A deep reinforcement learning-based approach for the residential appliances scheduling

Li, Sichen; Cao, Di; Huang, Qi; Zhang, Zhenyuan; Chen, Zhe; Blaabjerg, Frede; Hu, Weihao

Published in:
Energy Reports

DOI (link to publication from Publisher):
10.1016/j.egyr.2022.02.181

Creative Commons License
CC BY-NC-ND 4.0

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Li, S., Cao, D., Huang, Q., Zhang, Z., Chen, Z., Blaabjerg, F., & Hu, W. (2022). A deep reinforcement learning-based approach for the residential appliances scheduling. *Energy Reports*, *8*(5), 1034-1042. https://doi.org/10.1016/j.egyr.2022.02.181

Available online at www.sciencedirect.com

ScienceDirect

Energy Reports 8 (2022) 1034–1042

www.elsevier.com/locate/egyr

2021 The 2nd International Conference on Power Engineering (ICPE 2021), December 09–11, 2021, Nanning, Guangxi, China

A deep reinforcement learning-based approach for the residential appliances scheduling

Sichen Li[a], Di Cao[a,*], Qi Huang[a,b], Zhenyuan Zhang[a], Zhe Chen[c], Frede Blaabjerg[c], Weihao Hu[a]

[a] *School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China*
[b] *School of Energy, Chengdu University of Technology, Chengdu, China*
[c] *Department of Energy Technology, Aalborg University, Aalborg, Denmark*

Received 10 February 2022; accepted 17 February 2022
Available online xxxx

Abstract

This paper investigates the optimal real-time residential appliances scheduling of individual owner when participating in the demand response (DR) program. The proposed method is novel since we cast the optimization problem to an intelligent deep reinforcement learning (DRL) framework, which avoids solving a specific optimization model directly when facing dynamic operation conditions induced by the outdoor temperature, electricity price and resident's behavior. We consider the scheduling of power-shiftable, time-shiftable and deferrable appliances for the optimization of profit and satisfaction rate of resident. The optimization problem is first modeled as a Markov decision process and then solved by a model-free entropy-based DRL algorithm. Unlike traditional model-based methods which rely on accurate knowledge of parameters and physical models that are difficult to obtain in practice, the proposed method can develop real-time near-optimal control behavior by interacting with the environment and learning from data, which avoids the error caused by the simplification and assumption when building physical model. The proposed scheduling algorithm also achieves better tradeoff between the profit and the satisfaction rate than deterministic DRL algorithm owing to the introduction of the entropy term. Simulation results using real-world data demonstrate the effectiveness of the proposed method.

Peer-review under responsibility of the scientific committee of the 2021 The 2nd International Conference on Power Engineering, ICPE, 2021.

Keywords: Demand response; Residential appliances scheduling; Deep reinforcement learning

1. Introduction

With the deployment of smart meters, communication systems and intelligent controllers in modern power system, demand response has become an effective method to release the tension between electricity supply and demand and improve the system reliability. In general, demand response can be classified into incentive-based and

* Corresponding author.
E-mail address: caodi@std.uestc.edu.cn (D. Cao).

https://doi.org/10.1016/j.egyr.2022.02.181

Peer-review under responsibility of the scientific committee of the 2021 The 2nd International Conference on Power Engineering, ICPE, 2021.

price-based programs. To flatten the demand curve by providing electricity price that vary in time, price-based programs are the chief research content of this paper. In recent years, various price-based methods have been proposed in literature, which can be classified into three categories.

The first category is the optimization procedure based method. In [1], a stochastic dynamic programming based method is proposed for energy management of a smart home with plug-in electric vehicle energy storage. Ref. [2] formulated the optimization of energy use in a smart home as a mixed integer nonlinear programming problem, aiming to minimize the electricity cost and maximize the comfort of a resident at the same time. Authors in [3] utilized non-linear programming to schedule the appliances to flatten the demand curve. However, the methods mentioned above suffer from heavy computation burden and the "curse of dimensionality" [4].

The second category is the heuristics based method. Authors in [5] proposed an evolutionary algorithm based demand side management strategy for reshaping the load profiles of the smart grid. Due to the straightforward implementation of particle swarm optimization algorithm (PSO), Ref. [6] proposed an improved PSO based method for the optimization of appliances in an indeterminate environment in a residential energy management system. However, the heuristics based optimization methods are more suitable for deterministic cases, cannot react to the dynamic of the environment (e.g., indoor temperature, electricity price, appliances working state, etc.) [7]. In addition, methods mentioned above have to resolve an optimization problem when new states are encountered, which is time consuming.

In recent years, machine learning (ML) has become a hot research spot. By learning powerful knowledge from historical data, ML based methods can deal with the uncertainty and dynamic of environment, and have been successfully applied to fields like image processing [8], speech recognition [9] and optimization and control [10]. Among various machine learning methods, reinforcement learning (RL) is most suitable for the residential demand side management problem. In an RL model, the residential appliances are scheduled by the controller composed of RL algorithm, which can evolve continuously and improve their performance by utilizing their past experiences. Ref. [11] proposed effective management strategy for household electricity usage based on RL algorithm. The electricity cost minimization problem is first modeled as a Markov decision process (MDP), then the Q-learning algorithm is used to solve the MDP. Simulation results showed the effectiveness of proposed method. The Q-learning algorithm is also applied to obtain the optimal incentive rate based on the prediction of the electricity price and load demand provided by the deep neural network in [12]. However, the input state (time information, appliances working states, etc.) and the output action (the residential appliances power) of Q-learning algorithm need to be discretized. The naive discretization of state and action spaces would cause information loss since both the state signal (time information, appliances working states, etc.) and output action (the residential appliances power) are continuous value. This would lead to the suboptimal solution to the residential demand side management problem. At the same time, the computation complexity would increase sharply and the training is hard to converge when the state and action are discretized with a finer granularity.

To solve the problems mentioned above, neural networks are used for fitting the action value function in Q-learning algorithm due to its strong non-linear fitting capability. This fall into the research field named deep reinforcement learning (DRL), which combines the non-linear fitting capability of neural network and the decision making ability of RL. DRL employ DNN to approximate the action value function, can take the continuous signal as the input and provide action in continuous domain. This avoid the information loss during training and can provide better solution than Q-learning algorithm to problems with continuous state and action domain in real-world scenarios. The DRL-based methods have been widely adopted in the various fields to solve optimization problem, such as the optimization of distribution network work [13], the management of electric vehicle charging [14], and power system related studies [15].

In this paper, the physical properties of different kinds of appliances and the resident's power consuming behaviors is considered to simulate the power consumption in one household. Then, the DRL method is used to schedule different kinds of appliances to minimize the electricity cost and dissatisfaction of resident. The rest of this paper is organized as follows. Section 2 describes the mathematical model of the residential appliances. Then, the residential appliances scheduling problem is formulated as MDP. In Section 3, the soft actor–critic algorithm is illustrated in detail. Numerical simulation results are discussed in Section 4. Then Section 5 concludes this paper.

2. Problem modeling

The resident employs the home energy management system (HEMS) to schedule the appliances to minimize the electricity cost and the dissatisfaction level of resident. The control framework is shown in Fig. 1. In this section,

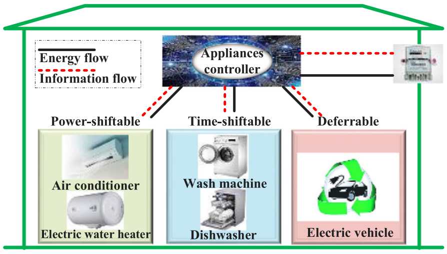

Fig. 1. The framework of residential appliances.

the mathematical model of power-shiftable, time-shiftable, and deferrable appliances are first illustrated, followed by the formulation of Markov decision process.

2.1. Power-shiftable appliances

The power-shiftable appliances, such as air conditioner (AC) and electric water heater (EWH), can adjust the working power continuously within the predefined range. In this paper, the two power-shiftable appliances, i.e. AC and EWH are considered.

(1) AC: The dynamic of the indoor temperature under AC is modeled by [16,17]:

$$\theta_{t+1}^{AC} = \alpha\theta_t^{AC} + (1-\alpha)\left(\theta_t^{out} - RP_t^{AC}\right), \alpha = \exp\left(-\frac{\Delta t}{CR}\right), 0 \le P_t^{AC} \le P_{\max}^{AC} \tag{1}$$

where the θ_t^{AC} (°C) represents the indoor air temperature at time slot t, θ_t^{out} (°C) denotes the ambient temperature at time slot t, α is the inertia factor of the AC, R (°C/kW) is the thermal resistance, P_t^{AC} (kW) is the power consumption of AC at time slot t, $P_{\max}^{AC}$ (kW) denotes the maximum power of AC, $\triangle t$(hour) means interval of a time slot, and C (kWh/°C) is the thermal capacity. Of which, C and R are the parameters of house, the both parameters are related to the insulation level, volume, walls, and surface [17].

(2) EWH: The temperature of hot water in EWH water tank changes with time [18,19]:

$$\theta_{t+1}^{EWH} = \beta\theta_t^{EWH} + (1-\beta)\left(G\theta_t^{out} + B_t\theta_{water}^{EWH} + Q_t\right)R', \beta = \exp\left(-\frac{\Delta t}{ZR'}\right), G = \frac{SA}{TR}, B_t = C_p \cdot F_t \cdot \rho_{water}, \tag{2a}$$

$$R' = \frac{1}{G+B_t}, Q_t = 3600P_t^{EWH} \triangle t, Z = C_p \cdot vol \cdot \rho_{water}, 0 \le P_t^{EWH} \le P_{\max}^{EWH} \tag{2b}$$

where the θ_t^{EWH} (°C) represents the water temperature in the EWH at time slot t, θ_t^{out} (°C) is the ambient temperature at time slot t, θ_{water}^{EWH} (°C) denotes the inlet cold water temperature, β is the inertia factor of the EWH, SA (m^2) represents the tank surface area, TR $\left(\text{hour} \cdot \text{m}^2 \cdot ^\circ \text{C/kJ}\right)$ is the tank insulation thermal resistance, C_p (kJ/ (°C · kg)) denotes the specific heat of water, F_t (L/hour) represents water flow rate at time slot t, ρ_{water} (kg/L) means density of water, P_t^{EWH} (kW) is the power consumption of EWH at time slot t, $P_{\max}^{EWH}$ (kW) denotes the maximum power of EWH, $\triangle t$ (hour) means interval of a time slot, and vol (L) represents the volume of the tank.

2.2. Time-shiftable appliances

The time-shiftable appliances can be dispatched from the peak to the off-peak to reduce the cost of power consumption. This paper considers two time-shiftable appliances: wash machine (WM) and dishwasher (DW). Both appliances have two operating points, "on" and "off":

$$P_t^{WM/DW} = \begin{cases} \partial_t^{WM/DW} P_{\max}^{WM/DW} \\ 0 \end{cases}, \partial_t^{WM/DW} = \begin{cases} 1, & \text{if } t_{start}^{WM/DW} \le t < t_{end}^{WM/DW} \text{ and } \varepsilon_t^{WM/DW} < \varepsilon_{req}^{WM/DW} \\ 0, & \text{otherwise} \end{cases} \tag{3a}$$

where $P_t^{WM/DW}$ represents the power consumption of WM/DW at time slot t, $P_{\max}^{WM/DW}$ denotes the maximum power of WM/DW, $\partial_t^{WM/DW}$ represents the working state of WM/DW,$\varepsilon_t^{WM/DW}$ is the working time up to time t of WM/DW, $t_{start}^{WM/DW}$ is the time to start work of the WM/DW, $t_{end}^{WM/DW}$ is the end working time of the WM/DW, and the $\varepsilon_{req}^{WM/DW}$ means the required working time of WM/DW.

2.3. Deferrable appliances

The deferrable load means that when the customer satisfaction is lower than the threshold, it will stop consuming power, and the load will be transferred to the time when the customer satisfaction is higher than the threshold. The customer satisfaction can be determined by factors such as electricity cost, etc. This paper only considers electric vehicle (EV) as the deferrable appliance. The dynamic of EV battery is modeled by

$$E_{t+1} = E_t + P_t^{EV} \cdot \triangle t,\ E_{\min} \le E_t \le E_{\max}, \tag{4a}$$

$$-P_{\max}^{EV} \le P_t^{EV} \le P_{\max}^{EV},\ if\ t \in \left[t_{arr}, t_{dep}\right],\ P_t^{EV} = 0, \text{otherwise} \tag{4b}$$

where the E_t is the EV battery energy at time slot t, P_t^{EV} denotes the charging/discharging power at time slot t, $\triangle t$ (h) means interval of a time slot, $E_{\min}$ and $E_{\max}$ are minimum and maximum battery storage energy, respectively, $P_{\max}^{EV}$ is the maximum power of EV, t_{arr} means EV arrives home at time t_{arr}, and t_{dep} represents EV arrives home at time t_{dep}.

2.4. Markov decision process formulation

In this paper, the problem of residential appliances scheduling is modeled as the MDP of finite time steps, which is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{F})$, where:

(1) $\mathcal{S}$ is the state space. The states of three types appliances can be defined as follows:

(a) Power-shiftable Appliances: In this paper, the state of one AC at time slot t is defined as: $s_t^{AC} = \left(\partial_t^{AC}, \theta_t^{AC}, \theta_{set}^{AC}\right)$ where the ∂^{AC} represents the working state of AC, 1 is "on", 0 is "off", and the θ_{set}^{AC} denotes the setpoint temperature of the AC thermostat at time slot t.

Similar to AC, the state of one EWH is defined as: $s_t^{EWH} = \left(\partial_t^{EWH}, \theta_t^{EWH}, \theta_{set}^{EWH}\right)$.

(b) Time-shiftable Appliances: The state of one WM/DW at time slot t is defined as: $s_t^{WM/DW} = \left(\partial_t^{WM/DW}, t, \varepsilon_t^{WM/DW}, t_{end}^{WM/DW}, \varepsilon_{req}^{WM/DW}\right)$ where t denotes the time information.

(c) Deferrable Appliances: The state of one EV at time slot t can be defined as : $s_t^{EV} = \left(\partial_t^{EV}, t, \kappa_t^{EV}, t_{dep}, \kappa_{\max}^{EV}\right)$, $\partial_t^{EV} = \begin{cases} 1, & \text{if} t_{arr} \le t < t_{dep} \\ 0, & \text{otherwise} \end{cases}$ where the ∂_t^{EV} is the working state of EV, t is the time information, κ_t^{EV} is the battery storage energy of EV at time slot t, and $\kappa_{\max}^{EV}$ represents the maximum battery storage energy.

In this paper, the state at each time slot t $s_t \in S$ consists of the electricity price $\mathcal{P}_t$, and the states of three types appliances:

$$s_t = \left(\mathcal{P}_t, s_t^{AC}, s_t^{EWH}, s_t^{WM}, s_t^{DW}, s_t^{EV}\right) \tag{5}$$

(2) $\mathcal{A}$ is the action space. The action at time slot t $a_t \in A$ is a set of actions of the appliances:

$$a_t = \left(P_t^{AC}, P_t^{EWH}, P_t^{WM}, P_t^{DW}, P_t^{EV}\right) \tag{6}$$

(3) $\mathcal{R}$ is the reward function. The objective of the residential appliances scheduling considers both the profit and satisfaction of resident.

(a) Power-shiftable Appliances: The reward function of AC at time slot t can be defined as: $r_t^{AC} = \mu^{AC}\psi_t^{AC} + \left(1-\mu^{AC}\right)\varphi_t^{AC}$ where $\varphi_t^{AC} = \begin{cases} \left(\theta_t^{AC} - \left(\theta_{set}^{AC} - \mathcal{T}^{AC}\right)\right)^2, \text{if} \theta_t^{AC} < \left(\theta_{set}^{AC} - \mathcal{T}^{AC}\right) \\ 0 \quad, \text{otherwise} \\ \left(\theta_t^{AC} - \left(\theta_{set}^{AC} + \mathcal{T}^{AC}\right)\right)^2, \text{if} \theta_t^{AC} > \left(\theta_{set}^{AC} + \mathcal{T}^{AC}\right) \end{cases}$, $\psi_t^{AC} = P_t^{AC}\mathcal{P}_t$ represents the electricity consumption of AC, φ_t^{AC} is the resident's dissatisfaction level with θ_t^{AC}, the $\mu^{AC} \in [0, 1]$ is a parameter which balances the ψ_t^{AC} and φ_t^{AC}, $\mathcal{T}^{AC}$ means the acceptable range of indoor temperature under the action of AC for resident, and the $\mathcal{P}_t$ is the electricity price.

And the reward function of EWH is defined as: $r_t^{EWH} = \mu^{EWH}\psi_t^{EWH} + \left(1-\mu^{EWH}\right)\varphi_t^{EWH}$ where the ψ_t^{EWH} and φ_t^{EWH} are similar to ψ_t^{AC} and φ_t^{AC}, respectively.

(b) Time-shiftable Appliances: The reward function of WM/DW is defined as: $r_t^{WM/DW} = \begin{cases} \partial_t^{WM/DW}\psi_t^{WM/DW}, & \text{if} \quad t \neq t_{end}^{WM/DW} \\ \varphi_t^{WM/DW}, & \text{if} \quad t = t_{end}^{WM/DW} \end{cases}$ where the $\psi_t^{WM/DW} = P_t^{WM/DW}\mathcal{P}_t$ denotes the electricity consumption of WM/DW, and the $\varphi_t^{WM/DW} = \left(\varepsilon_t^{WM/DW} - \varepsilon_{req}^{WM/DW}\right)^2$ is the resident's dissatisfaction level with WM/DW.

(c) Deferrable Appliances: The reward function of EV is defined as: $r_t^{EV} = \begin{cases} \partial_t^{EV}\psi_t^{EV}, & \text{if}\varphi_t^{EV} = 0 \\ \partial_t^{EV}\varphi_t^{EV}, & \text{otherwise} \end{cases}$ where $\psi_t^{EV} = P_t^{EV}\mathcal{P}_t$ denotes the electricity consumption of EV, $\kappa_{\min}^{EV}$ represents the minimum battery storage energy and the $\varphi_t^{EV} = \begin{cases} (\kappa_{\max}^{EV} - \kappa_t^{EV})^2, & \text{if } \kappa_t^{EV} > \kappa_{\max}^{EV} \text{ or } t = t_{dep} \\ (\kappa_{\min}^{EV} - \kappa_t^{EV})^2, & \text{if } \kappa_t^{EV} < \kappa_{\min}^{EV} \\ 0, & \text{otherwise} \end{cases}$ is the resident's dissatisfaction level with EV, which considers three situations that may cause dissatisfaction, one is overcharge, the second one is overdischarge, and the third is that the battery is not full at the time of leaving home.

Then, the immediate reward the agent obtains at time slot t r_t is defined as:

$$r_t\left(s_t, a_t\right) = r_t^{AC} + r_t^{EWH} + r_t^{WM} + r_t^{DW} + r_t^{EV} \tag{7}$$

(4) $\mathcal{F}$ is the transition function. $s_{t+1} = \mathcal{F}(s_t, a_t)$ denotes the probability distribution that the environment transfers to next state when action a_t is executed under state s_t.

One MDP is composed of a finite number of time steps. At each time slot, the residential appliances controller decides the power of above-mentioned appliances a_t, then obtains an immediate reward r_t and residential appliances controller transfer to the next state s_{t+1}. The reward function of the controller is to learn a policy $\pi(a_t|s_t)$ to maximize the discounted cumulative reward from start state s_1 onward: $R_1\left(s_t, a_t\right) = \sum_{t=1}^{T}\gamma^{(t-1)}r\left(s_t, a_t\right)$ where a_t is sampled from policy π, and $\gamma \in [0, 1]$ is the reward discount factor, which is introduced to balance the future rewards and the immediate reward. Considering that the policy $\pi(a_t|s_t)$may be stochastic, the objective function of the controller is defined as $\mathcal{O} = \max\left(\mathbb{E}_{a\sim\pi}\left[R_1\left(s_t, a_t\right)\right]\right)$.

3. Proposed method

Different from standard RL, which naïve maximize the expected return, the reward function of SAC is to maximize a trade-off between expected return and entropy $\mathcal{H}$ [20]:

$$R_1^{\mathcal{H}}\left(s_t, a_t\right) = r(s_1, a_1) + \sum_{t=2}^{T}\gamma^{(t-2)}\left(r\left(s_t, a_t\right) + \eta\mathcal{H}\left(\pi(a_t|s_t)\right)\right) \tag{8}$$

where the temperature parameter η is used for balancing the exploration and exploitation during the training process. Note that, when $\eta = 0$, $R_1^{\mathcal{H}} = R_1$. Considering the entropy regularization, the Bellman equation for $Q_{\mathcal{H}}^{\pi}$ and $V_{\mathcal{H}}^{\pi}$ are defined as [20]:

$$Q_{\mathcal{H}}^{\pi}\left(s_t, a_t\right) = \mathbb{E}_{s_{t+1}\sim\text{env}}\left[r(s_t, a_t) + \gamma V_{\mathcal{H}}^{\pi}\left(s_{t+1}\right)\right], V_{\mathcal{H}}^{\pi}\left(s_t\right) = \mathbb{E}_{a_t\sim\pi}\left[Q_{\mathcal{H}}^{\pi}\left(s_t, a_t\right) - \eta\log\pi\left(a_t|s_t\right)\right] \tag{9}$$

SAC consists of a state value network $V_{\mathfrak{B}}\left(s_t\right)$, a target state value network $V_{\widetilde{\mathfrak{B}}}\left(s_t\right)$, an action-value network $Q_{\mathfrak{X}}\left(s_t, a_t\right)$, and a policy network $\pi_{\mathfrak{S}}\left(a_t|s_t\right)$ where $\mathfrak{B}$, $\widetilde{\mathfrak{B}}$, $\mathfrak{X}$ and $\mathfrak{S}$ denote the parameters of these networks, respectively. Therein, the $V_{\mathfrak{B}}\left(s_t\right)$, $V_{\widetilde{\mathfrak{B}}}\left(s_t\right)$, and $Q_{\mathfrak{X}}\left(s_t, a_t\right)$ are regarded as the *critic* part, and the $\pi_{\mathfrak{S}}\left(a_t|s_t\right)$ is the *actor* part. In SAC, *actor* part is used for making decision, and *critic* part is used for approximating Eq. (9).

As described above, $V_{\mathfrak{B}}\left(s_t\right)$ and $Q_{\mathfrak{X}}\left(s_t, a_t\right)$ are used for approximating the Eq. (9), respectively. Therefore, the value network $V_{\mathfrak{B}}\left(s_t\right)$ can be optimized by minimizing:

$$\mathcal{J}_V\left(\mathfrak{B}\right) = \mathbb{E}_{s_t\sim\mathcal{Z}}\left[0.5\left(V_{\mathfrak{B}}\left(s_t\right) - \mathbb{E}_{a_t\sim\pi_{\mathfrak{S}}}\left[\begin{array}{c} Q_{\mathfrak{X}}\left(s_t, a_t\right) - \\ \eta\log\pi_{\mathfrak{S}}\left(a_t|s_t\right) \end{array}\right]\right)^2\right] \tag{10}$$

where $\mathcal{Z}$ represents the replay buffer which can reuse of previously collected $\left(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{F}\right)$ data for efficiency and stabilize the training process [21]. And the parameters of the target network $\widetilde{\mathfrak{B}}$ are updated by slowly tracking $\mathfrak{B}$: $\widetilde{\mathfrak{B}} = \tau * \mathfrak{B} + (1-\tau) * \widetilde{\mathfrak{B}}$ where $\tau < < 1$.

Table 1. The parameters of power-shiftable appliances.

Appliance	Parameters
AC1	$C_1 = 10$ kWh/°C, $R_1 = 2$ °C/kW, $P_{\max}^{AC1} = 15$ kW, $\theta_{set}^{AC1} = 24$°C, $\mathcal{T}^{AC1} = 3$ °C, $\mu^{AC1} = 0.4$
AC2	$C_2 = 6$ kWh/ °C, $R_2 = 3.33$ °C/kW, $P_{\max}^{AC2} = 9$ kW, $\theta_{set}^{AC2} = 24$ °C, $\mathcal{T}^{AC2} = 2$ °C, $\mu^{AC2} = 0.4$
EWH	$\theta_{water}^{EWH} = 15$ °C, $SA = 2.5$ m^2, $vol = 150$ L, $TR = 0.75$ hour $\cdot$ m^2 $\cdot$° C/kJ, $\theta_{set}^{EWH} = 55$ °C, $C_p = 4.2$ kJ/ (°C $\cdot$ kg), $\mathcal{T}^{EWH} = 4$ °C, $\rho_{water} = 1$ kg/L, $P_{\max}^{EWH} = 4.5$ kW, $\mu^{EWH} = 0.3$ $F_{base} = 3$ L/hour, $\Delta F_t \sim \mathcal{N}(0, 1)$ or $\Delta F_t = 0$
WM/DW	$P_{\max}^{WM} = 2$ kW, $t_{start}^{WM} = t_{arr}$, $t_{end}^{WM} = 0$, $\varepsilon_{req}^{WM} = 60\,\text{min}$, $P_{\max}^{DW} = 2.5$ kW, $t_{start}^{DW} = 0$, $t_{end}^{DW} = 9$, $\varepsilon_{req}^{DW} = 30\,\text{min}$
EV	$t_{arr} \sim \mathcal{U}(15, 16, 17, 18)$, $t_{dep} \sim \mathcal{U}(7, 8, 9, 10)$, $P_{\max}^{EV} = 6$ kW, $\kappa_{\max}^{EV} = 24$ kWh, $\kappa_{\min}^{EV} = 1$ kWh, $\kappa_{t_{arr}}^{EV} \sim \mathcal{N}(9.6, 0.2)$

The $Q_{\mathfrak{X}}(s_t, a_t)$ is optimized as $\mathcal{J}_Q(\mathfrak{X}) = \mathbb{E}_{s_t, a_t \sim \mathcal{Z}}\left[0.5\left(Q_{\mathfrak{X}}(s_t, a_t) - y\right)^2\right]$, $y = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \text{env}}\left[V_{\mathfrak{F}}(s_{t+1})\right]$. The *actor* part is updated in the direction that maximize the *critic* value. Thus, the goal of policy network $\pi_{\mathfrak{S}}(a_t|s_t)$ is to maximize the follow equation $\mathcal{J}_\pi(\mathfrak{S}) = \mathbb{E}_{s_t \sim \mathcal{Z}}\left[Q_{\mathfrak{X}}(s_t, a_t) - \eta \log \pi_{\mathfrak{S}}(a_t|s_t)\right]$ where a_t is sample from the $\pi_{\mathfrak{S}}(a_t|s_t)$ which is differentiable wrt $\mathfrak{S}$ via the reparametrization trick [20].

4. Experimental results

4.1. Experimental setup

The real-time electricity price data [22] of 2017 PJM and typical ambient temperature [17] are used in this study. The real-time electricity price data are divided into training and test set. Training set contains data for the first 200 days of 2017 and the test data includes the 201th day to 300th day of 2017. For case studies, there are three power-shiftable appliances: AC$_1$, AC$_2$ and EWH, two time-shiftable appliances: WM and DW; and one deferrable appliance: EV are considered for case studies. The epoch starts at the time that the EV arrives and ends when the time pass 144 time slots. One-time slot Δt is assumed to be 10 min. The parameters of above-mentioned appliances are shown in Table 1.

This paper assumes that all power-shiftable appliances work the whole day, WM starts working when the EV arrives home and stops at 0:00; DW starts working at 0:00 and stops at 9:00; and EV arrival time t_{arr} and departure time t_{dep} follow the uniform distribution $t_{arr} \sim \mathcal{U}(15, 16, 17, 18)$ and $t_{dep} \sim \mathcal{U}(7, 8, 9, 10)$, respectively. Normally, the thermal capacity C ranges from 0.015 to 0.065 kWh/°C per square meter, and the thermal resistance R can be selected approximately from 0.001 to 0.003 kW/°C per square meter of floor space [16]. Supposing that every room in a house has the same parameters such as wall material and thickness, etc. Under this assumption, the 0.04 kWh/°C/m^2 and 0.002 kWh/°C/m^2 [16] are selected for the calculation of C_1, C_2 and R_1, R_2, respectively. Assuming the floor area of the room affected by AC$_1$ is 250 m^2, C_1 can be calculated as: $\left(0.04\ \text{kWh/°C/m}^2\right) \times \left(250\ \text{m}^2\right) = 10$ kWh/°C, and the R_1 equal to $\left(\left(0.002\ \text{kWh/°C/m}^2\right) \times \left(250\ \text{m}^2\right)\right)^{-1} = 2$ °C/kW. Assuming AC$_2$ works in a room of 150 m^2, C_2 and R_2 can be calculated in the same way. In this model, the water flow rate of EWH is calculated by $F_t = F_{base} + \Delta F_t$, where F_{base} is a fixed value and ΔF_t is sampled from normal distribution or equal to 0. In general, the water flow rate changes frequently when water is used frequently. Therefore, this paper assumes that ΔF_t from t_{arr} to 0:00, 7:00 to 9:00, and 11:00 to 13:00 follows the normal distribution, otherwise $\Delta F_t = 0$.

4.2. Performance evaluation

The performance of proposed method in one epoch which starts at 15:00, October 12, 2017 are shown in Fig. 2. Fig. 2(a) is the electricity price diagram. Fig. 2(b) is the EV battery storage diagram. It can be observed that EV is charged when electricity price is low and EV discharges when the electricity price is high. Figs. 2(c) and 2(d) are the WM and DW power diagram, respectively. It can be observed from the figure that WM and DW both choose low price period to finish the task. Fig. 2(e) is the ambient temperature diagram. Figs. 2(f) and 2(g) are the indoor temperature under AC1 and AC2, respectively. Fig. 2(h) is the water temperature. Note that the electricity price (Fig. 2(a)) is high on edges (15:00∼23:00 and 5:00∼15:00) and low in the center (23:00∼5:00), so the indoor temperature which affected by AC should be high at the edges (15:00∼23:00 and 5:00∼15:00) and low in the center (23:00∼5:00), and the water temperature should be low at the edges (15:00∼23:00 and 5:00∼15:00) and

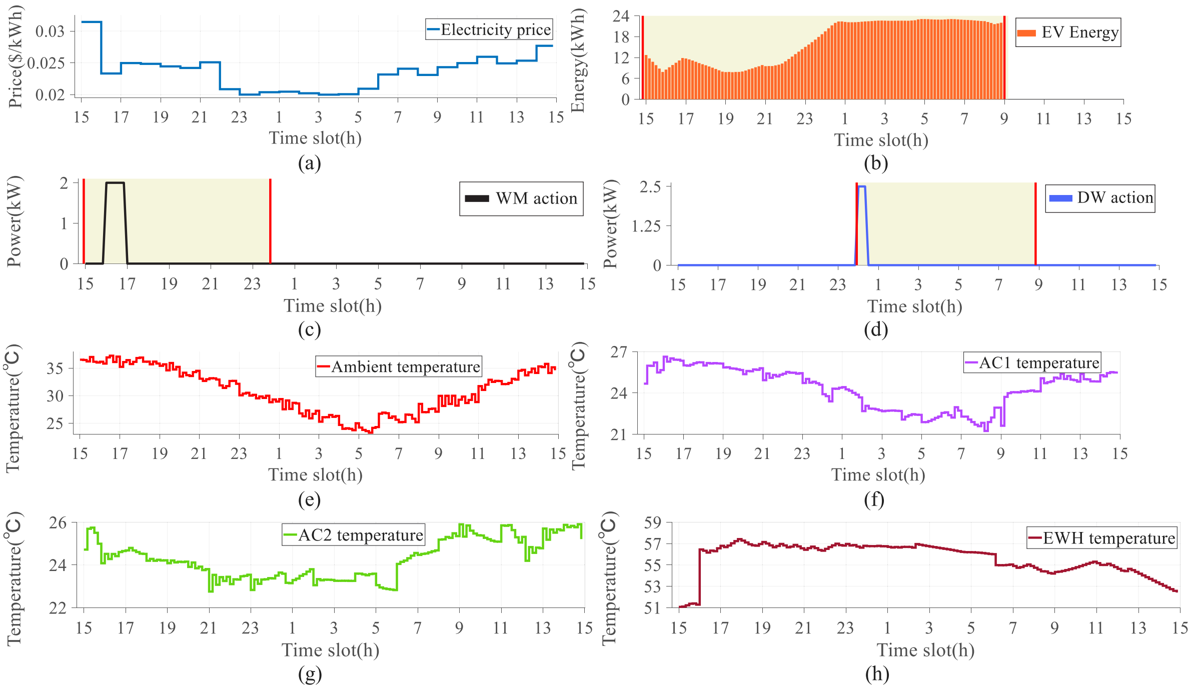

Fig. 2. Residential appliances scheduling.

high in the center (23:00∼5:00). Only in this way can resident obtain economic benefits under the electricity price trend in Fig. 2(a).

Next, the comparative tests are carried out among various benchmark methods to evaluate the performance of the proposed method. The benchmark solutions and proposed method have the same start state, ambient temperature, electricity price and flow rate. For the uncontrolled strategy, power-shiftable appliances work in maximum power operating condition $P_{\max}^{AC/EWH}$ when $\theta_t^{AC/EWH} > \theta_{set}^{AC/EWH} + \mathcal{T}^{AC/EWH}$, or $P^{AC/EWH} = 0$ when $\theta_t^{AC/EWH} < \theta_{set}^{AC/EWH} - \mathcal{T}^{AC/EWH}$; otherwise keep the same power as the previous time slot [19]. Under the uncontrolled strategy, WM/DW start to work once they receive an assignment. The EV will charge immediately once it arrives home until the battery is full. The Deep Deterministic Policy Gradient (DDPG) method [23] is also an *actor-critic* based DRL method. But unlike SAC method, DDPG method is based on deterministic policy, while SAC method is based on entropy-regularized stochastic policy. The Twin Delayed Deep Deterministic policy gradient method (TD3) [24] which is an improvement version of DDPG can address function approximation error in *actor-critic* methods, and has better effect than DDPG in robot control scene. The DDPG and TD3 methods share the same structure of neuron network with the proposed method. For the theoretical-limit strategy, all the variables ($\varDelta F_t$, t_{arr}, t_{dep}, and $\kappa_{t_{arr}}^{EV}$), the electricity price and ambient temperature of current epoch are assumed to be known in advance. Note that the theoretical-limit strategy cannot be achieved in the practice due to the randomness of the all variables ($\varDelta F_t$, t_{arr}, t_{dep}, $\kappa_{t_{arr}}^{EV}$, the electricity price and ambient temperature).

In the proposed model, the evaluation metric can be divided into two part: one is electricity cost:

$$\psi_t = \psi_t^{AC_1} + \psi_t^{AC_2} + \psi_t^{EWH} + \psi_t^{WM} + \psi_t^{DW} + \psi_t^{EV} \tag{11}$$

the other is dissatisfaction level:

$$\varphi_t = \varphi_t^{AC_1} + \varphi_t^{AC_2} + \varphi_t^{EWH} + \varphi_t^{WM} + \varphi_t^{DW} + \varphi_t^{EV} \tag{12}$$

The simulation results of proposed method and benchmark methods are shown in Fig. 3. Note that the data of Figs. 3(a) and 3(b) are normalized for better visualization of the performance of various methods. Fig. 3(a) is a comparison of the cumulative electricity cost of the four methods, it can be observed that the cost of the proposed method (i.e. red line) is 82.3% of uncontrolled strategy, the DDPG (yellow line), TD3 (purple line) and theoretical-limit strategy (green line) are 80.1%, 77.6% and 52% of uncontrolled strategy, respectively.

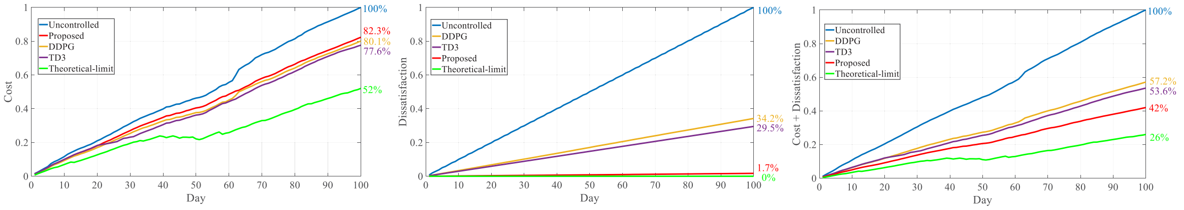

Fig. 3. Performance evaluation (a) Cumulative economic cost in test set; (b) Cumulative dissatisfaction level in test set; (c) Cumulative economic cost and dissatisfaction level in test set.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Though the theoretical-limit strategy has the best performance, the theoretical-limit strategy is difficult to implement in practice. Next to theoretical-limit strategy is the DDPG strategy, followed by proposed method. This study considers the economic benefits as well as the dissatisfaction level of resident. The cumulative dissatisfaction level of the four methods are shown in Fig. 3(b). It can be observed from the figure that the proposed method (i.e. red line) is 1.7% of uncontrolled strategy, the DDPG (yellow line), TD3 (purple line) and theoretical-limit strategy (green line) are 34.2%, 29.5% and 0% of uncontrolled strategy, respectively. There exists large difference between DDPG and proposed method when taking into account the resident comfort. As mentioned in Section 4 *C*, the DDPG and TD3 strategy are deterministic policy but the proposed method is a based on entropy-regularized stochastic policy which can achieve better performance in complex environment [20]. Although the DDPG and TD3 strategy takes the economic benefits into account, the dissatisfaction level of resident is ignored. Different with DDPG and TD3 strategy, the proposed method take into account two metrics at the same time due to the proposed method has a stronger ability to explore the environment and it is always moving towards the goal of optimizing the two metrics in the training process, rather than just learning the strategy of improving economic efficiency and ignoring the user dissatisfaction like DDPG and TD3 strategy. Combining electricity cost and dissatisfaction level metric, uncontrolled strategy is 0.5*(100% + 100%) = 100%, DDPG is 0.5*(80.1% + 34.2%) = 57.2%, TD3 is 0.5*(77.6% + 22.5%) = 53.6%, proposed method is 0.5*(82.3% + 1.7%) = 42% and theoretical-limit is 0.5*(52% + 0%) = 26%. The smaller the sum of the two metrics is, the better the performance of the method achieves. The combination of the two metrics are show in Fig. 3(c). The results demonstrate the effectiveness of the proposed method for these residential appliances scheduling problem.

5. Conclusion

In this paper, the residential appliances scheduling problem is formulated as a MDP from the perspective of resident considering the randomness of ambient temperature, electricity price, flow rate, and commuting behavior. Then proposed method is applied to solve the MDP. The proposed method can develop a control policy using data obtained by interacting with the environment. Experimental results demonstrate that the proposed method can reduce the electricity cost and the dissatisfaction level of resident. The proposed method avoids the dependence on the physical model and comparative results demonstrates that it can achieve better results than the deterministic DRL algorithm.

Acknowledgment

This work was supported by the National Key Research and Development Program of China under Grant 2018YFE0127600.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Wu X, Hu X, Yin X, Moura S. Stochastic optimal energy management of smart home with PEV energy storage. IEEE Trans Smart Grid 2016;9(3):2065–75.

[2] Moghaddam AA, Monsef H, Kian AR. Optimal smart home energy management considering energy saving and a comfortable lifestyle. IEEE Trans Smart Grid 2015;6(1):324–32.

[3] Rupanagunta P, Baughman ML, Jones JW. Scheduling of cool storage using non-linear programming techniques. IEEE Trans Power Syst 1995;10(3):1279–85.

[4] Ahrarinouri M, Rastegar M, Seifi AR. Multi-agent reinforcement learning for energy management in residential buildings. IEEE Trans Ind Inf 2020.

[5] Logenthiran T, Srinivasan D, Shun TZ. Demand side management in smart grid using heuristic optimization. IEEE Trans Smart Grid 2012;3(3):1244–52.

[6] Huang Y, Wang L, Guo W, Kang Q, Wu Q. Chance constrained optimization in a home energy management system. IEEE Trans Smart Grid 2018;9(1):252–60.

[7] Huang Ting, Liu Derong. A self-learning scheme for residential energy system control and management. Neural Comput Appl 2013;22(2):259–69.

[8] Ren W, et al. Deep video dehazing with semantic segmentation. IEEE Trans Image Process 2019;28(4):1895–908.

[9] Padi B, Mohan A, Ganapathy S. Towards relevance and sequence modeling in language recognition. IEEE/ACM Trans Audio Speech Lang Process 2020;28:1223–32.

[10] Lambert NO, Drew DS, Yaconelli J, Levine S, Calandra R, Pister KSJ. Low-level control of a quadrotor with deep model-based reinforcement learning. IEEE Robot Autom Lett 2019;4(4):4224–30.

[11] Liu Y, Yuen C, Ul Hassan N, Huang S, Yu R, Xie S. Electricity cost minimization for a microgrid with distributed energy resource under different information availability. IEEE Trans Ind Electron 2015;62(4):2571–83.

[12] Lu R, Hong SH. Incentive-based demand response for smart grid with reinforcement learning and deep neural network. Appl Energy 2019;236:937–49.

[13] Cao D, Hu W, Zhao J, et al. A multi-agent deep reinforcement learning based voltage regulation using coordinated PV inverters. IEEE Trans Power Syst 2020;35(5):4120–3.

[14] Li S, Hu W, Cao D, et al. Electric vehicle charging management based on deep reinforcement learning. J Mod Power Syst Clean Energy 2021. http://dx.doi.org/10.35833/MPCE.2020.000460.

[15] Cao D, Hu W, Zhao J, et al. Reinforcement learning and its applications in modern power and energy systems: A review. J Mod Power Syst Clean Energy 2020;8(6):1029–42.

[16] Callaway, Duncan S, et al. Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy. Energy Convers Manage 50(5):1389–400.

[17] Ucak C, Caglar R. The effects of load parameter dispersion and direct load control actions on aggregated load. In: POWERCON '98. 1998 international conference on power system technology. proceedings (Cat. No. 98EX151), vol. 1. 1998. p. 280–4.

[18] Nehrir MH, Jia R, Pierre DA, Hammerstrom DJ. Power Management of Aggregate Electric Water Heater Loads by Voltage Control. In: 2007 IEEE Power Engineering Society General Meeting. Tampa, FL; 2007. p. 1–6.

[19] Li H, Wan Z, He H. Real-time residential demand response. IEEE Trans Smart Grid 2020.

[20] Haarnoja Tuomas, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. 2018.

[21] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG. Human-level control through deep reinforcement learning. Nature 2015;518:529–33.

[22] PJM. Zone comed. 2017, [Online]. Available: https://www.engieresources.com/.

[23] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. Comput Sci 2015.

[24] Fujimoto Scott, Van Hoof Herke, Meger David. Addressing function approximation error in actor-critic methods. 2018, arXiv preprint arXiv:1802.09477.