can learn these patterns effectively. In terms of electric load, the average error between the results obtained by the neural network and the two-layer model is 0.021 MWh, with end user B’s electricity demands curve being closest to the optimal value, and the average error being only 0.014 MWh. In terms of gas demands, the average error between the results obtained by the neural network and the two-layer model is 0.033 MWh, with user B being closest to the optimal value, and the average error being only 0.032 MWh. The detailed hourly errors are presented in Tables 1 and 2. In our study, the maximum average error in Figures 4 and 5 can be attributed to several factors. Firstly, it is important to note that the learning process in the multi-energy demand response is a complex task influenced by various uncertainties and dynamic factors. These factors can include variations in the energy supply, changes in user behavior, and fluctuations in energy prices. These results demonstrate that the proposed DRL-based framework can adaptively obtain the optimal pricing strategy and approximate the multiple energy demand curves of different users without any prior knowledge of the user parameters. Moreover, our proposed method can effectively learn the information about electricity prices from the perspective of a multi-energy microgrid. It is worth noting that the policy neural network does not explicitly predict the future electricity price but implicitly models it through the MDP. Figure 6 shows that the neural network will choose to purchase electricity when the electricity price is low, such as 0:00 to 3:00, and reduce the electricity purchase and increase the gas purchase when the electricity price is high.

![The training process of neural networks with deterministic data.](image)
during the day. In this way, the demand of gas turbine power generation is met, utilizing the complementary characteristics of the multi-energy microgrids.

![Figure 4](image1.png)  ![Figure 5](image2.png)

**Figure 4.** The comparison between the bi-level model and neural networks of electric demands.

**Figure 5.** The comparison between the bi-level model and neural networks of gas demands.

We note that the reason why the proposed method reaches the minimum power demand constraints between the 6th and 12th hour can be attributed to the preferences and welfare of user C during those hours. User C exhibits a relatively low preference for energy consumption during these specific time intervals, resulting in a lower overall welfare associated with energy usage.

In the proposed method, the reward signal is formulated based on 24 periods in a day, and the reward for each specific period contributes only a small proportion to the overall reward. Consequently, the policy network may not be strongly incentivized to explore fine-grained actions during these specific hours. As a result, the proposed method tends to prioritize minimizing the power demand during these periods to meet the minimum constraints, aligning with the relatively lower preference of user C.

5.2. Performance under Different Sources of Uncertainties

In this section, we conduct further training of the proposed approach under various uncertainties to investigate its effectiveness in regulating the end users’ demands in demand response. To introduce a realistic level of uncertainty, we consider a 5% level of uncertainty in the training process of the policy neural network. This uncertainty is reflected by setting the ratio of the variance of the truncated normal distribution of the PV output and electricity prices to the expected values at 5%.
Table 1. Hourly errors of electricity demands between the bi-level model and neural networks.

<table>
<thead>
<tr>
<th>Hour</th>
<th>Hourly Errors/MWh</th>
<th>User A</th>
<th>User B</th>
<th>User C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−0.01995</td>
<td>0.00736</td>
<td>0.03977</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>−0.0137</td>
<td>−0.01003</td>
<td>0.02102</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>−0.0222</td>
<td>−0.00241</td>
<td>0.00904</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>−0.00194</td>
<td>−0.01694</td>
<td>0.00954</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>−0.0235</td>
<td>0.00734</td>
<td>0.04359</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>−0.00671</td>
<td>−0.02221</td>
<td>0.0244</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.01693</td>
<td>0.02661</td>
<td>0.04964</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>−0.03101</td>
<td>0.01371</td>
<td>0.04179</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.00419</td>
<td>0.02728</td>
<td>0.02286</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>−0.00189</td>
<td>−0.01182</td>
<td>0.02071</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>−0.00484</td>
<td>0.03804</td>
<td>0.03464</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.02852</td>
<td>−0.02331</td>
<td>0.04536</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.00133</td>
<td>−0.00603</td>
<td>0.06999</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>−0.03729</td>
<td>−0.02082</td>
<td>0.04002</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>−0.00575</td>
<td>−0.00348</td>
<td>0.01835</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.00032</td>
<td>−0.00389</td>
<td>0.00596</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>−0.04472</td>
<td>0.01054</td>
<td>−0.00604</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.0447</td>
<td>−0.00917</td>
<td>0.03792</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.06453</td>
<td>−0.02099</td>
<td>0.03089</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.0575</td>
<td>−0.00857</td>
<td>0.01788</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.0094</td>
<td>−0.01341</td>
<td>0.02726</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.03182</td>
<td>0.01796</td>
<td>0.02183</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.01263</td>
<td>0.00296</td>
<td>0.00786</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>−0.03201</td>
<td>−0.00719</td>
<td>−0.01414</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Hourly errors of gas demands between the bi-level model and neural networks.

<table>
<thead>
<tr>
<th>Hour</th>
<th>Hourly Errors/MWh</th>
<th>User A</th>
<th>User B</th>
<th>User C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>−0.04213</td>
<td>0.08414</td>
<td>−0.07802</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>−0.05686</td>
<td>−0.02308</td>
<td>0.05061</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>−0.00436</td>
<td>0.04866</td>
<td>0.01414</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>−0.00637</td>
<td>−0.01083</td>
<td>0.00128</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.0179</td>
<td>−0.09156</td>
<td>−0.01309</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>−0.05397</td>
<td>0.06944</td>
<td>−0.00448</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.0506</td>
<td>−0.00771</td>
<td>0.06495</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>−0.05449</td>
<td>0.00174</td>
<td>0.03304</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.00331</td>
<td>0.0000</td>
<td>0.0112</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>−0.04153</td>
<td>0.0289</td>
<td>0.00322</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>−0.04124</td>
<td>−0.01603</td>
<td>−0.01264</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.02022</td>
<td>−0.05587</td>
<td>−0.05077</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.0734</td>
<td>0.02723</td>
<td>0.00992</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>−0.04404</td>
<td>−0.05814</td>
<td>0.05534</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>−0.04055</td>
<td>−0.01046</td>
<td>−0.00012</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.00345</td>
<td>0.01429</td>
<td>−0.05413</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>−0.00654</td>
<td>−0.03245</td>
<td>−0.03731</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>−0.03167</td>
<td>0.01615</td>
<td>−0.03698</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.00649</td>
<td>0.01492</td>
<td>−0.09224</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.06967</td>
<td>−0.02087</td>
<td>0.04356</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>−0.00887</td>
<td>0.06686</td>
<td>0.01799</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>−0.01169</td>
<td>0.05959</td>
<td>−0.05212</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.0365</td>
<td>−0.02661</td>
<td>−0.02633</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>−0.05421</td>
<td>−0.03495</td>
<td>−0.02088</td>
<td></td>
</tr>
</tbody>
</table>
Figure 6. The net purchase of electricity and gas of the multi-energy microgrids.

Figure 7 provides valuable insights into the performance of the policy neural network during training under uncertainties. During the initial 2000 epochs, the performance of the network is comparable to that achieved with the deterministic data. The performance ratio fluctuates between 0.90 and 0.70, indicating the network's ability to adapt to uncertainties and maintain reasonable performance levels. From the 2000th to the 4000th epoch, the performance of the policy neural network shows rapid improvement, with the ratio reaching approximately 0.92.

Figure 7. The training process of neural networks under uncertainties.

Compared to the training process using the deterministic data, the training process considering uncertainties exhibits two notable differences. Firstly, after the 4000th epoch, although the performance of the neural network still fluctuates to some extent, the overall trend remains upward, and the performance ratio ultimately converges to around 0.94. This demonstrates the network's capability to learn and adapt to uncertain conditions, gradually improving its performance. Secondly, due to the presence of uncertainties, the policy neural
network encounters some large spikes during training. However, it is important to note that as the number of training epochs increases, the network successfully learns to avoid unreasonable actions, leading to the elimination of these spikes.

Furthermore, an interesting observation can be made regarding the range of ratios obtained under uncertainties compared to the deterministic data scenario. The wider range of ratios observed in the uncertain environment can be attributed to the fact that uncertainties introduce additional entropy to the value function in the Soft Actor-Critic (SAC) algorithm, thereby promoting exploration. It is important to note that as training progresses and the number of epochs increases, the fluctuations in the ratios gradually converge. This convergence indicates that the policy neural network adapts and learns to navigate the uncertain environment, eventually settling into a more stable and optimal performance. This behavior highlights the capability of the proposed approach to effectively handle uncertainties and improve its performance over time through exploration and learning.

By investigating the effectiveness of the bi-level model and policy neural network in regulating end users’ demands under multiple sources of uncertainties, we gain valuable insights into the adaptability of the proposed approach, as shown in Figures 8 and 9. Our results reveal that the policy neural network can effectively learn the preferences of different users’ electricity and gas demands, closely approximating the optimal curve calculated by the bi-level model. Specifically, in terms of electricity demands, the neural network achieves an average error of 0.037 MWh compared to the bi-level model, while for gas demands, it yields an average error of 0.056 MWh. These results validate the effectiveness and adaptability of the proposed approach, demonstrating its capability to handle random fluctuations and uncertainties.

![Figure 8. The comparison between the bi-level model and neural networks of electric demands under uncertainties.](image)

In the final analysis, the generalization ability of the proposed DRL-based approach is tested. The parameters of the policy neural network are first fixed after training, and the level of uncertainties is then adjusted to evaluate the neural network’s generalization ability. Specifically, 100 scenarios are generated for each uncertainty level, and the results are shown in Figure 10. The green portion of the figure represents the probability density distribution of the ratio under a given level of uncertainty, while the black rectangle represents the range of 25% and 75% quantiles, and the white square represents the median. It can be observed that as the level of uncertainty increases gradually from 0% to 25%, the ratio as a whole remains at a good level. When the uncertainty level is not higher than 12.5%, the median ratio is higher than 92%, which indicates that the policy neural network has a strong generalization performance.
In addition to evaluating the performance of the proposed approach, we conducted an analysis of the DRP’s total daily profits in the context of multi-energy demand response. In the deterministic environment, where the PV output and electricity prices are fixed but unknown, the DRP’s total profits were found to amount to USD 407.28. Comparatively, the optimal solution obtained from the bilevel model resulted in total profits of USD 427.57. The ratio between these two profit values is calculated to be 95.3%, indicating the effectiveness of the proposed approach in achieving significant profitability for the DRP.

To further assess the financial implications and profitability of the DRP in a stochastic environment, we introduced a 5% uncertainty level. The results, as shown in Table 3, demonstrate that the DRP’s total profits with the proposed approach amounted to USD 370.24, while the profits based on the optimal solution reached USD 405.42. It is important to note that the stochastic environment introduces additional uncertainties, resulting in a reduction in profits for both approaches. Nevertheless, the ratio between the profits obtained with the proposed approach and the optimal solution remains notably high, at 91.2%.

Figure 9. The comparison between the bi-level model and neural networks of gas demands under uncertainties.

Figure 10. Performance test of the proposed DRL-based approach with different levels of uncertainties.
Table 3. DRPs's total daily profits under different environments.

<table>
<thead>
<tr>
<th>Environment</th>
<th>The Proposed Approach</th>
<th>The Optimal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deterministic</td>
<td>$407.28</td>
<td>$427.57</td>
</tr>
<tr>
<td>Stochastic</td>
<td>$370.24</td>
<td>$405.42</td>
</tr>
</tbody>
</table>

These findings highlight the financial viability of the proposed approach even in the face of uncertainties. Despite the inherent variability in the stochastic environment, the DRP can still achieve substantial profits, with the proposed approach capturing a significant portion of the optimal profits. This analysis provides valuable insights into the economic implications and feasibility of implementing the proposed approach in real-world scenarios.

6. Conclusions

In this paper, we have presented a DRL-based approach for maximizing the profits of DRPs in the presence of unknown end user preferences and multiple sources of uncertainties. Our approach incorporates an integrated demand response model to determine the optimal strategies and employs a policy neural network to learn the optimal pricing strategy from the perspective of DRPs.

Through comprehensive performance evaluations under various uncertainty scenarios, including fluctuations in the PV output and electricity prices, we have demonstrated the effectiveness of the policy neural network in regulating end users’ demands. Remarkably, our approach nears the optimal curve computed by the bi-level model, requiring no prior information. Additionally, we have assessed the generalization ability of our approach, observing a strong performance across different levels of uncertainty. The findings of our study highlight the promising potential of the proposed DRL-based approach in enhancing the efficiency and effectiveness of demand response in real-world settings. By considering diverse uncertainties and incorporating end user preferences, this approach enables utilities and grid operators to better manage energy demand and supply, thereby contributing to the stability and sustainability of energy systems.

Looking ahead, future research can focus on further refining the DRL-based approach by exploring additional uncertainty factors and incorporating more complex multi-energy system dynamics. Additionally, investigations into scalability and deployment considerations in real-world scenarios and a competitive environment would be valuable for practical implementation.

Overall, our research contributes to advancing the field of demand response by providing a robust framework that addresses the key challenges and offers a promising avenue for optimizing energy management in dynamic and uncertain environments.

Author Contributions: Conceptualization, C.X. and Y.H.; methodology, C.X.; validation, C.X.; writing—original draft preparation, C.X.; writing—review and editing, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.