Model-Free Reinforcement Learning Economic Dispatch Algorithms for Price-Based Residential Demand Response Management System

Jun Li, Huaqing Li, *Senior Member, IEEE*, Tingwen Huang, *Fellow, IEEE*, Lifeng Zheng, Lianghao Ji, and Shen Yin, *Fellow, IEEE*

***Abstract*—It is projected that plug-in electric vehicles (PEVs) would steadily increase as household appliances. However, PEVs' high power consumption, stochastic usage patterns, and storage capacity will surely result in a rise in the elasticity of demand response and pose significant difficulties for price-based residential demand response management (PRDRM). This artcle aims to optimize a two-tier globally shared nonconvex PRDRM problem with local constraints and PEVs, known as social welfare: maximizing retailer profits and minimizing the combined residential costs. This is done by balancing residential electricity use with retail electricity prices in an unknown market environment. The proposed online/offline model-free reinforcement learning-based economic dispatch (MFRL-ED) methods can adaptively decide on the ideal retail price sequence by integrating the daily residential-retailer behavior model with the agent-environment interaction method, providing a basic MFRL-ED solution for PRDRM without a system identification step and an accurate load-retail model. Experiments show that MFRL-ED methods provide an effective class of PRDRM solutions.**

***Index Terms*—Economic dispatch, model -free reinforcement learning (MFRL), plug -in electric vehicles (PEVs), price-based residential demand response management (PRDRM).**

I. Introduction

ALONG with the rapid development of smart grids as a class of typical cyber-physical systems (CPSs) in the information era, home energy management systems (HEMS) are among the key technologies in the deployment of energy demand response [1]. The main goal of RDRM, which is a key component of HEMS, is to leverage changes in the energy use of loads to react to time-varying tariffs or "reward and penalty" incentives to achieve cost savings or other advantages [2]. However, developing effective RDRM strategies for households is extremely challenging due to stochasticity and elasticity of residential power consumption. Specifically, the timing and frequency of appliance turn-on and turn-off are uncertain and hard to be predicted due to residents' lifestyle routines. The complexity of RDRM is increased when the appliances are further classified as dispatchable and non-dispatchable on account of the transferability of their energy consumption. These make it difficult for RDRM to efficiently plan the timing of power demand in response to dynamic tariffs. In addition, for an efficient load operation, accurate appliance models and parameters need to be determined in time to model the power characteristics and operating dynamics of these appliances. However, expertise is not always available to the average household.

To solve the above-mentioned difficulties regarding RDRM, scholars have proposed a series of economic dispatch (ED) approaches. The earlier RDRM works mainly focus on minimizing the household's electricity cost. For example, [3] and [4] combine mixed-integer linear programming models with the demand response of appliances to reduce daily household energy consumption, but elasticity of appliance usage and dynamic electricity prices are not considered. Then, [5] proposes a robust optimization method to minimize the worst-case daily bill payment by considering the uncertainty of consumer behavior. To ensure the probabilistic satisfaction of appliance operating constraints, an opportunity constrained optimization model is developed in [6]. A Lyapunov optimization algorithm is applied to loads with heating, ventilation and air conditioning (HVAC) in [7]. Currently, there are two branches of RDRM: price-based RDRM (PRDRM) encourages loads to adjust their energy consumption according to time-based pricing mechanisms with common strategies such as real-time pricing [8] and time-of-use pricing (TOU) [9], while incentive-based RDRM (IB-RDRM) [10] provides incentives/penalties for loads to contribute/fail to reduce demand during peak periods [11]. PRDRM is more in line with residential electricity consumption habits and has been widely adopted in many countries [12], thus this article concentrates on PRDRM.

Manuscript received 25 January 2023; revised 10 June 2023; accepted 3 July 2023. Date of publication 12 July 2023; date of current version 29 August 2023. This work was supported by the National Natural Science Foundation of China under Grant 62173278. *(Corresponding author: Huaqing Li.)*

Jun Li, Huaqing Li, and Lifeng Zheng are with the Chongqing Key Laboratory of Nonlinear Circuits and Intelligent Information Processing, College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China (e-mail: jun_li2023@163.com; huaqingli@swu.edu.cn; zlf_swu@163.com).

Tingwen Huang is with Science Program, Texas A&M University at Qatar, Doha , Qatar (e-mail: tingwen.huang@qatar.tamu.edu).

Lianghao Ji is with the Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: lianghao.ji@gmail.com).

Shen Yin is with the Department of Mechanical and Industrial Engineering, Faculty of Engineering, Norwegian University of Science and Technology, 7033 Trondheim, Norway (e-mail: shen.yin@ntnu.no).

Digital Object Identifier 10.1109/TICPS.2023.3294874

From the perspective of benefits, the current ED research on PRDRM is divided into three parts. For the individual benefits, there is a preference for reducing electricity costs or other benefits to customers by choosing an appropriate pricing mechanism. For example, the work in [13] investigates the distributed generation scheduling problem considering the uncertainty of renewable energy sources and the different personality types of consumers. For the interest of the energy company, maximizing the company's benefits or minimizing the cost of generation is the pursuit [14]. And to meet the reasonable demands of social development, it becomes a trend to integrate the relative interests of the both, so maximizing social benefits becomes a new research hotspot. A distributed dual decomposition-based (DDB) approach [15] and its fast version [16] are proposed to maximize social welfare. From the perspective of more diverse energy options and management, the optimization and control of HVAC systems are considered in [17], [18]. However, there remain unsolved problems in the aforementioned efforts. 1) Requiring system identification steps, i.e., explicitly optimizing the model, predictor, and solver. Developing a model-based demand response strategy requires constructing a model and identifying parameters, and performance may degrade due to the inaccuracy of models. 2) The existing PRDRM works rely heavily on deterministic pricing models (e.g., TOU, real-time pricing) that do not reflect the uncertainty and flexibility of dynamic electricity markets. 3) The short-sightedness of the grid leads to a focus on the immediate response of loads to the current pricing strategy and an inability to predict the impact of all subsequent responses. Therefore, it is of significance to develop an approach based on the unknown residential environment model to solve the PRDRM problem in smart grids.

Deep reinforcement learning (DRL) has been widely used in the industry in recent years. It can overcome the above-mentioned problems by exploiting the end-to-end learning capability of neural networks (NN) and has achieved remarkable success in many complex decision-making applications such as distributed economic dispatch (DED) in smart grids [19]. As one of the energy scheduling problems, such model-free reinforcement learning (MFRL) algorithms have inspired researchers to investigate DRL-based PRDRM [20], [21]. In [22], a group smart home energy management scheme to minimize the energy cost and thermal discomfort of users is developed. Ref. [23] proposes a deep Q-network (DQN)-based demand response scheduling method for indoor air temperature control and thermal comfort management. The authors in [24] and [25] develop a DQN-based approach to optimize the charging scheduling of electric vehicles (EVs) in smart homes to minimize charging costs. In [26], an online building energy optimization method for scheduling timescales and time-shifted loads using the DQN method is proposed. As can be observed, the $Q-$framework is used by the majority of the aforementioned DRL-ED efforts to address demand response issues. However, due to the overestimation and the curse of dimensionality brought on by the structural defects of the $Q-$learning framework, they lack further modeling extensions and cross-sectional comparisons of convergence and performance of DRL.

We utilize MFRL-ED to plan the best retail price (RPs) in an ambiguous electricity market setting, drawing inspiration from the use of RL in energy scheduling. The contributions are summarized as follows.

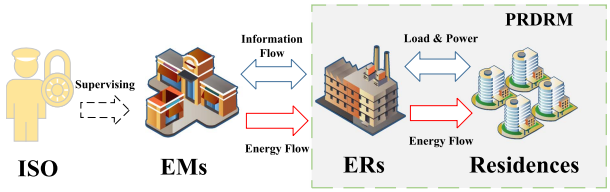

Fig. 1. Basic framework of RDRM.

1) Despite the highly stochastic usage patterns of PEVs, we effectively solve the optimization issue of globally shared nonconvex PRDRM with local constraints and nonsmooth terms. Additionally, in contrast to previous researches [24], [25] that just take into account charging, the load model with charging and discharging characteristics and complicated restrictions is well integrated in our PRDRM problem, and a detailed workflow of the DRL-ED-based charging and discharging mechanism is proposed.
2) The $Q-$learning framework is used for the majority of the existing researches of MFRL-ED algorithms [19], [20], [23], [24], [25], their overestimation is likely to lead to local solutions that are unable to resolve the dimensional disaster problem in continuous RP intervals. In order to address the aforementioned issues and provide the underlying MFRL-ED framework for the PRDRM problem, we propose the offline double DQN-based ED (DDQN-ED) algorithm and the online Actor-Critic-based ED (AC-ED) algorithm, respectively, and compare the superiority and inferiority among them. Simulation experiments verify their effectiveness and scenario applicability.
3) This study unifies both interests, in contrast to the RDRM problem in [13], [14], which exclusively focuses on business or personal interests. Taking into account a large number of residences sharing a single electricity retailer, we build a two-level optimization model and focus on only the social welfare goal through relative social value weighting, i.e., seeking a balance between the business interests of energy retailers and the individual benefits of users, which is consistent with the ideal pursuit of a harmonious economic community.

The rest of this article is organized as follows. Section II formulates the PRDRM problem. Section III presents a series of MFRL algorithms and the coupling model with PRDRM. Section IV demonstrates the effectiveness and advancement of the designed algorithms through simulation experiments.

II. Formulation of PRDRM

The fundamental framework of RDRM is depicted in Fig. 1, which features residences, electricity retailers (ERs), energy markets (EMs), and independent system operators (ISO). While EMs supply ERs with wholesale power, ERs are in charge of supplying retail electricity to homes in specific regions. ISO oversees the market's commercial functioning. We explore the PRDRM issue between retailers and residents by coordinating

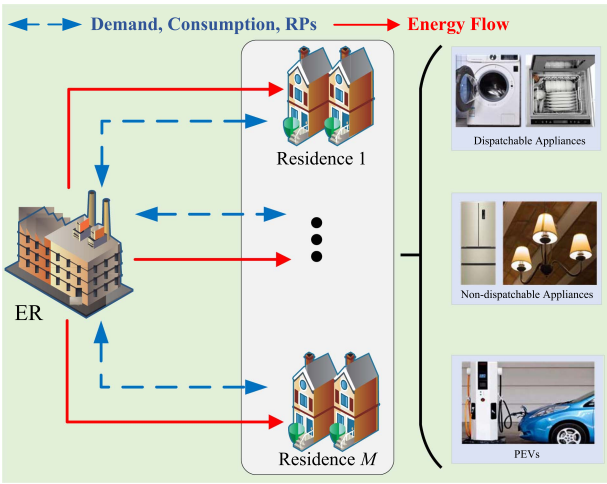

Fig. 2. Residences-retailer electricity transaction model for PRDRM.

RPs with residential daily electricity consumption using a variety of MFRL techniques, supposing that there is only one-way energy transfer between ERs and EMs.

A. Modeling of Residential Appliances

According to the transferable and abridged characteristics of the energy consumed by the loads, residential appliances can generally be divided into dispatchable (e.g., washing machines, dishwashers, etc.) and non-dispatchable ones (e.g., electric lights, refrigerators, etc.) [27]. As a class of dispatchable appliances with charging and discharging characteristics, PEVs are specifically considered in the PRDRM problem.

Fig. 2 depicts a model of power transmission and communication between the residences and the electricity retailer. The red solid lines indicate electrical wiring infrastructure, the retailer delivers electricity to a region of residences. PEV charging is managed by a smart but simple device installed in the user's home. The blue dotted line indicates the underlying communication and information system, a two-way information flow exists between the retailer and the residences. The ER receives the actual energy consumption of residents in the previous time slot and the expected energy demand in the current time slot, and then dynamically adjusts the RP strategy with business interests in mind, while customers actively change their energy consumption in line with the change in RP in the current time slot. Therefore, it can be assumed that the set of dispatchable appliances is $\mathcal{N}_d = \{1, \ldots, D\}$ (excluding PEVs), the set of non-dispatchable appliances is $\mathcal{N}_n = \{1, \ldots, N\}$, the set of PEVs is $\mathcal{N}_p = \{1, \ldots, P\}$, and the set of all appliances can be expressed as $\mathbb{N} = \mathcal{N}_d \cup \mathcal{N}_n \cup \mathcal{N}_p = \{1, \ldots, \mathbf{N}\}$.

Remark 1: Considering a region with the same electricity retailer, the star topology is naturally applied. However, it is more realistic that residences in different areas are free to choose multiple retailers at the same time, so PRDRM based on a network structure with multiple retailers and multiple residential areas needs to be developed using the distributed MFRL-ED technique, which is our later effort. And the set consisting of retailers should satisfy the minimum point coverage.

1) Dispatchable Appliances: Motivated by [29], the actual energy consumption of the dispatchable appliance $d \in \mathcal{N}_d$ is formulated as

$$E_{d,t} = R_{d,t}\left(1 + \delta_t\left(\frac{\rho_{d,t}}{\theta_t} - 1\right)\right) \tag{1}$$

where $\mathbb{T} = \{t | t = 1, \ldots, T\}$, T indicates the total time slots, $R_{d,t}$ (kWh) is the expected energy demand, $E_{d,t}$ (kWh) is the actual electricity consumption, $\rho_{d,t}$ $(\$/kWh)$ is the RP of electricity for the retailer's decision, and θ_t $(\$/kWh)$ is the wholesale price (WP) bought by the ER from EMs. Considering the profit of ER, there is $\rho_{d,t} \geq \theta_t$. $\delta_t < 0$ is the price elasticity coefficient, which shows the interrelationship between energy demand and RP.

In this day-based electricity trading model, the user sends signals $(R_{d,t}, E_{d,t})$ to the ER at time slot t, and the retailer gets the corresponding profit estimate from the feedback signals and makes an adjustment decision about the RP $\rho_{d,t+1}$. The essence of (1) shows that $R_{d,t}$ meets the maximum consumption of the residence, and that customers are willing to consume more electricity when WP θ_t is close to RP $\rho_{d,t}$, but they cannot accept the high retail price and thus consume less, this is in accordance with human intuitive thinking. Therefore, the demand error $(R_{d,t} - E_{d,t})$ can represent the happiness index of residents' electricity consumption. We denote this happiness characteristic by a quadratic function as follows:

$$C_{d,t} = \frac{1}{2}h_{d1}(R_{d,t} - E_{d,t})^2 + h_{d2}(R_{d,t} - E_{d,t}) \tag{2}$$

where h_{d1} $(\$/kWh^2)$ and h_{d2} $(\$/kWh)$ are happiness coefficients related to the appliance, respectively. The electric happiness error function $C_{d,t}$ describes a quadratic error between the actual energy consumption $E_{d,t}$ and expected energy demand $R_{d,t}$. Therefore, the closer $E_{d,t}$ is to $R_{d,t}$ (i.e., the smaller the error $C_{d,t}$) under a reasonable RP at time slot t, the higher the happiness of the residents.

However, when residents are limited in their incentive to consume energy by high RP, their happiness becomes lower. Then, the available range of the electrical happiness error function is constrained to

$$(R_{d,t} - E_{d,t}) \in [DE_d^{\min}, DE_d^{\max}] \tag{3}$$

where $DE_d^{\min}$ and $DE_d^{\max}$ are the minimum and maximum demand error bounds, respectively.

2) Non-Dispatchable Appliances: Non-dispatchable appliances satisfy the identity relation between energy demand $R_{n,t}$ and actual consumption $E_{n,t}$ in all time slots, i.e.,

$$R_{n,t} = E_{n,t} \tag{4}$$

3) PEVs: PEVs have all the characteristics of dispatchable appliances, and the actual energy consumption $E_{p,t}, p \in \mathcal{N}_p$ can be formulated as

$$E_{p,t} = R_{p,t}\left(1 + \delta_t\left(\frac{\rho_{p,t}}{\theta_t} - 1\right)\right),\ t \in \mathbb{T} \tag{5}$$

Note that $E_{p,t} < 0$ means discharging, while $E_{p,t} > 0$ denotes charging. Then the electrical happiness error function $C_{p,t}$ of

PEVs can be expressed as

$$C_{p,t} = \frac{1}{2}h_{p1}(R_{p,t} - E_{p,t})^2 + h_{p2}(R_{p,t} - E_{p,t}) \tag{6}$$

where the parameters are interpreted similarly to (1). $C_{p,t}$ indicates that the ERs cannot fully satisfy the PEV owners' willingness to charge. In addition, vehicles with on-board batteries require the corresponding limit of rated power for safety at every time slot, which satisfy that

$$E_{p,t} \in [-E_p^r, E_p^r] \tag{7}$$

where E_p^r denotes the rated power of the PEV p. Considering the charging/discharging characteristics of PEVs and the battery capacity, PEV p is subject to the following limitation:

$$B_p^{\min} \leqslant E_{p,t}^{SOC} \leqslant B_p^{\max} \tag{8}$$

where $E_{p,t}^{SOC} = E_p^0 + \sum_{\sigma=1}^{t} e_p E_{p,\sigma}$ denotes the state-of-charge (SOC) of PEVs, $B_p^{\min}$ and $B_p^{\max}$ represent the minimum and maximum electric quantity, respectively. E_p^0 denotes the initial energy level. e_p indicates the charging or discharging efficiency, which is associated with $E_{p,t}$. If $E_{p,t} > 0$, $e_p = e_p^c$, else $e_p = e_p^d$. We also consider the effect of frequent charging/discharging on the battery life and therefore quantify the cost of battery degradation [30]:

$$Deg_{p,t} = \upsilon \left| E_{p,t} \right| \tag{9}$$

where υ ($\$/kWh$) is the degradation coefficient. Due to ISO regulation, the RPs of electricity are subject to

$$\rho^{\min} \leq \rho_{n,t}, \rho_{p,t}, \rho_{d,t} \leq \rho^{\max} \tag{10}$$

with $n \in N_n, p \in N_p, d \in N_d$, where $\rho^{\min}$ and $\rho^{\max}$ are the bounds of RP.

Remark 2: Significant progress has been made in research on the monitoring and measurement of household load classifications [31]. For non-invasive methods, the load classification metering is based on the installation of meters at the entrance of the power line, so that the total characteristics of the appliances in the circuit can be collected and analyzed, and then the classes of appliances can be identified through complex processing methods such as signal processing, pattern recognition, and artificial neural networks, and then the classification metering of each type of load can be implemented [32], [33]. With the increasing maturity of smart grid technology and users' awareness of energy conservation, the implementation of the classification monitoring and metering of household appliances with different load types is promising to be popularized in the future.

B. Social Welfare

PRDRM is a two-level optimization problem in this article, where a minimized residential integrated cost to obtain the optimal actual energy consumption is expected from the user's point of view. From the perspective of the ER, maximizing the company's profit is the core business objective. These two wishes are considered together to show the relative social value of commercial profit and integrated cost of customers by maximizing social welfare [34].

The optimization problem of minimizing the comprehensive electricity cost of residents can be expressed as

$$\begin{aligned}\mathbf{EC} &= \min_{\mathbf{E}_t} \sum_{t=1}^{T} EC_t \\ &= \min_{\mathbf{E}_t} \sum_{t=1}^{T} \left[\sum_{n \in \mathcal{N}_n} E_{n,t}\rho_{n,t} + \sum_{d \in \mathcal{N}_d} (E_{d,t}\rho_{d,t} + C_{d,t}) \right. \\ &\quad \left. + \sum_{p \in \mathcal{N}_p} (E_{p,t}\rho_{p,t} + C_{p,t} + Deg_{p,t}) \right]\end{aligned} \tag{11}$$

where EC_t (\$) indicates the integrated cost at time slot t, and the energy consumption vector $\mathbf{E}_t$ contains all dispatchable, non-dispatchable appliances and PEVs.

For the benefit of the retailer, its profit maximization problem can be formulated as

$$\begin{aligned}\mathbf{EP} &= \max_{\mathcal{P}} \sum_{t=1}^{T} EP_t \\ &= \max_{\mathcal{P}} \sum_{t=1}^{T} \left[\left(\sum_{n \in N_n} E_{n,t}(\rho_{n,t} - \theta_t) \right) + \left(\sum_{d \in N_d} E_{d,t}(\rho_{d,t} - \theta_t) \right) \right. \\ &\quad \left. + \left(\sum_{p \in N_p} E_{p,t}(\rho_{p,t} - \theta_t) \right) \right]\end{aligned} \tag{12}$$

where EP_t (\$) represents the retailer profit at time slot t. $\mathcal{P}$ is the vector of electricity RPs consisting of three types of appliances $\{\rho_{n,t}, \rho_{p,t}, \rho_{d,t}\}, n \in \mathcal{N}_n, p \in \mathcal{N}_p, d \in \mathcal{N}_d$.

Based on (1), (4) and (5), the actual energy consumption $\mathbf{E}_t$ can be determined by $\mathcal{P}$, the two-tier optimization problem (11) and (12) can represent social welfare by parameter trade-offs, i.e., the PRDRM nonconvex optimization problem can be written as

$$\begin{aligned} \max \quad & \sum_{t=1}^{T} F_t(\mathcal{P}) = \sum_{t=1}^{T} \omega EP_t - (1-\omega)EC_t \\ \text{s.t.} \quad & (3), (7) - (8), (10) \\ & n, d, p \in \mathbb{N}, t \in \mathbb{T} \end{aligned} \tag{13}$$

where ω is the relative social value weight that balances commercial profits against residential energy consumption. When setting the weight ω to 0, we takes the residents' interests into account completely. And when ω is set to 1, only the interests of ER are considered.

Remark 3: The peak rebound problem (peak may shift to another time of the day [35]) may occur when the penetration of dwellings benefiting from PRDRM is high for a given social value weight ω. The reason behind load accumulation is the intuitive human behavior of using more appliances at the lowest RP [36], [37]. This causes ER to buy more power from EM to manage the load demand, which may not only result in financial loss to ER but also increase power imbalance, power loss in the network and grid instability leading to voltage violations and

overloading of transformers and distribution lines, etc. Therefore, considering the properties of the weight ω, we can bias the weights ω towards ER in time slots with low RP to ensure that the profits obtained by ER are stable, while the intuitive behavior of residents will autonomously reduce the use of more appliances in that time slot.

Problem (13) is a nonconvex optimization problem with globally shared objective function $F_t(\mathcal{P})$, local constraints and nonsmooth terms. However, the associated optimization algorithms, such as forward–backward splitting method [38], require not only regularity assumptions on the objective function, but also an accurate load-retail model. Considering the complex grid environment, this article proposes two types of MFRL-ED algorithms to optimize the objective (13).

III. MFRL-ED Algorithms for PRDRM

In this section, we introduce a class of MFRL-ED algorithms to address problem (13). Instead of concerning the complexity of objective functions and constraints with the construction of appliance models and accurate parameter predictions, the proposed online/offline RL algorithms exploit the transmission data from the grid to make RP-decisions based on policy exploration. The basic RL element consists of a five-tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}_t, \gamma \rangle$, corresponding to the retailer-resident electricity trading model as follows:

1) *State* ($\mathcal{S} = \{s_{i,1}, \ldots, s_{i,T}\}$): The energy demand $R_{i,t}$ and the actual electricity consumption $E_{i,t-1}$, $i \in \mathbb{N}$;
2) *Action* ($\mathcal{A} = \{a_{i,1}, \ldots, a_{i,\mathcal{M}}\}$): RPs $\rho_{i,t}, i \in \mathbb{N}$ determined by electricity retailer, $\mathcal{M}$ denotes the number of set after discretizing the RP interval $[\rho^{\min}, \rho^{\max}]$;
3) *Reward* ($\mathcal{R} = \{r_{i,1}, \ldots, r_{i,T}\}$): The social welfare $F_t(\mathcal{P})$;
4) *State transition function* ($\mathcal{T}_{i,t}$): The formulations (1), (4) and (5) related to RPs;
5) *Discount factor* ($\gamma \in [0, 1]$): The importance weight for future social welfare, associated with the convergence of cumulative rewards.

The historical observations $\{(s_{i,t-1}, a_{i,t-1}, r_{i,t-1}, s_{i,t})\}$ with $t = 1, \ldots, T$ are obtained from each episode, where $(s_{i,t-1}, a_{i,t-1}, r_{i,t-1}, s_{i,t})$ is called a transition. Additionally, there ought to be a greatest lower bound for the difference $\Delta\rho_{i,t}$ in actual transactions, which is another reason for discretizing the RP interval. We advise the readers to refer to [28] for the remaining concepts.

A. Q-Table-Based ED

The $Q-$Table-Based ED algorithm for the PRDRM is given in [29]. In general, the $Q-$table stores the $q-$values mapped by $\rho_{i,t}$ and $(R_{i,t}, E_{i,t-1})$. Its horizontal row represents a finite number of RPs generated after discretizing in the RP interval $[\rho^{\min}, \rho^{\max}]$, and the vertical row represents the time slots $\{1, \ldots, T\}$.

Remark 4: $Q-$table is not the only potential one for PRDRM, for example, the representation of the state space can be uniquely determined by the RPs $\{\rho_{d,t}, \rho_{p,t}, \rho_{n,t}\}$ and the energy demand $\{R_{d,t}, R_{p,t}, R_{n,t}\}$ instead of the time slots.

Based on the Bellman equation and the greedy strategy, the update formula for $q-$values can be expressed as

$$Q^k_{s_{i,t}a_{i,t}} = Q^{k-1}_{s_{i,t}a_{i,t}} + lr \cdot (F_t(\mathcal{P}) + \gamma \max_{a_{i,t+1}} Q^{k-1}_{s_{i,t+1}a_{i,t+1}} - Q^{k-1}_{s_{i,t}a_{i,t}}) \tag{14}$$

with $a_{i,t} \in \mathcal{A}$, $s_{i,t} \in \mathcal{S}$, $i \in \mathbb{N}$, where k denotes the episode index, and lr indicates the learning rate. When the $Q-$table converges, the optimal RP sequence can be obtained using the following target policy:

$$\pi^*_{s_{i,t}} = \arg\max_{a_{i,t}} Q^*_{s_{i,t}a_{i,t}} \tag{15}$$

where $\pi^*_{s_{i,t}}$ is also called greedy strategy. This $Q-$learning ED algorithm requires the creation of corresponding $Q-$table for each appliance in advance, and the growth in the number of appliances, discrete RPs and time slots all impose a significant burden on the storage and computation. Hence, $Q-$Table-Based ED algorithm has to take into account the appropriate discrete action space and residential area scale when solving the RDRM problem.

B. DQN-ED

We substitute the $Q-$value function approximation for the $Q-$table in order to address the dimensionality issue brought on by too many appliances and continuous RP intervals, which is usually constructed using deep neural networks (DNNs), i.e.,

$$\hat{Q}_{s_{i,t}a_{i,t}\alpha_{i,t}} \approx Q_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} \tag{16}$$

where $\alpha_{i,t}$ denotes the weight of the ith $Q-$network, $s_{i,t} = (R_{i,t}, E_{i,t-1})$, $a_{i,t} = \rho_{i,t}$, $i \in \mathbb{N}$, $t \in \mathbb{T}$. We simply need to concentrate on the convergence of the weights $\alpha_{i,t}$, much as the convergence of $q-$value in the $Q-$table. Ref. [39] uses the simplified DNN with three-layers to represent the linear approximation of the $q-$function. The $q-$function is an evaluation function for the pair $(R_{i,t}, E_{i,t-1})$ and $\rho_{i,t}$ based on the prediction of future social welfare $F_t(\mathcal{P})$. Then the target $q-$value can be indicated as

$$Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} = F_t(\mathcal{P}) + \gamma \max \hat{Q}^k_{s_{i,t+1}a_{i,t+1}\alpha^k_{i,t}} \tag{17}$$

The quadratic loss function of the $Q-$network can be expressed as

$$QL^k_{i,t} = \frac{1}{2}\left(Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} - \hat{Q}^k_{s_{i,t}a_{i,t}\alpha^k_{i,t}}\right)^2 \tag{18}$$

Thus the weight $\alpha_{i,t}$ can be updated iteratively using the gradient descent method. Based on (17)–(18), we present the online DQN-ED algorithm for PRDRM shown in Algorithm 1. The historical observation $\{\rho_{i,t}, F_t(\mathcal{P}), R_{i,t}, E_{i,t-1}\}$ after exploring is directly exploited by the $Q-$network, in which data correlation easily leads to obtaining local solutions. In addition, the max operation, although it can quickly bring the $q-$values closer to the possible optimization objectives, can easily be overdone and lead to overestimation problems.

Here, DDQN-ED is employed to achieve the elimination of the overestimation problem by decoupling the two steps of RP selection and calculation of target $q-$value. The social

Algorithm 1: Online DQN-ED for PRDRM.

Input: Learning rates lr, discount factor γ, maximum episodes index K, total time slots T, convergence threshold ξ.

Output: Convergent RP sequence $\{\rho^*_{i,1}, \rho^*_{i,2}, \ldots, \rho^*_{i,T}\}$, $i \in \mathbb{N}$.

1: Initialize: $Q-$network's weight $\alpha^0_{i,t} = rand(\cdot)$, time slot $t = 0$, episodes index $k = 0$.
2: **for all** $k = 1 : K$ or $|Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} - \hat{Q}^k_{s_{i,t}a_{i,t}\alpha^k_{i,t}}| > \xi$ **do**
3: **for all** $t = 1 : T$ **do**
4: Choose RPs $\{\rho_{i,1}, \ldots, \rho_{i,T}\}$ by using $\varepsilon-$greedy strategy as the behavioral policy;
5: Calculate the actual electricity consumption $E_{i,t}$ by $\mathcal{T}_{i,t}$;
6: Identify the estimated $q-$value $\hat{Q}^k_{s_{i,t}a_{i,t}\alpha^k_{i,t}}$;
7: Calculate social welfare $F_t(\mathcal{P})$ by (13);
8: **if** $t < T$ **then**
9: Calculate the current $q-$value $Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}}$ by (17);
10: **else**
11: Identify the current $q-$value $Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} = F_t(\mathcal{P})$;
12: **end if**
13: **end for**
14: Update the weight $\alpha^k_{i,t}$ by the gradient descent: $\alpha^{k+1}_{i,t} = \alpha^k_{i,t} - lr \cdot \nabla_{\alpha^k_{i,t}} QL^k_{i,t}$;
15: **end for**
16: Obtain the convergent RP sequence $\{\rho^*_{i,1}, \rho^*_{i,2}, \ldots, \rho^*_{i,T}\}$ by (15).

welfare expectation of DDQN is demonstrated to be unbiased estimation [40]. In the following, we will briefly construct the coupling of DDQN-ED with PRDRM.

DDQN-ED adopts two identical $Q-$network structures for each appliance: the current $Q-$network $\alpha^k_{i,t}$ and the target $Q-$network $\tilde{\alpha}^k_{i,t}$, which are responsible for the RP-decision and the RP estimation based on $q-$function, respectively. Based on (17), the target $q-$value can be expressed as

$$\begin{cases} Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} = F_t(\mathcal{P}) + \gamma \max \hat{Q}t_{s_{i,t+1},a'_{i,t+1},\tilde{\alpha}^k_{i,t}} \\ a'_{i,t+1} = \text{argmax}\hat{Q}c_{s_{i,t+1},a_{i,t+1},\alpha^k_{i,t}} \end{cases} \tag{19}$$

where $\hat{Q}t_{s_{i,t},a'_{i,t+1},\tilde{\alpha}^k_{i,t}}$ denotes the $q-$value of the target $Q-$network, and $\hat{Q}c_{s_{i,t+1},a_{t+1},\alpha^k_{i,t}}$ represents the estimated $q-$value of the current $Q-$network. Then the loss function of DDQN can be represented as

$$\begin{aligned} QL^k_{i,t} &= \frac{1}{2}\left|Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} - \hat{Q}c^k_{s_{i,t}a_{i,t}\alpha^k_{i,t}}\right|^2 \\ &= \frac{1}{2}\left|F_t(\mathcal{P}) + \gamma \max \hat{Q}t_{s_{i,t},a'_{i,t+1},\tilde{\alpha}^k_{i,t}} - \hat{Q}c_{s_{i,t+1},a_{i,t+1},\alpha^k_{i,t}}\right|^2 \end{aligned} \tag{20}$$

Using (19) and (20), the offline DDQN-ED algorithm with experience replay for PRDRM is presented in Algorithm 2.

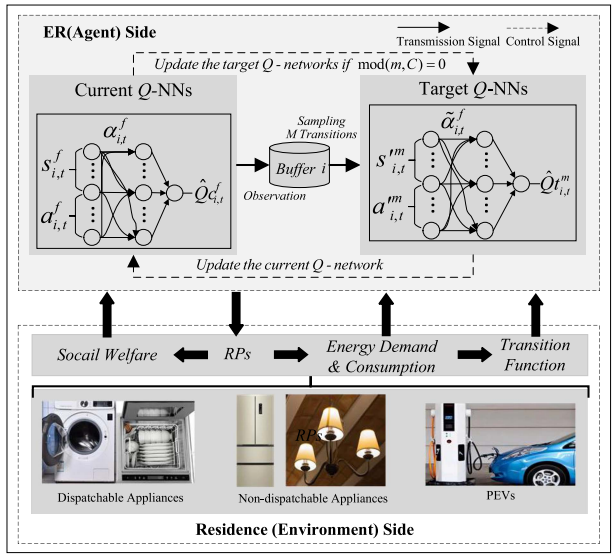

Fig. 3. Block diagram of DDQN with experience replay for PRDRM.

It can be seen that the ER interacts with the residences for storing the transitions in the experience replay buffer $\mathcal{D}$ with maximum buffer F based on the current $Q-$network $\alpha^k_{i,t}$. Then M transitions are sampled and removed from $\mathcal{D}$ to train the current $Q-$network, while the target $Q-$network weight $\tilde{\alpha}^k_{i,t}$ is updated at the fixed update period C. This delayed update reduces the parameter dependency between the target $Q-$network and the current $Q-$network. Therefore, DDQN-ED achieves the elimination of the overestimation problem by decoupling the two steps of the selection of the current action and the calculation of the target $q-$value. The detailed structures in Algorithm 2, such as the replay buffer and network models, are shown in Fig. 3. The experience replay buffer is $\mathcal{D} = \{\mathfrak{T}^f_i\}, f = 1, \ldots, F$ where the fth transition $\mathfrak{T}^f_i$ can be denoted as

$$\mathfrak{T}^f_i = \begin{pmatrix} s^f_{i,0} & \cdots & s^f_{i,t} & \cdots & s^f_{i,T-1} \\ a^f_{i,0} & \cdots & a^f_{i,t} & \cdots & a^f_{i,T-1} \\ F^f_{i,0} & \cdots & F^f_{i,t} & \cdots & F^f_{i,T-1} \\ s^f_{i,1} & \cdots & s^f_{i,t+1} & \cdots & s^f_{i,T} \end{pmatrix}$$

Noting that, when the capacity of replay buffer $\mathcal{D}$ is set to 1, Algorithm 2 changes to an online DDQN-ED. Moreover, we should make sure that each appliance has the same maximum buffer capacity F and update period C in order to maintain the algorithm iterating synchronously.

Remark 5: Since the target $q-$value is calculated through the target $Q-$network, the RP that maximizes the target $q-$value is originally selected according to the parameters of the target $Q-$network $Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}}$ in DQN, while the $q-$value calculated after selecting the optimal RP with the current $Q-$network $\hat{Q}c_{s_{i,t+1},a_{t+1},\alpha^k_{i,t}}$ of DDQN must be less than or equal to the original $q-$value. This approach reduces the overestimation to a certain extent and makes the $q-$value closer to the real value.

Remark 6: The iterative formulation (14) is derived from the Bellman equation and the greedy strategy. For the sake of clarity, we assume that the $q-$value is optimal, then we have the ideal

Algorithm 2: Offline DDQN-ED With Experience Replay for PRDRM.

Input: The predefined parameters similar to Algorithm 1 for the target $Q-$networks and the current $Q-$networks.

Output: Convergent RP sequence $\{\rho^*_{i,1}, \rho^*_{i,2}, \ldots, \rho^*_{i,T}\}$, $i \in \mathbb{N}$.

Initialize: The current $Q-$network's weight $\alpha^0_{i,t} = rand(\cdot)$, the target $Q-$network's weight $\tilde{\alpha}^0_{i,t} = \alpha^0_{i,t}$, replay buffer $\mathcal{D}$, maximum buffer capacity F, update period C, time slot $t = 0$, episodes index $k = 0$.
for all $k = 1 : K$ or $|Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} - \hat{Q}c^k_{s_{i,t}a_{i,t}\alpha^k_{i,t}}| > \xi$ **do**
  for all $f = 1 : F$ **do**
    for all $t = 1 : T$ **do**
      Identify the estimated $q-$value $\hat{Q}c^k_{s_{i,t}a_{i,t}\alpha^k_{i,t}}$;
      Choose the best RPs $\{\rho_{i,1}, \ldots, \rho_{i,T}\}$ by (15) with probability $1 - \varepsilon$, and select the random RPs with probability ε;
      Calculate the actual electricity consumption $E_{i,t}$ by $\mathcal{T}_{i,t}$;
      Calculate social welfare $F_t(\mathcal{P})$ by (13);
      Identify the energy demand $R_{i,t+1}$;
    end for
    Store the f-th transition into the experience replay buffer $\mathcal{D}$;
  end for
  Sample and remove $M(M \leq F)$ transitions from the replay buffer $\mathcal{D}$;
  for all $m = 1 : M$ **do**
    for all $t = 1 : T$ **do**
      if $t < T$ **then**
        Calculate the optimal action and the target $q-$value $Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}}$ by (19);
      else
        Identify $Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} = F_t(\mathcal{P})$;
      end if
      Update the current $Q-$network's weight: $\alpha^{k+1}_{i,t} = \alpha^k_{i,t} - lr \cdot \nabla_{\alpha^k_{i,t}} QL^k_{i,t}$;
    end for
    if $mod(m, C) = 0$ **then**
      Update the target $Q-$network's weight $\tilde{\alpha}^k_{i,t} = \alpha^k_{i,t}$;
    end if
  end for
end for
Obtain the convergent RP sequence $\{\rho^*_{i,1}, \rho^*_{i,2}, \ldots, \rho^*_{i,T}\}$ by (15).

optimal $q-$function $Q^*_{s_t,a_t} = \max_\pi E(r_t|s_t, a_t, \pi)$, where $E(\cdot)$ denotes the expectation. However, the optimal $q-$function should satisfy the Bellman equation: $Q^*_{s_t,a_t} = E_\pi(F_t(\mathcal{P}) + \gamma \max_{a_{t+1}} Q^{k-1}_{s_{t+1}a_{t+1}}|s_t, a_t)$. Hence, the overestimation problem can be attributed simply to the inequality: $E(\max(Q_{s_1,a_1}, Q_{s_2,a_2}, \ldots, Q_{s_T,a_T})) \geq \max(E(Q_{s_1,a_1}, Q_{s_2,a_2}, \ldots, Q_{s_T,a_T}))$. The result shows that fitting the $q-$function using the gradient descent yields larger estimated expectation. In summary, the overestimation problem is unavoidable for RL algorithms based on the $Q-$learning framework due to the adoption of greedy strategy to maximize cumulative rewards in unknown environments for maintaining efficient exploratory.

C. AC-ED

DQN-ED algorithms, which are a type of value-based learning algorithm, maximize social welfare expectations by selecting the best RPs based on a deterministic strategy. It is crucial to use stochastic policies for the PRDRM optimization problem to handle the continuous RP space and obtain more precise energy ED. Additionally, when compared to offline DDQN-ED algorithm, ER using online learning algorithms can maximize social welfare by adaptively and in real-time adjusting prices for real-time dispatch consideration. Therefore, this article proposes a policy-value-based learning algorithm for PRDRM, namely the Actor-Critic method, which approximates the policy distribution by adding a policy network to the $Q-$network

$$\pi^{\beta_{i,t}}_{s_{i,t},a_{i,t}} = \mathbf{Pr}(a_{i,t}|s_{i,t}, \beta_{i,t}) \approx \pi_{s_{i,t}}, t \in \mathbb{T} \tag{21}$$

which can be described by the parameter $\beta_{i,t}$, where $\mathbf{Pr}(\cdot)$ indicates the probability distribution. We formulate the policy network as the Actor network and the Critic network corresponds to the $Q-$network. In general, the input of the Actor network is a state vector and the output is an estimated action. The structure of the critic NN remains the same as that of the $Q-$network. Here, the outputs of the Actor-Critic network are represented as

$$\eta_{i,t} = \phi_a(s_{i,t}, \beta_{i,t}) \tag{22}$$

and

$$J_{i,t} = \phi_c(s_{i,t}, a_{i,t}, \alpha_{i,t}) \tag{23}$$

with $i \in \mathbb{N}, t \in \mathbb{T}$, where $\beta_{i,t}$ and $\alpha_{i,t}$ are the Actor NN's weight and Critic NN's weight, respectively, ϕ_a and ϕ_c are the activation functions, the estimated action $\eta_{i,t}$ is the output of the Actor network, and the action-value function $J_{i,t}$ is the output of the Critic network at the time slot t.

The temporal difference (TD) error for each appliance is given by

$$\mathcal{E}_{i,t} = F_t(\mathcal{P}) + \gamma J_{i,t+1} - J_{i,t} \tag{24}$$

The discount factor $\gamma = 0$ implies that a short-sighted learning algorithm focuses on immediate rewards and ignores the future state value. On the other extreme, $\gamma = 1$ indicates that the learning algorithm gives fair weight to the rewards of all time slots. Then the quadratic loss function for the Critic NN is defined as

$$PL_{i,t} = \frac{1}{2}\mathcal{E}^2_{i,t} \tag{25}$$

For the Actor network, we utilize TD error as the evaluation function. Then based on back propagation, the update formulas for Actor-Critic are represented as

$$\alpha^{k+1}_{i,t} = \alpha^k_{i,t} - lc \cdot \nabla PL_{i,t}(J^k_{i,t}) \cdot \nabla J^k_{i,t}(\alpha^k_{i,t}) \tag{26}$$

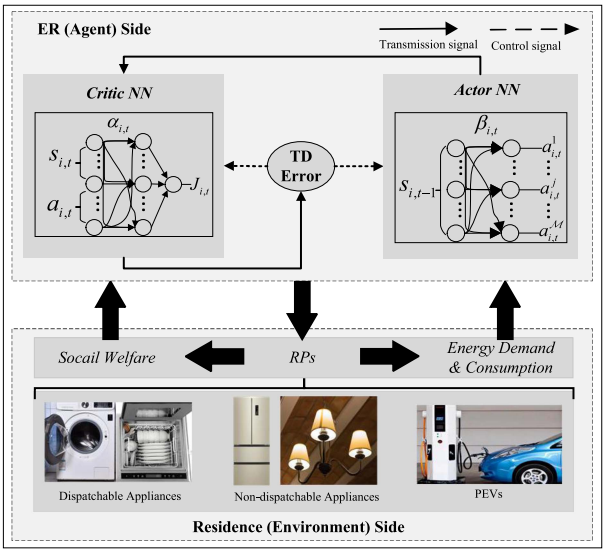

Fig. 4. Block diagram of Actor-Critic for PRDRM.

and

$$\beta_{i,t}^{k+1} = \beta_{i,t}^{k} - la \cdot \nabla \mathcal{E}_{i,t}^{k}(a_{i,t}) \cdot \nabla a_{i,t}(\beta_{i,t}^{k}) \tag{27}$$

where la and lc denote the learning rate of Actor and Critic, respectively. The details of which are illustrated in Fig. 4. Because we consider a discrete set of RPs, the Actor network adopts the softmax function to approximate the optimal policy. For details of the AC-ED algorithm, see Algorithm 3.

Remark 7: Algorithm 3 can be divided into three steps: 1) Calculate the pair $(s_{i,t}, a_{i,t})$ of the current time slot and the next time slot, respectively; 2) Identify the evaluation function $\mathcal{E}_{i,t}$ from the calculated action-value functions $J_{i,t}$ and $J_{i,t+1}$; 3) Update the AC-ED parameters $\alpha_{i,t}$ and $\beta_{i,t}$ using $\mathcal{E}_{i,t}$.

D. Charging-Discharging Mechanism in DRL-ED

The procedure of the battery charging and discharging mechanism is introduced here. The charging and discharging characteristics of the battery characterize the increase or decrease of the cumulative variable $\mathfrak{E}_{p,t}$ presenting the current battery capacity. We use the actual power consumption of each iteration, positive or negative, to represent charging or discharging, while the change in power consumption satisfies the corresponding rated power and capacity constraints (7)–(8). The highly random usage pattern of PEVs means that the SOC of the battery cannot be known in advance of each iteration, which greatly increases the challenge of studying PRDRMs with PEVs. Therefore the literature [24], [25] considers only charging.

In the designed DRL-ED programs, PEVs' RP $\rho_{p,t-1}^{k-1}$ at the ER end, the current power demand $R_{p,t}$ and a stochastic strategy (e.g., ϵ-greedy strategy) jointly decide and drive the PEVs' charging and discharging. The process is shown in Algorithm 4. It should be noted that we ignore the effect of battery degradation on battery capacity during charging and discharging.

Remark 8: It should be noted that PEVs' batteries are only used to supply electricity to households to ensure maximum social welfare (13). Residences can access not only the electricity provided by the ER, but also the PEV batteries to supplement the household electricity. When the households consume too much electricity, they may not necessarily get the maximum social welfare (affected by the user's electrical happiness error function (6)). Instead, supplementing electricity through PEV batteries may acquire higher social welfare.

Algorithm 3: Online AC-ED for PRDRM.

Input: Learning rates lc and la, discount factor γ, maximum episodes index K, total time slots T, convergence threshold ξ.

Output: Convergent RP sequence $\{\rho_{i,1}^{*}, \rho_{i,2}^{*}, \ldots, \rho_{i,T}^{*}\}$, $i \in \mathbb{N}$.

```
Initialize: The Actor network's weight β^0_{i,t} = rand(·), the Critic network's weight α^0_{i,t} = rand(·), time slot t = 0, episodes index k = 0.
for all k = 1 : K or Σ_{t=1}^{T} Σ_{n=1}^{N} |E_{i,t}| > ξ do
  for all t = 1 : T do
    Calculate the actual energy consumption E_{i,t} by T_{i,t};
    Calculate the RPs {ρ_{i,1}, ..., ρ_{i,T}} by (22);
    Obtain the action-value function J_{i,t} by (23);
    Identify the social welfare F_t(P) by (13);
    if t < T then
      Calculate the action-value function J_{i,t+1} by (23);
      Calculate the TD error E_{i,t} by (24);
    else
      Identify the action-value function J_{i,t+1} = 0 by (23);
      Calculate the TD error E_{i,t} by (24);
    end if
  end for
  Update the Critic network's weight α_{i,t} by (26);
  Update the Actor network's weight β_{i,t} by (27);
end for
Obtain the optimal RP sequence {ρ*_{i,1}, ρ*_{i,2}, ..., ρ*_{i,T}}.
```

IV. Simulations and Numerical Analyses

In this section, the effectiveness of the proposed MFRL algorithms is verified by several experiments. The algorithms are implemented by MATLAB R2014a on a desktop PC with i5-12400F CPU@2.50 GHz, 16 GB of RAM, and a 64-bit Windows 11 operating system.

A. Experimental Setup

We consider the energy demand response management problem for 6 dispatchable appliances $\{d1, d2, d3, d4, d5, d6\}$, 4 PEVs $\{p1, p2, p3, p4\}$ and 5 non-dispatchable appliances $\{n1, n2, n3, n4, n5\}$ in a whole day (24 time slots). The energy demand distribution for PEVs, non-dispatchable and dispatchable appliances (see Fig. 5) is referenced from San Diego Gas & Electric [8], and wholesale prices are set by the EM, with

Algorithm 4: Charging-Discharging Mechanism.

Input: RP $\rho_{p,t-1}^{k-1}$, current power demand $R_{p,t}$.
Output: Current charging or discharging status $E_{p,t}$ and PEVs' SOC $E_{p,t}^{SOC}$.

1: Initialize battery parameter: estimated power $\mathfrak{E}_{p,t}$, rated power E_p^r, minimum and maximum electric quantity $B_p^{\min}$ and $B_p^{\max}$, initial energy level E_p^0, charging or discharging efficiency $e_p \in \{e_p^c, e_p^d\}$, ϵ-greedy strategy (if increasing the exploration rate is necessary).
2: Identify RP $\rho_{p,t-1}^{k-1}$ based on deterministic or stochastic strategies.
3: Obtain the actual consumption $E_{p,t}^k$ by (5) at the iteration k.
4: Projecting $E_{p,t}^k$ onto the rated power interval, i.e., $E_{p,t}^k = Proj_{\mathcal{R}_p^r}(E_{p,t}^k), \mathcal{R}_p^r = [-E_p^r, E_p^r]$.
5: **if** $E_{p,t}^k > 0$ **then**
6: Calculate the estimated battery capacity $\mathfrak{E}_{p,t} = E_p^0 + \sum_{\sigma=1}^{t-1} e_p^c E_{p,\sigma} + E_{p,t}, t > 1$;
7: **if** $\mathfrak{E}_{p,t} > B_p^{\max}$ **then**
8: $E_{p,t} = E_{p,t} - (\mathfrak{E}_{p,t} - B_p^{\max})/e_p^c$;
9: $E_{p,t}^{SOC} = B_p^{\max}$;
10: **else**
11: $E_{p,t}^{SOC} = \mathfrak{E}_{p,t}$;
12: **end if**
13: **else**
14: Calculate the estimated battery capacity $\mathfrak{E}_{p,t} = E_p^0 + \sum_{\sigma=1}^{t-1} e_p^d E_{p,\sigma} + E_{p,t}, t > 1$;
15: **if** $\mathfrak{E}_{p,t} < B_p^{\min}$ **then**
16: $E_{p,t} = E_{p,t} - (\mathfrak{E}_{p,t} - B_p^{\min})/e_p^d$;
17: $E_{p,t}^{SOC} = B_p^{\min}$;
18: **else**
19: $E_{p,t}^{SOC} = \mathfrak{E}_{p,t}$;
20: **end if**
21: **end if**
22: Obtain current actual consumption $E_{p,t}$ and current power $\mathfrak{E}_{p,t}$.

data from Commonwealth Edison Company [42]. It can be seen that the energy demand of non-dispatchable appliances is significantly higher than that of dispatchable appliances and PEVs, and that the peak demand occurs in $12:00-16:00$ and $18:00-24:00$. In addition, PEVs have no demand for electricity at certain times of the day due to their stochastic usage patterns that increase energy demand elasticity. In order to ensure the proper functioning of the electricity market economy and to protect the reasonable demands of residents for normal electricity consumption, we must coordinate the retail pricing strategies of electricity retailers with the electricity consumption strategies of customers in an effort to maximize social welfare. The time-varying parameters and appliance-related parameters are listed in Tables I and Table II, respectively. Note that the discrete RPs have a gap of 0.1, and according to the price parameters in Table II, the RP interval is $[2.4, 6.7]$. Therefore the number of discrete actions is 44 (i.e., $\mathcal{M} = 44$).

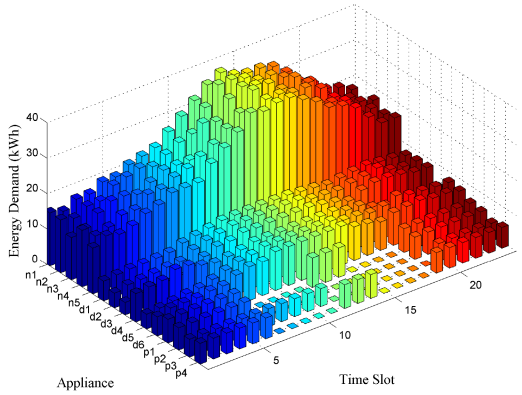

Fig. 5. Energy demands of all appliances over 24 Time slots.

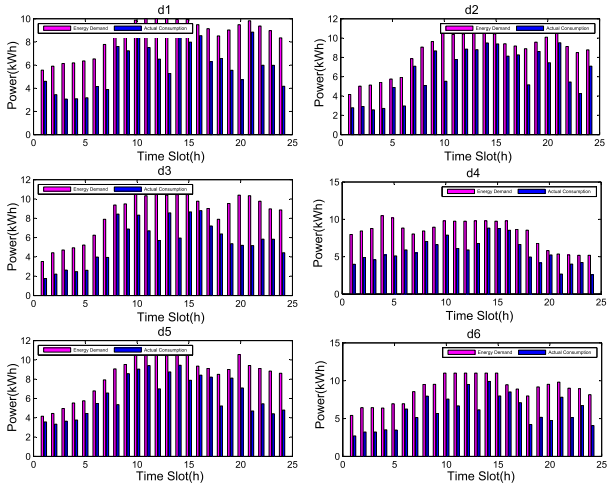

Fig. 6. Energy consumption for dispatchable appliances.

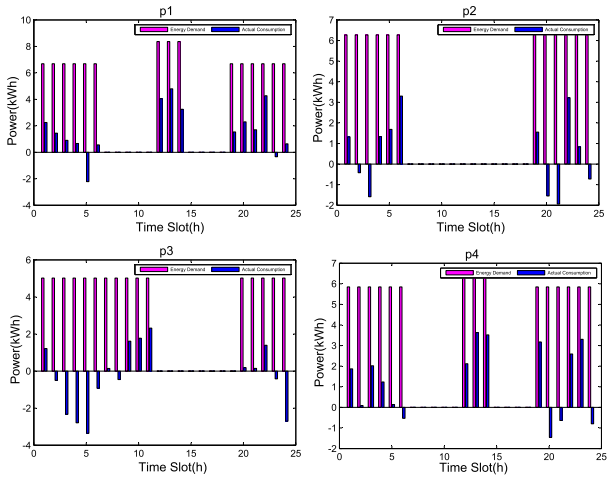

Fig. 7. Energy consumption for PEVs.

B. Validity Check

1) $Q-$Table-Based ED Results: We analyze the effectiveness of $Q-$learning methods based on $Q-$tables. Figs. 6 and Fig. 7 show the comparison of the energy demand and consumption of dispatchable appliances and PEVs for one day. Table III shows the specific daily RPs planning obtained by Algorithm 1.

TABLE I
TIME-VARYING PARAMETERS

Parameter	Time Slot	1	2	3	4	5	6	7	8	9	10	11	12
	θ_t	1.9	1.8	1.7	1.6	1.6	1.9	2.1	2.2	2.3	2.6	2.9	3.3
	δ_t	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.3	-0.5
	Time Slot	13	14	15	16	17	18	19	20	21	22	23	24
	θ_t	3.4	3.6	3.5	4.4	4.5	4	4	3.2	3.1	3.6	3	2.6
	δ_t	-0.5	-0.5	-0.3	-0.3	-0.7	-0.7	-0.7	-0.7	-0.7	-0.5	-0.5	-0.5

TABLE II
PARAMETERS OF APPLIANCES

	h_{i1}	h_{i2}	DE_i^{min}	DE_i^{max}	E_i^r	E_p^0	B_p^{min}	B_p^{max}	e_p^{charg}	$e_p^{discharg}$	ρ^{min}	ρ^{max}	ω	$lr(la=lc)$	ξ
$d1$	3	5	$0.1R_{d,t}$	$0.5R_{d,t}$	-	-	-	-	-	-	$1.5min\{\theta_t\}$	$1.5max\{\theta_t\}$	0.9	0.9	0.01
$d2$	4.5	5			-	-	-	-	-	-					
$d3$	5	5			-	-	-	-	-	-					
$d4$	4	5			-	-	-	-	-	-					
$d5$	5.5	5			-	-	-	-	-	-					
$d6$	6	5			-	-	-	-	-	-					
$p1$	3.5	4	$R_{p,t}-E_p^r$	$R_{p,t}+E_p^r$	9	10	5	98	0.9	0.9					
$p2$	3	4			9	12.5	5	95	0.95	0.95					
$p3$	4	4			7	5	5	95	0.95	0.85					
$p4$	4.5	4			8	13	5	90	0.9	0.8					

TABLE III
OPTIMAL RPS PLANNING FOR ALL APPLIANCES

t	$d1$	$d2$	$d3$	$d4$	$d5$	$d6$	$p1$	$p2$	$p3$	$p4$	$n1$	$n2$	$n3$	$n4$	$n5$	$Total(\$)$
1	3	4	6.6	5.6	2.8	5.8	3.5	4.4	2.8	3.6	6.6	4.2	6.6	5.5	5	70
2	4.3	4.3	4.8	4.3	3.3	5.2	3.9	5.7	4.5	5.1	3	5.8	4.7	5.6	2.5	67
3	4.7	5.1	4.2	4.4	3.2	6.6	4	6.3	6.1	2.8	5	4.4	5.4	3	4.7	69.9
4	5.2	5	4.6	6.1	3.3	4	3.8	3.3	6	3.2	2.8	2.4	3.8	5.1	2.7	61.3
5	5.5	2.4	5.6	6.6	2.8	5.5	6.1	3	6.6	4.2	5.7	3.5	5.6	6.2	6	75.3
6	4.2	5.7	4.2	4	3.1	2.4	5.1	2.4	5.5	6.2	6.5	6	3.4	4.8	5.4	68.9
7	6.4	2.4	6.2	4.3	3.3	4.9	4.1	6.3	5	5.5	5.1	6	6.1	5.8	6.1	77.5
8	2.5	5.4	2.9	3.4	5.2	3.4	5.3	6.2	6.3	6.5	6.2	3.7	6.6	6.3	2.9	72.8
9	3.9	2.4	4.4	4.3	2.4	5.4	5.6	5.2	3.6	3.1	2.4	6.1	3.5	5.7	5.8	63.8
10	3	6.7	4.4	4.3	3.8	5.3	6.3	3.9	4.3	6.2	6.7	5.4	3.3	4.9	6	74.5
11	5.3	5.3	6.3	6.5	3.9	6.7	4.6	5.3	4.2	4	4.2	6.3	5.1	3.6	6.2	77.5
12	5.6	4.3	6.3	5.9	5.5	4.2	4.1	4.6	3.8	5.2	4	6.1	4.6	6.1	6.5	76.8
13	6.6	4.5	4.6	5.5	4.5	6.4	3.7	4.8	6.2	3.9	5.8	3.9	5.2	5.8	4.5	75.9
14	4.2	4.2	6.7	3.6	4.2	3.6	5.4	4.3	3.8	4.4	5.6	4.5	5.7	3.9	6.4	70.5
15	5.8	4.3	5.5	4.5	6.4	6.7	5.5	4.6	6.3	3.6	4.7	6.5	4.7	4	5.8	78.9
16	5.6	6.4	4.5	6.3	4.7	4.9	4.8	6.6	4.8	5.3	4.9	5.6	5.1	6.1	5.4	81
17	6.5	4.6	5.8	6	4.8	5.8	6.2	6.7	4.5	5.7	5	5.8	4.6	5.2	5.9	83.1
18	5.3	6.4	5.1	6.4	6.2	6.7	4.8	6.7	4.3	5	5.7	5.9	4	4.6	6.7	83.8
19	6.2	4.5	6.5	6.2	4.5	6.5	5.8	5.8	5	4	5.9	4.2	5.7	5.6	6	82.4
20	6.2	4.4	6.3	3.2	4.7	5.6	3.6	6.4	3.7	6.3	6	4.3	6.3	3.6	6.6	77.2
21	3.4	3.7	6.6	5.7	6.1	4	3.8	6.4	3.5	5.4	4.4	5.7	4.1	6.1	5.4	74.3
22	6.2	6.5	6.5	5.3	6.5	6.7	3.6	4.6	4.9	5	6.2	6.1	4.1	4.5	3.7	80.4
23	5	6.5	5.1	4.1	6.4	4.5	6.7	5.7	5.6	3	5.8	3.8	6.6	6.1	4	78.9
24	5.9	3.6	5.8	6.3	4.9	6.1	4.7	5.9	6.7	5.9	4.4	5.6	5.5	5.5	4.2	81
$Total(\$)$	120.5	112.6	129.5	122.8	106.5	126.9	115	125.1	118	113.1	122.6	121.8	120.3	123.6	124.4	1802.7

It can be seen that the RPs strictly satisfy the price constraints. The overall trend of the RPs planning fluctuates over the 24 time slots, which is affected by social welfare and WPs. When the RP is too high, it is detrimental to the social welfare relative to the users. For example, $d1$ has been reducing energy consumption (from 8.9 (kWh) to 5.3 (kWh)) due to the increasing retail price (from 3 ($) to 6.6 ($)) in time slots $10-13$, and the retail price decreases rapidly (from 6.6 ($) to 4.2 ($)) under the influence of factors that favor the customer's interest such as the electrical happiness error function and the PEV cost, and subsequently the residence will tend to consume more electricity in the next time slot. Out of commercial interest to the retailer, this also affects relative social welfare. For example, when the RP is too small at time slot 9, and the retailer increases the RP significantly at time slot 10 for $\{d2, p4, n1\}$, while PEVs in time slots $2-6$ and $20-24$ actively discharge in response to excessive electricity prices. Combining Table III and Fig. 5, it can be observed that the RPs peak does not occur during demand peak hours (12:00-16:00 and 18:00-24:00, and the average retail price (77.24($)) at demand peak is smaller than the average price (79.62($)) at other time slots. This is because the goal of maximizing social welfare ensures that prices are set to benefit both retailers and customers, creating a virtuous circle between retail prices and actual electricity consumption to maintain a relatively balanced social welfare.

Fig. 8 depicts the convergence effect of $q-$values in the $Q-$table for the four PEVs over 24 time slots. The $q-$values are described by different color lines, which indicate stable

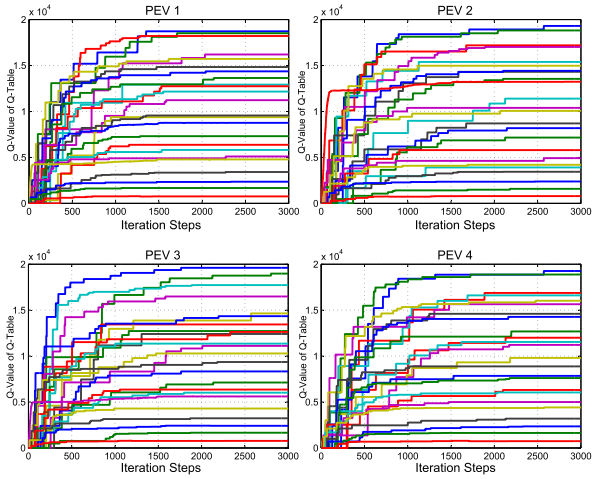

Fig. 8. $Q-$values of $Q-$tables from PEVs.

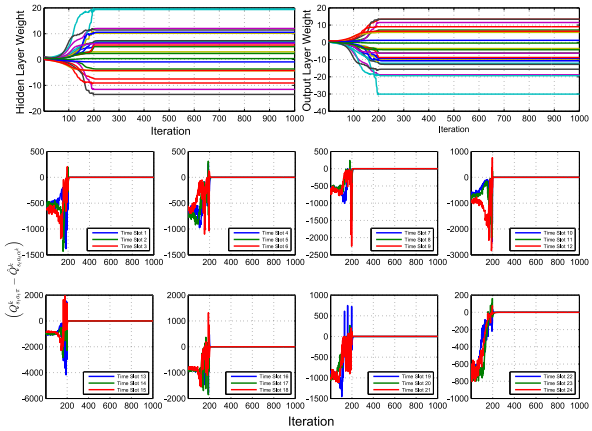

Fig. 9. Weights and convergence threshold of DQN for $p3$.

convergence of the algorithm. In order to maximize social welfare, the retailer constantly changes its electricity pricing strategies to gradually make $q-$values converge to their maximization, considering the stochastic electricity consumption patterns and storage capacity limitations of the PEVs.

2) DQN-ED Results: We combine deep learning with $Q-$learning to construct $Q-$networks instead of the evaluation role of $Q-$tables, and propose two algorithms: online DQN-ED and offline DDQN-ED, where the capacity of the experience buffer $\mathcal{D}$ is set to $F = 20$, and $M = 15$ transitions are sampled and removed from $\mathcal{D}$ at each iteration.

Fig. 9 represents the weights of hidden/output layers and the convergence of $(Q^k_{s_{i,t}a_{i,t}\pi_{s_{i,t}}} - \hat{Q}^k_{s_{i,t}a_{i,t}\alpha^k_{i,t}})$ over 24 time slots for $p3$, which demonstrates the effectiveness of the DQN-ED algorithms. Fig. 10 shows the evolution of the $q-$values of DQN and DDQN over 24 time slots, respectively. The DQN-ED algorithms coordinate the energy consumption of each appliance, and based on the RPs with instant feedback, make decisions on the expected actual energy consumption through the greedy strategy that eventually converges to the optimal.

Due to the complexity of the RDRM problem (e.g., the non-convexity of the optimization problem, the setting of parameters unique to the algorithm), the scheduling policies obtained by

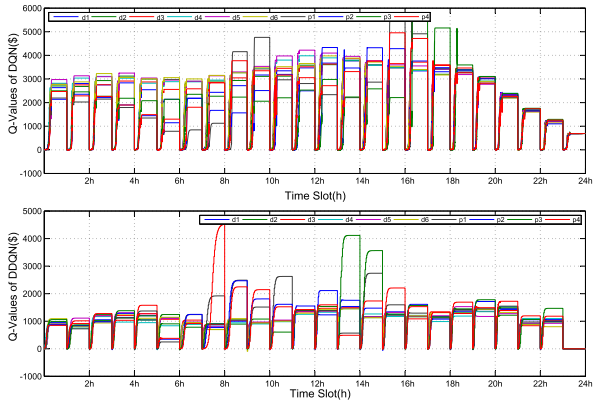

Fig. 10. $Q-$values over 24 time slots for DQN and DDQN.

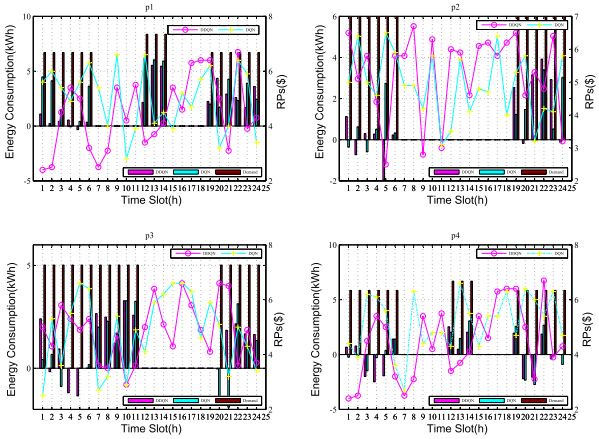

Fig. 11. Actual electricity consumption and RPs over 24 time slots for DDQN and DQN.

different learning algorithms are often distinct. Fig. 11 shows the actual energy consumption and retail price planning during 24 time slots for the four PEVs under both DDQN-ED and DQN-ED algorithms. As can be seen, although the two algorithms obtain different RPs and actual consumptions, the relative actual consumptions of the two algorithms determine the relative retail prices. In particular, when the two algorithms perform discharging ($E_{p,t} < 0$) and charging ($E_{p,t} > 0$) at a certain time slot, respectively, the RP of discharging tends to be smaller than the RP of charging. In addition, although the RPs are not identical under different algorithms, the variation trend of RP is similar at 24 time slots. We believe is due to the fact that both use the same behavioral policy ($\varepsilon-$greedy strategy) and target policy (greedy strategy).

3) AC-ED Results: For the discrete action space, the Actor policy network utilizes the softmax function to select the optimal action with the maximum probability principle for each appliance as follows:

$$\mathbf{Pr}(a^j_{i,t}|s_{i,t},\beta_{i,t}) = \frac{e^{\phi^j_a(s_{i,t},\beta_{i,t})}}{\sum_{\tau=1}^{\mathcal{M}} e^{\phi^\tau_a(s_{i,t},\beta_{i,t})}}, j \in \mathcal{M}, i \in \mathbb{N}$$

Fig. 12 represents the action probability output of the Actor network with 4 PEVs before and after training. Each component $\phi^j_a(s_{i,t},\beta_{i,t})$ of the feature vector corresponds to the probability

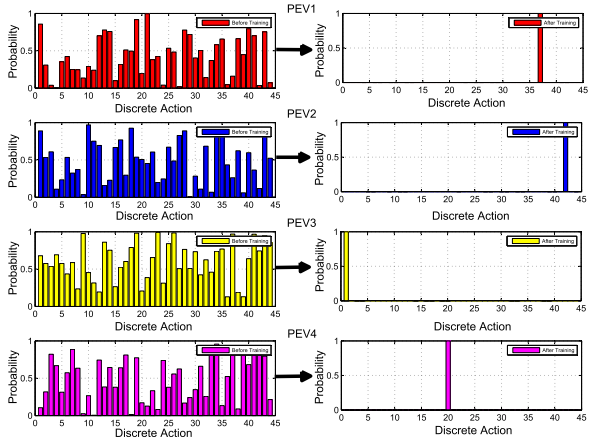

Fig. 12. Action probability of Actor NNs for PEVs.

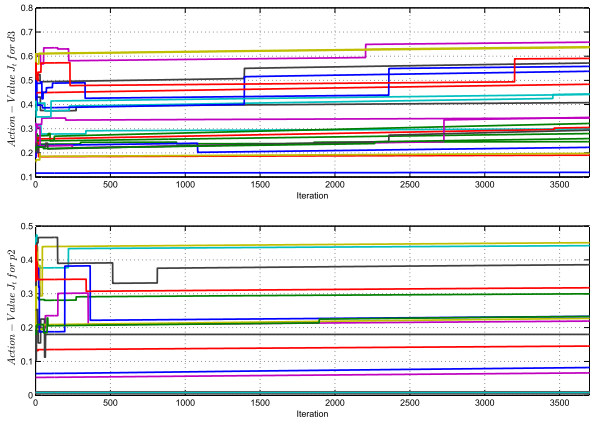

Fig. 13. Action-values of Actor networks.

distribution of 44 actions through the softmax function. All action probabilities are randomly initialized before training, and the evaluation function $\mathcal{E}_{i,t}$ is used to maximize the objective function to select the optimal action during training, and the proportion of that action probability in the action set is continuously increased while the probability of other actions is reduced, which will eventually converge to 1.

Fig. 13 represents the evolution of action values (23) for $d3$ and $p2$, respectively, where the different colors represent different time slots. The Critic networks finally converge due to the stability of the action values, demonstrating that the algorithm's efficacy is assured.

4) Comparisons: Fig. 14 shows the average level of social welfare for the proposed four algorithms after 10 trials. The following features can be observed. Since the $Q-$learning method is a class of value-based ($q-$value) algorithms, which tend to fall into overestimation by maximizing the $q-$value, the curves of the algorithms are characterized by large amplitude, high frequency, and fast convergence. The disadvantage is that it may not converge to optimality against nonconvex optimization objectives, so more potential solutions have to be explored and learned through $\epsilon-$greedy strategies. In contrast, the online AC-ED method is capable of both learning policy function (23) and evaluating current value function (22), as well as continuously optimizing policy network by TD error (24), so it has better stability and practicality. However, the high data

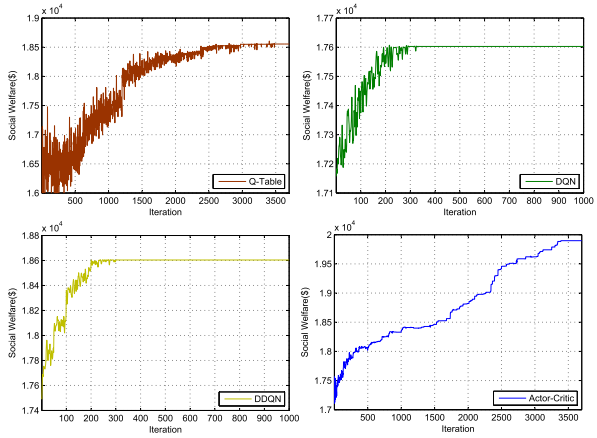

Fig. 14. Comparison of social welfare.

TABLE IV
COMPARISONS BETWEEN AVERAGE RP AND SOCIAL WELFARE

Algorithms	Total Social Welfare($)	Average RPs($)
Q-Table	18554.26	4.82
DQN	17602.53	4.99
DDQN	18604.05	5.00
Actor-Critic	19897.35	3.27

correlation caused by the online learning method directly affects the learning ability of the Actor-Critic network (some of the action values are updated slowly as can be seen in Fig. 13), resulting in the difficulty in obtaining the optimal social welfare quickly in Fig. 14. To decouple data correlation and fully explore policy distribution, we expect to study the offline learning based Actor-Critic framework and its variants applied to the PRDRM problem.

Table IV shows the comparison between total social welfare $(\sum_{t=1}^{T}\sum_{i=1}^{\mathbf{N}}\rho_{i,t})/(\mathbf{N}\times T)$ of the algorithms for all time slots. The AC-ED algorithm gets the highest social welfare, 13.04% higher than that obtained by DQN-ED, but its average daily RP is indeed 34.47% lower than that of DQN-ED. This shows that high social welfare combines corporate profits and user benefits, which characterizes the social well-being in electricity usage. In addition, as shown in the experiment, the disadvantage of $Q-$table-based ED method for discrete action space is only the calculation and storage of $q-$values, while the experimental results of social welfare do not show significant drawback compared with the DQN-ED algorithms.

V. CONCLUSION

In this article, we studied the PRDRM problem based on MFRL, which takes fully into account the charged and discharged PEV model that is gradually becoming popular among users. Firstly, three types of appliances (including dispatchable appliances, non-dispatchable appliances and PEVs) were mathematically modeled, and then, PRDRM was coupled with MFRL to reconcile the actual power consumption of appliances on the environment side with the retail price of electricity on the agent side through a series of MFRL algorithms (including Q-table-based framework, DQN framework and Actor-Critic framework). A systematic solution with long-view decision

capability was provided for real-time demand response in smart grids. Finally, simulation experiments validated the effectiveness of the proposed approach. Future work will apply the MFRL-based energy dispatching scheme to multi-carrier energy supply (i.e., gas and electricity) to achieve higher efficiency and lower operating costs.

REFERENCES

[1] H. R. Gholinejad, J. Adabi, and M. Marzband, "Hierarchical energy management system for home-energy-hubs considering plug-in electric vehicles," *IEEE Trans. Ind. Appl.*, vol. 58, no. 5, pp. 5582–5592, Sept./Oct. 2022.

[2] H. Hao, C. D. Corbin, K. Kalsi, and R. G. Pratt, "Transactive control of commercial buildings for demand response," *IEEE Trans. Power Syst.*, vol. 32, no. 1, pp. 774–783, Jan. 2017.

[3] N. G. Paterakis, O. Erdinç, A. G. Bakirtzis, and J. P. S. Catalão, "Optimal household appliances scheduling under day-ahead pricing and load-shaping demand response strategies," *IEEE Trans. Ind. Inf.*, vol. 11, no. 6, pp. 1509–1519, Dec. 2015.

[4] A. A. Moghaddam, H. Monsef, and A. Rahimi-Kian, "Optimal smart home energy management considering energy saving and a comfortable lifestyle," *IEEE Trans. Smart Grid*, vol. 6, no. 1, pp. 324–332, Jan. 2015.

[5] Y. Du, L. Jiang, Y. Li, and Q. Wu, "A robust optimization approach for demand side scheduling considering uncertainty of manually operated appliances," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 743–755, Mar. 2018.

[6] Y. Huang, L. Wang, W. Guo, Q. Kang, and Q. Wu, "Chance constrained optimization in a home energy management system," *IEEE Trans. Smart Grid*, vol. 9, no. 1, pp. 252–260, Jan. 2018.

[7] L. Yu, T. Jiang, and Y. Zou, "Online energy management for a sustainable smart home with an HVAC load and random occupancy," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 1646–1659, Mar. 2019.

[8] T. M. Aljohani, A. F. Ebrahim, and O. A. Mohammed, "Dynamic real-time pricing mechanism for electric vehicles charging considering optimal microgrids energy management system," *IEEE Trans. Ind. Appl.*, vol. 57, no. 5, pp. 5372–5381, Sept./Oct. 2021.

[9] W. Wu, Y. Lin, R. Liu, Y. Li, Y. Zhang, and C. Ma, "Online EV charge scheduling based on time-of-use pricing and peak load minimization: Properties and efficient algorithms," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 572–586, Jan. 2022.

[10] F. Wang et al., "Smart households' aggregated capacity forecasting for load aggregators under incentive-based demand response programs," *IEEE Trans. Ind. Appl.*, vol. 56, no. 2, pp. 1086–1097, Mar./Apr. 2020.

[11] R. Deng, Z. Yang, M. Chow, and J. Chen, "A survey on demand response in smart grids: Mathematical models and approaches," *IEEE Trans. Ind. Inf.*, vol. 11, no. 3, pp. 570–582, Jun. 2015.

[12] B. Shen, G. Ghatikar, Z. Lei, J. Li, G. Wikler, and P. Martin, "The role of regulatory reforms, market changes, and technology development to make demand response a viable resource in meeting energy challenges," *Appl. Energy*, vol. 130, pp. 814–823, Oct. 2014.

[13] H. Khorramdel, M. Gitizadeh, P. Siano, and S. Bakhtiari, "Evaluating the economic impact of users' personality on selection of demand response programs," *CSEE J. Power Energy Syst.*, vol. 9, no. 3, pp. 1036–1050, May 2023.

[14] A. Mohsenian-Rad, V. W. S. Wong, J. Jatskevich, R. Schober, and A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 320–331, Dec. 2010.

[15] R. Deng, Z. Yang, F. Hou, M. Y. Chow, and J. Chen, "Distributed realtime demand response in multiseller-multibuyer smart distribution grid," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2364–2374, Sep. 2015.

[16] Q. Dong, L. Yu, W. Song, J. Yang, Y. Wu, and J. Qi, "Fast distributed demand response algorithm in smart grid," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 2, pp. 280–296, Apr. 2017.

[17] L. Yu et al., "Multi-agent deep reinforcement learning for HVAC control in commercial buildings," *IEEE Trans. Smart Grid*, vol. 12, no. 1, pp. 407–419, Jan. 2021.

[18] D. Azuatalam, W. L. Lee, F. D. Nijs, and A. Liebman, "Reinforcement learning for whole-building HVAC control and demand response," *Energy AI*, vol. 2, pp. 1–18, 2020.

[19] D. Li, L. Yu, N. Li, and F. Lewis, "Virtual-action-based coordinated reinforcement learning for distributed economic dispatch," *IEEE Trans. Power Syst.*, vol. 36, no. 6, pp. 5143–5152, Nov. 2021.

[20] M. Ahrarinouri, M. Rastegar, and A. R. Seifi, "Multiagent reinforcement learning for energy management in residential buildings," *IEEE Trans. Ind. Inf.*, vol. 17, no. 1, pp. 659–666, Jan. 2021.

[21] Z. Zhang, Z. Chen, and W. -J. Lee, "Soft actor–critic algorithm featured residential demand response strategic bidding for load aggregators," *IEEE Trans. Ind. Appl.*, vol. 58, no. 4, pp. 4298–4308, Jul./Aug. 2022.

[22] A. Anvari-Moghaddam, A. Rahimi-Kian, M. S. Mirian, and J. M. Guerrero, "A multi-agent based energy management solution for integrated buildings and microgrid system," *Appl. Energy*, vol. 203, pp. 41–56, Oct. 2017.

[23] W. Valladares et al., "Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm," *Build. Env.*, vol. 155, pp. 105–117, May 2019.

[24] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[25] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, May 2020.

[26] E. Mocanu et al., "On-line building energy optimization using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 4, pp. 3698–3708, Jul. 2019.

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

[28] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, London, England: MIT Press, 2005.

[29] Y. Wan, J. Qin, X. Yu, T. Yang, and Y. Kang, "Price-based residential demand response management in smart grids: A reinforcement learning-based approach," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 1, pp. 123–134, Jan. 2022.

[30] L. Yu et al., "Deep reinforcement learning for smart home energy management," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 2751–2762, Apr. 2020.

[31] K. Mei, G. Cai, D. Luo, and B. Liang, "Research on electricity consumption classification andmeasurment method based on smart meter," *Cyber Secur. Data Governance*, vol. 39, no. 2, pp. 62–68, 2020, doi: 10.19358/j.issn.2096-5133.2020.02.012.

[32] Y. Liu, H. Li, L. Wang, and R. Hasan, "Non-intrusive load monitoring method based on convolutional neural network," *Elect. Meas. Instrum.*, vol. 59, no. 1, pp. 148–154, 2022, doi: 10.19753/j.issn1001-1390.2022.01.020.

[33] H. Geng, L. Liu, and X. Pang, "Non-intrusive residential electric load identification method based on artificial neural network," *J. Shenyang Inst. Eng. (Natural Sci.)*, vol. 15, no. 3, pp. 236–240, 2019, doi: 10.13888/j.cnki.jsie(ns).2019.03.010.

[34] R. Lu, S. H. Hong, and X. Zhang, "A dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach," *Appl. Energy*, vol. 220, pp. 220–230, Jun. 2018.

[35] C. Dewangan, S. Singh, S. Chakrabarti, and K. Singh, "Peak-to-average ratio incentive scheme to tackle the peak-rebound challenge in TOU pricing," *Electric Power Syst. Res.*, vol. 210, 2022, Art. no. 108048.

[36] B.-C. Lai, W.-Y. Chiu, and Y.-P. Tsai, "Multiagent reinforcement learning for community energy management to mitigate peak rebounds under renewable energy uncertainty," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 3, pp. 568–579, Jun. 2022.

[37] E. Salazar, M. Jurado, and M. Samper, "Reinforcement learning-based pricing and incentive strategy for demand response in smart grids," *Energies*, vol. 16, 2023, Art. no. 1466.

[38] P. Yi and L. Pavel, "An operator splitting approach for distributed generalized Nash equilibria computation," *Automatica*, vol. 102, pp. 111–121, 2019.

[39] P. Dai, W. Yu, G. Wen, and S. Baldi, "Distributed reinforcement learning algorithm for dynamic economic dispatch with unknown generation cost functions," *IEEE Trans. Ind. Inf.*, vol. 16, no. 4, pp. 2258–2267, Apr. 2020.

[40] E. Duryea, M. Ganger, and W. Hu, "Deep reinforcement learning with double Q-learning," *Intell. Control Automat.*, pp. 129–144, 2016.

[41] "Sdge.com. Home-san Diego gas & electric," [Online]. Available: https://www.sdge.com

[42] "Real-time hourly prices," Commonwealth Edison Company. [Online]. Available: https://rrtp.comed.com/live-prices/