Deep Reinforcement Learning for Smart Home Energy Management

Liang Yu, *Member, IEEE*, Weiwei Xie, Di Xie, Yulong Zou, *Senior Member, IEEE*, Dengyin Zhang, Zhixin Sun, Linghua Zhang, Yue Zhang, *Senior Member, IEEE*, and Tao Jiang, *Fellow, IEEE*

***Abstract*—In this article, we investigate an energy cost minimization problem for a smart home in the absence of a building thermal dynamics model with the consideration of a comfortable temperature range. Due to the existence of model uncertainty, parameter uncertainty (e.g., renewable generation output, nonshiftable power demand, outdoor temperature, and electricity price), and temporally coupled operational constraints, it is very challenging to design an optimal energy management algorithm for scheduling heating, ventilation, and air conditioning systems and energy storage systems in the smart home. To address the challenge, we first formulate the above problem as a Markov decision process, and then propose an energy management algorithm based on deep deterministic policy gradients. It is worth mentioning that the proposed algorithm does not require the prior knowledge of uncertain parameters and building the thermal dynamics model. The simulation results based on real-world traces demonstrate the effectiveness and robustness of the proposed algorithm.**

***Index Terms*—Deep reinforcement learning (DRL), energy cost, energy management, energy storage systems (ESSs), heating, ventilation, and air conditioning (HVAC) systems, smart home, thermal comfort.**

I. Introduction

As A NEXT-GENERATION power system, smart grid is typified by an increased use of information and communications technology (e.g., Internet of Things) in the generation, transmission, distribution, and consumption of electrical energy. In the smart grid environment, there are many opportunities for saving the energy cost of smart homes, which are evolved from traditional homes by adopting three components, i.e., the internal networks, intelligent controls, and home automations [1]. For example, dynamic electricity prices could be utilized to reduce energy cost by scheduling energy storage systems (ESSs) and thermostatically controllable loads intelligently. As one kind of thermostatically controllable loads, the heating, ventilation, and air conditioning (HVAC) systems consume about 40% of total energy in a household [2], which results in energy cost concerns for smart home owners. Since the primary purpose of HVAC systems is to maintain thermal comfort for the occupants, it is of great importance to optimize the energy cost of smart homes without sacrificing thermal comfort.

In this article, we investigate an energy optimization problem for a smart home with renewable energies, ESS, HVAC systems, and nonshiftable loads (e.g., televisions) in the absence of a building thermal dynamics model. To be specific, our objective is to minimize the energy cost of the smart home during a time horizon with the consideration of a comfortable indoor temperature range. However, it is very challenging to achieve the above aim due to the following reasons. First, it is often intractable to obtain accurate dynamics of indoor temperature, which can be affected by many factors [3]. Second, it is difficult to know the statistical distributions of all combinations of random system parameters (e.g., renewable generation output, power demand of nonshiftable loads, outdoor temperature, and electricity price). Third, there are temporally coupled operational constraints associated with the ESS and HVAC systems, which means that the current action would affect the future decisions. To address the above challenge, we propose a deep deterministic policy gradients (DDPGs)-based energy management algorithm, which can make decision about ESS charging/discharging power and HVAC input power simply based on the current observation information.

The main contributions of this article are summarized as follows.

1) We investigate an energy cost minimization problem for smart homes in the absence of a building thermal dynamics model with the consideration of a comfortable temperature range, energy exchange between the smart home and the utility grid, ESS charging/discharging, HVAC input power adjustment, and parameter uncertainties. Then, we reformulate the problem as a Markov decision process (MDP), where environment state, action, and reward function are designed.

Manuscript received September 24, 2019; revised November 10, 2019; accepted November 25, 2019. Date of publication December 3, 2019; date of current version April 14, 2020.This work was supported in part by the National Natural Science Foundation of China under Grant 61972214, Grant 61671253, Grant 61631020, Grant 91738201, Grant 61571241, Grant 61872423, Grant 61972208, Grant 61672299, and Grant 61771258; in part by the Natural Science Foundation of Jiangsu Province under Grant BK20171446; and in part by the Scientific Research Fund of Nanjing University of Posts and Telecommunications under Grant NY219062. *(Corresponding authors: Liang Yu; Tao Jiang.)*

L. Yu, W. Xie, D. Xie, Y. Zou, Z. Sun, and L. Zhang are with the Key Laboratory of Broadband Wireless Communication and Sensor Network Technology of Ministry of Education, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (e-mail: liang.yu@njupt.edu.cn).

D. Zhang is with the Jiangsu Key Laboratory of Broadband Wireless Communication and Internet of Things, School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China.

Y. Zhang is with the Department of Engineering, University of Leicester, Leicester LE1 7RH, U.K.

T. Jiang is with the Wuhan National Laboratory for Optoelectronics, School of Electronics Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: tao.jiang@ieee.org).

Digital Object Identifier 10.1109/JIOT.2019.2957289

2) We propose an energy management algorithm to jointly schedule the ESS and HVAC systems based on DDPG. Since the proposed algorithm makes decision simply based on the current environment state, it does not require prior knowledge of uncertain parameters and building the thermal dynamics model.
3) Extensive simulation results based on real-world traces show that the proposed algorithm can save energy cost by 8.10%–15.21% without sacrificing thermal comfort when compared with two baselines. Moreover, the robustness testing shows that the proposed algorithm has the potential of providing a more efficient and practical tradeoff between maintaining thermal comfort and reducing energy cost than an "optimal" strategy.

The remainder of this article is organized as follows. In Section II, we introduce related works. In Section III, the system model and problem formulation are given. Then, we propose a DDPG-based energy management algorithm in Section IV and its effectiveness is verified by simulation results in Section V. Finally, we make a conclusion and discuss the future work in Section VI.

II. Related Works

There have been many studies on energy cost and/or thermal comfort in smart homes. Due to the space limitation, we mainly focus on joint energy cost and thermal comfort management in smart homes [4]–[8]. The approaches proposed in these studies can be generally classified into two categories, i.e., model-based approaches and model-free-based approaches. To be specific, the model-based approaches are designed based on the model information about thermal dynamics of the environment [9], [10]. By contrast, the model-free-based approaches are designed without requiring the above-mentioned information.

A. Model-Based Approaches

Angelis *et al.* [4] presented a home energy management approach to minimize the energy cost related to task execution, energy storage, energy selling, and heat pump without violating the given comfortable temperature range and other constraints. Fan *et al.* [5] proposed an online home energy management scheme to minimize the energy cost associated with electric water heaters and HVAC systems with the consideration of indoor temperature ranges. Zhang *et al.* [6] developed a home energy management strategy to minimize energy cost related to the HVAC load and deferrable loads without violating the given comfortable temperature range. Pilloni *et al.* [7] proposed a quality-of-experience (QoE)-aware smart home energy management system (HEMS) to save energy cost while minimizing the annoyance perceived by the users. Yu *et al.* [8] proposed an online home energy management algorithm to minimize the sum of energy cost and thermal discomfort cost (here, thermal discomfort cost is the function of temperature deviation between indoor temperature and the comfortable temperature level). Franceschelli *et al.* [11] proposed a heuristic approach to optimize the peak-to-average power ratio of a large population of thermostatically controlled loads considering comfortable temperature ranges. Although some advances have been made in the above-mentioned works, their approaches need to model building thermal dynamics with the simplified mathematical models, e.g., equivalent thermal parameters (ETPs) model.

B. Model-Free-Based Approaches

Since it is very challenging to develop a building thermal dynamics model that is both accurate and efficient enough for HVAC control, some recent works have considered to use real-time data for HVAC control [12]–[14]. For example, Lu *et al.* [12] proposed an energy management scheme to minimize the sum of electricity cost and user dissatisfaction cost associated with wash machines and HVAC loads based on the multiagent reinforcement learning and artificial neural network approach. Ruelens *et al.* [13] proposed a residential demand response method to minimize energy cost with the consideration of temperature range based on batch reinforcement learning. Although the reinforcement learning-based methods in [12]–[14] do not require the prior knowledge of building thermal dynamics model. They are known to be unstable or even to diverge when a nonlinear function approximator (e.g., a neural network) is used to represent the action–value function [15]. To efficiently handle large and continuous state space, deep reinforcement learning (DRL) has been presented and shown successful in playing Atari and Go games [15]. Wei *et al.* [3] proposed a DRL-based method for building HVAC control, which can reduce energy cost while maintaining the desired indoor temperature range. Gao *et al.* [16] presented a DRL-based thermal comfort control method to minimize energy consumption and thermal discomfort. Zhang and Lam [17] conducted real-life implementation and evaluation of a DRL-based control method for a radiant heating system, which optimizes energy demand and thermal comfort. Valladares *et al.* [18] proposed a DRL-based thermal comfort and indoor air control algorithm. Wan *et al.* [19] proposed a DRL-based algorithm to minimize the energy cost of a smart home with battery energy storage. Although some model-free methods have been proposed in the above-mentioned studies, none of them can be applicable to the coordination between the ESS and HVAC systems in smart homes. To deal with this problem, we develop a DDPG-based energy management algorithm in this article.

III. System Model and Problem Formulation

The smart home considered in this article is shown in Fig. 1, where distributed generators, ESS, loads, and HEMS could be identified. Distributed generators could be solar panels or wind generators. ESS could be lead-acid batteries or lithium-ion batteries, which can reduce net-energy demand from main grids by storing excess renewable energies locally and are very important for implementing nearly zero-energy buildings in the future [20]. At present, ESS costs are very high (e.g., around 450\$/kWh), which means that installing ESS in a smart home is not very economical. However, ESS costs are dropping rapidly with the development of technology and are predicted to drop below 100\$/kWh within the next decade. As a

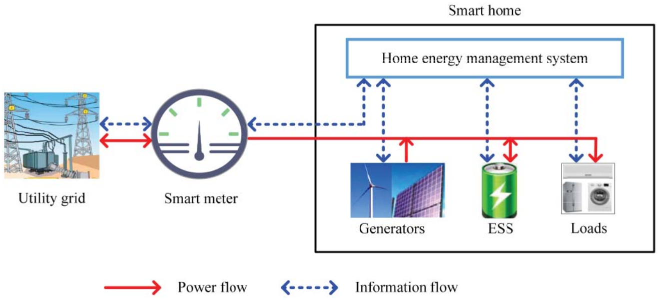


Fig. 1. Illustration of a smart home.

result, the profitability of adopting ESS will gradually increase. Therefore, we consider ESS in the model of the smart home. Loads in a smart home can be generally divided into several types, e.g., nonshiftable loads, shiftable and noninterruptible loads, and controllable loads [21]. To be specific, power demands of nonshiftable loads (e.g., televisions, microwaves, and computers) must be satisfied completely without delay. As for shiftable and noninterruptible loads (e.g., washing machines), their tasks can be scheduled to a proper time but cannot be interrupted. In contrast, controllable loads (e.g., HVAC systems, heat pumps, and electric water heaters) can be controlled to flexibly adjust their operation times and energy usage quantities by following some operational requirements, e.g., temperature ranges. In this article, we mainly focus on nonshiftable loads and thermostatically controlled loads [13]. As for thermostatically controlled loads, the HVAC systems are considered since they consume about 40% of the total energy in a smart home [2]. Suppose that the HEMS operates in slotted time, i.e., $t \in [1, T]$, where T is the total number of time slots. For simplicity, the duration of a time slot Δt is normalized to a unit time (e.g., one hour) so that power and energy could be used equivalently. In each time slot, the HEMS makes continuous decision on ESS charging/discharging power and HVAC input power according to a set of available information (e.g., renewable generation output, nonshiftable power demand, outdoor temperature, and electricity price), with the aim of minimizing the energy cost of the smart home while maintaining the comfortable temperature range in the absence of the building thermal dynamics model. In the following parts, models associated with the ESS and HVAC systems are provided. Then, an energy cost minimization problem is formulated. Next, we reformulate it as an MDP due to the difficulty in solving the minimization problem.

A. ESS Model

Let B_t be the stored energy in the ESS at time slot t. Then, the ESS storage dynamics model is given by

$$B_{t+1} = B_t + \eta_c c_t + \frac{d_t}{\eta_d} \quad \forall\ t \tag{1}$$

where $\eta_c \in (0, 1]$ and $\eta_d \in (0, 1]$ are the charging and discharging efficiency coefficients, respectively; and c_t and d_t are the ESS charging power and discharging power, respectively. Here, c_t and d_t are assigned with different signs (i.e., $c_t \geq 0$ and $d_t \leq 0$), which contributes to the design of action in Section III-F.

Since ESS cannot be charged above its capacity $B^{\max}$ or discharged below the minimal energy level $B^{\min}$, we have

$$B^{\min} \leq B_t \leq B^{\max} \quad \forall\ t. \tag{2}$$

Due to the existence of ESS charging and discharging rate limitations, we have

$$0 \leq c_t \leq c^{\max} \quad \forall\ t \tag{3}$$

$$-d^{\max} \leq d_t \leq 0 \quad \forall\ t \tag{4}$$

where $c^{\max}$ and $d^{\max}$ are the maximum charging and discharging power of the ESS, respectively.

To avoid the simultaneous ESS charging and discharging, we have

$$c_t \cdot d_t = 0 \quad \forall\ t. \tag{5}$$

B. HVAC Model

The HVAC system can be dynamically adjusted to maintain thermal comfort of the occupants in the smart home. Since thermal comfort depends on many factors (e.g., air temperature, mean radiant temperature, relative humidity, air speed, clothing insulation, and metabolic rate), its representation is very complex. In the existing studies, many modeling approaches and parameter measurement methods associated with thermal comfort have been developed [16], [22]–[28]. Similar to [3]–[6], this article uses a comfortable temperature range as the representation of thermal comfort for simplicity, i.e.,

$$T^{\min} \leq T_t \leq T^{\max} \quad \forall\ t \tag{6}$$

where $T^{\min}$ and $T^{\max}$ are the minimum and maximum comfort level, respectively.

In this article, we consider an HVAC system with inverter in the smart home, i.e., the HVAC system can adjust its input power e_t continuously [8]. Suppose $e^{\max}$ is the rating power of the HVAC system, then we have

$$0 \leq e_t \leq e^{\max} \quad \forall\ t. \tag{7}$$

C. Power Balancing

To keep the power balance in the smart home, the aggregated power supply should be equal to the served power demand. Then, we have

$$g_t + p_t - d_t = b_t + e_t + c_t \quad \forall\ t \tag{8}$$

where g_t, p_t, and b_t are the power drawn from the utility grid, renewable generation output, and nonshiftable power demand, respectively. If $g_t < 0$, it means that energy from the smart home will be sold to the utility grid. Otherwise, the smart home will purchase energy from the utility grid.

D. Cost Model

Let v_t and u_t be the buying and selling price of energy, respectively. Then, the energy cost of the smart home at time slot t can be calculated by

$$C_{1,t} = \left(\frac{v_t - u_t}{2}|g_t| + \frac{v_t + u_t}{2}g_t\right) \quad \forall\ t \tag{9}$$

where the intuition behind (9) is that just one variable g_t is needed to reflect the behavior of electricity buying or selling. For example, when $g_t \geq 0$, $C_{1,t} = v_t g_t$. For the case $g_t < 0$, $C_{1,t} = u_t g_t$.

It is well known that frequent discharging or charging would do harm to the lifetime of the ESS. To capture this phenomenon, the ESS depreciation cost at time slot t is introduced as follows [29]:

$$C_{2,t} = \psi(|c_t| + |d_t|) \quad \forall\ t \tag{10}$$

where ψ denotes the ESS depreciation coefficient in \$/kW.

E. Total Energy Cost Minimization Problem

Based on the above-mentioned models, we can formulate a total energy cost minimization problem as follows:

$$(\mathbf{P1})\ \ \min \sum_{t=1}^{T} \mathbb{E}\{C_{1,t} + C_{2,t}\} \tag{11a}$$

$$\text{s.t.}\ \ (1){-}(8) \tag{11b}$$

where the expectation operator $\mathbb{E}$ is taken over the randomness of the system parameters (i.e., renewable generation output p_t, nonshiftable power demand b_t, outdoor temperature T_t^{out}, and buying/selling electricity prices v_t/u_t) and the possibly random control actions (i.e., the amount of energy exchange between the smart home and the utility grid g_t, ESS charging/discharging power c_t/d_t, and HVAC input power e_t) at each time slot.

It is very challenging to solve **P1** due to the following reasons. First, it is often intractable to obtain accurate dynamics of indoor temperature T_t, which can be affected by many factors [3], e.g., building structure and materials, surrounding environment (e.g., ambient temperature, humidity, and solar radiation intensity), and internal heat gains from occupants, lighting systems, and other equipments. Second, it is very difficult to know the statistical distributions of all combinations of random system parameters. Third, there are temporally coupled operational constraints associated with the ESS and HVAC systems, which means that the current action would affect future decisions. To handle the "time-coupling" property, typical methods are based on dynamic programming [8], which suffers from "the curse of dimensionality" problem. In this article, we provide a way of solving **P1** without requiring the dynamics of indoor temperature and prior knowledge of random system parameters. In particular, we reformulate the above-mentioned sequential decision-making problem as an MDP problem. Then, we develop a DDPG-based energy management algorithm for the problem.

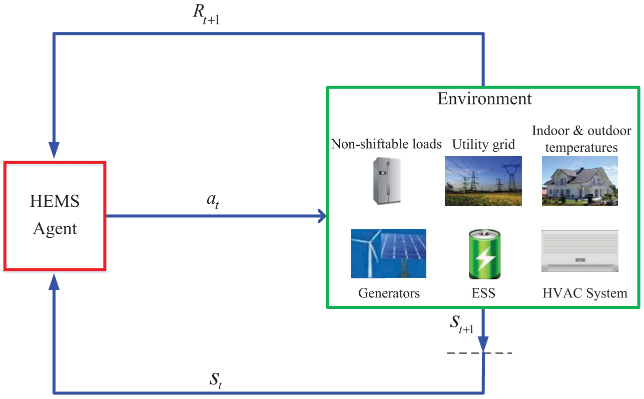


Fig. 2. Agent-environment interaction in the MDP.

F. MDP Formulation

In the smart home, the indoor temperature at the next time slot is only determined by the indoor temperature, HVAC power input, and environment disturbances (e.g., outdoor temperature and solar irradiance intensity) in the current time slot [6], [7], [30], [31]. Moreover, the ESS energy level at the next time slot just depends on the current energy level and current discharging/charging power according to (1), which is independent of previous states and actions. Thus, both of the ESS scheduling and HVAC control can be regarded as an MDP. In the following parts, we will formulate the sequential decision-making problem associated with the smart home energy management as an MDP. It is worth noting that the MDP formulation is an approximation description of the smart home energy management problem since some components of the environment state may not be Markovian in practice, e.g., renewable generation output and electricity price. According to the existing works [15], [32], even though the environment is not strictly MDP, the corresponding problem can still be solved by the reinforcement-learning-based algorithms empirically, which is also validated by the simulation results in this article. For the non-Markovian environment, many approaches could be adopted to improve the performance of reinforcement learning-based algorithms, e.g., approximate state [32], [33], recurrent neural networks [34], gated end-to-end memory policy networks [35], and eligibility traces [33].

A discounted MDP is formally defined as a five-tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the set of environment states and $\mathcal{A}$ is the set of actions. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is the transition probability function, which models the uncertainty in the evolution of states of the system based on the action taken by the agent [36]. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function and $\gamma \in [0, 1]$ is a discount factor. In this article, the agent denotes the learner and decision maker (i.e., HEMS agent), while the environment comprising many objects outside the agent (e.g., renewable generators, nonshiftable loads, ESS, the HVAC system, utility grid, and indoor/outdoor temperature). The interaction between the agent and the environment can be depicted by Fig. 2, where the HEMS agent observes environment state $\boldsymbol{s}_t$ and takes action $\boldsymbol{a}_t$. Then, environment state becomes $\boldsymbol{s}_{t+1}$ and the reward R_{t+1} is returned.

In the following parts, we will design the key components of the MDP, including environment state, action, and reward function.

1) Environment State: The environment state consists of seven kinds of information, i.e., renewable generation output p_t, nonshiftable power demand b_t, ESS energy level B_t, outdoor temperature T_t^{out}, indoor temperature T_t, buying electricity price v_t, and time slot index in a day t' ($t' = \text{mod}\ (t, 24)$). Since selling electricity price u_t is typically related to buying electricity price v_t (e.g., $u_t = \delta v_t$ [37]–[39], δ is a constant), u_t is not selected as a part of the environment state. For brevity, $\boldsymbol{s}_t$ is adopted to describe the environment state, i.e., $\boldsymbol{s}_t = (p_t,\ b_t,\ B_t, T_t^{\text{out}}, T_t, v_t, t')$.

2) Action: The aim of HEMS agent is to optimally decide the amount of energy exchange between the smart home and the utility grid (i.e., g_t), ESS charging power (i.e., c_t), ESS discharging power (i.e., d_t), and HVAC input power e_t. After c_t, d_t, and e_t are jointly decided, g_t can be known immediately according to (8). Therefore, the action of the MDP consists of ESS charging/discharging power c_t/d_t and HVAC input power e_t. Since adopting c_t and d_t simultaneously would complicate the design of the energy management algorithm, we use just one variable f_t, where the range of f_t is $[-d^{\max},\ c^{\max}]$. When $f_t \geq 0$, $c_t = f_t$, and $d_t = 0$. When $f_t \leq 0$, $c_t = 0$, and $d_t = f_t$. Therefore, constraints (3)–(5) could be guaranteed. To guarantee the feasibility of (1) and (2), $0 \leq c_t \leq \min\{c^{\max}, [(B^{\max} - B_t)/\eta_c]\}$ when $f_t \geq 0$, and $\min\{-d^{\max}, (B^{\min} - B_t)\eta_d\} \leq d_t \leq 0$ when $f_t \leq 0$. According to (6), the range of e_t is $[0,\ e^{\max}]$. When the indoor temperature T_t is lower than $T^{\min}$, e_t should be zero for avoiding further temperature deviation. Similarly, when $T_t > T^{\max}$, the feasible e_t should be non-negative. For brevity, $\boldsymbol{a}_t$ is used to describe the action, i.e., $\boldsymbol{a}_t = (f_t,\ e_t)$.

3) Reward: According to the MDP theory in [33], the transition of the environment state from $\boldsymbol{s}_{t-1}$ to $\boldsymbol{s}_t$ could be triggered by the execution of $\boldsymbol{a}_{t-1}$. Finally, the reward R_t will be obtained. Since the aim of the agent is to minimize the total energy cost while maintaining the comfortable temperature range, the corresponding reward consists of three parts, namely, the penalty for the energy consumption of the HVAC system, the penalty for ESS depreciation, and the penalty for temperature deviation. Since the energy cost of the HVAC system at slot $t-1$ is $C_{1,t-1}$, the first part of R_t can be represented by $-C_{1,t-1}(\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1})$. Similarly, the second part of R_t can be described by $-C_{2,t-1}(\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1})$. To maintain the comfortable temperature range, the third part of R_t can be computed by $-C_{3,t}(\boldsymbol{s}_t)$, where

$$C_{3,t}(\boldsymbol{s}_t) = \left(\left[T_t - T^{\max}\right]^+ + \left[T^{\min} - T_t\right]^+\right) \quad \forall\ t \tag{12}$$

which means that $C_{3,t} = 0$ if $T^{\min} \leq T_t \leq T^{\max}$. Otherwise, $C_{3,t} = T_t - T^{\max}$ if $T_t > T^{\max}$, and $C_{3,t} = T^{\min} - T_t$ if $T_t < T^{\min}$.

Taking three parts into consideration, the final reward function can be designed as follows:

$$R_t = -\beta\big(C_{1,t-1}(\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1}) + C_{2,t-1}(\boldsymbol{s}_{t-1}, \boldsymbol{a}_{t-1})\big) - C_{3,t}(\boldsymbol{s}_t)$$

where β denotes a positive weight coefficient in °C/$.

4) Action-Value Function: When jointly controlling the ESS and the HVAC system at time slot t, the HEMS agent intends to maximize the expected return it receives over the future. In particular, the return is defined as the sum of the discounted rewards [33], i.e., $R = \sum_{i=1}^{\infty} \gamma^{i-1} R_{t+i}$. Let $Q_\pi(\boldsymbol{s}, \boldsymbol{a})$ be the action-value function under a policy π (note that a policy is a mapping from states to probabilities of selecting each possible action), which represents the expected return if action $\boldsymbol{a}_t = \boldsymbol{a}$ is taken in state $\boldsymbol{s}_t = \boldsymbol{s}$ under the policy π. Then, the optimal action-value function $Q^*(\boldsymbol{s}, \boldsymbol{a})$ is $\max_\pi Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and can be calculated by the following Bellman optimality equation in a recursive manner, i.e.,

$$\begin{aligned} Q^*(\boldsymbol{s}, \boldsymbol{a}) &= \mathbb{E}\left[R_{t+1} + \gamma \max_{\boldsymbol{a}'} Q^*\big(\boldsymbol{s}_{t+1}, \boldsymbol{a}'\big) | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}\right] \\ &= \sum_{\boldsymbol{s}', \boldsymbol{r}} P\big(\boldsymbol{s}', \boldsymbol{r} | \boldsymbol{s}, \boldsymbol{a}\big)\left[\boldsymbol{r} + \gamma \max_{\boldsymbol{a}'} Q^*\big(\boldsymbol{s}', \boldsymbol{a}'\big)\right] \end{aligned}$$

where $\boldsymbol{s}' \in \mathcal{S}$, $\boldsymbol{r} \in \mathcal{R}$, $\boldsymbol{a}' \in \mathcal{A}$, and $P \in \mathcal{P}$.

To obtain $Q^*(\boldsymbol{s}, \boldsymbol{a})$, system state transition probabilities $P(\boldsymbol{s}', \boldsymbol{r} | \boldsymbol{s}, \boldsymbol{a})$ are required. Since indoor temperature in the smart home could be affected by many disturbances, it is difficult to accurately obtain state transition probabilities. To overcome this challenge, Q-learning methods could be used, which do not require the knowledge of state transition probabilities. To support the case with continuous system states, a function approximator could be adopted to estimate Q-function. When a neural network with weight θ is adopted as the nonlinear function approximator, we refer to it as Q-network. In [15], a deep Q-network (DQN) algorithm was proposed, which can use experience replay and target network to ensure the stability of the reinforcement learning methods when function approximators are adopted. However, DQN cannot be directly applied to the problem with continuous action spaces since it needs to discretize the action space and lead to an explosion of the number of actions. As a result, low computational efficiency, decreased performance, and the requirement of more training data would be incurred [16], [40].

IV. DDPG-Based Energy Management Algorithm

In this section, we first propose a DDPG-based energy management algorithm. Then, we analyze the computational complexity of the proposed algorithm.

A. Algorithmic Design

To solve the MDP problem defined in Section III-F, we propose a DDPG-based energy management algorithm. Different from DQN, DDPG is capable of dealing with continuous states and actions. For example, just two network outputs are needed to represent continuous actions in this article, which avoids the explosion of the number of actions. Since DDPG is a kind of actor–critic methods (i.e., methods that learn approximations to both policy function and value function), the actor network and critic network are incorporated, which are shown in Fig. 3. The input and output of the actor network are the environment state $\boldsymbol{s}_t$ and action $\boldsymbol{a}$, respectively. Then, $\boldsymbol{a}$ and $\boldsymbol{s}_t$ are adopted as the input of critic network, whose output is an action-value

Algorithm 1: Proposed Energy Management Strategy

```
Input: System state $S_t$, testing time slots $H_{\text{test}}$
Output: System decision $\boldsymbol{a}_t = (f_t, e_t)$ in each time slot
Load the weight of the actor network $\theta^{\mu}$ obtained by the Algorithm 2
for $t=1,2,\cdots,H_{test}$ do
    Select action $\boldsymbol{a}_t = \mu(\phi(\boldsymbol{s}_t)|\theta^{\mu})$
    Execute action $\boldsymbol{a}_t = (f_t, e_t)$ in smart home environment and observe next state $\boldsymbol{s}_{t+1}$ and reward $R_{t+1}$
end
```

Algorithm 2: Training Deep Neural Networks With DDPG

```
Input: Renewable generation output, nonshiftable power demand, outdoor temperature, electricity price
Output: The weights of actor network and critic network, i.e., $\theta^{\mu}$ and $\theta^{Q}$
1  Initialize memory $\mathcal{D}$ of size $N$
2  Initialize preprocess function $\phi(\boldsymbol{s}_t)$
3  Randomly initialize critic network $Q(\phi(\boldsymbol{s}), \boldsymbol{a}|\theta^{Q})$ and actor network $\mu(\phi(\boldsymbol{s})|\theta^{\mu})$ with weights $\theta^{Q}$ and $\theta^{\mu}$, respectively
4  Initialize target networks $Q'$ and $\mu'$ by copying: $\theta^{Q'} \Leftarrow \theta^{Q}, \theta^{\mu'} \Leftarrow \theta^{\mu}$
5  for $episode=1,2,\cdots,M$ do
6      Receive the initial environment state $\boldsymbol{s}_0$
7      for $t=0,2,\cdots,P-1$ do
8          Select action $\boldsymbol{a}_t = \mu(\phi(\boldsymbol{s}_t)|\theta^{\mu}) + \mathcal{N}_t$
9          Execute action $\boldsymbol{a}_t$ in smart home environment and observe next state $\boldsymbol{s}_{t+1}$ and reward $R_{t+1}$
10         Store $(\phi(\boldsymbol{s}_t),\ \boldsymbol{a}_t,\ R_{t+1},\ \phi(\boldsymbol{s}_{t+1}))$ in $\mathcal{D}$
11         Sample a random mini-batch of $K$ transitions $(\phi(\boldsymbol{s}_i),\ \boldsymbol{a}_i,\ R_{i+1},\ \phi(\boldsymbol{s}_{i+1}))$ from $\mathcal{D}$, $1 \le i \le K$
12         Set $y_i = R_{i+1} + \gamma Q'(\phi(\boldsymbol{s}_{i+1}),\ \mu'(\phi(\boldsymbol{s}_{i+1}) \mid \theta^{\mu'}) \mid \theta^{Q'})$
13         Update critic network by minimizing the loss:
14         $L = \frac{1}{K}\sum_{i=1}^{K} (y_i - Q(\phi(\boldsymbol{s}_i),\ \boldsymbol{a}_i|\theta^{Q}))^2$
15         Update actor policy using sampled policy gradient:
16         $\sum_{i=1}^{K} \frac{\nabla_a Q(\phi(\boldsymbol{s}),\boldsymbol{a}|\theta^{Q})|_{s=s_i,a=\mu(\phi(s_i))}}{K} \nabla_{\theta^{\mu}} \mu(\phi(\boldsymbol{s})|\theta^{\mu})|_{s_i}$
17         Update target networks:
18         $\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}$
19         $\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$
20     end
21 end
```

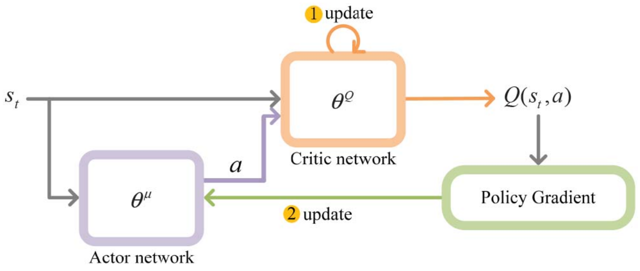


Fig. 3. Actor network and critic network in DDPG.

function [i.e., $Q(\boldsymbol{s}_t, \boldsymbol{a})$]. Next, the policy gradient can be computed and used to update the weight of the actor network. Before computing $Q(\boldsymbol{s}_t, \boldsymbol{a})$, the weight of the critic network should be updated based on two mechanisms, i.e., memory replay and target networks. More details will be introduced when explaining Algorithm 2.

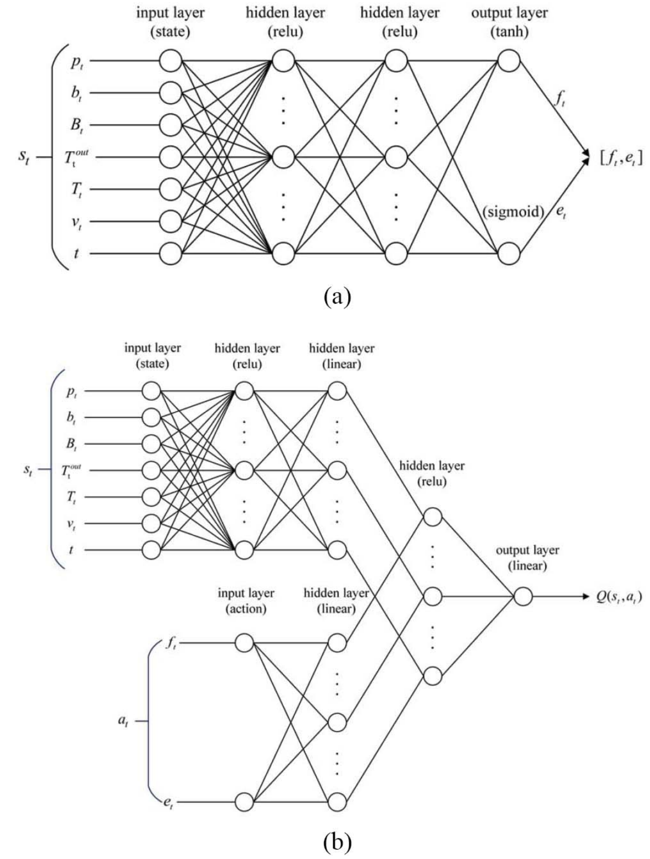
The proposed DDPG-based energy management algorithm can be found in Algorithm 1, where the key step is to load the weight of the actor network θ^{μ}, which is trained by Algorithm 2. In each time slot, the actor network selects an action on ESS charging/discharing power and HVAC input power according to the current environment state $\boldsymbol{s}_t$. Then, the action $\boldsymbol{a}_t$ is executed and the environment state becomes $\boldsymbol{s}_{t+1}$. Meanwhile, the reward R_{t+1} is obtained. In Algorithm 2, we first initialize a replay memory $\mathcal{D}$ with capacity N, which stores the transition tuple $(\boldsymbol{s}_t,\ \boldsymbol{a}_t,\ R_{t+1},\ \boldsymbol{s}_{t+1})$. Moreover, a preprocess function $\phi(\boldsymbol{s}_t)$ is introduced to facilitate the learning process by normalizing the input data. Specifically, each component in the environment state at time slot t (e.g., κ_t) should be normalized within the range [0, 1] using the following expression: $[(\kappa_t - \min_t \kappa_t)/(\max_t \kappa_t - \min_t \kappa_t)]$. Then, we randomly initialize the critic network $Q(\phi(\boldsymbol{s}), \boldsymbol{a}|\theta^{Q})$ and actor network $\mu(\phi(\boldsymbol{s})|\theta^{\mu})$ with weights θ^{Q} and θ^{μ}, respectively. Their architectures in the proposed energy management algorithm are described by Fig. 4, where there are two hidden layers in the actor network and four hidden layers in the critic network. Next, we initialize the weights of the target critic network $Q(\phi(\boldsymbol{s}), \boldsymbol{a}|\theta^{Q'})$ and target actor network $\mu(\phi(\boldsymbol{s})|\theta^{\mu'})$ by copying, i.e., $\theta^{Q'} \leftarrow \theta^{Q}$ and $\theta^{\mu'} \leftarrow \theta^{\mu}$. In each time slot of each episode, an action is selected based on the following expression in line 8, i.e.,

$$\boldsymbol{a}_t = \mu\big(\phi(\boldsymbol{s}_t)|\theta^{\mu}\big) + \mathcal{N}_t \tag{13}$$

where $\mathcal{N}_t$ is the exploration noise. In this article, we use the following way to introduce exploration noise, i.e.,

$$\boldsymbol{a}_t = \begin{cases} \mu(\phi(\boldsymbol{s}_t)|\theta^{\mu}), & \text{if } \omega_t > \xi_t \\ \big(U_{t,1}, U_{t,2}\big), & \text{if } \omega_t \le \xi_t \end{cases} \tag{14}$$

where ω_t, $U_{t,1}$, and $U_{t,2}$ are the random numbers, which follow uniform distributions with parameters (0, 1), $(-d^{\max}/\max\{c^{\max}, d^{\max}\},\ c^{\max}/\max\{c^{\max}, d^{\max}\})$, and (0, 1), respectively. $\xi_t = \max(\xi_t - \zeta * (\text{episode} - N/P), \xi_{\min})$, $\xi_0 = 1$, and $0 < \zeta < 1$. After $\boldsymbol{a}_t$ is obtained, it will be applied to ESS and the HVAC system. At the end of time slot t, the new state $\boldsymbol{s}_{t+1}$ and the reward R_{t+1} are returned from the environment. Then, the transition tuple $(\phi(\boldsymbol{s}_t),\ \boldsymbol{a}_t,\ R_{t+1},\ \phi(\boldsymbol{s}_{t+1}))$ will be stored in the memory for the training of actor and critic networks as shown in line 10. Next, K transitions are randomly sampled for training deep neural networks, i.e., actor network, critic network, target actor network, and target critic network. As shown in lines 12–14, $Q(\phi(\boldsymbol{s}_i),\ \boldsymbol{a}_i)$ and y_i generated by the critic network and target network are used to calculate mean square error loss. By minimizing the loss function, the weight



(a)

(b)

Fig. 4. Architectures of (a) actor network and (b) critic network.

of the critic network could be updated. Then, we can calculate the sampled policy gradient as shown in line 15, which is used to update the weight of the actor network. Finally, the weights of the target actor network and target critic network could be updated as shown in lines 17–19. Note that a small τ should be selected in order to improve the learning stability. Typically, $0 < \tau \ll 1$.

B. Algorithmic Computational Complexity

In Algorithm 1, it can be observed that the computational complexity of the proposed energy management algorithm depends on the number of testing slots H_{test}. Since simple calculations are carried out in Algorithm 1, its computational complexity can be described by $\mathcal{O}(H_{\text{test}})$. Given the fixed testing time horizon, a shorter duration of a time slot would result in a larger H_{test}. However, the time slot's duration cannot be selected arbitrarily in practice due to the following reasons. On the one hand, too long duration would result in the loss of many control opportunities of saving energy cost and maintaining a comfortable temperature range. On the other hand, too short duration may affect the training convergence of the DRL-based algorithms since the control actions taken by the DRL agent cannot take effect immediately in terms of environment states (e.g., indoor temperature) [17]. Therefore, the duration of a time slot should be selected appropriately in practice. In existing works, the typical duration of a time slot is several minutes or one hour (e.g., 15 min [3] and 1 h [17]), which is far greater than the computational time of the proposed energy management algorithm in a time slot. Therefore, the proposed energy management algorithm can be implemented in a real-time way.

TABLE I
MAIN PARAMETER SETTINGS

H_{test}	744 hours	Δt	1 hour
$B^{\max}$	$6kWh$	$B^{\min}$	$0.6kWh$
B_0	$1.2kWh$	$c^{\max}$	$3kW$
$d^{\max}$	$3kW$	$e^{\max}$	$2kW$
M	3000	P	24
K	120	N	24000
α_a	0.0001	α_c	0.001
N_a	300,600	N_c	300,600,600,600
τ	0.001	Optimizer	Adam

V. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed energy management algorithm. We first describe the simulation setup. Then, we describe the baselines used for performance comparisons. Finally, we provide simulation results about algorithmic convergence process, algorithmic performance under varying β, algorithmic effectiveness, and algorithmic scalability.

A. Simulation Setup

In simulations, we use real-world traces related to solar generation, nonshiftable power demand, outdoor temperature, and electricity price, which are extracted from Pecan Street database.[1] Note that such database is the largest real-world open energy database on the planet and includes the data related to home energy consumption and solar generation of the Mueller neighborhood in Austin, TX, USA. For simplicity, the cooling mode of a residential HVAC system is considered. Since summers in Austin are very hot,[2] we use the data during June 1, 2018 to August 31, 2018 for model training and testing. To be specific, the data in June and July are used to train the neural network models and the data in August are adopted for performance testing. Some important system parameters are configured as follows: $u_t = 0.9v_t$ [37], $\gamma = 0.995$, $\eta_{\text{c}} = \eta_{\text{d}} = 0.95$ [41], $\zeta = 0.0005$, $\xi_{\min} = 0.1$, $T^{\min} = 66.2$ °F(19 °C) [3], and $T^{\max} = 75.2$ °F(24 °C) [3], other parameter configurations are shown in Table I, where α_a and α_c denote the learning rate of actor network and critic network, respectively. In Table I, N_a and N_c denote the number of neurons in each hidden layer of the actor network and critic network, respectively. To simulate the environment, we adopt the following indoor temperature dynamics model for simplicity, i.e., $T_{t+1} = \varepsilon T_t + (1 - \varepsilon)(T_t^{\text{out}} - [\eta_{\text{hvac}}/A]e_t)$ [6], [7], [30], [31], where $\varepsilon = 0.7$ [42], $\eta_{\text{hvac}} = 2.5$ [30], and $A = 0.14$ kW/°F [30]. Note that the variant of the proposed energy management algorithm can be applicable to any indoor temperature dynamics model

[1] https://www.pecanstreet.org/

[2] https://en.wikipedia.org/wiki/Austin,_Texas#Climate

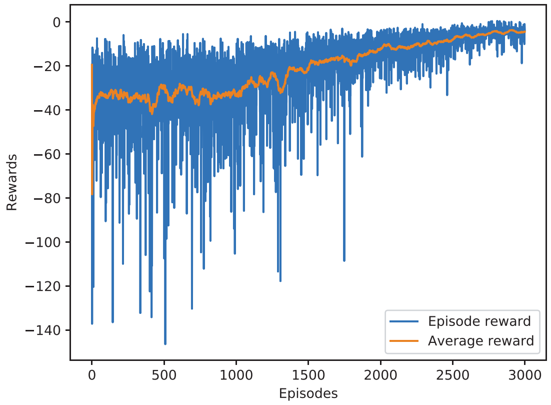


Fig. 5. Convergence process of Algorithm 2.

by incorporating more environment-related variables in system state, e.g., relative humidity and solar radiation intensity.

B. Baselines

To evaluate the performance of the proposed algorithm, we adopt three baselines as follows.

1) *Baseline1:* This scheme adopts ON/OFF policy [3] for building HVAC control but without considering the use of the ESS. Specifically, the HVAC system will be turned on if $T_i > T^{\max}$ and it will be turned off if $T_i < T^{\min}$.
2) *Baseline2:* This scheme uses the DDPG-based control policy in this article for HVAC control but without considering the use of the ESS, i.e., $c^{\max} = d^{\max} = 0$. Based on the performance comparison between *Baseline2* and the proposed algorithm, the energy cost saving caused by the use of the ESS can be known. Similarly, the energy cost saving incurred by the use of DDPG-based control policy can be obtained by comparing the performance of *Baseline2* with that of *Baseline1*.
3) *Baseline3:* This scheme intends to minimize the cumulative cost during the testing period H_{test} (i.e., $\sum_{t=1}^{H_{\text{test}}}(C_{1,t}+C_{2,t})$) with the consideration of constraints (1)–(8), assuming that all uncertainty system parameters and the dynamics model of indoor temperature can be known beforehand. Although the optimal solution of this scheme is not achievable in practice due to the existence of parameter and model uncertainties, it can provide the lower bound for the performance of the proposed algorithm when all constraints in **P1** are satisfied.

C. Simulation Results

1) Algorithmic Convergence Process: According to Algorithm 1, the proposed energy management algorithm needs to know the training result of Algorithm 2 before testing. In Fig. 5, the reward received during each episode generally increases. Since the minimum exploration probability $\xi_{\min}$ is 0.1 and system parameters (e.g., solar radiation power, nonshiftable power demand, outdoor temperature, and electricity price) are varying in each episode, the episode reward fluctuates within a small range. To show the changing trend of rewards more clearly, we provide the average value of the past 50 episodes. In Fig. 5, it can be found that the average reward generally increases and becomes more and more stable.

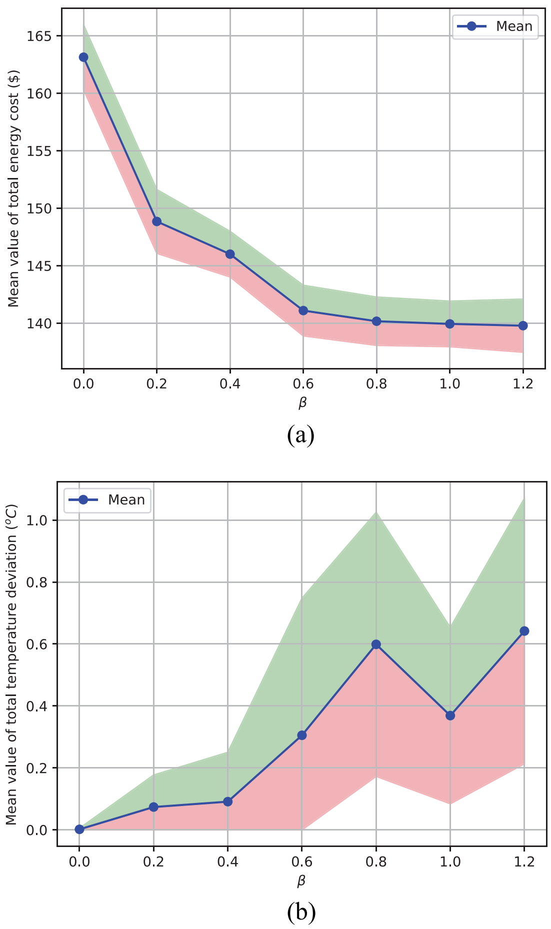


Fig. 6. Impact of β on the performance of the proposed algorithm. (a) Total energy cost. (b) Total temperature deviation.

2) Algorithmic Performance Under Varying β: Since many random number generators are adopted in the neural network initialization, mini-batch data collection for training, and action choice, the performance of the proposed algorithm is varying even the same system parameters are configured. To show the impact of β on the performance of the proposed algorithm more clearly, mean values of total energy cost (i.e., the sum of energy cost and ESS depreciation cost) and total temperature deviation with 95% confidence interval across 40 runs are considered and the corresponding results can be found in Fig. 6. It can be observed that the mean value of total energy cost and that of total temperature deviation generally decreases and increases with the increase of β, respectively. Such tendency is obvious since larger β results in more importance of energy cost and less importance of temperature deviation. By taking mean values of total energy cost and total temperature deviation into consideration, a proper value of β is 1 when the mean value of total temperature deviation is less than 1 °C.

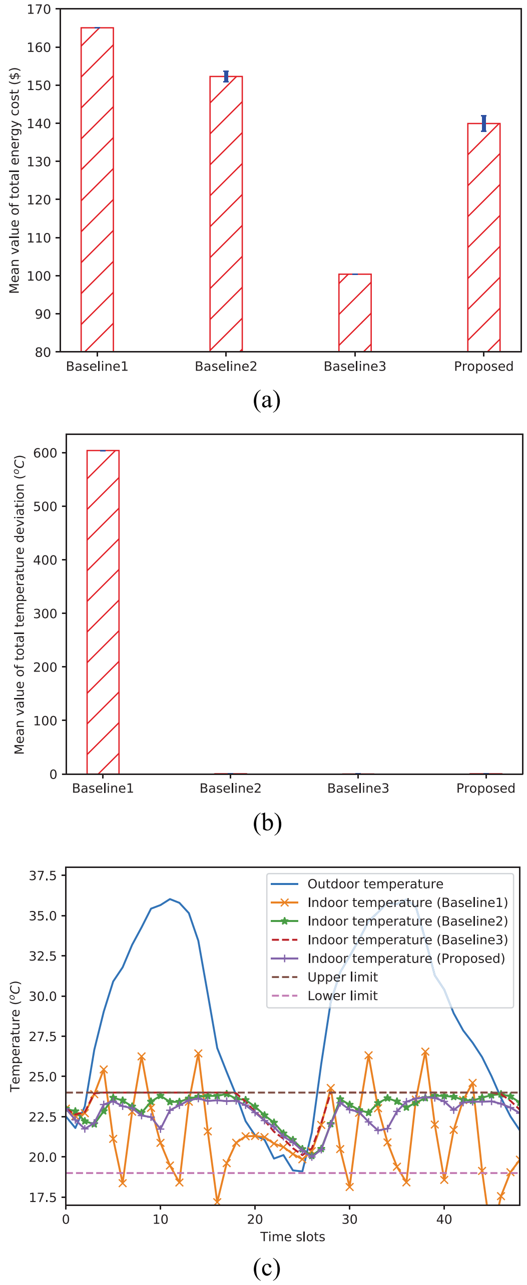


Fig. 7. Performance comparisons among three schemes ($\beta = 0.6$, 95% confidence interval across 40 runs is considered). (a) Mean value of total energy cost. (b) Mean value of total temperature deviation. (c) Indoor temperature.

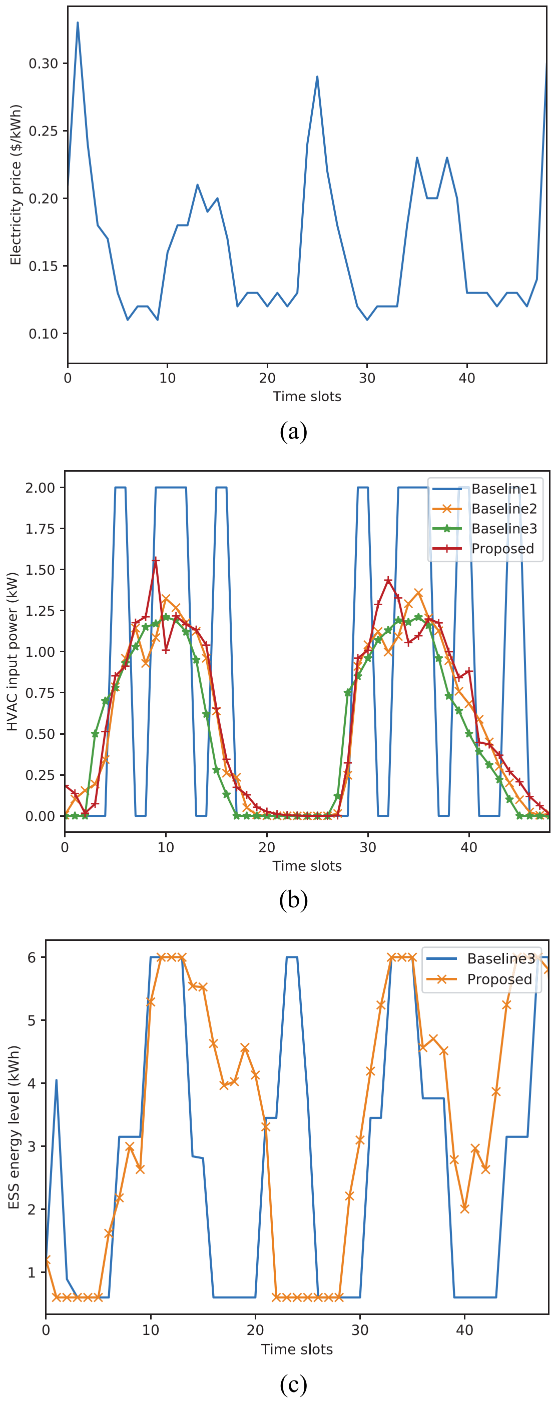


Fig. 8. Simulation results associated with the ESS and HVAC systems. (a) Price. (b) HVAC input power. (c) ESS energy level.

3) Algorithmic Effectiveness: Performance comparisons among four schemes are shown in Fig. 7, where the proposed energy management algorithm achieves better performance than *Baseline1* and *Baseline2*. To be specific, the proposed energy management algorithm can reduce the mean value of total energy cost by 15.21% and 8.10% when compared with *Baseline1* and *Baseline2*, respectively. Moreover, the mean value of total temperature deviation under the proposed algorithm is smaller than *Baseline1* and *Baseline2*, which can be illustrated by Fig. 7(b) and (c). Compared with *Baseline1*, *Baseline2* and the proposed algorithm could save energy cost by increasing/decreasing the HVAC input power when electricity price is low/high, which can be depicted by Fig. 8(a) and (b). Compared with *Baseline2*, the proposed algorithm could reduce the energy cost by charging/discharging ESS when electricity price is low/high, which can be shown in Fig. 8(a) and (c). Though *Baseline3* achieves the best performance, it requires all prior knowledge of uncertain system parameters and thermal dynamics model. Thus, *Baseline3* is just adopted for performance reference. By observing the performance gap between the proposed algorithm and *Baseline3*, it can be known that the potential of

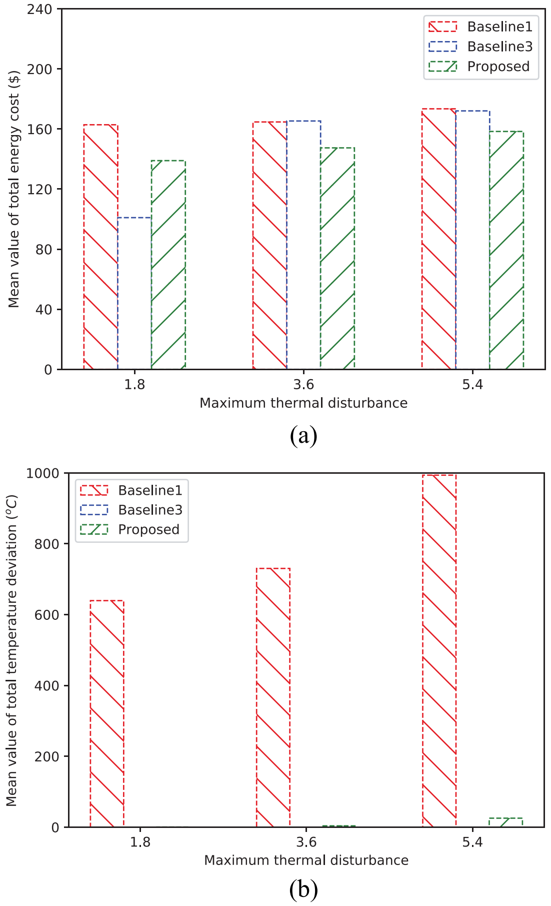


Fig. 9. Robustness of the proposed algorithm. (a) Mean value of total energy cost. (b) Mean value of total temperature deviation.

reducing the mean value of total energy cost is great. In future work, more training data and advanced DRL-based energy management algorithms would be adopted for reducing the performance gap.

4) Algorithmic Robustness: Note that the thermal dynamics model used in the above-mentioned simulations cannot capture thermal disturbances in practice, e.g., thermal disturbances from solar irradiance, lighting systems, and computers. Thus, we evaluate the robustness of the proposed algorithm when random thermal disturbance is introduced. To be specific, $T_{t+1} = \varepsilon T_t + (1-\varepsilon)(T_t^{\text{out}} - [\eta_{\text{hvac}}/A]e_t) + \epsilon_t$ [10], where the error item ϵ_t is assumed to follow a uniform distribution with parameters $[\vartheta_l, \ \vartheta_u]$°F. In this scenario, three cases are considered, i.e., $\vartheta_u = -\vartheta_l = 1.8, 3.6, 5.4$. In Fig. 9, it can be observed that the proposed algorithm achieves better performances than *Baseline1* under three cases. Compared with *Baseline3*, the proposed algorithm can save the total energy cost by up to 10% with a small increase of the total temperature violation. Moreover, unlike *Baseline3*, the proposed algorithm does not require any prior knowledge of all uncertain parameters and thermal dynamics model. Therefore, the proposed algorithm has the potential of providing a more efficient and practical tradeoff between maintaining thermal comfort and reducing energy cost than *Baseline3*.

VI. Conclusion

In this article, we proposed a DDPG-based energy management algorithm for a smart home to efficiently control the HVAC systems and ESS in the absence of a building thermal dynamics model, with the consideration of a comfortable temperature range and many parameter uncertainties. Extensive simulation results based on real-world traces showed the effectiveness and robustness of the proposed algorithm. In future work, more reasonable thermal comfort models and more types of controllable loads (e.g., electric vehicles and electric water heaters) will be incorporated. In addition, more opportunities of saving energy cost can be grasped by utilizing real-world occupant behavior information [43], which requires the adoption of more advanced deep neural network architectures/algorithms.

References

[1] S. Wu *et al.*, "Survey on prediction algorithms in smart homes," *IEEE Internet Things J.*, vol. 4, no. 3, pp. 636–644, Jun. 2017.

[2] A. Afram and F. Janabi-Sharif, "Effects of dead-band and set-point settings of on/off controllers on the energy consumption and equipment switching frequency of a residential HVAC system," *J. Process Control*, vol. 47, pp. 161–174, Nov. 2016.

[3] T. Wei, Y. Wang, and Q. Zhu, "Deep reinforcement learning for building HVAC control," in *Proc. 54th Annu. Design Autom. Conf.*, 2017, pp. 1–6.

[4] F. Angelis, M. Boaro, D. Fuselli, S. Squartini, F. Piazza, and Q. Wei, "Optimal home energy management under dynamic electrical and thermal constraints," *IEEE Trans. Ind. Informat.*, vol. 9, no. 3, pp. 1518–1527, Aug. 2013.

[5] W. Fan, N. Liu, and J. Zhang, "An event-triggered online energy management algorithm of smart home: Lyapunov optimization approach," *Energies*, vol. 9, no. 5, pp. 381–404, 2016.

[6] 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.

[7] V. Pilloni, A. Floris, A. Meloni, and L. Atzori, "Smart home energy management including renewable sources: A QoE-driven approach," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 2006–2018, May 2018.

[8] 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.

[9] M. Shad, A. Momeni, R. Errouissi, C. P. Diduch, M. E. Kaye, and L. Chang, "Identification and estimation for electric water heaters in direct load control programs," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 947–955, Mar. 2017.

[10] E. C. Kara, M. Bergés, and G. Hug, "Impact of disturbances on modeling of thermostatically controlled loads for demand response," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2560–2568, Nov. 2015.

[11] M. Franceschelli, A. Pilloni, and A. Gasparri, "A heuristic approach for online distributed optimization of multi-agent networks of smart sockets and thermostatically controlled loads based on dynamic average consensus," in *Proc. Eur. Control Conf.*, 2018, pp. 2541–2548.

[12] R. Lu, S. Hong, and M. Yu, "Demand response for home energy management using reinforcement learning and artificial neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6629–6639, Nov. 2019, doi: 10.1109/TSG.2019.2909266.

[13] F. Ruelens, B. J. Claessens, S. Vandael, B. D. 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.

[14] J. Vázquez-Canteli and Z. Nagy, "Reinforcement learning for demand response: A review of algorithms and modeling techniques," *Appl. Energy*, vol. 235, pp. 1072–1089, Feb. 2019.

[15] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–541, Feb. 2015.

[16] G. Gao, J. Li, and Y. Wen. (2019). *Energy-Efficient Thermal Comfort Control in Smart Buildings via Deep Reinforcement Learning*. Accessed: Sep. 1, 2019. [Online]. Available: https://arxiv.org/pdf/1901.04693v1.pdf

[17] Z. Zhang and K. P. Lam, "Practical implementation and evaluation of deep reinforcement learning control for a radiant heating system," in *Proc. 5th ACM Int. Conf. Syst. Built Environ.*, 2018, pp. 148–157.

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

[19] Z. Wan, H. Li, and H. He, "Residential energy management with deep reinforcement learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2018, pp. 1–7.

[20] A. I. Nousdilis, E. O. Kontis, G. C. Kryonidis, G. C. Christoforidis, and G. K. Papagiannis, "Economic assessment of lithium-ion battery storage systems in the nearly zero energy building environment," in *Proc. 20th Int. Symp. Elect. Apparatus Technol.*, 2018, pp. 1–6.

[21] M. Yousefi, A. Hajizadeh, and M. Soltani, "A comparison study on stochastic modeling methods for home energy management system," *IEEE Trans. Ind. Informat.*, vol. 15, no. 8, pp. 4799–4808, Aug. 2019, doi: 10.1109/TII.2019.2908431.

[22] B. Yang, X. Cheng, D. Dai, T. Olofsson, H. Li, and A. Meier, "Real-time and contactless measurements of thermal discomfort based on human poses for energy efficient control of buildings," *Build. Environ.*, vol. 162, pp. 1–10, Sep. 2019.

[23] X. Cheng, B. Yang, A. Hedman, T. Olofsson, H. Li, and L. Gool, "NIDL: A pilot study of contactless measurement of skin temperature for intelligent building," *Build. Environ.*, vol. 198, pp. 340–352, Sep. 2019.

[24] X. Cheng, B. Yang, T. Olofsson, G. Liu, and H. Li, "A pilot study of online non-invasive measuring technology based on video magnification to determine skin temperature," *Build. Environ.*, vol. 121, pp. 1–10, Aug. 2017.

[25] Y. Wang and Z. Lian, "A thermal comfort model for the non-uniform thermal environments," *Energy Build.*, vol. 172, pp. 397–404, Aug. 2018.

[26] W. Li, J. Zhang, T. Zhao, and R. Liang, "Experimental research of online monitoring and evaluation method of human thermal sensation in different active states based on wristband device," *Energy Build.*, vol. 173, pp. 613–622, Aug. 2018.

[27] L. Yang, Z. Zheng, J. Sun, D. Wang, and X. Li, "A domain-assisted data driven model for thermal comfort prediction in buildings," in *Proc. 9th ACM Int. Conf. Future Energy Syst.*, 2018, pp. 271–276.

[28] L. Yu, D. Xie, T. Jiang, Y. Zou, and K. Wang, "Distributed real-time HVAC control for cost-efficient commercial buildings under smart grid environment," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 44–55, Feb. 2018.

[29] H. Xu, X. Li, X. Zhang, and J. Zhang. (2019). *Arbitrage of Energy Storage in Electricity Markets With Deep Reinforcement Learning*. Accessed: Sep. 1, 2019. [Online]. Available: https://arxiv.org/pdf/1904.12232v1.pdf

[30] P. Constantopoulos, F. C. Schweppe, and R. C. Larson, "ESTIA: A real-time consumer control scheme for space conditioning usage under spot electricity pricing," *Comput. Oper. Res.*, vol. 18, no. 8, pp. 751–765, 1991.

[31] A. A. Thatte and L. Xie, "Towards a unified operational value index of energy storage in smart grid environment," *IEEE Trans. Smart Grid*, vol. 3, no. 3, pp. 1418–1426, Sep. 2012.

[32] Z. Zhang, A. Chong, Y. Pan, C. Zhang, S. Lu, and K. Lam, "Reinforcement learning in Markovian and non-Markovian environments," in *Proc. Build. Perform. Model. Conf. SimBuild ASHRAE IBPSA-USA*, 2018, pp. 1–8.

[33] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. London, U.K.: MIT Press, 2018.

[34] J. Schmidhuber, "Reinforcement learning in Markovian and non-Markovian environments," in *Proc. 3rd Int. Conf. Neural Inf. Process. Syst.*, 1990, pp. 500–506.

[35] J. Perez and T. Silander. (2017). *Non-Markovian Control With Gated End-to-End Memory Policy Networks*. Accessed: Sep. 1, 2019. [Online]. Available: https://arxiv.org/pdf/1705.10993v1.pdf

[36] S. Padakandla, K. J. Prabuchandran, and S. Bhatnagar. *Reinforcement Learning in Non-Stationary Environments*. Accessed: Sep. 1, 2019. [Online]. Available: https://arxiv.org/pdf/1905.03970.pdf

[37] L. Yu, T. Jiang, and Y. Cao, "Energy cost minimization for distributed Internet data centers in smart microgrids considering power outages," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 1, pp. 120–130, Jan. 2015.

[38] L. Yu, T. Jiang, and Y. Zou, "Distributed real-time energy management in data center microgrids," *IEEE Trans. Smart Grid*, vol. 9, no. 4, pp. 3748–3762, Jul. 2018.

[39] Y. Zhang, N. Gatsis, and G. B. Giannakis, "Robust management of distributed energy resources for microgrids with renewables," *IEEE Trans. Sustain. Energy*, vol. 4, no. 4, pp. 944–953, Oct. 2013.

[40] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 834–843.

[41] Y. Xu, L. Xie, and C. Singh, "Optimal scheduling and operation of load aggregator with electric energy storage in power markets," in *Proc. North Amer. Power Symp.*, 2010, pp. 1–7.

[42] R. Deng, Z. Zhang, J. Ren, and H. Liang, "Indoor temperature control of cost-effective smart buildings via real-time smart grid communications," in *Proc. IEEE Globecom*, 2016, pp. 1–6.

[43] S. Chen *et al.*, "Butler, not servant: A human-centric smart home energy management system," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 27–33, Feb. 2017.

Liang Yu (M'16) received the B.S. and M.S. degrees from Yangtze University, Jingzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in June 2014.

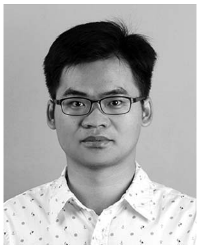
He is currently an Associate Professor with the Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests are the energy management of cyber-physical systems (e.g., smart grids, data centers, smart buildings, and unmanned aerial vehicles), cloud-fog computing, distributed optimization, and machine learning.

Weiwei Xie received the B.S. degree from the Tongda College of Nanjing University of Posts and Telecommunications, Yangzhou, China, in 2018, where she is currently pursuing the master's degree in information networking.

Her current research interests include smart grids, building energy management, and deep reinforcement learning.

Di Xie received the B.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016, where she is currently pursuing the master's degree in information networking.

Her current research interests include smart grids, distributed optimization, and machine learning.



Yulong Zou (SM'13) received the B.Eng. degree in information engineering from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, China, in July 2006, the first Ph.D. degree in electrical engineering from the Stevens Institute of Technology, Hoboken, NJ, USA, in May 2012, and the second Ph.D. degree in signal and information processing from NUPT in July 2012.

He is a Professor and the Doctoral Supervisor with NUPT. His research interests span a wide range of topics in wireless communications and signal processing, including the cooperative communications, cognitive radio, wireless security, and energy-efficient communications.

Prof. Zou was awarded the 9th IEEE Communications Society Asia–Pacific Best Young Researcher in 2014. He is serving (or served) as an Editor for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE COMMUNICATIONS LETTERS, *IET Communications*, and *China Communications*. He acted as a TPC Member for various IEEE sponsored conferences, e.g., IEEE ICC, IEEE GLOBECOM, IEEE WCNC, IEEE VTC, and IEEE ICCC.

Dengyin Zhang received the B.S., M.S., and Ph.D. degrees from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 1986, 1989, and 2004, respectively.

He is currently a Professor with the School of Internet of Things, Nanjing University of Posts and Telecommunications. He was a Visiting Scholar with Digital Media Lab, Umeå University, Umeå, Sweden, from 2007 to 2008. His research interests include signal and information processing, networking techniques, and information security.

Zhixin Sun received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1998.

From 2001 to 2002, he was a Postdoctoral with the School of Engineering, Seoul National University, Seoul, South Korea. He is currently a Professor and the Dean of the School of Modern Posts, Nanjing University of Posts and Telecommunications, Nanjing. His research interests include Internet of Things, big data analysis, blockchain techniques, and network security.

Linghua Zhang received the B.S. and M.S. degrees from Southeast University, Nanjing, China, in 1987 and 1990, respectively, and the Ph.D. degree from the Nanjing University of Posts and Telecommunications (NUPT), Nanjing, in 2005.

She is currently a Professor and the Doctoral Supervisor with NUPT. Her current research interests include intelligent signal processing in modern communications, routing, and energy saving technologies in wireless sensor networks.

Yue Zhang (M'06–SM'17) received the B.E. and M.E. degrees from the Beijing University of Post and Telecommunications, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree from Brunel University, Uxbridge, U.K., in 2008.

He is currently an Associate Professor with the Department of Engineering, University of Leicester, Leicester, U.K. His research interests are signal processing for 5G wireless and mobile systems, radio propagation model, and multimedia and wireless networks.

Dr. Zhang currently serves as an Associate Editor for the IEEE TRANSACTIONS ON BROADCASTING and IEEE ACCESS.

Tao Jiang (M'06–SM'10–F'19) received the Ph.D. degree in information and communication engineering from the Huazhong University of Science and Technology, Wuhan, China, in April 2004.

He is currently a Distinguished Professor with the Wuhan National Laboratory for Optoelectronics and School of Electronics Information and Communications, Huazhong University of Science and Technology, Wuhan. From August 2004 to December 2007, he worked in some universities, such as Brunel University, Uxbridge, U.K., and University of Michigan-Dearborn, Dearborn, MI, USA. He has authored or coauthored more 300 technical papers in major journals and conferences and 9 books/chapters in the areas of communications and networks.

Dr. Jiang served or is serving as the symposium technical program committee membership of some major IEEE conferences, including INFOCOM, GLOBECOM, and ICC. He was invited to serve as the TPC Symposium Chair for the IEEE GLOBECOM 2013, IEEEE WCNC 2013, and ICCC 2013. He served or is serving as an Associate Editor of some technical journals in communications, including IEEE NETWORK, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE INTERNET OF THINGS JOURNAL, and he is the Associate Editor-in-Chief of *China Communications*.