Mixed Deep Reinforcement Learning Considering Discrete-continuous Hybrid Action Space for Smart Home Energy Management

Chao Huang, *Member*, *IEEE*, Hongcai Zhang, *Member*, *IEEE*, Long Wang, *Member*, *IEEE*, Xiong Luo, *Senior Member*, *IEEE*, and Yonghua Song, *Fellow*, *IEEE*

***Abstract*—This paper develops deep reinforcement learning (DRL) algorithms for optimizing the operation of home energy system which consists of photovoltaic (PV) panels, battery energy storage system, and household appliances. Model-free DRL algorithms can efficiently handle the difficulty of energy system modeling and uncertainty of PV generation. However, discrete-continuous hybrid action space of the considered home energy system challenges existing DRL algorithms for either discrete actions or continuous actions. Thus, a mixed deep reinforcement learning (MDRL) algorithm is proposed, which integrates deep *Q*-learning (DQL) algorithm and deep deterministic policy gradient (DDPG) algorithm. The DQL algorithm deals with discrete actions, while the DDPG algorithm handles continuous actions. The MDRL algorithm learns optimal strategy by trial-and-error interactions with the environment. However, unsafe actions, which violate system constraints, can give rise to great cost. To handle such problem, a safe-MDRL algorithm is further proposed. Simulation studies demonstrate that the proposed MDRL algorithm can efficiently handle the challenge from discrete-continuous hybrid action space for home energy management. The proposed MDRL algorithm reduces the operation cost while maintaining the human thermal comfort by comparing with benchmark algorithms on the test dataset. Moreover, the safe-MDRL algorithm greatly reduces the loss of thermal comfort in the learning stage by the proposed MDRL algorithm.**

***Index Terms*—Demand response, deep reinforcement learning, discrete-continuous action space, home energy management, safe reinforcement learning.**

I. Introduction

DEMAND response (DR), which offers consumers the opportunity to change their consumption patterns in response to incentives or electricity prices to balance power demand and power supply, is considered as an integral part of smart grid [1]. The residential sector contributes greatly to the total consumption of electricity, e.g., the residential sector consumes 14.2% of total electricity consumption by 2019 in China [2]. Therefore, it is valuable to develop efficient DR programs for energy management in the residential sector.

In the residential sector, price-based DR programs including time-of-use (TOU) pricing program and real-time (RT) pricing program are most frequently studied [3], [4]. Within these DR programs, home energy management systems (HEMSs) are required to automatically make optimal scheduling of household appliances in response to electricity price signals. The application of renewable energies such as solar energy and wind energy in homes further complicates the development of HEMSs due to their nature of uncertainty [5]. Hence, a well-developed HEMS under a given DR program can provide positive effects such as improved human comfort level, reduced electricity cost, and reduced carbon emission by accommodating renewable energies.

The objectives of HEMSs are usually to minimize electricity cost and maximize human comfort [6]. However, the methods underlying HEMSs are different, including rule-based methods, model-based methods, and model-free methods. In [7], deterministic rules were applied for the management of household appliances. To improve the capacity in learning and adapting to occupant's pattern change, an adaptive rule-based technique was proposed for automatic control of air conditioner in [8]. In [9], an analytical rule-based approach was developed for combined heat and power residential energy system. Almost all rule-based methods use "if-then" rules, which are easy to implement. However, the settlement of rules highly depends on expert knowledge and these methods are less efficient for complex home energy

Manuscript received: June 27, 2021; revised: October 8, 2021; accepted: December 3, 2021. Date of CrossCheck: December 3, 2021. Date of online publication: January 14, 2022.

This work was supported by the National Natural Science Foundation of China (No. 62002016), the Science and Technology Development Fund, Macau S.A.R. (No. 0137/2019/A3), the Beijing Natural Science Foundation (No. 9204028), and the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515111165).

C. Huang was with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau S. A. R., China, he is currently with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, and he is also with Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China (e-mail: chao.huang@my.cityu.edu.hk).

H. Zhang (corresponding author) and Y. Song are with the State Key Laboratory of Internet of Things for Smart City and Department of Electrical and Computer Engineering, University of Macau, Macau S. A. R., China (e-mail: hczhang@um.edu.mo; yhsong@um.edu.mo).

L. Wang and X. Luo are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China, and they are also with Shunde Graduate School, University of Science and Technology Beijing, Foshan 528399, China (e-mail: lwang@ustb. edu. cn; xluo@ustb.edu.cn).

DOI: 10.35833/MPCE.2021.000394

systems with continuous changes in environmental conditions such as electricity price and renewable generation.

With model-based methods, a numerical model is required to characterize home energy system and an optimization problem is formulated considering the objective and system constraints [10], [11]. The operation of home energy system is optimized by solving the optimization problem. The main challenges of model-based methods lie in the modeling accuracy of energy system and prediction accuracy of unknown variables. In [12], a mixed-integer linear programming (MILP) based HEMS was developed for day-ahead optimal scheduling of household appliances including both thermostatically and non-thermostatically controlled appliances under hourly pricing DR program. Scenario-based stochastic programming was used for home energy management considering uncertainty of renewable energy and electrical vehicle availability in [13]. In [14], a stochastic model predictive control strategy was proposed for a residential building energy management. In [15], a game theory based strategy was developed for home energy management. With model-based methods, however, simplified thermal dynamic models are usually employed for modeling of thermostatically controlled loads, which deteriorate the modeling accuracy and quality of decisions.

With the advancement of artificial intelligence, model-free methods based on reinforcement learning (RL) have been developed for home energy management [16], [17]. To deal with the uncertainty of electricity price, a multi-agent deep Q-learning (DQL) algorithm was developed for scheduling of multiple home appliances in [18]. In [19], an actor-critic learning based load scheduling algorithm was proposed to reduce electricity cost of households and peak-to-average ratio in the aggregate load. In [20], a generalized actor-critic learning based optimal control method was developed to minimize the consumption cost of home users. A deep deterministic policy gradient (DDPG) based home energy management algorithm was developed in [21] for the control of heating, ventilation, and air conditioning (HVAC) system and energy storage system considering the uncertainty of electricity price, photovoltaic (PV) generation, and outdoor temperature. In addition to home energy management, DDPG was also applied for energy management in sensor networks with renewable generations [22].

The RL-based methods learn optimal decision-making strategy by iteratively interacting with the energy system, which do not require prior knowledge on the energy system [23]. This character of RL is valuable for complex energy systems composed of unknown variables, e. g., renewable generation, and dynamic processes that are difficult to model, e. g., thermal dynamic model for HVAC control. However, current RL-based methods mainly consider either discrete action space or continuous action space. Discrete action space usually includes "on/off" operation modes of household appliances such as washing machine and dish washer, while continuous action space is commonly reserved for the control of HVAC system and energy storage system. A simple way to dealing with discrete-continuous hybrid action space is to discretize continuous action in order to apply existing RL framework for discrete action space. In [24], a DQL algorithm was developed for optimal scheduling of dish washer, air conditioner, and electric vehicle, where discrete action space was used to model operation patterns of air conditioner and electric vehicle. However, the granularity of discretization of continuous action space significantly affects the performance of DQL. In [25], a DDPG-based strategy was proposed for residential multi-energy system management considering discrete-continuous hybrid action space where the discretization of continuous outputs from actor network was performed to derive discrete actions. However, the treatment of discrete actions as continuous ones may significantly improve the complexity of action space. With the above concerns, an RL-based method with capability in handling discrete-continuous hybrid action space is valuable for home energy management.

In many practical engineering problems, however, unsafe actions, which violate system constraints, can lead to system damages and high cost, especially during the learning stage [26], [27]. For the problem in this study, improper control of home appliances, i. e., the HVAC system, can give rise to high loss in human comfort. To handle the challenge from unsafe actions, two main trends for safe-RL were studied in [28]. The first trend lies in the modification of optimality criterion such as the worst-case criterion and risk-sensitive criterion instead of generally considered mean expected return. The second one lies in the modification of exploration process with external knowledge to avoid the actions that can lead the learning system to undesirable situations. In [29], a constrained cross-entropy-based RL method, which explicitly tracked its performance with respect to constraint satisfaction, was proposed for safety-critical applications. For RL-based energy management system, however, safety is seldom considered in published literature [30], [31].

This paper investigates deep reinforcement learning (DRL) based optimization algorithm for HEMS. The main contributions of the paper are outlined below.

1) The operation cost optimization problem of grid-connected home energy system including various household appliances, e. g., HVAC system, wash machine, dish washer, etc., renewable generation, and battery energy storage system (BESS) is formulated as a Markov decision process (MDP) without the prediction of unknown variables or thermal dynamic model. The operation modes of household appliances and BESSs constitute discrete-continuous hybrid action space for the MDP, which challenges existing RL algorithms for either discrete action space or continuous action space.

2) A mixed deep reinforcement learning (MDRL) algorithm that integrates DQL and DDPG is developed to solve the MDP. The proposed MDRL algorithm inherits the merits of DQL in handling discrete action space and takes advantages of DDPG in dealing with continuous action space. More precisely, the MDRL algorithm leverages the actor-critic framework as in the DDPG algorithm. The actor network with the proposed MDRL algorithm, however, receives discrete action and state as input and outputs continuous actions. The critic network evaluates the combination of dis-

crete action and continuous action for the given state. Similar to DQL, the optimal combination of discrete action and continuous action is determined by selecting the one that maximizes the Q-value. Meanwhile, to facilitate the training of the proposed MDRL algorithm, a special exploration policy is designed for discrete-continuous hybrid action space.

3) To avoid high loss of human thermal comfort with the HVAC system in the learning stage, a prediction model guided safe-MDRL algorithm is further proposed. In the safe-MDRL algorithm, an online prediction model is developed and applied to evaluate actions associated with the HVAC system to avoid severe violation of thermal constraints.

4) Simulation studies based on real data illustrate that the proposed MDRL algorithm can efficiently reduce operation cost while maintaining human thermal comforts compared with benchmark algorithms on the test dataset. Moreover, the safe-MDRL algorithm greatly reduces the loss of human thermal comfort in the learning stage by the MDRL algorithm.

The remainder of the paper is organized as follows. In Section II, the HEMS is introduced with mathematical formulations. In Section III, the optimization problem of HEMS is firstly formulated as an MDP, which is followed by the development of the proposed MDRL algorithm and its safe version. Simulation results are provided in Section IV, and conclusions are given in Section V.

II. HEMS

The HEMS considered in this paper is illustrated in Fig. 1.

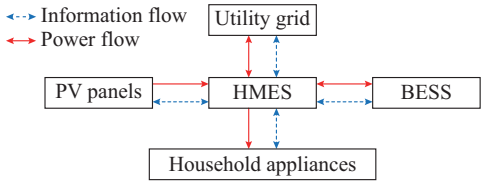

Fig. 1. Considered HEMS.

The home is equipped with PV panels, BESS, and household appliances. The household appliances can be generally classified into non-shiftable loads, shiftable and non-interruptible loads, and controllable loads in terms of their characteristics [32]. The non-shiftable loads, e.g., lighting, television, microwave, refrigerator, etc., which are essential to the home cannot be scheduled and their power demands should be satisfied without delay. The shiftable and non-interruptible loads such as washing machine, wash dryer, and dish washer can be scheduled to time slots of low electricity price. However, their operations cannot be interrupted and power demands are non-controllable. The controllable loads can be operated in a flexible manner in terms of operation time and power demand. Thermostatically controlled loads such as HVAC system and electric water heater are the most common controllable loads in a home, while the HVAC system consumes more energy than other loads [21]. Hence, this paper considers non-shiftable loads, shiftable and non-interruptible loads, and HVAC system in a smart home. The scheduling problem of home energy management is formulated in a discrete form where the scheduling horizon T is divided into a number of time slots $t \in [1, T]$ with equivalent time interval $\Delta T = 1$ hour in this paper. The HEMS makes decisions for optimal operation of electric loads. In this section, mathematical formulations associated with the home energy system will be investigated.

A. Shiftable and Non-interruptible Loads

Consider a set of N shiftable and non-interruptible loads. For each individual load n, $n = 1, 2, \ldots, N$, it is characterized by a tuple $\left(T_{n,ini}, T_{n,end}, T_{n,d}, P_n\right)$, where $T_{n,ini}$ and $T_{n,end}$ are the initial time and end time of working period, respectively; $T_{n,d}$ is the time slot required to complete the task; and P_n is the power demand. For shiftable and non-interruptible loads, there are two operation modes, i.e., "on" and "off". Power demand for all this kind of appliances in time slot t is obtained by:

$$P_{shift,t} = \sum_{n=1}^{N} x_{n,t} P_n \tag{1}$$

where $x_{n,t}$ is a binary decision variable for appliance n and 1/0 corresponds to "on/off", respectively. The operation of shiftable and non-interruptible loads should satisfy following constraints:

$$x_{n,t} = 0 \quad t < T_{n,ini} \tag{2}$$

$$x_{n,t} = 1 \quad t = T_{n,end} - T_{n,d} + 1, T_{n,t-1} = T_{n,d} \tag{3}$$

$$x_{n,t} = 1 \quad 0 < T_{n,t-1} < T_{n,d} \tag{4}$$

$$x_{n,t} = 0 \quad T_{n,t-1} = 0 \tag{5}$$

where $T_{n,t-1}$ is the remaining time slot required to complete the task at the end of time slot $t-1$ for appliance n satisfying $T_{n,t-1} = T_{n,t-2} - x_{n,t-1}$ and $T_{n,0} = T_{n,d}$. The constraint (2) ensures that the appliance should be "off" before initial time of the working period; the constraint (3) enforces the starting of the task to ensure the completion of the task in the working period; the constraint (4) ensures non-interruption of the task; and the constraint (5) enforces the appliance to be "off" once the task has been completed.

B. HVAC System

This paper considers an HVAC system that can adjust its input power continuously to maintain human thermal comforts.

$$0 \le P_{HVAC,t} \le P_{HVAC,\max} \tag{6}$$

where $P_{HVAC,t}$ and $P_{HVAC,\max}$ are the input power of the HVAC system at t and its maximum power, respectively.

Indoor air conditions such as air temperature, air speed, and relative humidity are essential for the determination of human thermal comfort level. To simplify the representation of human thermal comfort, human comfort temperature zone is considered as in [21], [33]:

$$K_{\min} \le K_{in,t} \le K_{\max} \tag{7}$$

where $K_{in,t}$ is the indoor temperature at t; and $[K_{\min}, K_{\max}]$ is the human comfort temperature zone. Indoor temperature depends on many factors including HVAC input power, outdoor temperature, and home thermal dynamics, which is dif-

ficult to model. However, thermal dynamic model for HVAC system is not required by the proposed MDRL/safe-MDRL algorithm because it can learn such dependence from experiences by trial-and-error. This demonstrates the advantage of model-free RL algorithm for HVAC system control.

C. BESS

Consider a BESS with the maximum capacity of $B_{\max}$. The dynamics of the BESS in terms of state of charge (SoC) is given by:

$$SoC_{t+1} = SoC_t + \frac{P_{B,t+1}\eta_B \Delta T}{B_{\max}} \tag{8}$$

where $SoC_t = B_t/B_{\max}$ is the level of available energy B_t with respect to BESS capacity; $P_{B,t+1}$ is the charging (if $P_{B,t+1} > 0$) or discharging (if $P_{B,t+1} < 0$) power; and η_B is the charging/discharging efficiency with $\eta_B = \eta_{B,c}$ for charging process and $\eta_B = 1/\eta_{B,d}$ for discharging process.

To sustain lifespan of the BESS, the following operation constraints are considered:

$$P_{B,\min}\eta_{B,d} \le P_{B,t} \le \frac{P_{B,\max}}{\eta_{B,c}} \tag{9}$$

$$SoC_{\min} \le SoC_t \le SoC_{\max} \tag{10}$$

where $P_{B,\min} < 0$ and $P_{B,\max} > 0$ are the limitations of charging and discharging power, respectively; and $SoC_{\min}$ and $SoC_{\max}$ are the minimum and maximum levels of SoC, respectively.

D. Energy Cost Minimization Problem

The home energy system exchanges energy with the utility grid to balance supply and demand:

$$P_{grid,t} = P_{non,t} + P_{shift,t} + P_{B,t} + P_{HVAC,t} - P_{PV,t} \tag{11}$$

where $P_{non,t}$, $P_{PV,t}$, and $P_{grid,t}$ are the power demand from non-shiftable loads, PV generation power, and power exchanged with utility grid, respectively. $P_{grid,t} > 0$ represents electricity purchased from the utility grid with TOU electricity price, while $P_{grid,t} \le 0$ represents surplus energy sold to the utility grid with fixed feed-in tariff (FT).

The operation cost of the home energy system for each time slot t is given by:

$$C_t = u_t P_{grid,t} \Delta T + v_B \left| P_{B,t} \right| \Delta T \tag{12}$$

where u_t is the electricity price; and v_B is the degradation cost coefficient of the BESS. In (12), the first term represents the electricity cost, while the second term represents the BESS degradation cost, which is proportional to charging/discharging power [34].

The objective of the scheduling problem is to minimize operation cost of the home energy system while maintaining human thermal comforts and satisfying constraints over scheduling horizon. Such optimization problem is summarized as:

$$\begin{cases} \min \sum_{t=1}^{T} C_t \\ \text{s.t. (1)-(12)} \end{cases} \tag{13}$$

Decision variables in (13) include $x_{n,t}$, $P_{HVAC,t}$, and $P_{B,t}$ for $t = 1, 2, \ldots, T$. It is a great challenge to solve the mixed-integer optimization problem due to the following difficulties. Firstly, due to the randomness of PV generation, power demand from non-shiftable loads, and outdoor temperature, it is difficult to make leading decisions. Secondly, indoor temperature is not only affected by input power of HVAC system but also highly depends on outdoor temperature and thermal properties of the home, while it is not easy to develop a proper model to describe such dependence. In this paper, DRL algorithms will be developed to solve the optimization problem without thermal dynamic model for HVAC system or prediction of unknown variables.

III. Safe-MDRL for Discrete-continuous Hybrid Action Space

RL is an area of machine learning concerned with how artificial agents take actions in an environment in order to maximize accumulative future rewards. The fundamental principle underlying RL is the MDP. In this section, the formulation of household sequential scheduling problem as an MDP will firstly be investigated, which is followed by the development of the MDRL algorithm and its safe version to solve the problem.

A. MDP

An MDP is usually defined by a 4-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is the state space consisting of a set of environment states; $\mathcal{A}$ is a set of actions called action space; $\mathcal{P}$: $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a function which determines the state transition probability considering environment uncertainty; and $\mathcal{R}$: $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function which returns immediate reward after state transition [35].

Considering the framework of MDP in Fig. 2, the agent represents the HEMS while the home energy system and other variables such as indoor/outdoor temperature constitute the environment. At each time slot t, the agent observes environment state s_t and takes action a_t following the proposed MDRL algorithm. With the execution of action a_t, the environment moves to a new state s_{t+1} and returns reward r_{t+1} associated with (s_t, a_t, s_{t+1}). Details on the MDP for the HEMS are as follows.

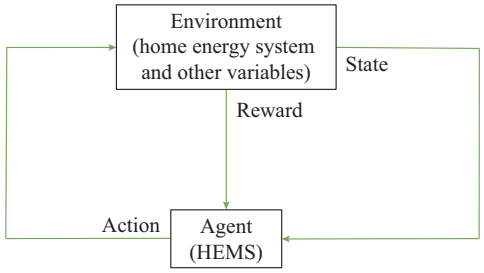

Fig. 2. Framework of MDP.

1) State: the state s_t is composed of information available at the end of time slot t, which reflects the status of components in home energy systems. It is defined by a high dimensional vector $\{h, SoC_t, K_{in,t}, T_{n,t}, P_{PV,t}, P_{PV,t-1}, P_{non,t}, P_{non,t-1}, K_{out,t}, K_{out,t-1}\}$, where h denotes hour of day for time slot t. Lagged values of PV generation, non-shiftable loads, and outdoor tempera-

ture $K_{out,t}$ are considered to capture their patterns of variation.

2) Action: the agent receives state s_t at the end of time slot t and takes control actions $a_t=\{x_{1,t+1},x_{2,t+1},\ldots,x_{N,t+1},P_{B,t+1},P_{HVAC,t+1}\}$ following a policy. The action vector determines the operation of the home energy system for time slot $t+1$. It is noticeable that the action vector consists of both discrete action and continuous action. To ensure non-violation of SoC constraints, $P_{B,t+1}$ should be bounded to $[0,\min\{(SoC_{\max}-SoC_t)B_{\max}/(\Delta T\eta_{B,c}),P_{B,\max}\}]$ for charging process and to $[\max\{(SoC_{\min}-SoC_t)B_{\max}\eta_{B,d}/\Delta T,P_{B,\min}\},0]$ for discharging process.

3) State transition: the transitions of SoC_t and $T_{n,t}$ have been discussed in Section II. The transitions of state features including PV generation, non-shiftable loads, and outdoor temperature are random, while indoor temperature depends not only on actions but also on outdoor temperature and home thermal properties. The values of these features indexed at $t+1$ will be taken from observations. The developed DRL algorithms will learn their correlations from the training data to make optimal decisions.

4) Reward: the objective of the HEMS is to minimize operation costs while maintaining human thermal comforts considering constraints. Hence, the reward consisting of operation cost and penalty for temperature deviation from comfort zone is given by:

$$r_{t+1}=-C_{t+1}-\beta\Delta K_{in,t+1} \tag{14}$$

where $\Delta K_{in,t+1}=\max\{0,K_{in,t+1}-K_{\max}\}+\max\{0,K_{\min}-K_{in,t+1}\}$; and β is a parameter which balances the operation cost and penalty for temperature deviation.

5) State-action value function: the goal of the agent in RL is to construct an optimal policy π^* that maximizes accumulated discounted rewards in the future, i. e., $R_t=\sum_{i=1}^{\infty}\lambda^{i-1}r_{t+i}$ [36]. The discounted factor $\lambda\in[0,1]$ balances the importance between immediate reward and future rewards. Let $Q_\pi(s,a)$ denote state-action value function under a policy π that estimates the expected accumulated discounted rewards R_t by taking action $a_t=a$ in state $s_t=s$ following the policy π, i. e., $Q_\pi(s,a)=\mathbb{E}_\pi\left(R_t\middle|s_t=s,a_t=a\right)$. The optimal policy π^* can be derived from the optimal Q-values by selecting the action leading to the highest Q-value with the given state, i. e., $Q^*(s,a)=\max_\pi Q_\pi(s,a)$. Moreover, the Q-value can be derived from Bellman equation in a recursive manner as in (15) [37], which sets the foundation of RL.

$$Q(s,a)=\mathbb{E}\left(r_{t+1}+\lambda\max_{a'}Q\left(s_{t+1},a'\right)\middle|s_t=s,a_t=a\right) \tag{15}$$

where $Q(s,a)$ is the state-action value; $\mathbb{E}$ is a function that returns expected value; and a' is the action to be taken at the following time step.

From the above analysis, this paper develops a DRL-based algorithm for one-step ahead control of the home energy system based on currently available information. The underlying principle of using currently available measurements of PV generation, outdoor temperature, and non-shiftable loads instead of their predictions is that these values are highly temporally correlated and their temporal evolution can be learned by the proposed MDRL algorithm. Moreover, the dependence of indoor temperature variation on controlled HVAC power, outdoor temperature, and building thermal property is also learned from experiences by trail-and-error in the learning stage. Hence, the proposed MDRL algorithm does not need thermal dynamic model for HVAC system or prediction of unknown variables.

B. MDRL Algorithm

For the existing DRL algorithms, most of them require action space to be either discrete or continuous. For instance, DQL as well as its variants are applicable for discrete action space; while DDPG is widely used for continuous action space. To handle the discrete-continuous hybrid action space with the HEMS, an MDRL algorithm that integrates DQL and DDPG is developed.

Let $a_d\in A_d$ and $a_c\in A_c$ denote the discrete action and continuous action, respectively, where A_d and A_c denote the discrete action space and continuous action space, respectively. The discrete-continuous hybrid action is represented by $a=\{a_d,a_c\}$. Then Bellman equation becomes:

$$Q\left(s,a_d,a_c\right)=\mathbb{E}\left(r_{t+1}+\lambda\max_{a'_d,a'_c}Q\left(s_{t+1},a'_d,a'_c\right)\middle|s_t=s,a_{d,t}=a_d,a_{c,t}=a_c\right) \tag{16}$$

where $a_{d,t}$ and $a_{c,t}$ are the discrete and continuous actions at time slot t, respectively; a'_d and a'_c are the discrete and continuous actions to be taken at the following time slot, respectively.

If $a'^*_d=\arg\max_{a'_d}Q\left(s,a'_d,a_c\right)$ holds, (16) can be re-written as:

$$Q\left(s,a_d,a_c\right)=\mathbb{E}\left(r_{t+1}+\lambda\max_{a'_c}Q\left(s_{t+1},a'^*_d,a'_c\right)\middle|s_t=s,a_{d,t}=a_d,a_{c,t}=a_c\right) \tag{17}$$

It is noticeable that the right side of (17) deals with continuous action only, which can be efficiently handled by actor-critic framework. Similar to DDPG, a deep critic network $Q\left(s,a_d,a_c;\theta\right)$ is deployed to approximate state-action value function while a deterministic deep policy network $\mu\left(s,a_d;\varphi\right)$ is used to generate continuous action $a_c=\mu\left(s,a_d;\varphi\right)$, where θ and φ are the corresponding network parameters including weights and biases.

The illustration of networks of MDRL algorithm is depicted in Fig. 3. In this way, the optimal discrete action can be easily reached by searching the discrete action space, i. e., $a^*_d=\arg\max_{a_d\in A_d}Q\left(s,a_d,\mu\left(s,a_d;\varphi\right);\theta\right)$. The selection of discrete action corresponding to the highest Q-value is identical to DQL. Hence, the proposed MDRL algorithm inherits the merits of both DDPG and DQL. To facilitate the search of optimal discrete action, the constraints in (2)-(5) associated with state s can be used to depress discrete action space into $A_d(s)\subset A_d$. Thereby, the proposed MDRL agorithm always

satisfies the constraints associated with shiftable and non-interruptible loads and will not cause any discomfort.

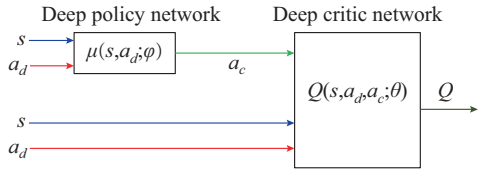

Fig. 3. Illustration of networks of MDRL algorithm.

Similar to DDPG, the critic network parameter θ is optimized by minimizing the squared loss L_θ in (18) with gradient descent methods [37].

$$L_\theta = \frac{1}{2}\mathbb{E}\left(Q\left(s_t, a_{d,t}, a_{c,t};\theta\right) - y_t\right)^2 \tag{18}$$

where $y_t = r_{t+1} + \lambda \max_{a_{d,t+1}\in A_d\left(s_{t+1}\right)} Q\left(s_{t+1}, a_{d,t+1}, \mu\left(s_{t+1}, a_{d,t+1};\varphi\right);\theta\right)$ is the target Q-value. To optimize the actor network parameter, the basic idea is to adjust φ in the direction of the performance gradient $\nabla_\varphi Q\left(s_t, a_{d,t}, \mu\left(s_t, a_{d,t};\varphi\right);\theta\right)$ that boosts Q-value. With the application of chain rule, the performance gradient can be decomposed into gradient of state-action value function with respect to continous actions and gradient of policy with respect to policy parameters, which results in policy gradient $\nabla_\varphi J$ for the update of policy parameters considering state distribution $\rho^\mu(s)$ [38].

$$\nabla_\varphi J = \mathbb{E}_{s\sim\rho^\mu}\left(\nabla_\varphi \mu\left(s, a_d;\varphi\right)\cdot\nabla_{a_c} Q\left(s, a_d, a_c;\theta\right)\Big|_{a_c=\mu\left(s,a_d;\varphi\right)}\right) \tag{19}$$

In DRL, the balance between exploration and exploitation is critical to train an efficient agent for decision-making. To facilitate the training of deep networks considering discrete-continuous hybrid action space, a special exploration policy in (20) which integrates the ε-greedy policy for DQL and the policy by adding Gaussian noise $\mathcal{N}\left(0,\delta^2 I\right)$ into the actions from actor network for DDPG is developed.

$$a_t = \begin{cases} a_{d,t} \text{ uniformly sampled from } A_d\left(s_t\right) \\ a_{c,t} \text{ uniformly sampled from } A_c \end{cases} \quad rand \le \varepsilon$$
$$\begin{cases} a_{d,t} = \arg\max_{a_{d,t}\in A_d\left(s_t\right)} Q\left(s_t, a_{d,t}, \mu\left(s_t, a_{d,t};\varphi\right);\theta\right) \\ a_{c,t} = \mu\left(s_t, a_{d,t};\varphi\right) + N\left(0,\delta^2 I\right) \end{cases} \quad \text{otherwise} \tag{20}$$

To handle the challenges caused by temporal correlation of samples for network optimization in DRL, experience reply is considered [36], [37]. Tuples $(s_t, a_t, s_{t+1}, r_{t+1})$ are stored in a reply buffer $\mathcal{M}$ with size of M, where the oldest ones are dropped when the buffer is full. At each time step, a mini-batch of B tuples are uniformlly sampled for the update of networks.

To stabilize the learning process, target networks are introduced for actor network and critic network, denoted as $\mu'\left(s, a_d;\varphi'\right)$ and $Q'\left(s, a_d, a_c;\theta'\right)$, respectively, to evaluate the target Q-value [37]. The parameters of target networks are updated with soft update strategy in (21).

$$\begin{cases} \theta' \leftarrow \tau\theta' + (1-\tau)\theta \\ \varphi' \leftarrow \tau\varphi' + (1-\tau)\varphi \end{cases} \tag{21}$$

where $\tau \ll 1$ ensures slow change of target network parameters, and consequently improves the stability of learning process. Procedures for the training of networks are summarized in Algorithm 1, which include the initialization of networks and the main loop of training process. In the main loop, each day constitutes an episode. In each time slot of an episode, the agent receives state s_t and selects action a_t according to the exploration policy in (20). With the executation of action, the state moves to s_{t+1} and the reward r_{t+1} is obtained. The tuples $(s_t, a_t, s_{t+1}, r_{t+1})$ are then stored in the replay buffer. Next, B tuples uniformly sampled from the replay buffer are used to update θ and φ based on sampled mean squared loss and policy gradient. This is followed by soft update of target networks.

Algorithm 1: training for MDRL

1. Initialize the actor network and the critic network with random weights φ and θ, respectively
2. Initialize the target networks by copying $\theta' \leftarrow \theta$ and $\varphi' \leftarrow \varphi$
3. Initialize the buffer $\mathcal{M}$
4. **for** $e = 1{:}E$
5. Obtain the initial state s_0 from a random day with random SoC and K_{in}
6. **for** $t = 0{:}23$ **do**
7. Select action $a_t = \left\{a_{d,t}, a_{c,t}\right\}$ according to the exploration policy in (20)
8. Execute action a_t, observe reward r_{t+1}, and move to next state s_{t+1}
9. Store tuple $\left(s_t, a_t, s_{t+1}, r_{t+1}\right)$ in $\mathcal{M}$
10. Sample B tuples $\left(s_b, a_b, s_{b+1}, r_{b+1}\right)$ for $b = 1, 2, ..., B$ from $\mathcal{M}$
11. Obtain target Q-values: $y_b = r_{b+1} + \lambda \max_{a_{d,b+1}} Q\left(s_{b+1}, a_{d,b+1}, \mu\left(s_{b+1}, a_{d,b+1};\varphi'\right);\theta'\right)$
12. Update θ by minimizing the loss $L_\theta = \frac{1}{2B}\sum_{b=1}^{B}\left(Q\left(s_b, a_{d,b}, a_{c,b};\theta\right) - y_b\right)^2$
13. Update φ with the sampled policy gradient: $\frac{1}{B}\sum_{b=1}^{B}\left\{\nabla_\varphi \mu\left(s, a_d;\varphi\right)\Big|_{s=s_b, a_d=a_{d,b}} \nabla_{a_c} Q\left(s, a_d, a_c;\theta\right)\Big|_{s=s_b, a_d=a_{d,b}, a_c=\mu\left(s,a_d;\varphi\right)}\right\}$
14. Soft update of target networks
15. **end for**
16. **end for**

C. Safe-RL

The fundamental idea of safe-RL is to develop a prediction model for action evaluation where safe actions are executed by the system while unsafe actions are modified to satisfy safe constraints. In this paper, indoor temperature is expected to stay in comfort zone with well-controlled HVAC input power. Thereby, unsafe actions refer to those that will lead to violation of constraints on indoor temperature. To ensure thermal comfort, an indoor temperature prediction model $f_{K_{in}}$ based on multilayer perception (MLP) is developed for HVAC input power evaluation.

$$K_{in,t+1} = f_{K_{in}}\left(K_{in,t}, K_{out,t+1}, P_{HVAC,t+1}\right) + e \tag{22}$$

The model in (22) predicts indoor temperature from the most influential factors including lagged indoor temperature, outdoor temperature, and HVAC input power. The term e captures modeling error due to unconsidered weather conditions such as wind speed and humidity as well as uncertainty associated with thermal dynamic process.

Since leading outdoor temperature $K_{out,t+1}$ is usually unknown at time slot t, a probabilistic outdoor temperature prediction model f_{GPR} based on Gaussian process regression [39] is developed.

$$\left\{\bar{K}_{out,t+1}, \delta_{out,t+1}\right\} = f_{GPR}\left(K_{out,t}, K_{out,t-1}, \sin\left(2\pi\frac{h}{24}\right), \cos\left(2\pi\frac{h}{24}\right)\right) \quad (23)$$

The model in (23) predicts the mean value $\bar{K}_{out,t+1}$ and standard deviation $\delta_{out,t+1}$ of outdoor temperature from its lagged values and temporal information h. Outdoor temperature illustrates the diurnal cycle that the sine and cosine functions are used to capture temporal periodicity. The input features are contained in the state s_t, hence, outdoor temperature prediction model is simplied as:

$$\left\{\bar{K}_{out,t+1}, \delta_{out,t+1}\right\} = f_{GPR}\left(s_t\right) \quad (24)$$

With (22) and (24), it is easy to construct outdoor temperature prediction interval $\left[K^{low}_{out,t+1}, K^{up}_{out,t+1}\right]$ and indoor temperature prediction interval $\left[K^{low}_{in,t+1}, K^{up}_{in,t+1}\right]$:

$$K^{low}_{out,t+1} = \bar{K}_{out,t+1} - \eta\delta_{out,t+1} \quad (25)$$

$$K^{up}_{out,t+1} = \bar{K}_{out,t+1} + \eta\delta_{out,t+1} \quad (26)$$

$$K^{low}_{in,t+1} = f_{K_{in}}\left(K_{in,t}, K^{low}_{out,t+1}, P_{HVAC,t+1}\right) \quad (27)$$

$$K^{up}_{in,t+1} = f_{K_{in}}\left(K_{in,t}, K^{up}_{out,t+1}, P_{HVAC,t+1}\right) \quad (28)$$

where η is a parameter which controls the confidence level that actual outdoor temperature falls in the constructed interval.

The safety-checking function f_{sc} in Algorithm 2 is developed for action evaluation and modification associated with the HVAC system. The idea of Algorithm 2 is that the input power is modified if $K^{low}_{in,t+1}$ is greater than the upper limit of comfort temperature zone or $K^{up}_{in,t+1}$ is lower than the lower limit of comfort temperature zone; otherwise, modification is not required.

Algorithm 2: $\tilde{a}_{c,t} = f_{sc}\left(s_t, a_{c,t}, f_{K_{in}}, f_{GPR}\right)$

Step 1: obtain outdoor temperature prediction interval $\left[K^{low}_{out,t+1}, K^{up}_{out,t+1}\right]$ with (24)-(26)

Step 2: obtain $K_{in,t}$ from s_t and $\{P_{B,t+1}, P_{HVAC,t+1}\}$ from $a_{c,t}$

Step 3: obtain indoor temperature prediction interval $\left[K^{low}_{in,t+1}, K^{up}_{in,t+1}\right]$ with (27) and (28)

Step 4: **if** $K^{low}_{in,t+1} > K_{\max} + \rho$ **then**
$P_{HVAC,t+1} = P_{HVAC,t+1} - \alpha$
Go to *Step 3*

Step 5: **else if** $K^{up}_{in,t+1} < K_{min} - \rho$ **then**
$P_{HVAC,t+1} = P_{HVAC,t+1} + \alpha$
Go to *Step 3*

Step 6: **end if**

Step 7: output $\tilde{a}_{c,t} = \{P_{B,t+1}, P_{HVAC,t+1}\}$

The parameter α ($\alpha > 0$ for heating system and $\alpha < 0$ for cooling system) denotes the moving step of HVAC input power and the parameter ρ compensates modeling errors. In Algorithm 2, outdoor temperature prediction model f_{GPR} is trained offline while indoor temperature prediction model $f_{K_{in}}$ is trained and renewed online in accordance with learning process. The output $\tilde{a}_{c,t}$ will be applied for home energy system control.

IV. Simulation Results

A. Simulation Setup

1) Home energy system: normalized PV generation and outdoor temperature obtained from National Renewable Energy Laboratory (NREL), USA [40], are considered for simulation studies. Simulated hourly residential loads based on Building America House Simulation Protocols is used to represent non-shiftable loads [41]. Dish washer and washing machine are considered to represent shiftable and non-interruptible loads. This paper considers electrical HVAC system for heating in cold winter. To simplify the simulation study, a mathematical model in (29) is used to simulate the dynamics of indoor temperature [42], [43].

$$K_{in,t+1} = \omega K_{in,t} + (1-\omega)\left(K_{out,t+1} + \frac{\eta_{HVAC}}{\xi}P_{HVAC,t+1}\right) \quad (29)$$

where $\omega = 0.93$ [43], $\eta_{HVAC} = 2.5$ [43], and $\xi = 0.14$ [21] are the factor of air inertial, coefficient of HVAC performance, and thermal conductivity, respectively. Comfort temperature zone is considered to be [66.2 °F, 75.2 °F] or [19 ℃, 24 ℃] as in [21].

Outdoor temperature prediction model is trained on the data from December 2011 to February 2012. The MDRL algorithm and safe-MDRL algorithm are trained on the data from December 2012 to January 2013 and tested on data in February 2013. The parameters for the home energy system are listed in Table I and TOU electricity prices are given in Table II.

TABLE I
Parameters for Home Energy System

Component	Parameter	Value
PV	$P_{PV,r}$	5.6 kW
BESS	$(B_{\max}, v_B)$	(12 kWh, 0.01 \$/kWh)
	$(P_{B,\min}, P_{B,\max})$	(–4 kW, 4 kW)
	$(SoC_{\min}, SoC_{\max})$	(0.1, 0.9)
	$(\eta_{B,d}, \eta_{B,c})$	(0.98, 0.98)
HVAC	$(P_{HVAC,\max}, \beta)$	(4 kW, 0.7 \$/°F)
Dish washer	$(T_{n,ini}, T_{n,end}, T_{n,d}, P_n)$	(08:00, 22:00, 2 hours, 1.2 kW)
Washing machine	$(T_{n,ini}, T_{n,end}, T_{n,d}, P_n)$	(07:00, 22:00, 3 hours, 1.5 kW)
Grid	FT	0.067 \$/kWh

The profiles of PV generation and outdoor temperature in February 2013 are illustrated in Fig. 4. As can be observed from Fig. 4, PV generation and outdoor temperature illus-

trate significant fluctuations, which imposes great challenge to derive optimal actions.

TABLE II
TOU ELECTRICITY PRICES

Period	Price ($/kWh)
00:00-06:00	0.067
06:00-08:00; 12:00-15:00; 22:00-24:00	0.140
08:00-12:00; 15:00-22:00	0.250

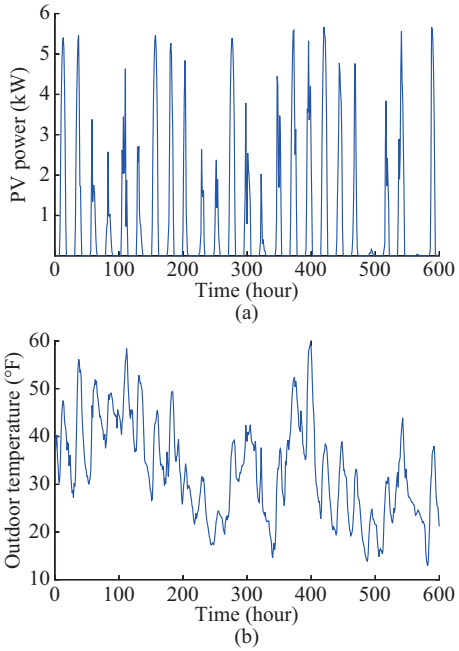

Fig. 4. Profiles of PV generation and outdoor temperature. (a) PV generation. (b) Outdoor temperature.

2) DRL algorithms: deep neural networks consisting of input layer, hidden layers, and output layer are considered. Rectified linear unit (ReLU) activation function is used for hidden layers of both actor network and critic network; while hyperbolic tangent activation function and linear activation function are used for the output layers of actor network and critic network, respectively. Adam optimizer [44] is deployed for the training of deep networks. Critical parameters of deep networks including the number of hidden layers and the number of hidden neurons, parameters associated with the optimizer, and parameters in Algorithm 1 are given in Table III.

TABLE III
CRITICAL PARAMETERS OF DEEP NETWORKS

Module	Parameter	Value
Actor or critic	Number of hidden layers	2
	Number of hidden neurons	(128, 64)
Optimizer	Learning rate	10^{-4} (actor), 10^{-3} (critic)
Algorithm 1	(λ, τ, E, M, B)	(0.995, 10^{-3}, 10^{4}, 10^{4}, 240)

To facilitate the training of deep neural networks, states are normalized into [0, 1]. The outputs from actor network are in [−1, 1] and should be mapped into the range of continuous action space. For the exploration policy in (20), parameters ε and δ decay with training episode as $\varepsilon_e = \max(0.1, 1 - e/E)$ and $\delta_e = \max(0.01, 1 - e/E)$. Indoor temperature prediction model is represented by an MLP with one hidden layer. There are three neurons in the hidden layer. Hyperbolic tangent activation function and linear activation function are used for hidden layer and output layer, respectively. Parameters associated with safe-MDRL in Algorithm 2 are set as $\rho = 0.1(K_{\max} - K_{\min})$ and $\alpha = 0.01P_{HVAC,\max}$.

B. Benchmark Algorithms

This paper considers the following benchmark algorithms to illustrate the effectiveness of the proposed MDRL/safe-MDRL algorithm for home energy management with discrete-continuous hybrid action space.

1) B1: the "on/off" operation modes are considered by this benchmark algorithm. With this benchmark algorithm, the shiftable and non-interruptible load is switched "on" at its initial working time and maintains "on" until the completion of the task. The HVAC system is turned "on" with the maximum power if $K_{in,t} < K_{\min}$ and turned "off" if $K_{in,t} > K_{\max}$; otherwise, it maintains its operation mode. However, this benchmark algorithm does not consider BESS.

2) B2: an algorithm based on MILP is developed for the scheduling of home energy system supposing that all the information including PV generation, outdoor temperature, non-shiftable loads, and home thermal dynamics are known. This is an ideal case that sets the lower limit in energy cost while keeping thermal comforts.

3) DDPG algorithm: classical DDPG algorithm is applied for the home energy system control where discretization is used to derive the decisions for shiftable and non-interruptible loads. The studies in [21], [25] have illustrated that DDPG algorithm outperforms DQL for continuous control in home energy management. Hence, DQL is not considered in this study. The comparison of this benchmark algorithm against B1 will illustrate the advantage of the BESS in reducing energy cost. More importantly, based on performance comparison between the proposed MDRL algorithm and this benchmark algorithm, the merits of the proposed MDRL algorithm in handling discrete-continuous hybrid action space can be observed.

C. Simulation Results

The objective of simulation study is twofold: ① through the comparison between the proposed MDRL algorithm and its safe version to illustrate the effectiveness of the safe-MDRL algorithm in reducing the loss of human thermal comfort in the learning stage; and ② through the comparison among all the applied algorithms to illustrate the merits of the proposed MDRL algorithm and its safe version in home energy management in terms of operation cost and satisfaction of human comforts on the test dataset. To verify their robustness, the DDPG algorithm, the MDRL algorithm, and the safe-MDRL algorithm are executed for 5 independent runs.

1) To illustrate the effectiveness of the safe-MDRL algo-

rithm in reducing the loss of human thermal comfort thereby in improving rewards, average episode rewards over 5 runs by the proposed MDRL algorithm and the safe-MDRL algorithm during the training process are depicted in Fig. 5. For the first few thousands of episodes, the agent of the proposed MDRL algorithm is in its early learning stage with large probability of taking inappropriate action, which gives rise to low rewards with significant fluctuations. The reward gradually increases with the growing number of training episodes and finally converges with slight oscillations due to randomness associated with the exploration policy and the random environment such as PV generation, outdoor temperature, and non-shiftable loads. With the safe-MDRL algorithm, safety checking procedures in Algorithm 2 are activated after few dozens of episodes (60 episodes) to obtain sufficient data for online training of indoor temperature prediction model. Compared with the proposed MDRL algorithm, the reward is greatly improved with much smaller oscillations by the safe-MDRL algorithm even at the early training stage. This demonstrates the effectiveness of safe-MDRL in improving rewards in the learning stage.

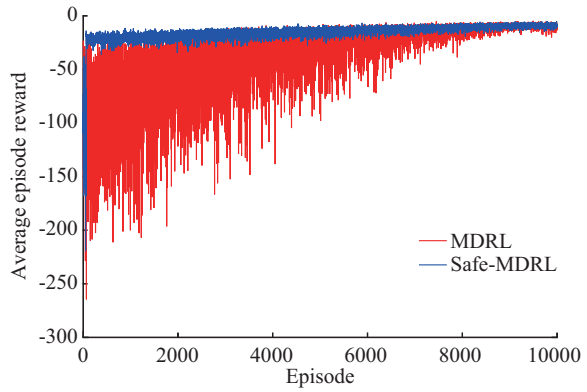

Fig. 5. Average episode rewards during training process.

To further illustrate the effectiveness of the safe-MDRL algorithm in maintaining thermal comforts thereby in improving rewards, average episode operation cost (including electricity cost and battery degradation cost) and temperature deviation from comfort zone for the first 2500, 5000, 7500, and 10000 episodes over 5 runs are reported in Table IV.

TABLE IV
AVERAGE EPISODE OPERATION COST AND TEMPERATURE DEVIATION FROM COMFORT ZONE

Iteration	Cost ($)		Temperature deviation (°F)	
	MDRL	Safe-MDRL	MDRL	Safe-MDRL
2500	11.53	11.63	80.77	14.64
5000	11.14	11.22	59.95	11.32
7500	10.64	10.70	44.65	9.03
10000	10.16	10.20	34.17	7.14

It can be observed that both the MDRL algorithm and the safe-MDRL algorithm improve decision quality in term of operation cost and thermal comfort with increasing number of training episodes. The difference in operation cost between the MDRL algorithm and its safe version is minor. The safe-MDRL algorithm reduces temperature deviation from comfort zone by almost 80% compared with the proposed MDRL algorithm and, thereby greatly improves rewards.

2) The statistics (mean value and standard deviation) over 5 runs on average daily operation cost and temperature deviation from comfort zone by the proposed algorithms and benchmark algorithms on the test dataset are presented in Table V.

TABLE V
STATISTICS ON AVERAGE DAILY OPERATION COSTS AND TEMPERATURE DEVIATION FROM COMFORT ZONE

Algorithm	Cost ($)	Temperature deviation (°F)
B1	10.93	5.503
B2	6.51	0
DDPG (mean)	8.73	0.108
DDPG (standard)	0.19	0.114
MDRL (mean)	8.11	0.058
MDRL (standard)	0.16	0.032
Safe-MDRL (mean)	8.13	0.042
Safe-MDRL (standard)	0.14	0.033

From Table V, it can be observed that the MDRL algorithm and the safe-MDRL algorithm outperform classical DDPG algorithm and B1 with reduced operation cost and improved human thermal comforts. More precisely, the MDRL algorithm saves operation cost by 25.8% and 7.1% against B1 and the DDPG algorithm, respectively. The outstanding performance of the MDRL algorithm over the DDPG algorithm can be explained by following factors: ① the treatment of discrete action as continuous action augments and complicates decision space; and ② the discretization of outputs from actor network to derive discrete actions impairs decision quality. The comparison between the MDRL algorithm and the safe-MDRL algorithm illustrates that neither of them dominates the other on the blind test dataset. The MDRL algorithm slightly outperforms its safe version in terms of cost; conversely, the safe-MDRL algorithm performs better on temperature violation. The comparison of MDRL/safe-MDRL algorithm against B1 also illustrates that the application of BESS and advanced optimization methods can reduce home energy cost and improve human thermal comforts. However, there is a gap on the cost between the MDRL/safe-MDRL algorithm and B2 due to randomness of PV generation, outdoor temperature, and non-shiftable loads which are difficult to be exactly captured by deep networks with the MDRL/safe-MDRL algorithm. The B2 provides theoretical optimal decisions supposing that accurate predictions of PV generation, outdoor temperature, and non-shiftable loads are available before decision-making. However, such assumption does not hold in practice. With the MDRL/safe-MDRL algorithm, the artificial agent strives to make leading decisions based on current observations. However, the random nature of PV generation and outdoor temperature makes it difficult to exactly capture their temporal evolution.

Hence, it's not surprising that a gap on the cost between the MDRL/safe-MDRL algorithm and B2 can be observed.

Temperature deviations from comfort zone are observed with the DDPG algorithm, the MDRL algorithm, and the safe-MDRL algorithm. This is because indoor temperature dynamic model in (29) considers the impact of uncertainty of outdoor temperature on indoor temperature. At the end of time slot t when the decision on $P_{HVAC,t+1}$ is issued, outdoor temperature $K_{out,t+1}$ is actually unknown. The proposed MDRL/safe-MDRL algorithm learns to handle the challenge, however, it cannot be fully addressed in the extreme cases where large variation of outdoor temperature occurs.

Figure 6 illustrates simulation results obtained by the proposed algorithms and benchmark algorithms. It can be observed that the indoor temperature, SoC of BESS, HVAC input power, and grid power obtained by the DDPG algorithm, MDRL algorithm, and safe-MDRL algorithm generally capture the trend of the results obtained by B2.

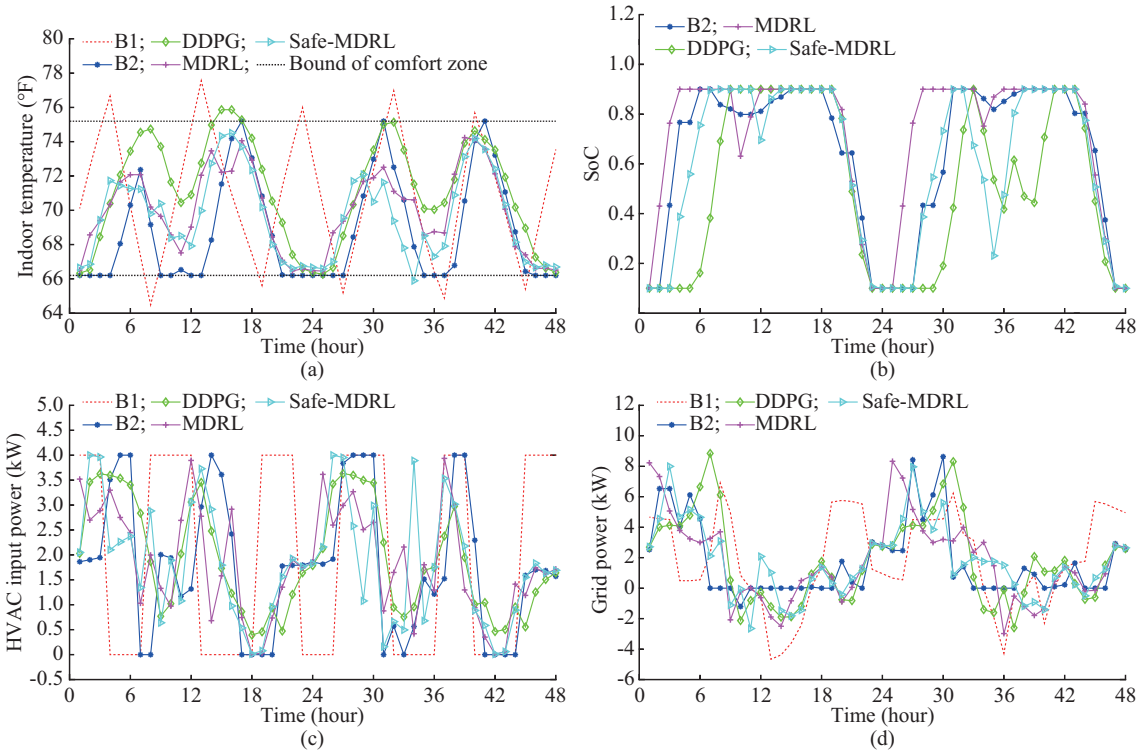

Fig. 6. Illustration of simulation results. (a) Indoor temperature. (b) SoC of BESS. (c) HVAC input power. (d) Grid power.

From Fig. 6(a), it can be observed that the indoor temperature obtained by the DDPG algorithm, MDRL algorithm, and safe-MDRL algorithm generally lies in the comfort zone while large temperature deviation obtained by B1 is observed. From Fig. 6(b), it can be observed that the BESS is charged during valley hours (the 1th-6th hour and the 25th-30th hour) when electricity price is low and is discharged during peak hours in the morning (the 9th-12th hour and the 33th-36th hour) when electricity price is high. During flat hours in the middle of the day when PV generation is high and electricity price is moderate, the BESS is charged again. In the late afternoon and early evening (the 19th-22th hour and the 43th-46th hour), the BESS is discharged to provide energy. From Fig. 6(c), it can be observed that the HVAC system operates at high power during valley hours and flat hours and its power is greatly reduced during peak hours. The PV system generates power in the daytime and its power generation usually arrives at peak in the middle of the day. The home energy system can make use of PV generation in the daytime considering that the power drawn from the grid is much lower in the daytime than in the evening and in some hours the surplus energy is sold to the grid, as illustrated in Fig. 6(d). With the above analysis, it is reasonable to say that the BESS and HVAC system take advantage of TOU electricity price and PV generation to reduce the operation cost of the home energy system while maintaining the human thermal comfort.

V. Conclusion

In this paper, a novel DRL-based algorithm is developed for home energy management under TOU pricing program. The operation modes of various household appliances constitute discrete-continuous hybrid action space, which challenges the existing RL frameworks for either discrete action space or continuous action space. The proposed MDRL algorithm integrates DQL and DDPG where the DQL deals with discrete action space and the DDPG handles continuous action space. To reduce the loss of human thermal comfort during the learning stage with the MDRL algorithm, a safe version (safe-MDRL algorithm) which deploys a prediction model to guide the exploration of the MDRL algorithm is further developed.

To verify the effectiveness of the MDRL algorithm in cost saving for home energy management and the safe-MDRL al-

gorithm in reducing the loss of human thermal comfort in the learning stage, simulation studies based on real data are conducted. The results illustrate that the MDRL algorithm can efficiently handle the challenges from discrete-continuous hybrid action space for the existing RL frameworks. Meanwhile, the MDRL algorithm reduces the operation cost while keeping human thermal comforts by comparing with benchmark algorithms including classical DDPG on the test dataset. Simulation results also illustrate that the safe-MDRL algorithm can greatly reduce the loss of human thermal comforts in the learning stage.

REFERENCES

[1] F. Zeng, Z. Bie, S. Liu *et al*., "Trading model combining electricity, heating, and cooling under multi-energy demand response," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 1, pp. 133-141, Jan. 2020.

[2] National Energy Administration. (2020, Jan.). The electricity consumption by the whole society. [Online]. Available: http://www.nea.gov.cn/2020-01/20/c_138720877.htm

[3] S. Xu, X. Chen, J. Xie *et al*., "Agent-based modeling and simulation for the electricity market with residential demand response," *CSEE Journal of Power and Energy Systems*, vol. 7, no. 2, pp. 368-380, Mar. 2021.

[4] F. Luo, W. Kong, G. Ranzi *et al*., "Optimal home energy management system with demand charge tariff and appliance operational dependencies," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 4-14, Jan. 2020.

[5] X. Wang, Y. Liu, J. Zhao *et al*., "A hybrid agent-based model predictive control scheme for smart community energy system with uncertain DGs and loads," *Journal of Modern Power Systems and Clean Energy*, vol. 9, no. 3, pp. 573-584, May 2021.

[6] S. Althaher, P. Mancarella, and J. Mutale, "Automated demand response from home energy management system under dynamic pricing and power and comfort constraints," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1874-1883, Jul. 2015.

[7] T. Yoshihisa, N. Fujita, and M. Tsukamoto, "A rule generation method for electrical appliances management systems with home EoD," in *Proceedings of the 1st IEEE Global Conference on Consumer Electronics 2012*, Tokyo, Japan, Oct. 2012, pp. 248-250.

[8] A. Keshtkar, S. Arzanpour, and F. Keshtkar, "Adaptive residential demand-side management using rule-based techniques in smart grid environments," *Energy and Buildings*, vol. 133, pp. 281-294, Dec. 2016.

[9] M. J. Sanjari, H. Karami, and H. B. Gooi, "Analytical rule-based approach to online optimal control of smart residential energy system," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 4, pp. 1586-1597, Aug. 2017.

[10] Y. Huang, L. Wang, W. Guo *et al*., "Chance constrained optimization in a home energy management system," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 252-260, Jan. 2018.

[11] T. Molla, B. Khan, B. Moges *et al*., "Integrated optimization of smart home appliances with cost-effective energy management system," *CSEE Journal of Power and Energy Systems*, vol. 5, no. 2, pp. 249-258, Jun. 2019.

[12] N. G. Paterakis, O. Erdinc, A. G. Bakirtzis *et al*., "Optimal household appliances scheduling under day-ahead pricing and load-shaping demand response strategies," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 6, pp. 1509-1519, Dec. 2015.

[13] M. Shafie-Khah and P. Siano, "A stochastic home energy management system considering satisfaction cost and response fatigue," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 2, pp. 629-638, Feb. 2018.

[14] M. Yousefi, A. Hajizadeh, M. N. Soltani *et al*., "Predictive home energy management system with photovoltaic array, heat pump, and plug-in electric vehicle," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 1, pp. 430-440, Jan. 2021.

[15] A. Mondal, S. Misra, and M. S. Obaidat, "Distributed home energy management system with storage in smart grid using game theory," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1857-1866, Sept. 2017.

[16] Q. Wei, D. Liu, and G. Shi, "A novel dual iterative *Q*-learning method for optimal battery management in smart residential environments," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 4, pp. 2509-2518, Apr. 2015.

[17] M. N. Faqiry, L. Wang, and H. Wu, "HEMS-enabled transactive flexibility in real-time operation of three-phase unbalanced distribution systems," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 6, pp. 1434-1449, Nov. 2019.

[18] R. Lu, S. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6629-6639, Nov. 2019.

[19] S. Bahraini, V. Wong, and J. Huang, "An online learning algorithm for demand response in smart grid," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4712-4725, Sept. 2018.

[20] Q. Wei, Z. Liao, and G. Shi, "Generalized actor-critic learning optimal control in smart home energy management," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 6614-6623, Oct. 2021.

[21] L. Yu, W. Xie, D. Xie *et al*., "Deep reinforcement learning for smart home energy management," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2751-2762, Apr. 2020.

[22] C. Qiu, Y. Hu, Y. Chen *et al*., "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8577-8588, Oct. 2019.

[23] D. Cao, W. Hu, J. Zhao *et al*., "Reinforcement learning and its applications in modern power and energy systems: a review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, pp. 1029-1042, Nov. 2020.

[24] E. Mocanu, D. Mocanu, P. Nguyen *et al*., "On-line building energy optimization using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3698-3708, Jul. 2019.

[25] Y. Ye, D. Qiu, X. Wu *et al*., "Model-free real-time autonomous control for a residential multi-energy system using deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3068-3082, Jul. 2020.

[26] M. Sun, I. Konstantelos, and G. Strbac, "A deep learning-based feature extraction framework for system security assessment," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5007-5020, Sept. 2019.

[27] H. Zhao, J. Zhao, J. Qiu *et al*., "Cooperative wind farm control with deep reinforcement learning and knowledge-assisted learning," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 11, pp. 6912-6921, Nov. 2020.

[28] J. Garcia and F. Fernandez, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, pp. 1437-1480, Aug. 2015.

[29] M. Wen and T. Ufuk, "Constrained cross-entropy method for safe reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 66, no. 7, pp. 3123-3137, Jul. 2021.

[30] L. Yu, Y. Sun, Z. Xu *et al*., "Multi-agent deep reinforcement learning for HVAC control in commercial buildings," *IEEE Transactions on Smart Grid*, vol. 12, no. 1, pp. 407-419, Jan. 2021.

[31] Y. Gao, W. Wang, J. Shi *et al*., "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 5357-5369, Nov. 2020.

[32] X. Xu, Y. Jia, Y. Xu *et al*., "A multi-agent reinforcement learning-based data-driven method for home energy management," *IEEE Transactions on Smart Grid*, vol. 11, no. 4, pp. 3201-3211, Jul. 2020.

[33] D. Zhang, S. Li, M. Sun *et al*., "An optimal and learning-based demand response and home energy management system," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1790-1801, Jul. 2016.

[34] H. Li, A. T. Eseye, J. Zhang *et al*., "Optimal energy management for industrial microgrids with high-penetration renewables," *Protection and Control of Modern Power Systems*, vol. 2, no. 1, p. 12, Apr. 2017.

[35] K. Arulkumaran, M. P. Deisenroth, M. Brundage *et al*. (2017, Aug.). A brief survey of deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1708.05866v2

[36] V. Mnih, K. Kavukcuoglu, D. Silver *et al*., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529-533, Feb. 2015.

[37] T. Lillicrap, J. Hunt, A. Pritzel *et al*. (2015, Sept.). Continuous control with deep reinforcement learning. [Online]. Available: https://arxiv.org/abs/1509.02971

[38] D. Silver, G. Lever, N. Heess *et al*., "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, Jun. 2014, pp. 387-395.

[39] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge: MIT Press, 2006.

[40] National Renewable Energy Laboratory. (2021, Mar.). PVDAQ. [Online]. Available: http://maps.nrel.gov/pvdaq

[41] E. Wilson. (2014, Nov.). Commercial and residential hourly load pro-

files for all TMY3 locations in the United States. [Online]. Available: https://data.openei.org/submissions/153
[42] N. Lu, "An evaluation of the HVAC load potential for providing load balancing service," *IEEE Transactions on Smart Grid*, vol. 3, no. 3, pp. 1263-1270, Sept. 2012.
[43] Y. Hong, J. Lin, C. Wu *et al*., "Multi-objective air-conditioning control considering fuzzy parameters using immune clonal selection programming," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1603-1610, Dec. 2012.
[44] D. P. Kingma and J. Ba. (2014, Dec.). Adam: a method for stochastic optimization. [Online]. Available: https://arxiv.org/abs/1412.6980

Chao Huang received the B.Eng. degree in electrical engineering and automation from Harbin Institute of Technology, Harbin, China, in 2011, the M.S. degree in intelligent transport system from University of Technology of Compiegne, Compiegne, France, in 2013, and the Ph.D. degree in systems engineering and engineering management from City University of Hong Kong, Hong Kong, China, in 2017. He is currently an Associate Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, China. From 2019 to 2021, he was a Post-doctoral Research Fellow with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macao S.A.R, China, under UM Macau Talent program. His research interests include data mining, computational intelligence, and energy informatics.

Hongcai Zhang received the B.S. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2013 and 2018, respectively. He is currently an Assistant Professor with the State Key Laboratory of Internet of Things for Smart City and Department of Electrical and Computer Engineering, University of Macau, Macao S.A.R, China. In 2018-2019, he was a Postdoctoral Scholar with the Energy, Controls, and Applications Lab at University of California, Berkeley, USA, where he also worked as a Visiting Student Researcher in 2016. His current research interests include Internet of Things for smart energy, optimal operation and optimization of power and transportation systems, and grid integration of distributed energy resources.

Long Wang received the M.S. degree in computer science with distinction from University College London, London, UK, in 2014, and the Ph.D. degree in systems engineering and engineering management from City University of Hong Kong, Hong Kong, China, in 2017. He is currently an Associate Professor with the University of Science and Technology Beijing, Beijing, China. He is an Associate Editor for IEEE Access and an Academic Editor for PLoS One. His research interests include machine learning, computational intelligence, and computer vision.

Xiong Luo received the Ph.D. degree in computer applied technology from Central South University, Changsha, China, in 2004. He is currently a Professor with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China. He has published extensively in his areas of interest in several journals such as IEEE Transactions on Industrial Informatics and IEEE Transactions on Human-Machine Systems. His research interests include machine learning, cloud computing, and computational intelligence.

Yonghua Song received the B.E. and Ph.D. degrees from the Chengdu University of Science and Technology, Chengdu, China, and the China Electric Power Research Institute, Beijing, China, in 1984 and 1989, respectively, all in electrical engineering. He was awarded D.Sc. by Brunel University, London, UK, in 2002, honorary D.Eng. by University of Bath, Bath, UK, in 2014, and honorary D.Sc. by University of Edinburgh, Edinburgh, UK, in 2019. From 1989 to 1991, he was a Post-doctoral Fellow at Tsinghua University, Beijing, China. He then held various positions at Bristol University, Bristol, UK; University of Bath; and John Moores University, Liverpool, UK, from 1991 to 1996. In 1997, he was a Professor of power systems at Brunel University, where he was a Pro-vice Chancellor for Graduate Studies since 2004. In 2007, he took up a Pro-vice Chancellorship and Professorship of electrical engineering at the University of Liverpool, Liverpool, UK. In 2009, he joined Tsinghua University as a Professor of electrical engineering and an Assistant President and the Deputy Director of the Laboratory of Low-carbon Energy. During 2012 to 2017, he worked as the Executive Vice President of Zhejiang University, Hangzhou, China, as well as Founding Dean of the International Campus and Professor of electrical engineering and higher education of the university. Since 2018, he became Rector of the University of Macau, Macao S.A.R, China, and the Director of the State Key Laboratory of Internet of Things for Smart City. He was elected as the Vice President of Chinese Society for Electrical Engineering (CSEE) and appointed as the Chairman of the International Affairs Committee of the CSEE in 2009. In 2004, he was elected as a Fellow of the Royal Academy of Engineering, UK. In 2019, he was elected as a Foreign Member of the Academia Europaea. His current research interests include smart grid, electricity economics, and operation and control of power systems.