Real-Time Residential Demand Response

Hepeng Li, *Student Member, IEEE*, Zhiqiang Wan, *Student Member, IEEE*, and Haibo He, *Fellow, IEEE*

***Abstract*—This paper presents a real-time demand response (DR) strategy for optimal scheduling of home appliances. The uncertainty of the resident's behavior, real-time electricity price, and outdoor temperature is considered. An efficient DR scheduling algorithm based on deep reinforcement learning (DRL) is proposed. Unlike traditional model-based approaches, the proposed approach is model-free and does not need to know the distribution of the uncertainty. Besides, unlike conventional RL-based methods, the proposed approach can handle both discrete and continuous actions to jointly optimize the schedules of different types of appliances. In the proposed approach, an approximate optimal policy based on neural network is designed to learn the optimal DR scheduling strategy. The neural network based policy can directly learn from high-dimensional sensory data of the appliance states, real-time electricity price, and outdoor temperature. A policy search algorithm based upon trust region policy optimization (TRPO) is used to train the neural network. The effectiveness of our proposed approach is validated by simulation studies where the real-world electricity price and outdoor temperature are used.**

***Index Terms*—Demand response, deep reinforcement learning, smart home, trust region policy optimization.**

NOMENCLATURE

Indices

c	Index of critical appliance
d	Deferrable appliance index
r	Regulatable appliance index
i	Iteration of the TRPO algorithm
l	Index of hidden layer of the policy network
n	Index of smart appliance
t	Index of time slot.

Variables

Δ t	Interval of a time slot (hour)
t_{α}^{n}	Task starting time of the appliance n
t_{β}^{n}	Task deadline of the appliance n
s_{t}^{n}	Operating state of the appliance n at time slot t
o_{t}^{n}	Task status of the nth appliance at time slot t
ρ_{t}^{n}	Task progress of the nth appliance at time slot t
τ_{t}^{n}	Task attribute of the nth appliance at time slot t
s_t	The state of the MDP at time step t
a_t	The action of the MDP at time step t
r_t	The reward of the MDP at time step t
u_{t}^{d}	Binary control variable of deferrable appliance d
P_{t}^{AC}	Power consumption of the AC at time slot t (kW)
P_{t}^{EWH}	Power consumption of the EWH at time slot t (kW)
P_{t}^{EV}	EV charging/discharging power at time slot t (kW)
P_{t}^{g}	Total power consumption at time slot t (kW)
T_{t}^{AC}	Indoor air temperature at time slot t (°C)
T_{t}^{out}	Outdoor air temperature at time slot t (°C)
T_{t}^{EWH}	Water temperature in the EWH at time slot t (°C)
E_{t}^{EV}	EV battery energy at time slot t (kWh)
F_t	Water flow rate at time slot t (L/hour)
ΔF_t	Random variation in water flow rate (L/hour)
Υ_t	The real-time electricity price at time slot t
ς	Coefficient of inclining block rate (IBR) price
θ	The parameters of the policy network
ϑ	The parameters of the value network.

Constants

N	Number of smart appliances
N_c	Number of critical appliances
N_d	Number of deferrable appliances
K^d	Required operating time slots of appliance d
ξ	Inertia factor of the AC
ε	Inertia factor of the EWH
η^{AC}	Thermal conversion efficiency of the AC
G_h	Thermal conductivity of the house (kW/°C)
TR	Thermal resistance of the EWH (hour·m^2·°C/kJ)
SA	Surface area of the tank (m^2)
vol	Volumn of the tank (L)
d_w	Density of water (kg/L)
C_p	Specific heat of water (kJ/(°C·kg))
η_{ch}^{EV}	EV Charging efficiency
η_{dis}^{EV}	EV Discharging efficiency
P_{max}^{c}	Maximum power of critical appliance c (kW)
P_{max}^{d}	Maximum power of deferralbe appliance d (kW)
P_{max}^{AC}	Maximum power of AC (kW)
P_{max}^{EWH}	Maximum power of EWH (kW)
P_{max}^{EV}	Maximum charging/discharging power of EV (kW)
T_{set}^{AC}	Setpoint temperature of the AC thermostat (°C)
T_{set}^{EWH}	Setpoint temperature of EWH thermostat (°C)
T_{cold}^{EWH}	Inlet cold water temperature (°C)
ΔT_{thes}^{AC}	Threshold of the indoor temperature deviation (°C)
ΔT_{thes}^{EWH}	Threshold of the water temperature deviation (°C)
E_{max}^{EV}	Capactity of the EV battery (kWh)

Manuscript received October 29, 2019; revised January 30, 2020; accepted February 25, 2020. Date of publication March 3, 2020; date of current version August 21, 2020. This work was supported by the National Science Foundation under Grant ECCS 1917275. Paper no. TSG-01628-2019. *(Corresponding author: Haibo He.)*

The authors are with the Department of Electrical, Computer and Biomedical Engineering, University of Rhode Island, South Kingstown, RI 02881 USA (e-mail: hepengli@uri.edu; zwan@ele.uri.edu; haibohe@uri.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2020.2978061

SoC_{max} Maximum SoC of the EV battery
SoC_{min} Minimum SoC of the EV battery.

I. Introduction

HOME energy management system (HEMS) plays a significant role in deploying DR programs in the residential sector [1]. However, it is challenging to develop an efficient HEMS strategy for residential consumers due to the existence of randomness in residential environments. Specifically, affected by residents' living activities, the operational time and duration of home appliances are usually uncertain and difficult to forecast. The uncertainty makes it difficult for a HEMS to plan DR schedules effectively to respond to dynamic electricity prices. Besides, to efficiently operate DR appliances, accurate appliance models and parameters should be determined by domain experts to model the power characteristics and operational dynamics of these appliances. However, expert knowledge is not always available in ordinary households.

Traditionally, DR management of residential appliances is formulated as an optimization problem where the customers' electricity cost is minimized. Early studies, such as [2], [3], applied mix integer linear programming model to day-ahead scheduling of DR appliances to reduce the electricity cost of a household. However, they did not consider the randomness of the operational time of the appliances and the electricity prices. To handle the randomness, Du *et al.* [4] proposed a robust optimization approach to minimize the worst-case daily bill payment by considering the consumer's behavior uncertainty. Chen *et al.* [5] developed a scenario-based stochastic optimization approach to deal with the uncertainty in DR prices via Monte-Carlo simulation. In [6], Shafie-khah and Siano considered the uncertainty of EV's availability and solar Photovoltaics (PV) generation in a stochastic model to minimize the electricity cost. In [7], Huang *et al.* proposed a chance-constrained optimization model to ensure the probabilistic satisfaction of the operational constraints of appliances. In [8], Li *et al.* proposed a rolling horizon optimization approach to minimize the cost payment considering the uncertainty in renewable generation and electricity consumption. In [9], Yu *et al.* proposed a Lyapunov optimization algorithm to minimize the energy cost and thermal discomfort cost for online energy management of a sustainable smart home with a heating, ventilation, and air conditioning (HVAC) load. These aforementioned works are model-based, which require an explicit optimization model, a predictor, and a solver. Developing a model-based DR strategy requires to construct the models and determine the parameters. This process requires detailed domain knowledge, and the performance may deteriorate due to model inaccuracy.

Learning-based approaches that relax the requirement of an explicit model have attracted much attention in recent years. For instance, Wen *et al.* [10] proposed an RL approach for optimal scheduling of appliances in a residence. Ruelens *et al.* [11] designed a batch RL-based DR strategy for optimal control of thermostatic loads, and a special focus on electric water heater (EWH) was studied by [12]. Ahmed *et al.* [13] developed a heuristic algorithm based on binary backtracking search to optimize the energy usage of smart home appliances. However, these works did not consider the uncertainty of residents' behavior and real-time electricity prices. Ahmed *et al.* [14] considered the uncertainty of residents' behavior by using an unsupervised approach. A Hidden Markov Model (HMM) is developed to estimate the probability of each living activity and the operational time of corresponding appliances. Keerthisinghe *et al.* [15] considered the uncertainty of household consumption and PV generation by using an approximate dynamic programming (ADP) approach. A computationally efficient DR strategy based on temporal-difference learning is proposed. Bahrami *et al.* [16] considered the game between the consumers' DR strategies and utility's real-time pricing process by proposing an actor-critic RL approach for online optimal scheduling of DR appliances. Lu *et al.* [17] proposed an hour-ahead DR algorithm for HEMS based on multi-agent RL to optimize both shiftable appliances and air conditioners (ACs) considering the uncertainty in future prices. In general, these aforementioned methods do not require explicit models of the appliances, but they still need to know the distribution knowledge of the uncertainty or use hand-crafted features for agent learning.

Deep RL (DRL) techniques overcome the issue by taking advantage of the end-to-end learning ability of deep neural networks and have achieved significant success in many complex decision-making applications [18], [19]. The success has inspires many researchers to develop DRL based approaches for real-time residential DR. For example, Anvari-Moghaddam *et al.* [20] proposed an multi-agent home energy management scheme to minimize the energy cost and user's thermal discomfort where Bayesian RL and dual-iterative Q-learning are used for optimal battery bank scheduling. Valladares *et al.* [21] proposed a deep Q-learning (DQN)-based DR scheduling method for indoor air temperature control and thermal comfort management. Wan *et al.* [22] developed a DQN-based method to optimize the charging scheduling of an electric vehicle (EV) in a smart home to minimize the charging cost, and the charging constraint by departure was studied in [23]. Mocanu *et al.* [24] proposed an on-line building energy optimization method for scheduling of time-scaling and time-shifting loads using DQN and deterministic policy gradient (DPG) approaches. Yu *et al.* [25] proposed a DRL algorithm based on deep DPG (DDPG) to optimize the schedules of continuously controlled appliances, such as energy storage and HVAC systems. While these works have encouragingly applied DRL techniques to real-time residential DR problems, their approaches can only handle either discrete or continous control actions.

In a smart home, some appliances, such as washing machine, require discrete control actions (e.g., on/off control), while some others, such as EV, require continuous control actions. To solve real-time scheduling of different types of appliances, a DRL approach based on trust region policy optimization (TRPO) is proposed in this paper. Unlike traditional DRL approaches that can only handle either discrete or continuous control actions, the proposed TRPO-based approach can deal with both discrete and continuous actions to jointly optimize the schedules of all kinds of appliances. Besides, the proposed approach can directly learn from

high-dimensional sensory data of the smart home and does not require the distribution knowledge of the uncertainty.

The proposed DRL-based approach aims to find a real-time DR scheduling strategy to minimize the electricity cost of a household and maximize the resident's thermal comfort. Considering the uncertainty of the resident's behavior, electricity price, and outdoor temperature, we formulate the real-time DR management problem as a Markov decision process (MDP) with unknown transition probability. To solve the MDP, a DRL approach based on TRPO is designed, which directly learns from raw observation data of the appliance states, real-time electricity price, and outdoor temperature. Finally, simulation studies using real-world electricity price data, and outdoor temperature data are performed to verify the effectiveness of the proposed approach. Compared to the aforementioned works, the main contributions of the paper are as follows:

(1) A systematic simulation model considering the physical properties of different kinds of appliances and the resident's activities is built to simulate the electricity consumption in a household. Specifically, three kinds of appliances are considered in the simulation model, including deferrable appliances, regulatable appliances, and critical appliances.

(2) An MDP model with unknown transition probability is developed to formulate the real-time DR management problem, where all kinds of appliances are integrated and optimized jointly. The uncertainty of the resident's activities, real-time DR electricity price, and outdoor temperature are taken into account to formulate realistic scenarios.

(3) A model-free approach based on DRL is proposed to solve the real-time DR scheduling. Specifically, a neural network is designed to approximate the DR scheduling policy and trained by TRPO. The designed neural network can generate both discrete and continuous DR actions for jointly scheduling of all different types of appliances.

The rest of the paper is organized as follows. Section II gives the simulation model of the home appliances. Section III models the real-time DR scheduling problem as an MDP. Section IV elaborates the proposed DRL-based approach. Cases studies are given in Section V. Finally, Section VI draws the conclusions.

II. Modeling of Home Appliances

Consider a smart home wherein a set of N smart appliances are equipped (Fig. 1). For each appliance $n \in \{1, \ldots, \mathrm{N}\}$, its operational state s_t^n is represented by

$$\mathrm{s}_t^n = \left(\mathrm{o}_t^n, \rho_t^n, \tau_t^n\right), \quad \forall t \tag{1}$$

where $\mathrm{o}_t^n \in \{0, 1\}$ denotes the operating status and its value is 1 if the appliance operates with a task or 0 if otherwise; ρ_t^n measures the task progress; τ_t^n is an attribute variable of the appliance. Next, we formulate the state s_t^n for each category of appliances.

A. Deferrable Appliances

Assume a deferrable appliance d needs continuous operation of K^d time slots to fulfill a task. Denoting the starting time and deadline of the task by t_α^d and t_β^d ($\mathrm{t}_\beta^d > \mathrm{t}_\alpha^d + \mathrm{K}^d$), respectively, the state s_t^d of the appliance is defined as

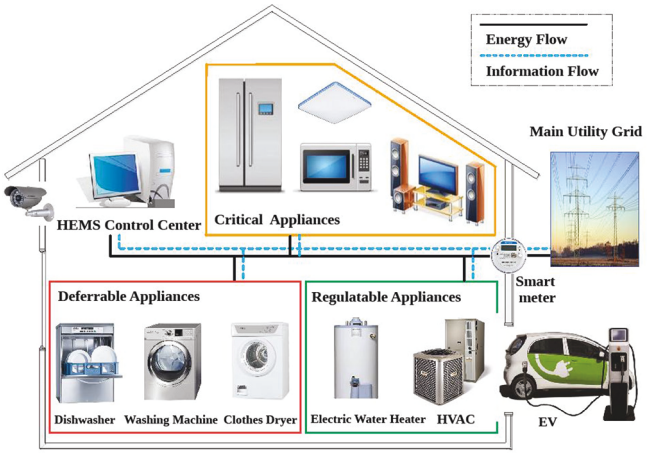


Fig. 1. Smart house that is equipped with three kinds of appliances: deferrable appliances, regulatable appliances, and critical appliances.

$$\left(\mathrm{o}_t^d, \rho_t^d, \tau_t^d\right) = \begin{cases} \left(1, \sum_{k=\mathrm{t}_\alpha^d}^{t-1} \mathrm{u}_k^d / \mathrm{K}^d, t_\beta^d - t\right), & t \in \left[\mathrm{t}_\alpha^d, \mathrm{t}_\beta^d\right] \\ (0, 0, 0), & \text{otherwise} \end{cases} \tag{2}$$

where the binary variable u_t^d determines whether to carry out the task ($\mathrm{u}_t^d = 1$) or not ($\mathrm{u}_t^d = 0$); the task process ρ_t^d measures how much the task has been fulfilled so far, and the attribute variable τ_t^d is the remaining time slots $t_\beta^d - t$ to the deadline.

The control variable u_t^d is constrained by

$$\mathrm{u}_t^d = \begin{cases} 1, & \text{if } \mathrm{u}_{t-1}^d = 1 \text{ and } 0 < \rho_t^d < 1, \ \forall t, & \text{(3a)} \\ 1, & \text{if } \rho_t^d = 0 \text{ and } t = \mathrm{t}_\beta^d - \mathrm{K}^d & \text{(3b)} \\ 0, & \text{if } t \notin \left[\mathrm{t}_\alpha^d, \ \mathrm{t}_\beta^d\right] & \text{(3c)} \end{cases}$$

where Eq. (3a) forces the appliance to operate continuously, Eq. (3b) constrains it to finish the task by deadline, and Eq. (3c) restricts the action u_t^d to 0 when the appliance is OFF.

B. Regulatable Appliances

For regulatable appliances, the power consumption is continuously adjustable. In this paper, we consider three regulatable appliances, i.e., AC, EWH, and EV.

1) AC: The state $\mathrm{s}_t^{\mathrm{AC}}$ of the AC is defined as

$$\left(\mathrm{o}_t^{\mathrm{AC}}, \rho_t^{\mathrm{AC}}, \tau_t^{\mathrm{AC}}\right) = \left(1, \mathrm{T}_t^{\mathrm{AC}} - \mathrm{T}_{\mathrm{set}}^{\mathrm{AC}}, \mathrm{T}_{\mathrm{set}}^{\mathrm{AC}}\right), \ \forall t. \tag{4}$$

The indoor temperature $\mathrm{T}_t^{\mathrm{AC}}$ is modeled as [1], [9],

$$\mathrm{T}_{t+1}^{\mathrm{AC}} = \xi \cdot \mathrm{T}_t^{\mathrm{AC}} + (1 - \xi)\left(\mathrm{T}_t^{\mathrm{out}} - \eta^{\mathrm{AC}} \cdot \mathrm{P}_t^{\mathrm{AC}} \, \Delta t / \mathrm{G_h}\right), \tag{5a}$$

$$0 \le \mathrm{P}_t^{\mathrm{AC}} \le \mathrm{P}_{\max}^{\mathrm{AC}}. \tag{5b}$$

2) EWH: The state of EWH $\mathrm{s}_t^{\mathrm{EWH}}$ is defined by

$$\left(\mathrm{o}_t^{\mathrm{EWH}}, \rho_t^{\mathrm{EWH}}, \tau_t^{\mathrm{EWH}}\right) = \left(1, \mathrm{T}_t^{\mathrm{EWH}} - \mathrm{T}_{\mathrm{set}}^{\mathrm{EWH}}, \mathrm{T}_{\mathrm{set}}^{\mathrm{EWH}}\right), \forall t. \tag{6}$$

The dynamics of the water temperature is modeled by [26]

$$\mathrm{T}_{t+1}^{\mathrm{EWH}} = \varepsilon \, \mathrm{T}_t^{\mathrm{EWH}} + (1 - \varepsilon)\left(\mathrm{W} \, \mathrm{T}_t^{\mathrm{AC}} + \mathrm{B}_t \, \mathrm{T}_{\mathrm{cold}}^{\mathrm{EWH}} + \mathrm{Q}_t\right) \mathrm{R}' \tag{7a}$$

$$\varepsilon = \exp\left(-\frac{\Delta t}{\mathrm{R'Z}}\right),\ \mathrm{R}' = \frac{1}{\mathrm{W}+\mathrm{B}_t},\ \mathrm{Z} = \mathrm{vol}\cdot \mathrm{d_w}\cdot \mathrm{C_p} \tag{7b}$$

$$\mathrm{W} = \frac{\mathrm{SA}}{\mathrm{TR}},\ \mathrm{B}_t = \mathrm{F}_t\cdot \mathrm{d_w}\cdot \mathrm{C_p},\ \mathrm{Q}_t = 3600\,\mathrm{P}_t^{\mathrm{EWH}}\,\Delta t \tag{7c}$$

$$0 \le \mathrm{P}_t^{\mathrm{EWH}} \le \mathrm{P}_{\max}^{\mathrm{EWH}}. \tag{7d}$$

3) EV: Assuming the EV arrives home at $\mathrm{t}_\alpha^{\mathrm{EV}}$ and departs at $\mathrm{t}_\beta^{\mathrm{EV}}$, we define its state $\mathrm{s}_t^{\mathrm{EV}}$ as

$$\left(\mathrm{o}_t^{\mathrm{EV}}, \rho_t^{\mathrm{EV}}, \tau_t^{\mathrm{EV}}\right) = \begin{cases} (1, \mathrm{SoC}_t, t), & t \in \left[\mathrm{t}_\alpha^{\mathrm{EV}}, \mathrm{t}_\beta^{\mathrm{EV}}\right], \\ (0,0,0), & \text{otherwise}, \end{cases} \tag{8}$$

The dynamics of EV battery is modeled by

$$\mathrm{SoC}_{t+1} = \begin{cases} \mathrm{SoC}_t + \eta_{\mathrm{ch}}^{\mathrm{EV}}\cdot \mathrm{P}_t^{\mathrm{EV}}\,\Delta t/\,\mathrm{E}_{\max}^{\mathrm{EV}}, & \mathrm{P}_t^{\mathrm{ev}} \ge 0, \\ \mathrm{SoC}_t + 1/\eta_{\mathrm{dis}}^{\mathrm{EV}}\cdot \mathrm{P}_t^{\mathrm{EV}}\,\Delta t/\,\mathrm{E}_{\max}^{\mathrm{EV}}, & \mathrm{P}_t^{\mathrm{ev}} < 0 \end{cases} \tag{9a}$$

$$\mathrm{SoC}_{\min} \le \mathrm{SoC}_t \le \mathrm{SoC}_{\max}. \tag{9b}$$

The charging/discharging power is contrained by

$$-\mathrm{P}_{\max}^{\mathrm{EV}} \le \mathrm{P}_t^{\mathrm{EV}} \le \mathrm{P}_{\max}^{\mathrm{EV}}, \text{if } t \in \left[\mathrm{t}_\alpha^{\mathrm{EV}}, \mathrm{t}_\beta^{\mathrm{EV}}\right], \tag{10a}$$

$$\mathrm{P}_t^{\mathrm{EV}} = 0, \text{ otherwise}. \tag{10b}$$

C. Critical Appliances

Critical appliances do not participate in DR. Assuming a critical appliance c operates in the interval $[\mathrm{t}_\alpha^c, \mathrm{t}_\beta^c]$, its state s_t^c is defined by

$$\left(\mathrm{o}_t^c, \rho_t^c, \tau_t^c\right) = \begin{cases} \left(1, t-\mathrm{t}_\alpha^c, t\right), & t \in \left[\mathrm{t}_\alpha^c, \mathrm{t}_\beta^c\right]. \\ (0,0,0), & \text{otherwise} \end{cases} \tag{11}$$

The power consumption of the appliance is calculated by

$$\mathrm{P}_t^c = \begin{cases} \mathrm{P}_{\max}^c, & t \in \left[\mathrm{t}_\alpha^c, \mathrm{t}_\beta^c\right], \\ 0, & \text{otherwise}. \end{cases} \tag{12}$$

III. Problem Formulation

In this section, the real-time DR scheduling is formulated as an MDP. In the MDP, the real-time DR scheduling problem is represented by a 5-tuple (S, A, Pr, R, γ), where S is the set of states, A is the set of actions; $Pr(\mathrm{s}'|\mathrm{s},\mathrm{a}) \to [0,1]$ denotes the transition probability from $\mathbf{s}$ to $\mathbf{s}'$ taking action a; $R(\mathrm{s},\mathrm{a},\mathrm{s}') \to \mathbb{R}$ is the reward function; γ is a discount factor, which balances the importance between the immediate reward and future rewards. Next, we model the components of the MDP in the following subsections.

A. The State

The state s_t of the smart home at time step t is defined as

$$\mathrm{s}_t = \left(\mathrm{s}_t^1, \ldots, \mathrm{s}_t^{\mathrm{N}}, \Upsilon_{t-T+1}, \ldots, \Upsilon_t, \mathrm{T}_{t-T+1}^{\mathrm{out}}, \ldots, \mathrm{T}_t^{\mathrm{out}}\right), \tag{13}$$

which encapsulates: 1) the states $\mathrm{s}_t^1, \ldots, \mathrm{s}_t^N$ of all appliances at time step t, 2) the real-time electricity prices $\Upsilon_{t-T+1}, \ldots, \Upsilon_t$ over the past T time steps, and 3) the outdoor temperature $\mathrm{T}_{t-T+1}^{\mathrm{out}}, \ldots, \mathrm{T}_t^{\mathrm{out}}$ over the past T time steps.

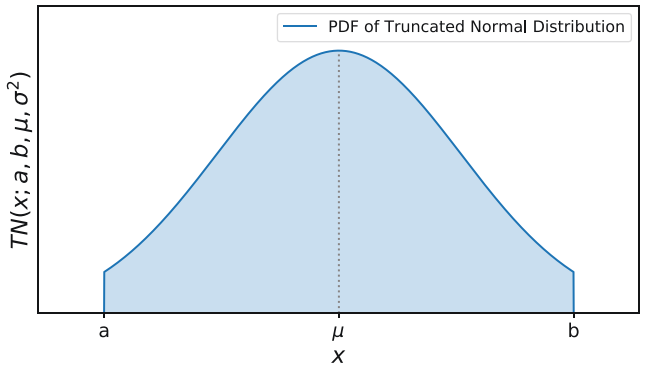


Fig. 2. Probability density function of a truncated normal distribution.

B. The Action

Given the state s_t at time step t, an action a_t is determined to control the DR appliances, which is defined as

$$\mathrm{a}_t = \left(\mathrm{u}_t^1, \ldots, \mathrm{u}_t^{\mathrm{N_d}}, \mathrm{P}_t^{\mathrm{AC}}, \mathrm{P}_t^{\mathrm{EWH}}, \mathrm{P}_t^{\mathrm{EV}}\right), \forall t, \tag{14}$$

where $\mathrm{u}_t^1, \ldots, \mathrm{u}_t^{\mathrm{N_d}}$ are binary control variables of the deferrable appliances; $\mathrm{P}_t^{\mathrm{AC}}, \mathrm{P}_t^{\mathrm{EWH}}, \mathrm{P}_t^{\mathrm{EV}}$ are continuous control variables of the regulatable appliances.

C. The State Transition Probability

Following the action a_t, the system state changes from s_t to s_{t+1} at the time step $t+1$ with the probability $P(\mathrm{s}_{t+1}|\mathrm{s}_t, \mathrm{a}_t)$.

For the state s_t^n of appliance $n \in \{1, \ldots, \mathrm{N}\}$, the transition probability is affected by resident's behavior. Since resident's behavior is random, it is unknown when an appliance is triggered to carry out a task in advance. To model the randomness, we assume the task starting time t_α^n of appliance n follows a truncated normal distribution $\mathcal{TN}(\mathrm{t}_\alpha^n; \mu, \sigma^2, a, b)$ [16], whose probability density function (PDF) is (Fig. 2)

$$f\left(\mathrm{t}_\alpha^n\right) = \frac{1}{\delta}\frac{\phi\left(\frac{\mathrm{t}_\alpha^n - \mu}{\delta}\right)}{\Phi\left(\frac{b-\mu}{\delta}\right) - \Phi\left(\frac{a-\mu}{\delta}\right)}, \tag{15}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$ is the PDF of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function. Then, the probability of appliance n being triggered with a task at time step $t+1$ can be calculated by [16]

$$Pr\left\{\mathrm{o}_{t+1}^n = 1 | \mathrm{o}_t^n = 0\right\} = \frac{\int_t^{t+1} f\left(\mathrm{t}_\alpha^n\right) d\mathrm{t}_\alpha^n}{1 - \int_a^t f\left(\mathrm{t}_\alpha^n\right) d\mathrm{t}_\alpha^n}. \tag{16}$$

After being triggered, the appliance is scheduled to carry out its task and the state transition is calculated according to the models formulated in Section II. Once the task is finished, the appliance is turned off at the time step t_β^n. We assume that t_β^n is a random variable following a truncated normal distribution, and the probability $Pr\{\mathrm{o}_{t+1}^n = 0 | \mathrm{o}_t^n = 1\}$ of being turned off at time step $t+1$ can be calculated in similar way as Eq. (16).

We also consider the randomness of resident's hot water demand. In our study, we use the average residential hot water demand profile in [26] as a base value $\mathrm{F}_t^{\mathrm{Base}}$ of the hot water

flow rate and its true value is modeled as $\mathrm{F}_t = \mathrm{F}_t^{\mathrm{Base}} + \Delta\mathrm{F}_t$, where $\Delta\mathrm{F}_t$ is a random variable following normal distribution.

For the electricity prices $\Upsilon_{t-T+1}, \ldots, \Upsilon_t$ and outdoor temperature $\mathrm{T}^{\mathrm{out}}_{t-T+1}, \ldots, \mathrm{T}^{\mathrm{out}}_t$ in s_t, the state transition is influenced b many random factors. Constructing an explicit joint probability distribution for such a multivariate random variable is challenging. To handle this problem, a deep learning method is designed to implicitly learn the transition probability from real-world data samples of the real-time electricity price and outdoor temperature.

D. The Reward Function

From the resident's perspecitve, we model the reward as

$$\mathrm{r}_t = R(\mathrm{s}_t, \mathrm{a}_t, \mathrm{s}_{t+1}) = \mathrm{I}_t^{\mathrm{comf}} - \mathrm{C}_t^{\mathrm{elec}} - \mathrm{E}_t^{\mathrm{range}}, \ \forall t, \tag{17}$$

where $\mathrm{I}_t^{\mathrm{comf}}$ is an index of resident's thermal comfort measured in \$, $\mathrm{C}_t^{\mathrm{elec}}$ is the electricity cost measured in \$, and $\mathrm{E}_t^{\mathrm{range}}$ reflects the EV range anxiety in \$.

1) Thermal Comfort: The resident's thermal comfort index $\mathrm{I}_t^{\mathrm{comf}}$ is calculated by [3]

$$\begin{aligned}\mathrm{I}_t^{\mathrm{comf}} = {} & \mathrm{w}_1 \exp\Big\{\min\Big(0, \Delta\,\mathrm{T}^{\mathrm{AC}}_{\mathrm{thes}} - \Big|\mathrm{T}^{\mathrm{AC}}_{\mathrm{set}} - \mathrm{T}^{\mathrm{AC}}_t\Big|\Big)\Big\} \\ & + \ \mathrm{w}_2 \exp\big\{\min\big(0, \Delta\,\mathrm{T}^{\mathrm{EWH}}_{\mathrm{thes}} - \big|\mathrm{T}^{\mathrm{EWH}}_{\mathrm{set}} - \mathrm{T}^{\mathrm{EWH}}_t\big|\big)\big\},\end{aligned} \tag{18}$$

where the thermal comfort is measured based on the deviation $|\,\mathrm{T}^{\mathrm{AC}}_{\mathrm{set}} - T^{\mathrm{AC}}_t|$. When the deviation $|\,\mathrm{T}^{\mathrm{AC}}_{\mathrm{set}} - \mathrm{T}^{\mathrm{AC}}_t\,|$ is smaller than the threshold $\Delta\,\mathrm{T}^{\mathrm{AC}}_{\mathrm{thes}}$, the thermal comfort value reaches its maximum. If the deviation becomes larger than the threshold, the thermal comfort value decreases. The weigting factors w_1 and w_2 measured in \$/°C are introduced to map the comfort terms into money.

2) Electricity Cost: The electricity cost is calculated by

$$\mathrm{C}_t^{\mathrm{elec}} = \mathrm{P}_t^{\mathrm{g}} \cdot \Delta t \cdot price_t, \tag{19a}$$

$$\mathrm{P}_t^{\mathrm{g}} = \sum_{c=1}^{\mathrm{N_c}} \mathrm{P}_t^c + \sum_{d=1}^{N_{\mathrm{d}}} \mathrm{u}_t^d\, \mathrm{P}^d_{\max} + \mathrm{P}_t^{\mathrm{AC}} + \mathrm{P}_t^{\mathrm{EWH}} + \mathrm{P}_t^{\mathrm{EV}}. \tag{19b}$$

The electricity price $price_t$ is determined based on a real-time pricing (RTP) scheme combined with the inclining block rate (IBR) [27],

$$price_t = \begin{cases} \Upsilon_t, & \text{if } \ 0 \le \mathrm{P}_t^{\mathrm{g}} \le \mathrm{P}^{\mathrm{g}}_{\max} \\ \varsigma \cdot \Upsilon_t, & \text{if } \ \mathrm{P}_t^{\mathrm{g}} > \mathrm{P}^{\mathrm{g}}_{\max}. \end{cases} \tag{20}$$

where the resident is charged with the RTP price Υ_t if the total power consumption $\mathrm{P}_t^{\mathrm{g}}$ is smaller than the threshold $\mathrm{P}^{\mathrm{g}}_{\max}$; or a higher IBR price, i.e., $\varsigma \cdot \Upsilon_t$ and $\varsigma > 1$, is applied if the total power $\mathrm{P}_t^{\mathrm{g}}$ exceeds the threshold $\mathrm{P}^{\mathrm{g}}_{\max}$.

3) Range Anxiety: The range anxiety measures the resident's fear that the EV has insufficient energy to reach its destination, which is measured by,

$$\mathrm{E}_t^{\mathrm{range}} = \mathrm{w}_3\big(\mathrm{E}_t^{\mathrm{EV}} - \mathrm{E}^{\mathrm{EV}}_{\max}\big)^2, \ t = \mathrm{t}_\beta^{\mathrm{EV}}, \tag{21}$$

where $\mathrm{E}_t^{\mathrm{EV}} - \mathrm{E}^{\mathrm{EV}}_{\max}$ represents the uncharged battery energy when the EV departs at $\mathrm{t}_\beta^{\mathrm{EV}}$, and the squared term measures the range anxiety in \$/kWh^2 [22]. The weighting factor w_3 measured in \$/kWh^2 is introduced to map the range anxiety into money.

E. The Objective Function

The objective is to find the optimal DR policy π^* that maximizes the expectation of the discounted cumulative rewards over a horizon of T time steps,

$$\max_{\pi \in \Pi} J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{T-1} \gamma^t\, \mathrm{r}_t\right] \tag{22}$$

where $0 < \gamma \le 1$ is the discount factor, Π is the set of all policies, and $\mathbb{E}_{\tau\sim\pi}$ denotes the expected value over the trajectory $\tau = (\mathrm{s}_0, \mathrm{a}_0, \mathrm{s}_1, \ldots)$ following the policy π.

IV. Proposed Approach

The formulated MDP model has both discrete and continuous high-dimensional actions. Traditional RL algorithms have difficulties in dealing with such problem due to the curse of the dimensionality. To solve the MDP, we propose a novel TRPO-based policy optimization approach. In the proposed approach, a neural network based stoachstic policy is degined to approximate the optimal policy $\pi^*(\mathrm{a}_t \,|\, \mathrm{s}_t)$. The neural network based policy can generate both discrete and continuous actions from the observation of the appliance states, electricity price, and outdoor temperature. To optimize the NN-based policy, the TRPO algorithm is introduced to train the NN-based policy.

A. Neural Network-Based Policy

Since the formulated MDP problem contains both discrete and continuous actions

$$\mathrm{a}_t = \Big(\underbrace{\mathrm{u}_t^1, \ldots, \mathrm{u}_t^{\mathrm{D}}}_{\text{discrete}}, \underbrace{\mathrm{P}_t^{\mathrm{AC}}, \mathrm{P}_t^{\mathrm{EWH}}, \mathrm{P}_t^{\mathrm{EV}}}_{\text{continuous}}\Big), \forall t, \tag{23}$$

we use the following probability distribution

$$\pi(\mathrm{a}_t|\mathrm{s}_t) = \begin{cases} \mathcal{B}(p_d(\mathrm{s}_t)), & \text{if } \mathrm{a}_t \in \big\{\mathrm{u}_t^1, \ldots, \mathrm{u}_t^{\mathrm{N_d}}\big\}, \\ \mathcal{N}\big(\mu_r(\mathrm{s}_t), \sigma_r^2\big), & \text{otherwise}, \end{cases} \tag{24}$$

to approximate the optimal policy. When the action is discrete, the approximate policy $\pi(\mathrm{a}_t|\mathrm{s}_t)$ is a Bernoulli distribution $\mathcal{B}(p_d(\mathrm{s}_t))$, where $p_d(\mathrm{s}_t)$ represents the probability of switching ON the deferrable appliance d to carry out its task, i.e., $p(\mathrm{u}_t^d = 1|\mathrm{s}_t)$. When the action is continuous, the approximate policy $\pi(\mathrm{a}_t|\mathrm{s}_t)$ is a Gaussian distribution $\mathcal{N}(\mu_r(\mathrm{s}_t), \sigma_r^2)$, where $\mu_r(\mathrm{s}_t)$ and σ_r are the mean and standard deviation for the regulatable appliance r, respectively.

To determine the parameters $p_d(\mathrm{s}_t)$, $\mu_r(\mathrm{s}_t)$, σ_r of the approximate policy $\pi(\mathrm{a}_t|\mathrm{s}_t)$, a neural network, which is referred to as policy network, is designed to learn these parameters. The architecture of the policy network is presented in Fig. 3. As shown in Fig. 3, the inputs of the policy network are the states $\mathrm{s}_t^1, \ldots, \mathrm{s}_t^N$ of the appliances, the past T-step electricity price $\Upsilon_{t-T+1}, \ldots, \Upsilon_t$, and the past T-step outdoor temperature $\mathrm{T}^{\mathrm{out}}_{t-T+1}, \ldots, \mathrm{T}^{\mathrm{out}}_t$. The outputs are the probability $p_d(\mathrm{s}_t)$ of the Bernoulli distribution $\mathcal{B}$ for the discrete actions, and the mean $\mu_r(\mathrm{s}_t)$ and logarithmic standard deviation $\log(\sigma_r)$ of the Gaussian distribution $\mathcal{N}$ for the continuous actions,

$$\begin{aligned} p_d(\mathrm{s}_t;\theta) &= \mathrm{Sigmoid}([\mathbf{W}]_d \cdot f(\mathrm{s}_t) + [\mathbf{B}]_d), \\ \mu_r(\mathrm{s}_t;\theta) &= [\mathbf{W}]_{r+\mathrm{N_d}} \cdot f(\mathrm{s}_t) + [\mathbf{B}]_{r+\mathrm{N_d}}, \\ \log\sigma_r(\theta) &= [\mathbf{W}_\sigma]_r, \end{aligned} \tag{25}$$

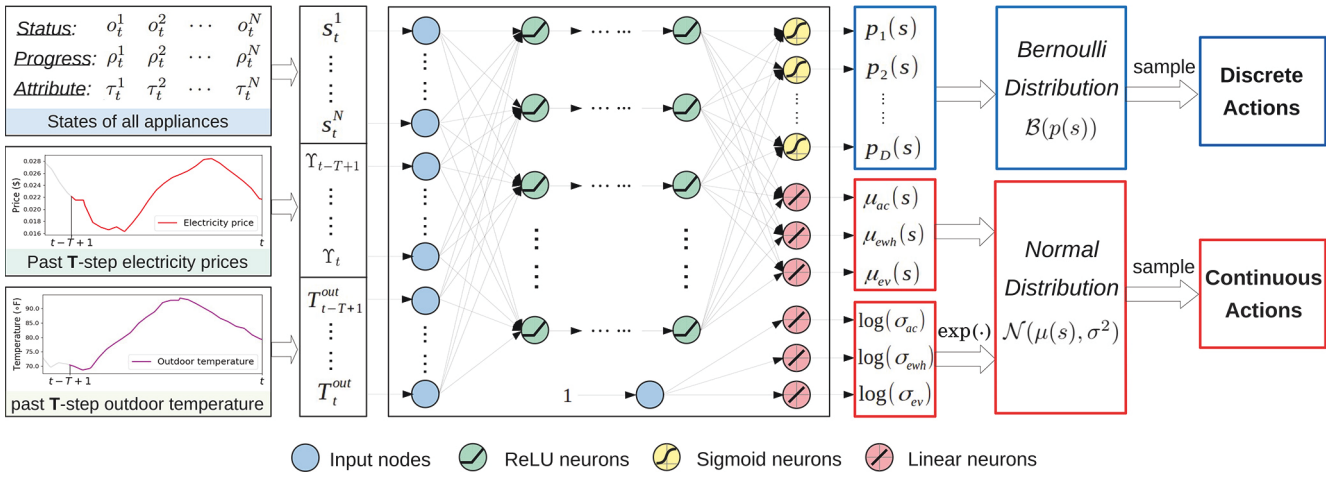


Fig. 3. The overall architecture of the policy network. The inputs of the policy network are the states of all appliances $s_t^1, \ldots, s_t^N$, the past T-step electricity price $\Upsilon_{t-T+1}, \ldots, \Upsilon_t$, and the past T-step outdoor temperature $T_{t-T+1}^{out}, \ldots, T_t^{out}$. The neural network extracts features from this information; then it outputs the probabilities $p_d(s), d = 1, \ldots, N_d$ of the Bernoulli distribution and the mean values $\mu_{AC}(s), \mu_{EWH}(s), \mu_{EV}(s)$ and logarithmic standard deviations $\log\sigma_{AC}, \log\sigma_{EWH}, \log\sigma_{EV}$ of the normal distribution. By sampling from the Bernoulli distribution, the neural network based policy generates discrete actions for the deferrable appliances. By sampling from the normal distribution, it generates continuous actions for the regulatable appliances.

where $\mathbf{W}, \mathbf{B}, \mathbf{W}_\sigma \in \theta$ are the output layer's weight and bias matrices of the policy network, respectively; $[\mathbf{W}]_i$ denotes the ith row of the matrix $\mathbf{W}$; $\text{Sigmoid}(x) = 1/(1+e^{-x})$ is the sigmoid function; $f(s_t)$ is the latent feature extracted by the hidden layers from the input s_t, which is calculated by

$$\begin{aligned} f(\mathrm{s}_t) &= \text{ReLU}\ (\mathbf{W}_L \cdot v_L(\mathrm{s}_t) + \mathbf{B}_L), \\ v_{l+1}(s_t) &= \text{ReLU}\ (\mathbf{W}_l \cdot v_l(\mathrm{s}_t) + \mathbf{B}_l),\ \ l = 1, \ldots, \mathrm{L}\text{-}1, \\ v_1(\mathrm{s}_t) &= \mathrm{s}_t, \end{aligned} \tag{26}$$

where $\mathbf{W}_l, \mathbf{B}_l \in \theta, l = 1, \ldots, L$ denote the lth hidden layer's weight and bias matrices, respectively; $\text{ReLU}(x) = \max(0, x)$ is the Rectified Linear Units function.

B. Policy Optimization

To optimize the approximate policy, we need to find the best parameters θ^* of the policy network to maximize the objective $J(\pi)$,

$$\theta^* = \max_\theta J(\pi_\theta). \tag{27}$$

To this end, we use a policy search method based on TRPO [28], which iteratively updates the parameters $\theta^0 \rightarrow \theta^1 \rightarrow \cdots$ of the policy network. The TRPO algorithm is explained as follows.

Let $\pi_{\theta^{i+1}}$ and π_{θ^i} denote two different policies and $\alpha = D_{\text{KL}}^{\max}(\theta^i||\theta^{i+1}) = \max_{\mathrm{s}} D_{\text{KL}}(\pi_{\theta^i}(\cdot|\mathrm{s})||\pi_{\theta^{i+1}}(\cdot|\mathrm{s}))$ denote the maximum KL divergence of π_{θ^i} and $\pi_{\theta^{i+1}}$. Schulman *et al.* prove that, when α is sufficiently samll, the objectives $J(\pi_{\theta^{i+1}})$ and $J(\pi_{\theta^i})$ meet the following inequation [28]

$$\begin{aligned} J\left(\pi_{\theta^{i+1}}\right) \geq J\left(\pi_{\theta^i}\right) &+ \sum_s \rho_{\pi_{\theta^i}}(s) \sum_a \pi_{\theta^{i+1}}(\mathrm{a}|\mathrm{s}) A_{\pi_{\theta^i}}(\mathrm{s}, \mathrm{a}) \\ &- \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha \end{aligned} \tag{28}$$

where $\rho_{\pi_{\theta^i}}(s)$ is the stationary distribution of the state s following the policy π_{θ^i}; $\epsilon = \max_{\mathrm{s,a}} A_{\pi_{\theta^i}}(\mathrm{s}, \mathrm{a})$; $A_\pi(\mathrm{s}, \mathrm{a})$ is the advantage function,

$$A_\pi(\mathrm{s}, \mathrm{a}) = V_\pi(\mathrm{s}') + R\left(\mathrm{s}, \mathrm{a}, s'\right) - V_\pi(\mathrm{s}), \tag{29}$$

where $V_\pi(\mathrm{s}) = \mathbb{E}_{\mathrm{a}_t, \mathrm{s}_{t+1}, \ldots}[\sum_{l=0}^{\infty} \gamma^l \mathrm{r}_{t+l} | \mathrm{s}_t = \mathrm{s}]$ is the value function following the policy π.

The inequation (28) is important because it provides a lower bound of the performance $J(\pi_\theta)$ when updating the policy network from θ^i to θ^{i+1}. In order to obtain an improved policy, we can maximize the lower bound, i.e., the right part of the inequation (28) with resepect to θ^{i+1}

$$\begin{aligned} &\max_{\theta^{i+1}}\ L_{\theta^i}\left(\theta^{i+1}\right) - C \cdot D_{\text{KL}}^{\max}\left(\theta^i \| \theta^{i+1}\right), \\ &L_{\theta^i}\left(\theta^{i+1}\right) = J\left(\pi_{\theta^i}\right) + \sum_{\mathrm{s}} p_{\pi_{\theta^i}}(\mathrm{s}) \sum_{\mathrm{a}} \pi_{\theta^{i+1}}(\mathrm{a}|\mathrm{s}) A_{\pi_{\theta^i}}(\mathrm{s}, \mathrm{a}), \end{aligned} \tag{30}$$

where $C = 4\epsilon\gamma/(1-\gamma)^2$. The KL-divergence $D_{\text{KL}}^{\max}(\theta^i||\theta^{i+1})$ can be viewed as a penalty term to prevent large step updates.

Schulman *et al.* [28] prove that the update rule (30) is guaranteed to generate a monotonically nondecreasing sequence of policies $J(\pi_{\theta^0}) \leq J(\pi_{\theta^1}) \leq \cdots$. Since it is difficult to determine the best value for the penalty coefficient C, we can restrict the KL divergence by a trust region δ and maximize the following surrogate objective [28]

$$\begin{aligned} &\max_{\theta^{i+1}}\ \sum_{\mathrm{s}} \rho_{\pi_{\theta^i}}(\mathrm{s}) \sum_a \pi_{\theta^{i+1}}(\mathrm{a}|\mathrm{s}) A_{\pi_{\theta^i}}(\mathrm{s}, \mathrm{a}) \\ &s.t.\ \ D_{\text{KL}}^{\max}\left(\theta^i||\theta^{i+1}\right) \leq \delta. \end{aligned} \tag{31}$$

In practice, the parameter update method (31) can be approximately solved by using Monte Carlo simulation,

$$\begin{aligned} &\max_{\theta^{i+1}}\ \mathbb{E}_{\mathrm{s}\sim\rho_{\pi_{\theta^i}}, \mathrm{a}\sim\ \pi_{\theta^i}} \frac{\sum_a \pi_{\theta^{i+1}}(\mathrm{a}|\mathrm{s})}{\sum_{\mathrm{a}} \pi_{\theta^i}(\mathrm{a}|\mathrm{s})} A_{\pi_{\theta^i}}(\mathrm{s}, \mathrm{a}) \\ &s.t.\ \ \mathbb{E}_{\mathrm{s}\sim\rho_{\pi_{\theta^i}}} D_{\text{KL}}\left(\pi_{\theta^i}(\cdot||\mathrm{s})||\pi_\theta(\cdot||\mathrm{s})\right) \leq \delta. \end{aligned} \tag{32}$$

Algorithm 1 Training of the Neural Network-Based Policy

Inputs: Initialized θ^0, ϑ^0, I, δ, β, $\mathcal{D}$.
for i = 0, I **do**
  for d = 1, D **do**
    Set time step $t \rightarrow 0$;
    Reset the state s_t of the smart home in Eq. (13);
    while $t < T$ **do**
      Sample an action a_t according to $\pi_{\theta^i}(\mathrm{a}|\mathrm{s}_t)$;
      Clip the action a_t by its nearest feasible value;
      Observe the next state s_{t+1};
      Calculate the reward r_t according to Eq. (17);
      Set $t \rightarrow t+1$;
    end while
    Store the trajectory $\tau_d = (\mathrm{s}_0, \mathrm{a}_0, \mathrm{r}_0, \mathrm{s}_1, \dots)$ in $\mathcal{D}$;
  end for
  Calculate the sample estimate of $A_{\pi_{\theta^i}}(\mathrm{s}, \mathrm{a})$;
  Sovle the problem (32);
  Update the policy network θ^i to the solution of (32);
  Update the value network ϑ^i by (33);
end for
Output: Optimal policy π_{θ^I}.

To solve the constrained optimization problem (32), the conjugate gradient algorithm is used, followed by a backtracking line search, as suggested by [28]. To calculate the advantage function $A_{\pi_\theta}(\mathrm{s}, \mathrm{a})$, we approximate the value function $V_{\pi_\theta}(\mathrm{s})$ by a neural network. The neural network has the same architecture as the policy network except that the dimensionality of its out is 1. We refer to this neural network as *value network* and denote it as $V_{\pi_\theta}(\mathrm{s}; \vartheta)$, where ϑ are the parameters of the value network. The value network is trained by gradient descent

$$\vartheta^{i+1} = \vartheta^i + \beta \bigtriangledown_{\vartheta^i} \mathbb{E}_{\substack{\mathrm{s}\sim\rho_{\pi_{\theta^i}} \\ \mathrm{a}\sim\pi_{\theta^i} \\ \mathrm{s}'\sim P}} \times \left[\left(V_{\pi_{\theta^i}}\left(\mathrm{s}; \vartheta^i\right) \right) - \sum_{t=0}^{\infty} \gamma^t R\left(\mathrm{s}, \mathrm{a}, \mathrm{s}'\right) \right]^2 \quad (33)$$

where β is a step size parameter. The training procedure of the neural networks is summarized in Algorithm 1.

V. Case Studies

A. Experimental Setup

For case studies, we consider three deferrable appliances: a dishwasher (DW), a washing machine (WM) and a clothes dryer (CD), three regulatable appliances: an AC, an EWH and an EV, and five critical appliances: a refrigerator, a hairdryer, a vaccum, a laptop, a television, and lights. We map one day into $T = 144$ time slots and each time slot has $\Delta t = 10$ minutes. The DR scheduling starts at each day's 8:00 a.m. The parameters of the three kinds of appliances are listed in Tables I, II, and III, respectively. In general, CD should starts to work after WM finishes the task. In real-world scenarios, the customer could allow the operation of CD to be delayed for a while but the delay should not be long. In our study, we constrain the CD to be activated once the WM finishes its task at time step t, i.e.,

$$\mathrm{t}_\alpha^{\mathrm{CD}} = t, \text{ if } \rho_{t-1}^{\mathrm{WM}} < 1 \text{ and } \rho_t^{\mathrm{WM}} = 1, \quad (34)$$

TABLE I
Operational Parameters of the Deferrable Appliances

Appliance	Power	Starting Time t_α	Deadline t_β	K^d
DW	1.5kW	$\mathcal{TN}(63, 3^2, 60, 66)$	$\mathcal{TN}(90, 3^2, 87, 93)$	3
WM	0.7kW	$\mathcal{TN}(12, 6^2, 9, 15)$	$\mathcal{TN}(54, 6^2, 48, 60)$	6
CD	1.2kW	$\mathrm{t}_\alpha^{\mathrm{CD}}$ (Eq.(34))	$\mathrm{t}_\alpha^{\mathrm{CD}} + 8$	5

TABLE II
Operational Parameters of the Regulatable Appliances

Appliance	Parameters
AC	$\mathrm{T}_{\mathrm{set}}^{\mathrm{AC}} = 24°\mathrm{C}$, $\Delta\mathrm{T}_{\mathrm{thre}}^{\mathrm{AC}} = 2°\mathrm{C}$, $\mathrm{P}_{\mathrm{max}}^{\mathrm{AC}} = 2.5\mathrm{kW}$ $\xi = 0.968$, $\eta^{\mathrm{AC}} = 1.0$, $\mathrm{G_h} = 7.27\text{e-}3$ kW/°C
EWH	$\mathrm{T}_{\mathrm{set}}^{\mathrm{EWH}} = 52°\mathrm{C}$, $\Delta\mathrm{T}_{\mathrm{thre}}^{\mathrm{EWH}} = 3°\mathrm{C}$, $\mathrm{T}_{\mathrm{cold}}^{\mathrm{EWH}} = 15°\mathrm{C}$ $\mathrm{P}_{\mathrm{max}}^{\mathrm{EWH}} = 4.5\mathrm{kW}$, SA=2.238m^2, $\mathrm{d_w} = 1.0$kg/L $\mathrm{C_p} = 4.1867$kJ/(°C·kg), TR=0.73hour·m^2·°C/kJ vol=150L, $\Delta\mathrm{F}_t^{\mathrm{EWH}} \sim \mathcal{N}(0, 4.42^2)$L/hour
EV	$\mathrm{P}_{\mathrm{max}}^{\mathrm{EV}} = 6.0\mathrm{kW}$, $\mathrm{E}_{\mathrm{max}}^{\mathrm{EV}} = 24\mathrm{kWh}$, $\mathrm{SoC}_{\mathrm{min}} = 0.1$ $\mathrm{t}_\alpha^{ev} \sim \mathcal{TN}(60, 6^2, 42, 78)$, $\mathrm{SoC}_{\mathrm{max}} = 1.0$ $\mathrm{t}_\beta^{ev} \sim \mathcal{TN}(141, 3^2, 138, 144)$, $\eta_{\mathrm{ch}}^{\mathrm{EV}} = \eta_{\mathrm{dis}}^{\mathrm{EV}} = 0.98$

TABLE III
Operational Parameters of the Critical Appliances

Appliance	Power	Starting Time t_α	Duration $t_\beta - t_\alpha$
Refrigerator	0.2 kW	0	144
Hairdryer	1.0 kW	$\mathcal{TN}(75, 3^2, 72, 78)$	1
Vaccum	1.5 kW	$\mathcal{TN}(42, 6^2, 36, 48)$	$\mathcal{TN}(4, 2^2, 2, 6)$
Laptop	0.1 kW	$\mathcal{TN}(72, 3^2, 69, 75)$	$\mathcal{TN}(18, 4^2, 14, 22)$
TV	0.1 kW	$\mathcal{TN}(63, 3^2, 60, 66)$	$\mathcal{TN}(24, 4^2, 20, 28)$
Lights	0.2 kW	$\mathcal{TN}(54, 6^2, 48, 60)$	$\mathcal{TN}(30, 3^2, 27, 33)$

where $\rho_{t-1}^{\mathrm{WM}} < 1$ and $\rho_t^{\mathrm{WM}} = 1$ mean that the WM's task is in progress at time step $t-1$ but finishes at time step t. The allowable delay for the operation of CD is restricted to 30 minutes (3 time slots), i.e., $\mathrm{t}_\beta^{\mathrm{CD}} = \mathrm{t}_\alpha^{\mathrm{CD}} + \mathrm{K}^{\mathrm{CD}} + 3$.

For the proposed method, the policy network has three layers of 128 ReLU neurons. The output layer consists of 4 sigmoid neurons and 6 linear neurons. The value network has the same structure as the policy network except that the output dimensionality is 1. The policy network and value network are orthogonally initialized. The weighting factors $\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3$ are set to $\mathrm{w}_1 = \mathrm{w}_2 = 0.01$, $\mathrm{w}_3 = 0.1$. For the RPT-IBR price, we use $\mathrm{P}_{\mathrm{max}}^{\mathrm{g}} = 8$ kWh and $\varsigma = 1.4423$ [4]. Other parameters are summarized in Table IV. The training is conducted on a computer with Intel Core i7-4790 CPU @ 3.60 GHz × 8. The code is written in Python and run with TensorFlow 1.12.

B. Training and Test Datasets

The proposed approach is trained using a training set, and then evaluated on a different test set. In the training set, real-world data of electricity price [21] and outdoor temperature [22] from Jul. 1 to Aug. 31 in 2016 are used. Besides, a set of simulation data of the appliance working time $[\mathrm{t}_\alpha^n, \mathrm{t}_\beta^n]$ and hot water flow rate $\Delta\mathrm{F}_t^{\mathrm{EWH}}$ are generated by sampling from the distributions in Tables I, II, and III. In the test set, the data of

TABLE IV
HYPERPARAMETERS USED IN OUR EXPERIMENTS

Notion	Value	Description
I	5000	maximum iterations
D	100	trajectory buffer size
γ	0.995	reward discount factor
δ	0.01	trust-region of the KL-divergence
β	0.001	learning stepsize of value network

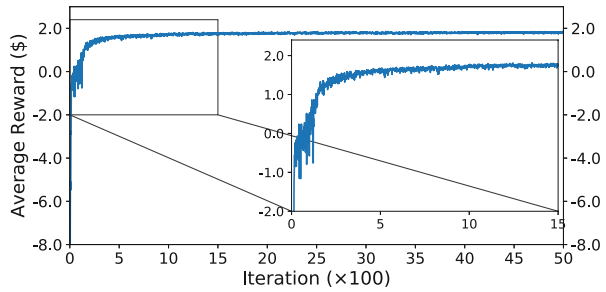


Fig. 4. Average rewards at each iteration during the training process.

electricity price [21] and outdoor temperature [22] at the same period in 2017 is used. Besides, a different set of simulation data of the appliance working time and hot water flow rate are generated for test. Note that the test set has never been shown to the proposed method during the training.

C. Benchmark Methods

1) Without DR: In this scenario, the deferrable appliances operate immediately once they have a task. The EV is charged with the maximum charging power as soon as it arrives home and never discharged. The AC operates with its maximum power when $\mathrm{T}_t^{\mathrm{ac}} \geq \mathrm{T}_{\mathrm{set}}^{\mathrm{ac}} + \Delta\mathrm{T}_{\mathrm{thre}}^{\mathrm{ac}}$, or minimum power when $\mathrm{T}_t^{\mathrm{ac}} \leq \mathrm{T}_{\mathrm{set}}^{\mathrm{ac}} - \Delta\mathrm{T}_{\mathrm{thre}}^{\mathrm{ac}}$; otherwise, the operating power remains the same as the power in the last time step. The EWH operates in a similar way as the AC.

2) Perfect Information Optimum (PIO): In the benchmank, we assume that the future electricity price, outdoor temperature, hot water flow rate and operational time of each appliance can be perfectly predicted. The DR scheduling is formulated as a deterministic optimization problem and solved by SCIP [31]. Note that this benchmark provides a limit for the performance but it cannot be reached due to the existence of randomness.

3) Model Predictive Control (MPC): The MPC forecasts the future electricity price, outdoor temperature, hot water flow rate and operational time of each appliance at each time step t for a receding horizon $[t, T)$. Based on the forecasts, a optimization problem is solved to derive the DR schedules and only the first step's schedule is executed. We assume the MPC knows the distribution of each appliance's operational time and predicts it by drawing a sample from the corresponding distribution. For the electricity price, outdoor temperature, and hot water flow rate, the forecast data are generated by using the actual value plus a bias. The bias is sampled from the normal distribution $\mathcal{N}(0, \sigma_\tau^2)$ truncated by $[-0.15\sigma_\tau, 0.15\sigma_\tau]$, where the standard deviation σ_τ is 15 percent of the actual value of the corresponding vairable for $\tau \in [t, T)$.

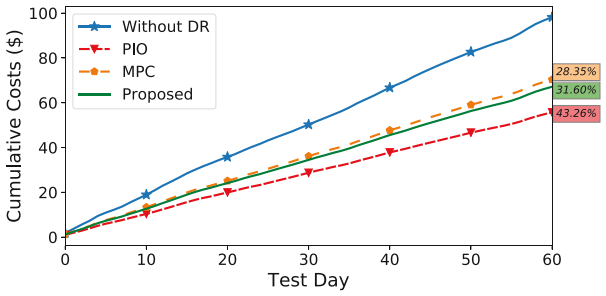


Fig. 5. Cumulative electricity costs on the test days.

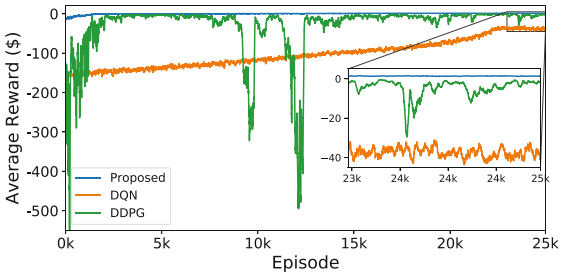
D. Simulation Results

Fig. 4 shows the average rewards of the proposed approach during the training process. It can be observed that the average rewards increase quickly at the begining and converge around 1.8 after 1500 iterations. Fig. 5 compares the cumulative electricity costs over the test days. The percentage terms on the right axis denote the cost reduction ratio of the corresponding approach compared to the *Without DR* benchmark. From this figure, we can observe that the proposed approach reduces the electricity cost by 31.60% while the MPC method only reduces the electricity cost by 28.35%. It is worth noting that we assume that the distribution of the working time of esch appliance is known for MPC. Moreover, the cost reduction ratio of the proposed model is only 11.66% less than that of the *PIO* policy. The comparison result illustrates that the proposed approach is effective for learning a real-time DR scheduling strategy to minimize the electricity cost of the household.

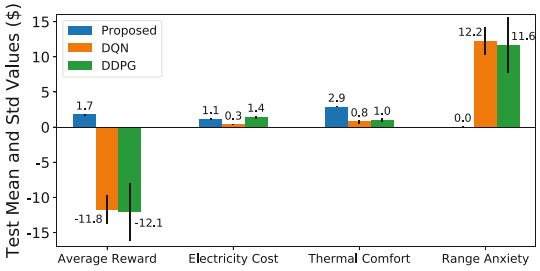
E. Comparison With Other DRL Approaches

The proposed approach is benchmarked against two widely used DRL methods, i.e., deep Q-learning [20] (DQN) and deep deterministic policy gradient [25] (DDPG). In order to apply DQN, a Q-network with 3 hidden layers of 128 ReLU neurons is used to approximate the Q-function. In addition, the action space $a_t = (u_t^{\mathrm{DW}}, u_t^{\mathrm{WM}}, u_t^{\mathrm{CD}}, P_t^{\mathrm{AC}}, P_t^{\mathrm{EWH}}, P_t^{\mathrm{EV}}), \forall t$ is discretized into $2 \times 2 \times 2 \times 2 \times 2 \times 3 = 96$ different choices. To apply DDPG, an actor network with 3 hidden layers of 128 ReLU neurons is used to learn the optimal action a_t. Since the control variables $u_t^{\mathrm{DW}}, u_t^{\mathrm{WM}}, u_t^{\mathrm{CD}}$ in a_t are binary, we need to map the corresponding output of the actor network into binary values. Specifically, if the output of the actor network is less than 0.5, the binary action is set to 0. Otherwise, it is set to 1. A critic network with the same architecture as the actor is used to approximate the optimal value function.

From Fig. 6(a) we can observe that the proposed approach demonstrates faster learning and higher rewards than the DQN and DDPG. In addition, the proposed method achieves better performance on the test set as shown in Fig. 6(b). Specifically, the average reward of the proposed approach is 1.7 but those of the DQN and DDPG are only -11.8 and -12.1, repectively. Although the DQN achieves a small electricity cost, it leads to low thermal comfort and high EV range anxiety, and so does the DDPG. This means that the DQN and DDPG cannot well-control the AC and EWH to maintain the indoor temperature

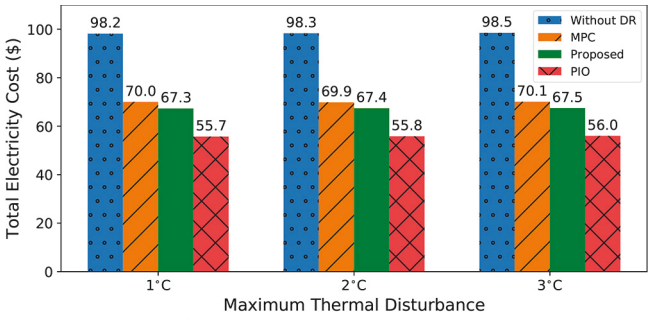


(a) Average rewards during the training process

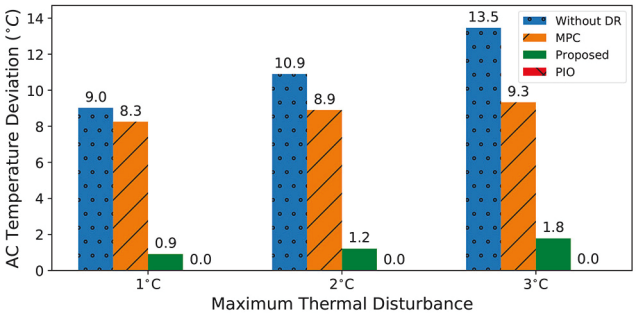


(b) Distribution of learned components in the reward on the test.

Fig. 6. Comparison of the DQN, DDPG, and the proposed method.



(a) Total electricity cost on the test days.



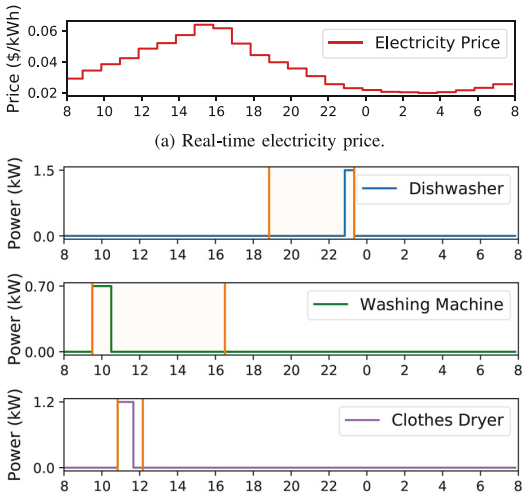
(b) Average of daily indoor temperature deviation from comfortable levels.

Fig. 7. The robustness of the proposed algorithm.

and hot water temperature in comfortable levels, and also fail to fully-charge the EV upon departure.

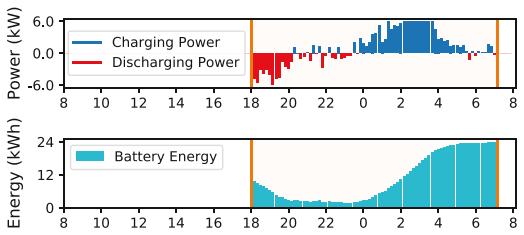
F. Algorithmic Robustness

In real-world, the indoor temperature is subject to many factors, such as solar irradiance, human activities, and computers. The AC dynamics model used in the previous studies cannot capture the thermal disturbances. To evaluate the robustness of the proposed method when thermal disturbance is considered, we use a modified model, $\mathrm{T}^{\mathrm{AC}}_{t+1} = \xi \cdot \mathrm{T}^{\mathrm{AC}}_{t} + (1-\xi)(\mathrm{T}^{\mathrm{out}}_{t} - \eta^{\mathrm{AC}} \cdot \mathrm{P}^{\mathrm{AC}}_{t}\, \Delta t / \mathrm{G_h}) + \omega_t$, where the distrubance term ω_t is assumed to

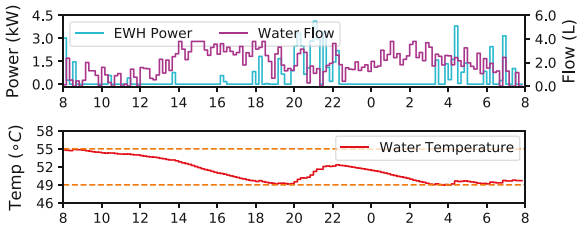


(a) Real-time electricity price.

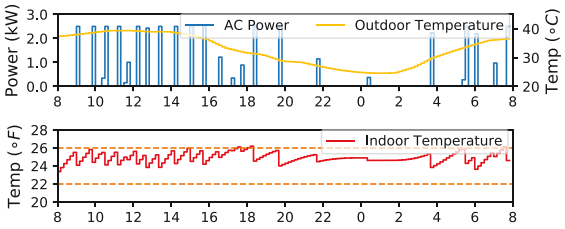
(b) Deferrable appliances. The orange region between two orange lines in each subfigure denotes the working time of the appliance.



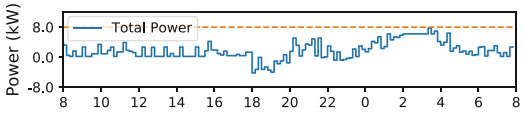
(c) EV. The orange region between two orange lines in each subfigure denotes the time when the EV is home.



(d) EWH. Throughout the scheduling, the water temperature is controlled within $[49°\mathrm{C}, 55°\mathrm{C}]$, which is the comfortable range defined by the resident.



(e) AC. Throughout the scheduling, the indoor temperature is mostly maintained within $[22°\mathrm{C}, 26°\mathrm{C}]$, which is the comfortable range defined by the resident.



(f) The power consumption at each time step is controlled below $\mathrm{P}^{\mathrm{g}}_{\max} = 8\mathrm{kW}$ to avoid being charged by the IBR price.

Fig. 8. Scheduling results of the proposed approach on a test day.

follow a uniform distribution parameterd by $[\upsilon_l, \upsilon_u]$°C. In our study, three cases, i.e., $\upsilon_u = -\upsilon_u = 1, 2, 3$, are considered. From Fig. 7(a) we can observe that the proposed method has lower electricity cost than the *Without DR* and *MPC* under the three cases. From Fig. 7(b) we can observe that compared to *Without DR* and *MPC*, the proposed approach can well maintain the indoor temerpature in the comfortable range with a small daily average deviation for three cases.

G. Scehdules of Appliances

To demonstrate the effectiveness of the proposed approach, we present the DR scheduling results on a test day in Fig. 8. It can be observed from Fig. 8(b) that each of the deferrable appliances is scheduled to operate during the periods when the prices are relatively low in its working time. Moreover, as shown in Fig. 8(c), the EV is discharged during the period 18:00-00:00, when the electricity price is relatively high. When the price becomes low in the period 0:00-6:00, the EV is charged. When the EV departs, the EV battery is adequately charged. For the EWH, we can observe from Fig. 8(d) that the water temperature is controlled within the comfortable range [49°C, 55°C] over the scheduling horizon. For the AC, we can observe from Fig. 8(e) that the indoor air temperature is also well-maintained in the comfortable range [22°C, 26°C]. Besides, as shown in Fig. 8(f), the total power consumption at each time step is controlled below the threshold $P^g_{max} = 8$ kW to avoid being charged by the IBR price.

VI. Conclusion

Focusing on the issue of residential DR, we have proposed a DRL strategy for optimal scheduling of smart appliances considering the uncertainty of resident's behavior, real-time electricity price, and outdoor temperature. The proposed approach is model-free and does not require the distribution of the uncertainty. In particular, the proposed DRL approach can handle both discrete and continuous actions, which makes it effective for scheduling all kinds of appliances. Through evaluation with real-world data, we have verified the effectiveness of the proposed approach in learning to optimize the appliance schedules in a smart house. Comparison results have demonstrated that the proposed DRL method can achieve better performance than the benchmarks.

Although the proposed approch has advantages, it takes time to train the nerual network. In our study, it takes about 9.1 hours to finish the overall training process. To apply the proposed method in the real building energy management system, we perform offline training. During the online implementation, we do not update the neural network model. It takes about 1.1 ms for the well-trained offline model to generate one schedule in the real-time implementation.

References

[1] D. Zhang, S. Li, M. Sun, and Z. O'Neill, "An optimal and learning-based demand response and home energy management system," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 1790–1801, Jul. 2016.

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

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

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

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

[6] M. Shafie-khah and P. Siano, "A stochastic home energy management system considering satisfaction cost and response fatigue," *IEEE Trans. Ind. Inf.*, vol. 14, no. 2, pp. 629–638, Feb. 2018.

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

[8] S. Li, J. Yang, W. Song, and A. Chen, "A real-time electricity scheduling for residential home energy management," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2602–2611, Apr. 2019.

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

[10] Z. Wen, D. O'Neill, and H. Maei, "Optimal demand response using device-based reinforcement learning," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2312–2324, Sep. 2015.

[11] F. Ruelens, B. J. Claessens, S. Vandael, B. De Schutter, R. Babuška, and R. Belmans, "Residential demand response of thermostatically controlled loads using batch reinforcement learning," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2149–2159, Sep. 2017.

[12] F. Ruelens, B. J. Claessens, S. Quaiyum, B. De Schutter, R. Babuška, and R. Belmans, "Reinforcement learning applied to an electric water heater: From theory to practice," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3792–3800, Jul. 2018.

[13] M. S. Ahmed, A. Mohamed, T. T. N. Khatib, H. Shareef, R. Z. Homod, and J. A. Ali, "Real time optimal schedule controller for home energy management system using new binary backtracking search algorithm," *Energy Build.*, vol. 138, pp. 215–227, Mar. 2017.

[14] N. Ahmed, M. Levorato, and G. P. Li, "Residential consumer-centric demand side management," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4513–4524, Sep. 2018.

[15] C. Keerthisinghe, G. Verbič, and A. C. Chapman, "A fast technique for smart home management: ADP with temporal difference learning," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3291–3303, Jul. 2018.

[16] S. Bahrami, V. W. S. Wong, and J. Huang, "An online learning algorithm for demand response in smart grid," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4712–4725, Sep. 2018.

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

[18] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[19] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

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

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

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

[23] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Trans. Smart Grid*, early access, doi: 10.1109/TSG.2019.2955437.

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

[25] L. Yu *et al.*, "Deep reinforcement learning for smart home energy management," *IEEE Internet Things J.*, early access, doi: 10.1109/JIOT.2019.2957289.

[26] M. H. Nehrir, R. Jia, D. A. Pierre, and D. J. Hammerstrom, "Power management of aggregate electric water heater loads by voltage control," in *Proc. IEEE Power Eng. Soc. Gen. Meeting*, Tampa, FL, USA, Jun. 2007, pp. 1–6.
[27] Y. F. Du, L. Jiang, Y. Li, and Q. Wu, "A robust optimization approach for demand side scheduling considering uncertainty of manually operated appliances," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 743–755, Mar. 2018.
[28] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, vol. 37, 2015, pp. 1889–1897.
[29] Ameren Illinois. *Day-Ahead and Historical RTP/HSS Prices*. Accessed: Jan. 31, 2019. [Online]. Available: https://www.ameren.com/account/retail-energy
[30] Kaggle. *Historical Hourly Weather Data 2012–2017*. Accessed: Jul. 18, 2019. [Online]. Available: https://www.kaggle.com/selfishgene/historical-hourly-weather-data
[31] A. Gleixner *et al.* (Jul. 2018). *The SCIP Optimization Suite 6.0*. [Online]. Available: https://scip.zib.de/

Hepeng Li (Student Member, IEEE) received the B.S. degree in information and computing science, and the M.S. degree in control theory and control engineering from the Northeastern University, Shenyang, China, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the School of Electrical, Computer and Biomedical Engineering, University of Rhode Island, South Kingstown, RI, USA. His research interests include microgrids, demand response, deep reinforcement learning, and cyber physical system. From 2014 to 2019, he was an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences, Shenyang, China. He was a recipient of the Best Paper Award in the IEEE Power & Energy Society General Meeting in 2018.

Zhiqiang Wan (Student Member, IEEE) received the B.S. degree from the Harbin Institute of Technology, Harbin, China, in 2012, and the M.S. degree from the School of Electrical and Electronics Engineering, Huazhong University of Science and Technology in 2015. He is currently pursuing the Ph.D. degree with the School of Electrical, Computer and Biomedical Engineering, University of Rhode Island, South Kingstown, RI, USA. His current research interests include deep learning, deep reinforcement learning, and cyber-physical system, with a particular interest in smart grid applications. He was a recipient of the URI Graduate Student Research & Scholarship Excellence Award in the Life Sciences, Physical Sciences, and Engineering in 2019, the Best Paper Award in the IEEE Power & Energy Society General Meeting in 2018, and the Best Paper Award in the IEEE 11th International Conference on Power Electronics and Drive Systems in 2015.

Haibo He (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Huazhong University of Science and Technology, China, in 1999 and 2002, respectively, and the Ph.D. degree in electrical engineering from Ohio University in 2006. He is currently the Robert Haas Endowed Chair Professor with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island. His current research interests include computational intelligence, machine learning, data mining, and various applications. He received the IEEE International Conference on Communications Best Paper Award in 2014, the IEEE CIS Outstanding Early Career Award in 2014, and the National Science Foundation CAREER Award in 2011. He was the General Chair of the IEEE Symposium Series on Computational Intelligence in 2014. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.