A Deep Reinforcement Learning Based Approach for Home Energy Management System

Hepeng Li, Zhiqiang Wan and Haibo He
Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island
Kingston, RI 02881 USA, Email: hepengli@uri.edu, zwan@ele.uri.edu and haibohe@uri.edu

***Abstract*—Home energy management system (HEMS) enables residents to actively participate in demand response (DR) programs. It can autonomously optimize the electricity usage of home appliances to reduce the electricity cost based on time-varying electricity prices. However, due to the existence of randomness in the pricing process of the utility and resident's activities, developing an efficient HEMS is challenging. To address this issue, we propose a novel home energy management method for optimal scheduling of different kinds of home appliances based on deep reinforcement learning (DRL). Specifically, we formulate the home energy management problem as an MDP considering the randomness of real-time electricity prices and resident's activities. A DRL approach based on proximal policy optimization (PPO) is developed to determine the optimal DR scheduling strategy. The proposed approach does not need any information on the appliances' models and distribution knowledge of the randomness. Simulation results verify the effectiveness of our proposed approach.**

***Index Terms*—demand response; deep reinforcement learning; energy management; proximal policy optimization**

I. Introduction

THE availability of bidirectional communication, advanced metering infrastructure, and real-time electricity price in smart grid makes it possible for smart appliances to participate in demand response (DR) programs [1]. By scheduling the electricity usage of smart appliances from on-peak hours to off-peak hours, HEMS contributes to peak shaving and valley filling and helps to reduce the electricity cost of smart home residents. However, due to the existence of randomness in the real-time electricity prices and resident's activities, efficiently managing the electricity usage of home appliances is challenging.

To solve this problem, numerous approaches in the literature have been developed over the past few years. For instance, Huang *et al.* [2] proposed a chance constrained programming model to optimize the operational schedules of shiftable and thermal appliances to minimize the electricity cost considering the uncertainty of electricity prices and demand. Chen *et al.* [3] developed a scenario-based stochastic optimization approach to deal with the uncertainty in the real-time electricity prices via Monte-Carlo simulation. Du *et al.* [4] proposed a robust optimization approach to minimize the worst-case daily bill payment by considering the randomness of the resident's activity.

In addition to the aforementioned model-based solutions, learning-based approaches have attracted much attention in recent years. For instance, Keerthisinghe *et al* [5] proposed an approximate dynamic programming (ADP) method for DR scheduling of appliances considering the uncertainty of home demand and PV generation. Bahrami *et al.* [6] proposed an online learning algorithm for optimal scheduling of DR appliances based on the actor-critic approach. In [7], Ruelens *et al.* developed a batch RL approach for optimal control of an electric water heater (EWH). Wen *et al.* [8] formulated the optimal DR scheduling of appliances as a device-based MDP model and a Q-learning algorithm was employed to solve the problem. Lu *et al.* [9] developed a multi-agent RL approach to make optimal DR schedules for different home appliances in a decentralized manner. These learning-based approaches facilitate the design of a HEMS since they do not need an explicit system model and an optimizer to solve for the optimal DR schedules. However, the aforementioned approaches use simple approximators, such as look-up tables or linear functions, to learn the optimal DR strategies. The limited approximation capability makes these approaches difficult to handle complicated nonlinearity in the time-varying electricity prices and resident's activities in real-world. Therefore, the performance of these approaches may deteriorate.

In this paper, a DRL-based approach for optimal HEMS is proposed to overcome the disadvantage. DRL approaches can utilize deep neural networks (DNN) to learn complicated nonlinear mapping from high-dimensional observations of a system for decision-making. DRL approaches have achieved great success in many applications, such as games, [10] robots [11] and smart grids [12]. In this paper, we formulate the optimal energy management of home appliances as an MDP. The aim is to minimize the electricity cost of a household considering the uncertainty of the real-time electricity prices and resident's activities. A DRL algorithm based on PPO [13] is adopted to solve the MDP to obtain the optimal DR scheduling strategy.

The contributions of this paper include two aspects. First, we formulate the optimal home energy management of different kinds of appliances as an MDP where the randomness of the real-time electricity prices and resident's activities are taken into account. Second, a DRL approach that does not need the appliance models and distribution knowledge of the randomness is adopted to learn the optimal DR strategy.

II. Problem Formulation

We consider a smart home with three types of appliances, including the critical, shiftable and controllable appliances.

978-1-7281-3103-0/20/$31.00 ©2020 IEEE

The operation of these appliances is managed by a HEMS, which receives hour-ahead electricity price from the utility. Next, we model the electricity usage of these appliances and formulate the DR scheduling of the appliances as an MDP.

A. Modeling of the Appliances

1) Critical Appliances: Critical appliances cannot be scheduled for DR. If a critical appliance $c = 1, ..., C$ is required to operate in the period $[t_c^\alpha, t_c^\beta]$, it should operate immediately and its power consumption should be equivalent to the demand

$$P_c^C(t) = P_{c,R}^C,\ t \in [t_c^\alpha, t_c^\beta], \tag{1}$$

where $P_{c,R}^C$ is rated power of the appliance c.

2) Shiftable Appliances: Shiftable appliances can defer their demand to off-peak hours. For the shiftable appliance $s = 1, ..., S$, its power consumption at time step t should equal to

$$P_s^S(t) = b_s^S(t) P_{s,R}^S,\ t \in [t_s^\alpha, t_s^\beta], \tag{2}$$

where $b_s^S(t)$ is a binary control variable determining whether to operate the appliance or not; $P_{s,R}^S$ denotes the rated power; $[t_s^\alpha, t_s^\beta]$ represents the scheduling window.

The control variable $b_d^S(t)$ should be constrained by

$$b_s^S(t) = 1, \text{ if } \sum_{\tau=t_s^\alpha}^{t} b_s^S(\tau) = 0, t_s^\beta - t = K_s^S, \tag{3a}$$

$$b_s^S(t) = 1, \text{ if } b_s^S(t-1) = 1 \text{ and } \sum_{\tau=t_s^\alpha}^{t} b_s^S(\tau) < K_s. \tag{3b}$$

where K_s is the required time slots to meet the energy demand of the appliance s. Eq. (3a) makes sure that the energy demand s is satisfied in the scheduling window; Eq. (3b) ensures that the operation of the shiftable appliance s is not interrupted.

3) Controllable Appliances: The power consumption of a controllable appliance can be flexibly controlled or regulated. A typical controllable appliance is EV. Consider that an EV arrives home at time step t_{ev}^α and departs at t_{ev}^β. Then, the EV can be formulated by

$$E(t+1) = E(t) + P^{ev}(t)\Delta t,\ t \in [t_{ev}^\alpha, t_{ev}^\beta - 1], \tag{4a}$$

$$E_{min} \le E(t) \le E_{max},\ t \in [t_{ev}^\alpha, t_{ev}^\beta - 1], \tag{4b}$$

where $E(t)$ is the remaining energy of the EV battery at time step t, which is bounded in (4b) by its minimum energy E_{min} and maximum energy E_{max}, respectively; $P^{ev}(t)$ is the charging/discharging power, which is positive when the EV is charging or negative when the EV is discharging. The charging/discharging power $P^{ev}(t)$ is constrained by

$$P^{ev}(t) \in [-P_{dis,max}^{ev}, P_{ch,max}^{ev}], \tag{5}$$

where $P_{ch,max}^{ev}$ and $P_{dis,max}^{ev}$ are the maximum charging and discharging power of the EV battery, respectively.

To ensure the EV is fully charged when its departs home, the remaining energy of the EV battery should satisfy

$$E(t) = E_{max},\ t = t_{ev}^\beta. \tag{6}$$

B. MDP Formualtion of the HEMS

1) States: The states of the smart home is defined as

$$\begin{aligned} s(t) = (&R(t-23), \ldots, R(t), E(t), \\ &B_1^S(t), ..., B_S^S(t), T_1^S(t), ..., T_S^S(t)), \end{aligned} \tag{7}$$

where $R(t-23), \ldots, R(t)$ denote the past 24-hour's electricity prices; $E(t)$ is the remaining energy of the EV battery; $B_1^S(t), ..., B_S^S(t)$ represents the operational states of the shiftable appliances and $T_1^S(t), ..., T_S^S(t)$ denote the remaining time slots to the corresponding deadline. Here $B_s^S(t) = \sum_t b_s^S(t-1)/K_s^S$ represents how many percents of the energy demand of the appliance s have been fulfilled and $T_1^S(t) = t_s^\beta - t, s = 1, \ldots, S$.

2) Actions: The actions are defined by

$$a(t) = (b_1^S(t), ..., b_S^S(t), P^{ev}(t)), \tag{8}$$

which include the binary control variables $b_1^s(t), \ldots, b_S^S(t)$ of the shiftable appliances and the charging/discharging power $P^{ev}(t)$ of the EV battery.

3) State Transition Probability: The transition probability of the states is assumed to be unknown by the HEMS.

4) Rewards: The reward is formulated as

$$\begin{aligned} r(t) = &- R(t) \left[\sum_{c=1}^{C} P_c^C(t) + \sum_{d=1}^{D} P_d^S(t) + P^{ev}(t) \right] \cdot \Delta t, \\ &- w(E(t_{ev}^\beta) - E_{max})^2, \end{aligned} \tag{9}$$

where the first term is the electricity cost at time step t; the second term is the EV range anxiety, which is the square of the unfulfilled energy to fully charge the EV at the departure time [12]. The coefficient w is a weighting parameter.

5) Objective: The HEMS aims to find an optimal control policy π^* to maximize the discounted cumulative rewards over one day,

$$\max_{\pi \in \Pi}\ J^\pi = \mathbb{E}_\pi \left[\sum_{t=1}^{T} \gamma^{t-1} r(t) \right], \tag{10}$$

where the policy $\pi(a|s) \in [0, 1] : s \to P(a)$ is a probability of choosing the action a when the system state is s, Π is the set of all feasible policies, and $0 < \gamma \le 1$ is the discount factor.

III. Deep Reinforcement Learning Solution

It is intractable to exactly solve the MDP (10) due to the unknown transition probability and the curse of the dimensionality. However, we can approximately search for the optimum in a set Π_θ of parameterized policies π_θ by ascending the objective $J(\pi_\theta)$ with respect to the parameters θ,

$$\theta^{i+1} = \theta^i + \alpha \bigtriangledown_\theta J(\pi_\theta),\ i \to 1, 2, \ldots. \tag{11}$$

For the formulated DR scheduling problem, we propose a DNN to approximate the policy π and optimize the DNN parameters based on the PPO algorithm.

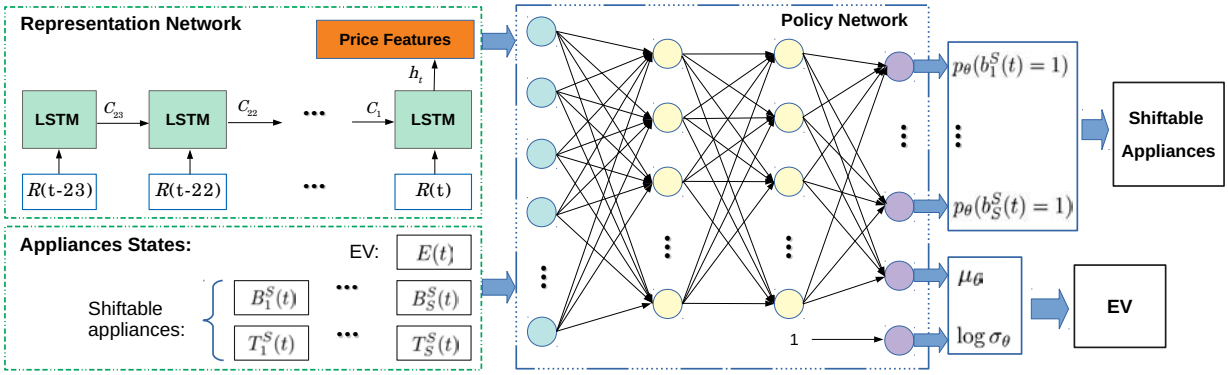

Fig. 1. Overall architecture of the designed DNN-based stochastic policy. The input of the DNN-based policy is the system state $s(t)$ of the smart home, including the past 24-hour's electricity prices $R(t-23),\dots,R(t)$, the EV battery energy $E(t)$, and the operational states $B_1^S(t),...,B_S^S(t),T_1^S(t),...,T_S^S(t)$ of the shiftable appliances. The output is the distribution parameters p_θ, μ_θ and $\log\sigma_\theta$ of the policy π_θ. The DNN consists of a representation network and a policy network. The representation network uses an LSTM to extract temporal features from the past 24-hour's electricity prices. The policy network is a MLP, which approximates the optimal distribution parameters p_θ, μ_θ and $\log\sigma_\theta$ of the policy π_θ based on the features of the past 24-hour's electricity prices and the operational states of the DR appliances.

A. DNN-based Stochastic Policy

We search for the optimal policy π within a mixed probability distribution π_θ parameterized by θ as follows

$$\pi_\theta(a|s)=\begin{cases}\mathcal{B}(p_\theta), & \text{if } a\in p_\theta(b_S^S(t)=1),\dots,b_S^S(t)],\\ \mathcal{N}(\mu_\theta,\sigma_\theta), & \text{if } a=P^{ev}(t),\end{cases} \tag{12}$$

where π_θ is a Bernoulli distribution $\mathcal{B}(p_\theta)$ if the action a is a binary control variable; or π is a normal distribution $\mathcal{N}(\mu_\theta,\sigma_\theta)$ if the action a is a continuous vairble, i.e. $P^{ev}(t)$. A DNN is proposed to learn the optimal distribution parameters p_θ, μ_θ and $\log\sigma_\theta$ of the policy π_θ.

Fig. 2 shows the architecture of the proposed DNN. The input of the DNN is the system state $s(t)$ of the smart home, including the operational state of all DR appliances and the past 24-hour's electricity prices. The output is the distribution parameters p_θ, μ_θ and $\log\sigma_\theta$ of the policy π_θ. The DNN consists of a representation network and a policy network. The representation network uses an LSTM to extract temporal features from the past 24-hour's electricity prices. Specifically, the electricity price $x_\tau=R(\tau)$ at each time step $\tau\in\{t-23,...,t\}$ are processed by a LSTM cell

$$f_\tau=\sigma(W_f\cdot[h_{\tau-1},x_\tau]+b_f), \tag{13a}$$
$$i_\tau=\sigma(W_i\cdot[h_{\tau-1},x_\tau]+b_i), \tag{13b}$$
$$o_\tau=\sigma(W_o\cdot[h_{\tau-1},x_\tau]+b_o), \tag{13c}$$
$$\widetilde{C}_\tau=\sigma(W_C\cdot[h_{\tau-1},x_\tau]+b_C), \tag{13d}$$
$$C_\tau=f_\tau*C_{\tau-1}+i_\tau*\widetilde{C}_\tau, \tag{13e}$$
$$h_\tau=o_\tau*\tanh(\widetilde{C}_\tau), \tag{13f}$$

where x_τ and h_τ are the input and output of the LSTM cell, respectively; C_τ is the state of the cell; f_τ, i_τ and o_τ are the forget gate, input gate and output gate, respectively; σ and $\tanh$ denote the sigmoid and hyperbolic tangent function, respectively. Each LSTM cell fuses its input x_τ with the output $h_{\tau-1}$ of its predecessor and selectively passes the fused information to its successor. The last LSTM cell outputs the feature h_t extracted by the LSTM.

Then, the feature h_t is concatenated with the states of the shiftable appliances and the EV and fed into the policy network. The policy network is a multilayer perceptron (MLP):

$$\vec{v}_0=[h_t;s(t)], \tag{14a}$$
$$\vec{v}_j=\max(0,W_j\cdot\vec{v}_{j-1}+b_j),\ j=1,\dots,J-1 \tag{14b}$$
$$p_\theta=\sigma(W_p\cdot\vec{v}_J+b_p), \tag{14c}$$
$$\mu_\theta=W_\mu\cdot\vec{v}_J+b_\mu,\ \log\sigma_\theta=W_\sigma. \tag{14d}$$

which outputs the distribution parameters p_θ, μ_θ and $\log\sigma_\theta$ of the policy π_θ.

B. Proximal Policy Optimization

To train the designed DNN-based policy, the policy gradient method (11) is adopted. Motivated by the PPO algorithm, we use a surrogate objective $L^{CLIP}(\theta)\approx J(\pi_\theta)$ to calculate the policy gradient. The surrogate objective $L^{CLIP}(\theta)$ is calculated by

$$L^{CLIP}(\theta)=\mathbb{E}_t\left[\min(\delta_t(\theta)\hat{A}_t,\text{clip}(\delta_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t)\right] \tag{15}$$

where $\delta_t(\theta)$ denotes the probability ratio $\delta_t(\theta)=\frac{\pi_\theta(a(t)|s(t))}{\pi_{\theta^i}(a(t)|s(t))}$; $\hat{A}_t$ is the sample estimate of the advantage function $A_{\pi_{\theta^i}}=V_{\pi_{\theta^i}}(s(t+1))+r(t)-V_{\pi_{\theta^i}}(s(t))$; ϵ is a hyperparameter. $V_{\pi_\theta}(s(t))=\mathbb{E}_{a(t+1),s(t+1),...}[\sum_{l=0}^{\infty}\gamma^l r(t+l)]$ denotes the value function, which is approximated by a DNN. The DNN shares the same archtecture and parameters as the policy network. Therefore, the value function is approximated by $V_{\pi_\theta}(s(t))=W_V\cdot\vec{v}_J+b_V$, and (W_V,b_V) are network weights. The loss function for the value function approximation is

$$L^{VF}(\theta)=\mathbb{E}_t[V_{\pi_\theta}(s(t))-\sum_{l=0}^{\infty}\gamma^l r(t)]^2.$$

Overall, the parameters θ of the proposed DNN is updated by maximizing the surrogate objective $L^{CLIP}(\pi_\theta)$ and minimizing the loss function $L^{VF}(\theta)$ as follows,

$$\begin{aligned}L^{CLIP\_VF}(\theta)&=L^{CLIP}(\theta)-c\cdot L^{VF}(\theta),\\ \theta^{i+1}&=\theta^i+\alpha\bigtriangledown_\theta L^{CLIP\_VF}(\theta)\end{aligned} \tag{16}$$

where c is a coefficient. Alg. 1 presents the training algorithm of the proposed DNN-based policy.

Algorithm 1 Training process of the proposed DNN-based policy

Initialize: Number of iterations I; sample buffer $\mathcal{F}_{\mathcal{N}}$;
for i = 1, I **do**
  for n = 1, N **do**
    Initialize the state $s(0)$ of the smart home;
    for t = 0, T-1 **do**
      Sample $a(t)$ according to π_{θ^i};
      Calculate reward $r(t)$ and observe $s(t+1)$;
    end for
    Calculate the sample estimates $\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_T$;
    Store $A^n = (\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_T)$ in $\mathcal{F}_{\mathcal{N}}$;
  end for
  Optimize $L^{CLIP\_VF}(\theta)$ with respecte to θ with K epochs and minibatch size M by gradient descent;
  Update $\theta_i \longleftarrow \theta$;
end for

IV. Case Studies

In this section, we evaluate the performance of the proposed DRL-based approach on a smart home. We consider four shiftable appliances, i.e., dishwasher, washing machine, cloth dryer, and stove and one controllable appliance, i.e. the EV. The refrigerator, TV, and lights are modeled as critical appliances. We divide one day into $T = 96$ time slots and each time slot represents $\Delta t = 15$ minutes. For each appliance, we assume that the scheduling window $[t_\alpha, t_\beta]$ changes in different days due to the randomness of the resident's activities. Specifically, the starting time t_α and end time t_β are selected uniformly from the bounds presented in Table I [4], [6]. For the EV, we consider a Nissan Leaf with the maximum battery capacity $E_{max} = 24$ kWh and minimum battery energy $E_{min} = 2.4$ kW. The battery SoC of the EV when it arrives home is randomly chosen from the normal distribution $\mathcal{N}(0.5, 0.1)$ bounded by $[0.4, 0.6]$ [12].

The real-world hourly electricity price from [14] is used to train and test the proposed approach. The one-year data in 2017 are used for training and the one-year data in 2018 are used for performance evaluation. The representation network outputs a 16-dimension feature vector. The policy network has three hidden layers and each layer has 64 ReLU neurons. In the output layer, there are six units. Four of them output the probability p_θ for the shiftable appliances and two of them output μ_θ and $\log \delta_\theta$ for the EV scheduling. All network weights are randomly initialized. The weight in the reward (9) is $w = 0.01$. The discounted factor γ is set to 0.995. The minibatch size, training epochs and the size of the samples buffer $\mathcal{F}_{\mathcal{N}}$ in each iteration of training is $M = 2400$, $K = 4$, $N = 100$, respectively. The stepsize parameters α and c are set to $3e-4$ and 0.5, respectively. The hyperparameter ϵ is set to 0.05. The number of iterations is $I = 500$. The experiment is conducted on a personal computer with four i5-6300U CPU. The code is written in Python and run with TensorFlow1.12.

Fig. 2 shows the episode cumulative rewards of the proposed approach during the training process. As shown in the figure, the cumulative rewards increase quickly after the start of the training. From episode 20,000, the cumulative rewards

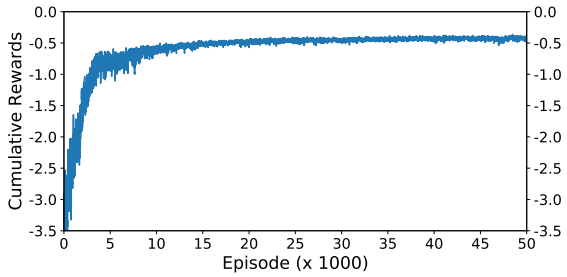

Fig. 2. Episode returns during the training process.

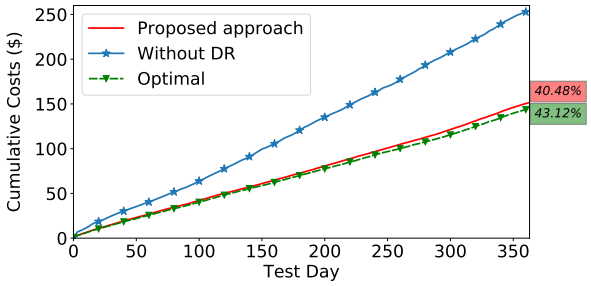

Fig. 3. Cumulative electricity cost curves over the 365 test days.

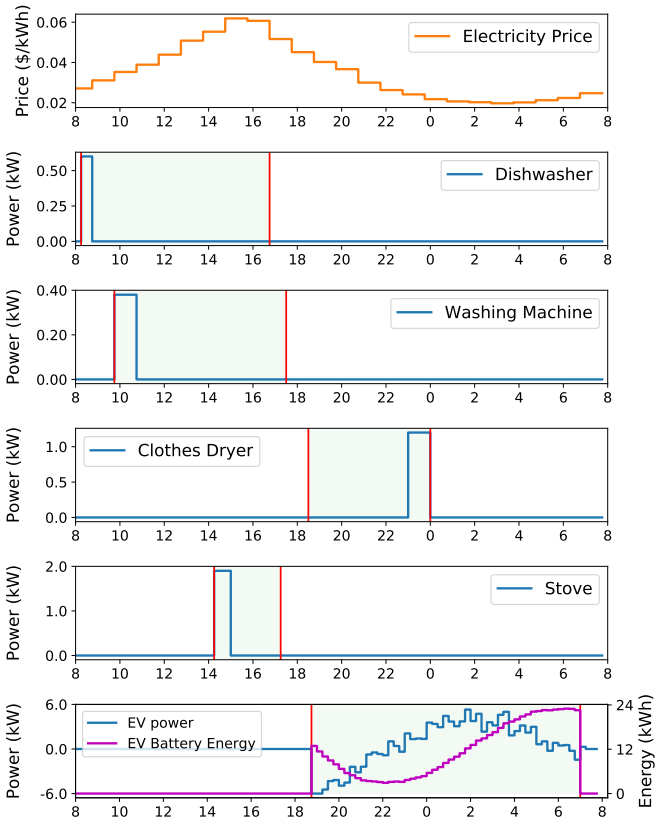

Fig. 4. DR scheduling results of the proposed approach on a test day. The green region in each subfigure indicates the scheduling window of the corresponding appliance.

TABLE I
OEPRATIONAL SPECIFICATIONS OF THE HOME APPLIANCES

		Critical			Shiftable				Controllable
Device		Refrigerator	TV	Lights	Dish Washer	Washing Machine	Clothes Dryer	Stove	EV
Power Rating (kW)		0.2	0.1	0.2	0.6	0.38	1.2	1.9	-6∼6
Task Demand (time slots)		24 hr	-	-	2	4	4	3	-
Scheduling Time	t_α	-	[40, 43]	[36, 39]	[0, 4]	[0, 4]	[20, 23]	[24, 27]	[32, 43]
(time slots)	t_β	-	[56, 59]	[56, 59]	[32, 35]	[16, 19]	[40, 47]	[36, 37]	[92, 95]

gradually stabilize and converge around -0.4 at the end of the training. This result demonstrates that the proposed approach succeeds in learning to increase the cumulative rewards.

To assess the performance of the proposed approach on the test dataset, we compare the proposed DRL-based approach with two benchmarks including the *without DR* policy and the *theoretical optimal* policy. For the *without DR* policy, all the shiftable and controllable appliances are scheduled to operate as soon as they are "switched on" to carry out a task. For the *theoretical optimal* policy, it is assumed that the future 24-hour's electricity prices and the scheduling windows of each shiftable appliance and the EV are known in advance. The DR scheduling problem is solved as a deterministic optimization problem via the optimization toolbox SCIP [15]. It should be noted that *theoretical optimal* policy provides a theoretical bound of the performance and it cannot be achieved in practice due to the randomness. Fig. 3 compares the cumulative electricity cost curves over the 365 test days. It can be seen that the proposed approach reduces the accumulative electricity cost by 40.48% when compared to the without DR policy. Moreover, the performance of the proposed model is close to that of the *theoretical optimal* policy. This comparison verifies the effectiveness of the proposed approach.

To further demonstrate the effectiveness of the proposed approach, we present the DR scheduling results on a test day in Fig. 4. The green regions in the subfigures indicate the scheduling windows of the shiftable appliances and the EV, respectively. As shown in Fig. 4, the shiftable appliances are scheduled to operate in the periods when the prices are low in their scheduling windows. In addition, the EV is discharged when the electricity price is high at 18:00-22:00 and charged when the price is low at 0:00-4:00. When the EV departs, the EV battery is adequately charged. These results demonstrate that the proposed approach can learn to optimize the operation of the DR appliances to save electricity cost.

V. CONCLUSION

In this paper, the home energy management problem is formulated as an MDP. The aim is to optimize the operational schedules of different kinds of appliances to minimize the electricity cost. The randomness of real-time electricity prices and resident's activities are taken into consideration. A DRL approach based on PPO is developed to solve the problem. The proposed approach does not need any information about the appliance models and distribution knowledge of the randomness. A DNN is designed to learn the optimal DR policy. The DNN-based policy can directly learn from raw sensory data of the system and output the DR controls of the shiftable appliances and the EV. Experimental results demonstrate that the proposed approach outperforms the benchmark methods.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation under grant ECCS 1917275.

REFERENCES

[1] B. Zhou, W. Li, K. W. Chan, Y. Cao, Y. Kuang, X. Liu, X. Wang, "Smart home energy management systems: Concept, configurations, and scheduling strategies", Renewable and Sustainable Energy Reviews, 2016.

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

[3] Z. Chen, L. Wu, and Y. Fu, "Real-time price-based demand response management for residential appliances via stochastic optimization and robust optimization," IEEE Transactions on Smart Grid, vol. 3, no. 4, pp. 1822-1831, Dec 2012.

[4] Y. Du and L. Jiang and Y. Li and Q. H. Wu, "A robust optimization approach for demand side scheduling under energy consumption uncertainty of manually operated appliances," IEEE Transactions on Smart Grid, vol. 9, no. 2, pp. 743-755, March 2018.

[5] C. Keerthisinghe, G. Verbic and A. C. Chapman, "A Fast Technique for Smart Home Management: ADP With Temporal Difference Learning," IEEE Transactions on Smart Grid, vol. 9, no. 4, pp. 3291-3303, July 2018.

[6] S. Bahrami, V. W. S. Wong and J. Huang, "An Online Learning Algorithm for Demand Response in Smart Grid," in IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 4712-4725, Sept. 2018.

[7] F. Ruelens, B. J. Claessens, S. Vandael, B. D. Schutter, R. Babuka, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," IEEE Transactions on Smart Grid, vol. 9, no. 4, pp. 3792-3800, July 2018.

[8] Z. Wen, D. ONeill, and H. Maei, "Optimal demand response using device-based reinforcement learning," IEEE Transactions on Smart Grid, vol. 6, no. 5, pp. 2312-2324, Sept 2015.

[9] R. Lu, S. H. Hong and M. Yu, "Demand Response for Home Energy Management using Reinforcement Learning and Artificial Neural Network," in IEEE Transactions on Smart Grid.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," Nature, vol. 518, no. 7540, pp. 529533, 2015.

[11] Z. Wan, C. Jiang, M. Fahad, Z. Ni, Y. Guo and H. He, "Robot-Assisted Pedestrian Regulation Based on Deep Reinforcement Learning," in IEEE Transactions on Cybernetics.

[12] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," IEEE Transactions on Smart Grid, pp. 11, 2019.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, "Proximal Policy Optimization Algorithms", ArXiv, 2017.

[14] "Day-ahead and historical RTP/HSS prices," Ameren Illinois. [Online]. Available: https://www.ameren.com/account/retail-energy.

[15] A. Gleixner, M. Bastubbe, L. Eifler, T. Gally, G. Gamrath, R. L. Gottwald, G. Hendel, C. Hojny, T. Koch, M. E. Lubbecke, S. J. Maher, M. Miltenberger, B. Muller, M. E. Pfetsch, C. Puchert, D. Rehfeldt, F. Schlosser, C. Schubert, F. Serrano, Y. Shinano, J. M. Viernickel, M. Walter, F. Wegscheider, J. T. Witt, J. Witzig, "The SCIP Optimization Suite 6.0", July 2018.