sensors

MDPI

Article

Energy Management of Smart Home with Home Appliances, Energy Storage System and Electric Vehicle: A Hierarchical Deep Reinforcement Learning Approach

Sangyoon Lee and Dae-Hyun Choi *

School of Electrical and Electronics Engineering, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 156-756, Korea; sangyoon1207@naver.com

* Correspondence: dhchoi@cau.ac.kr; Tel.: +82-2-820-5101

Received: 9 March 2020; Accepted: 9 April 2020; Published: 10 April 2020

Abstract: This paper presents a hierarchical deep reinforcement learning (DRL) method for the scheduling of energy consumptions of smart home appliances and distributed energy resources (DERs) including an energy storage system (ESS) and an electric vehicle (EV). Compared to Q-learning algorithms based on a discrete action space, the novelty of the proposed approach is that the energy consumptions of home appliances and DERs are scheduled in a continuous action space using an actor–critic-based DRL method. To this end, a two-level DRL framework is proposed where home appliances are scheduled at the first level according to the consumer's preferred appliance scheduling and comfort level, while the charging and discharging schedules of ESS and EV are calculated at the second level using the optimal solution from the first level along with the consumer environmental characteristics. A simulation study is performed in a single home with an air conditioner, a washing machine, a rooftop solar photovoltaic system, an ESS, and an EV under a time-of-use pricing. Numerical examples under different weather conditions, weekday/weekend, and driving patterns of the EV confirm the effectiveness of the proposed approach in terms of total cost of electricity, state of energy of the ESS and EV, and consumer preference.

Keywords: home energy management; deep reinforcement learning; smart home appliance; energy storage system; electric vehicle

1. Introduction

Approximately 30 percent of the United States' total energy consumption comes from the residential sector, and the amount of the residential energy consumption is expected to grow owing to increased use of home appliances (e.g., air conditioners (ACs) and washing machines (WMs)) and modern electronic devices [1]. Thus, an efficient and economical methodology for residential energy management is required to reduce the electricity bill of consumers and keep the efficiency of their appliances. Furthermore, as distributed energy resources (DERs) (e.g., a rooftop solar photovoltaic (PV), a residential energy storage system (ESS), and an electric vehicle (EV)) become integrated into an individual residential house via an advanced metering infrastructure with smart meters for reliable residential grid operations, the complexity of the residential energy management increases. The aforementioned challenges require more intelligent systems, i.e., home energy management systems (HEMSs), through which an electric utility or a third party provides consumers with an efficient and economical control of home appliances.

A primary goal of HEMS is to reduce the electricity bill of consumers while satisfying their comforts and preferences. To achieve this goal, HEMSs perform the following two functions:

(1) real-time monitoring of the energy usage of consumers using smart meters; (2) scheduling of the optimal energy consumption of home appliances. To implement this second function, an HEMS algorithm is generally formulated as a model-based optimization problem. Recently, numerous studies have been published on the development of HEMS optimization algorithms [2–12]. These studies address the scheduling of the energy consumption for home appliances and DERs, while maintaining the consumer's comfort level using mixed-integer nonlinear programming (MINLP) [2], the load scheduling using mixed-integer linear programming (MILP) for single and multiple households [3,4], robust optimization for scheduling of home appliances to resolve the uncertainty of consumer behavior [5], and distributed HEMS architectures consisting of local and global HEMSs [6]. A quickly distributed HEMS algorithm was developed for a large number of households using the MINLP approach with a nonconvex relaxation [7]. Based on an emerging technology for ESSs and EVs, a home energy consumption model under the control of the ESS was presented in [8]. A model predictive control-based HEMS algorithm was proposed using the prediction of the EV state [9]. An HEMS optimization model considering both ESS and EV was formulated for a single household [10,11] based on their bi-directional operation and multiple households with a renewable energy facility [12]. In addition, many studies proposed the methods to evaluate and preserve the consumer comfort during the HEMS process. In [13], a quality of experience (QoE)-aware HEMS was developed where the QoE-aware cost saving appliance scheduling and the QoE-aware renewable source power allocation are conducted to schedule the operation of the controllable loads based on the consumer preferences and the available renewable energy sources. A new demand management scheme based on the operational comfort level (OCL) of consumer was proposed to minimize the peak-to-average ratio while maximizing the OCL of consumers [14]. A score-based HEMS method was presented to maintain the total household power consumption below a certain limit by scheduling various household loads based on the consumer comfort level setting [15]. A QoE-aware HEMS algorithm was presented to reduce the peak load and electricity cost while satisfying the consumer comfort and QoE with a fixed threshold [16]. More recently, an advanced QoE-aware HEMS method considering renewable energy sources and EVs was developed for adaptively varying the QoE threshold [17]. Compared to the method with the fixed OoE threshold in [16], a fuzzy logic controller was designed to dynamically adjust the QoE threshold for optimizing the consumer QoE in [17]. A recent work on HEMS was summarized in [18].

However, the aforementioned optimization-based HEMS methods were executed according to deterministic equations to illustrate the operation characteristics of home appliances and DERs (e.g., consecutive operation time intervals of WMs and state of energy (SOE) dynamics of ESS) as well as consumer's comfort level (e.g., preferred indoor temperature using indoor temperature dynamics). Consequently, the model-based HEMS optimization approach suffers from two limitations. First, the characteristics for the operation of appliances/DERs and consumer preference are expressed through approximated unrealistic equations with fixed parameters, thereby leading to an inaccurate energy consumption schedule. Second, an optimization method including a large number of decision variables could significantly increase the computation complexity and could not scale well with a greater number of houses. Furthermore, the solution resulting from model-based optimization may not always be guaranteed and often diverges owing to a smaller feasible region with a large number of operational constraints for the HEMS optimization problem. To address the aforementioned limitations, we propose a data-driven approach that leverages model-free reinforcement learning (RL) to calculate the optimal schedule of home energy consumption.

Recently, data-driven approaches based on various machine learning (ML) methods have gained popularity owing to their more efficient residential energy management. In [19,20], methods to predict the generation output of a PV system accurately during the day for efficient energy management of buildings were presented. These methods employed an artificial neural network (ANN) and a deep neural network (DNN), respectively. DNN methods were also used in [21–23] for load forecasting to minimize the energy usage of buildings and households. More recently, RL has received attention as

a promising ML method for the energy management of buildings and homes. A pioneering study on RL-based energy management is Google DeepMind, which was developed using RL and proved to decrease the electricity bill by cooling the data center by approximately 40%. Another RL-based method, referred to as Q-learning, was applied to HEMS problems. It was integrated with the ANN module for estimating the consumer's comfort level, maintaining the energy efficiency of household appliances [24,25] and predicting the pricing in real time [26]. A new demand response strategy for HEMS was proposed through the combination of Q-learning and fuzzy reasoning, which reduces the number of state-action pairs and fuzzy logic for reward functions [27]. Furthermore, methods based on deep reinforcement learning (DRL) such as Deep Q-Network, and policy gradient, were applied for energy management of building [28,29] based on both discrete and continuous action spaces. A holistic DRL method for the energy management of commercial buildings was presented in [30] where Heating, Ventilation, and Air conditioning (HVAC) system, lighting, blind, and window systems are controlled to achieve energy savings within the buildings' occupants comfort in terms of thermal, air quality, and illumination conditions. To resolve the limit of model-free DRL methods such as low sample efficiency, a model-based RL method was developed for building HVAC control that trains the system dynamics using neural networks [31]. Based on the trained system dynamics, the operation of the HVAC system was managed by model predictive control to minimize both the energy cost and the indoor temperature constraints violation. The DRL approach was also applied to data centers with servers, which aim to minimize the energy used for moving air and on-demand cooling in the data centers through the control of the temperature and relative humidity of air supplied to the server [32].

Recent studies addressed on energy management systems for buildings and households using an RL-based method. However, to the best of the authors' knowledge, no study has presented an DRL-based algorithm that considers the continuous operations of heterogeneous home appliances and DERs according to the consumer's comfort and preferences. In prior studies [24–26], the operations of home appliances and DERs were scheduled using a simple Q-learning method based on an unrealistic discrete action space. Furthermore, prior studies [28,29] focused on the scheduling of the energy consumption of buildings without considering the operation characteristics of home appliances in detail.

In this study, we propose a two-level DRL framework that employs an actor–critic method where the controllable home appliances (WM and AC) are scheduled at the first level according to the consumer's preferred appliance scheduling and comfort level. The ESS and EV are scheduled at the second level to cover the aggregated WM and AC loads that are calculated at the first level along with the fixed load of the uncontrollable appliances. The proposed two-level scheme is motivated by the interdependent operation between the home appliances at the first level and the ESS/EV at the second level. If the proposed algorithm is executed in a single-level framework, the optimal policy for charging and discharging actions of the ESS and EV is independently determined without considering the energy consumption schedule of aggregated home appliances, thereby degrading the performance of the algorithm. Figure 1 presents the conceptual system model for the proposed two-level DRL-based HEMS that employs an actor–critic method, along with the data classification associated with the utility company, weather station, and consumer. The main contributions of this study are summarized as follows:

- We present a two-level distributed DRL model for optimal energy management of a smart home consisting of a first level for WM and AC, and a second level for ESS and EV. In such a model, the energy consumption scheduling at the second level is based on the aggregated energy consumption scheduled at the first level to determine the better policy of charging and discharging actions for the ESS and EV.
- Compared to the existing method using Q-learning in a discrete action space, we propose a hierarchical DRL in a continuous action space with the following two scheduling steps: (i) the controllable appliances including WM and AC are scheduled at the first level according to the

consumer's preferred appliance scheduling and comfort level; (ii) ESS and EV are scheduled at the second level, thereby resulting in optimal cost of electricity for a household.

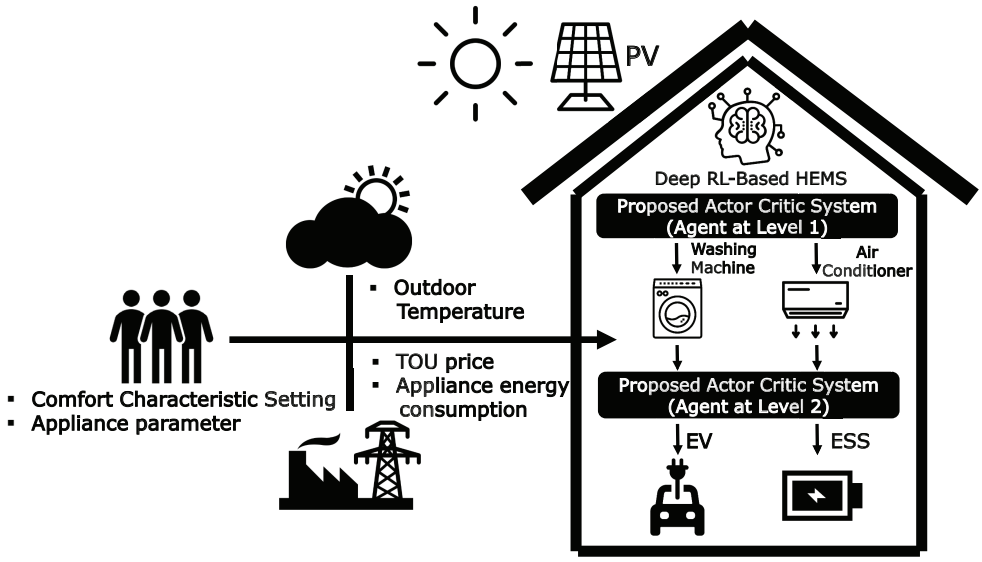

Figure 1. Conceptual architecture of the proposed deep reinforcement learning (DRL)-based home energy management system (HEMS) algorithm.

The simulation results confirmed that the proposed HEMS algorithm can successfully schedule the energy consumptions of multiple home appliances and DERs using a single DRL structure under various consumer preferences. In addition, through various case studies and comparative analysis, we evaluated the impact of different weather and driving patterns of the EV with different initial SOEs on the proposed algorithm. Furthermore, we verified that charging and discharging of the ESS and EV significantly contribute to the reduction of the cost of electricity.

The remainder of this paper is organized as follows. Section 2 introduces the various types of smart home appliances and the traditional HEMS optimization approach; it also provides an overview of the RL methodology. Section 3 presents the formulation of the proposed DRL-based HEMS algorithm based on the actor–critic method. The numerical examples for the proposed HEMS algorithm are reported in Section 4, and the conclusions are given in Section 5.

2. Background

2.1. Types of Smart Home Appliances

We consider the following types of smart home appliances in a single household where an HEMS automatically schedules their energy consumption under the time-of-use (TOU) pricing:

1. Uncontrollable appliance ($\mathcal{A}^{uc}$): An HEMS cannot manage the energy consumption scheduling of uncontrollable appliances such as televisions, personal computers, and lighting. Thus, an uncontrollable appliance is assumed to follow fixed energy consumption scheduling.
2. Controllable appliance ($\mathcal{A}^{c}$): It is an appliance for which the energy consumption scheduling is calculated by the HEMS. According to its operation characteristics, the controllable appliance is categorized into a reducible appliance ($\mathcal{A}^{c}_{r}$) and shiftable appliance ($\mathcal{A}^{c}_{s}$). A representative example of a reducible appliance is an air conditioner whose energy consumption can be curtailed to reduce the cost of electricity. By the contrast, under TOU pricing, the energy consumption scheduling of a shiftable appliance can be moved from one time slot to another to minimize the cost of electricity. A shiftable appliance has two types of load: (i) a non-interruptible load

($\mathcal{A}_s^{c,NI}$), and (ii) an interruptible load ($\mathcal{A}_s^{c,I}$). A shiftable appliance with an interruptible load can be interrupted at any time. For example, the HEMS must stop the discharging process and start the charging process of the ESS instantly when the PV power generation is greater than the load demand. However, the operation period of a shiftable appliance with a non-interruptible load must not be terminated by the HEMS. For example, a washing machine must finish a washing cycle prior to drying.

2.2. Traditional HEMS Optimization Approach

A conventional HEMS method that calculates the optimal operating scheduling of home appliances and DERs is formulated in terms of the following constrained multi-objective optimization problem described in the following section.

2.2.1. Objective Function

The objective function (1) for the HEMS optimization problem comprises two terms, each of which includes different decision variables (E_t^{net}, T_t^{in}) [25]:

$$\min_{E_t^{\text{net}}, T_t^{\text{in}}} \underbrace{\sum_{t \in \mathcal{T}} \pi_t E_t^{\text{net}}}_{J_1\left(E_t^{\text{net}}\right)} + \underbrace{\epsilon \sum_{t \in \mathcal{T}} \left| T_t^{\text{in}} - T^{\text{set}} \right|}_{J_2\left(T_t^{\text{in}}\right)}. \tag{1}$$

The first term $J_1\left(E_t^{\text{net}}\right)$ represents the total cost of electricity that is calculated under TOU pricing π_t and E_t^{net}, which is the net energy consumption accounting for the energy consumption of the controllable/uncontrollable appliances and the predicted PV generation output. The second term $J_2\left(T_t^{\text{in}}\right)$ represents the total penalty related to the cost of the consumer's discomfort. Here, discomfort is defined as a deviation of the consumer's preferred temperature T^{set} from the indoor temperature T_t^{in}. ϵ is a penalty for the cost of the consumer's discomfort. A larger ϵ yields a smaller $J_2\left(T_t^{\text{in}}\right)$, thereby offering decreasing discomfort to the consumer at the expense of less energy saving. The value of ϵ can be tuned by the HEMS operator to maintain the consumer's preferred comfort level at the cost of a higher electricity bill. The equality and inequality constraints for the HEMS optimization problem are illustrated in the following subsections.

2.2.2. Net Energy Consumption

Equation (2) expresses the constraint on the net energy consumption that represents the difference between the total consumption of all home appliances ($\mathcal{A} = \mathcal{A}_r^c \bigcup \mathcal{A}_s^{c,NI} \bigcup \mathcal{A}_s^{c,I} \bigcup \mathcal{A}^{uc}$) and the predicted PV generation output E_t^{PV} at time t. In Equation (3), the total energy consumption of all appliances in Equation (2) is decomposed into four different types of consumptions corresponding to (i) reducible appliances ($a \in \mathcal{A}_r^c$), (ii) shiftable appliances with a non-interruptible load ($a \in \mathcal{A}_s^{c,NI}$), (iii) shiftable appliances with an interruptible load ($a \in \mathcal{A}_s^{c,I}$), and (iv) uncontrollable appliances ($a \in \mathcal{A}^{uc}$) [25]:

$$E_t^{\text{net}} = \sum_{a \in \mathcal{A}} E_{a,t} - \widehat{E}_t^{\text{PV}} \tag{2}$$

$$\sum_{a \in \mathcal{A}} E_{a,t} = \sum_{a \in \mathcal{A}_r^c} E_{a,t} + \sum_{a \in \mathcal{A}_s^{c,NI}} E_{a,t} + \sum_{a \in \mathcal{A}_s^{c,I}} \left(E_{a,t}^{\text{ch}} - E_{a,t}^{\text{dch}} \right) + \sum_{a \in \mathcal{A}^{uc}} E_{a,t}. \tag{3}$$

2.2.3. Operation Characteristics of Controllable Appliances

For a reducible appliance $a \in \mathcal{A}_r^c$, (4) expresses the constraint for the indoor temperature dynamics of a reducible appliance (e.g., an AC) at time t (T_t^{in}), which is expressed in terms of T_{t-1}^{in} at time $t-1$, the predicted outdoor temperature at time $t-1$ ($\widehat{T}_{t-1}^{\text{out}}$), the energy consumption of the reducible appliances ($E_{a,t}$), and the environmental parameters (α, β) characterizing the indoor

thermal condition [7]. Equation (5) illustrates the range of consumer's preferred indoor temperatures. The energy consumption capacity for the reducible appliances is limited according to (6):

$$T_t^{\text{in}} = T_{t-1}^{\text{in}} + \alpha(\widehat{T}_{t-1}^{\text{out}} - T_{t-1}^{\text{in}}) + \beta E_{a,t} \tag{4}$$

$$T^{\min} \leq T_t^{\text{in}} \leq T^{\max} \tag{5}$$

$$E_a^{\min} \leq E_{a,t} \leq E_a^{\max}. \tag{6}$$

Equations (7), (8), and (9) guarantee the consumer's preferred operation of shiftable appliances with a non-interruptible load $a \in \mathcal{A}_s^{c,NI}$ (e.g., a WM) with the binary decision variable $b_{a,t}^{c,NI}$ in different situations: (i) for a stopping period in which ω_s^{pref} and ω_f^{pref} are the consumer's preferred starting and finishing time (7), respectively; (ii) for an operation period of L_a hours during a day in (8); and (iii) for a consecutive operation period of L_a hours in (9). The energy consumption capacity for the shiftable appliances with a non-interruptible load is calculated using (10):

$$b_{a,t}^{c,NI} = 0, \qquad t \in [1, \omega_s^{\text{pref}}) \cup (\omega_f^{\text{pref}}, T] \tag{7}$$

$$\sum_{t=\omega_s^{\text{pref}}}^{\omega_f^{\text{pref}}} b_{a,t}^{c,NI} = L_a \tag{8}$$

$$\sum_{t=p}^{p+L_a-1} b_{a,t}^{c,NI} \geq (b_p^{c,NI} - b_{p-1}^{c,NI})L_a, \ \ \forall p \in (\omega_s^{\text{pref}}, \omega_f^{\text{pref}} - L_a + 1) \tag{9}$$

$$E_{a,t} = b_{a,t}^{c,NI} E_a^{\max}. \tag{10}$$

Equation (11) presents the operational dynamics of the SOE for the ESS and EV ($a \in \mathcal{A}_s^{c,I}$) at current time instant t in terms of the SOE at previous time instant t-1, the charging and discharging efficiency, i.e., η_a^{ch} and η_a^{dch}, and the charging and discharging energy, i.e., $E_{a,t}^{\text{ch}}$ and $E_{a,t}^{\text{dch}}$, respectively [11]. Equation (12) represents the SOE capacity constraint for the ESS and EV. Equations (13) and (14) present the limits of the charging ($E_{a,t}^{\text{ch}}$) and discharging ($E_{a,t}^{\text{dch}}$) energies of the ESS and EV, respectively, where $b_{a,t}^{c,I}$ represents the binary decision variable that determines the charging and discharging status of the ESS and EV:

$$SOE_{a,t} = SOE_{a,t-1} + \eta_a^{\text{ch}} E_{a,t}^{\text{ch}} - \frac{E_{a,t}^{\text{dch}}}{\eta_a^{\text{dch}}} \tag{11}$$

$$SOE_a^{\min} \leq SOE_{a,t} \leq SOE_a^{\max} \tag{12}$$

$$E_a^{\text{ch,min}} b_{a,t}^{c,I} \leq E_{a,t}^{\text{ch}} \leq E_a^{\text{ch,max}} b_{a,t}^{c,I} \tag{13}$$

$$E_a^{\text{dch,min}} (1 - b_{a,t}^{c,I}) \leq E_{a,t}^{\text{dch}} \leq E_a^{\text{dch,max}} (1 - b_{a,t}^{c,I}). \tag{14}$$

Notably, the constraints (11)–(14) for the EV remain true in $t \in [\omega^{\text{arr}}, \omega^{\text{dep}}]$, whereas $E_{a,t}^{\text{ch}}$ and $E_{a,t}^{\text{dch}}$ become zero in $t \notin [\omega^{\text{arr}}, \omega^{\text{dep}}]$. In addition, when the EV departs from home at $t = \omega^{\text{dep}}$, the SOE of the EV must be larger than the consumer preferred SOE SOE_a^{pref}

$$SOE_{a,t} \geq SOE_a^{\text{pref}}. \tag{15}$$

2.3. Reinforcement Learning Methodology

2.3.1. Reinforcement Learning

RL is an ML method that addresses a problem in a specific environment with the objective of maximizing a numerical reward. This learning process is applied to various types of general and special engineering problems. In the RL framework, while an agent interacts with an environment, it learns a particular type of action depending on the state of the environment and conveys the learned action to the environment. The environment then returns a reward along with its new state to the agent. This learning process continues until the agent maximizes the total cumulative rewards received from the environment.

A policy is defined in terms of the procedure through which the agent acts from a specific state. The main objective of the agent is to find an optimal policy that maximizes the agent's cumulative reward in the environment. In our study, we consider that the environment is characterized by a Markov decision process, in which the change of the agent's next state depends only on the current state, along with the action chosen in the current state ignoring all previous states and actions.

In this study, the value function is selected as $Q(s_t, a_t)$, namely the Q-value, which is written in terms of a pair of state s_t and action a_t at a discrete time t. By using the Q-value, the agent's main objective is to achieve the maximum Q-value at every time step t. Q-learning is one of the basic RL methods to find the optimal policy v^* in decision-making problems. The general Q-learning process computes and updates the Q-value $Q(s_t, a_t)$ to achieve the maximum total rewards using the following Bellman equation:

$$Q^*_{v^*}(s_t, a_t) = r(s_t, a_t) + \gamma \max Q(s_{t+1}, a_{t+1}) \tag{16}$$

where, based on the optimal policy v^*, the optimal Q-value $Q^*_{v^*}(s_t, a_t)$ is obtained by the summation of the present reward $r(s_t, a_t)$ and the maximum discounted future reward $\gamma \max Q(s_{t+1}, a_{t+1})$.

In general, a discounting factor $\gamma \in [0, 1]$ is used to explain the relative importance of the current and future rewards. As the discounting factor γ decreases, the agent becomes short-sighted because it increasingly focuses on the current reward. However, a larger γ enables the agent to focus increasingly on the future reward and thus becomes far-sighted. The value of γ can be tuned by the system operator to balance the current and future rewards.

2.3.2. Actor–Critic Method

The actor–critic method is an extension of the policy gradient method that can improve the stability and reduce the variance of the gradient when the optimal solution of the algorithm converges [33]. If the value of a certain state of the agent is known, the corresponding Q-value can be calculated and applied to the REINFORCE method (Algorithm 1) to calculate the gradient of policy network parameters and renew the agent's policy network, thereby leading to a better cumulative result by increasing the probability for the agent's action. In the actor–critic method, the agent can use an additional network to judge the goodness of the action the agent selects in a certain state. The policy network that returns the probability of the agent's action is called actor network, whereas the network that returns the evaluation value of the agent's action is called a critic network. The policy gradient method is suitable for handling the problem with a continuous action space; however, this method may have a poor convergence performance. An additional critic network in the actor–critic method can resolve the convergence issue in the policy gradient method. In our study, the actor and critic networks share a common body network owing to an effective convergence consideration.

The actor network returns the probability of the action that the agent selects in a particular state and the critic network returns the numerical future value that the agent would obtain in the terminal state. The critic network updates the function that distinguishes between the action and value, whereas the policy network updates its parameters in the direction suggested by the critic network. In this study, the parameters in the actor network are updated by the REINFORCE method (Algorithm 1) and

the parameters in the critic network are updated by a linear temporal difference (TD) method [34]. The TD method directly learns from episodes of experience, and the present and guessed values. This is known to be a useful method to solve the Markov decision process (MDP) problem. In our study, we select the linear value function approximation for applying the TD method to the critic network. The actor–critic method updates the parameters of the actor and critic networks to minimize the TD error, which encodes the difference between the value function of the present state and target value function. Algorithm 2 illustrates the actor–critic approach with the TD method. In Algorithm 2, ∂, θ, and ω represent the TD error from the Q-value, the parameter of the actor network, and the parameter of the critic network, respectively. ξ_θ and ξ_ω are the learning rates of the actor network and critic networks in the algorithm, respectively.

Algorithm 1: REINFORCE method

Initialize the weights of network θ randonmly
for $t = 1, N$ **do**
  ▷ Store each episode $\{s_1, a_1, r_2, \ldots, s_N, a_N, r_{N+1}\} \sim \pi_\theta$
  for $t = 1, N$ **do**
    ▷ Apply the stochastic gradient (SG) method for updating the weights
    $\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) Q(s_t, a_t)$
  end
end
Return θ

Algorithm 2: Actor-critic method

Initialize the state s of an agent and the weights of the actor and critic networks θ and ω, respectively, randomly
Sample action $a \sim \pi_\theta$
Linear value function approximation $Q_w(s, a) = \phi(s, a)$
for $t = 1, N$ **do**
  ▷ Sample reward $r = R_s^a$ and next state $\hat{s}$
  ▷ Select action $\hat{a}$
  ▷ $\partial = r + \gamma Q_w(\hat{s}, \hat{a}) - Q_w(s, a)$
  ▷ $\theta = \theta + \xi_\theta \nabla_\theta \log \pi_\theta(s, a) Q_{s,a}$
  ▷ $w \leftarrow w + \xi_\omega \partial \phi(s, a)$
  ▷ $a \leftarrow \hat{a}$, $s \leftarrow \hat{s}$
end

3. Proposed Method for DRL-Based Home Energy Management

In this section, we propose a hierarchical two-level DRL framework that employs the actor–critic method to schedule the day-ahead optimal energy consumption of a single household with smart home appliances and DERs within the consumer's preferred appliance scheduling and comfort range. In the proposed framework, the energy consumption scheduling problem is decomposed into a two-level DRL problem corresponding to the energy consumption scheduling of i) WM and AC at the first level and ii) ESS and EV at the second level. A detailed illustration of the state, action, and reward for each level of the proposed actor–critic method is provided in the following two subsections. In addition, the superscript of the variables for the state/action spaces and the reward function represents the first and second level, respectively.

3.1. Energy Management Model for WM and AC: Level 1

3.1.1. State Space

We consider the situation in which the proposed DRL-based HEMS algorithm performs optimal day-ahead scheduling of appliances with a 1-h scheduling resolution. For $\forall t = 1, \ldots, 24$, the state space of the agent for WM and AC at the first level is defined as follows:

$$\mathcal{S}^{(1)} = \{t, \pi_t, \widehat{T}_t^{\text{out}}, T_{t-1}^{\text{in}}\} \tag{17}$$

where the states t, π_t, $\widehat{T}_t^{\text{out}}$, and T_{t-1}^{in} denote the scheduling time of the WM and AC, TOU price, and outdoor and indoor temperatures, respectively, at time t.

3.1.2. Action Space

The optimal action for the first level relies on the environment of the agent including the present state, as defined in Section 3.1.1. The action space at the first level is defined as follows:

$$\mathcal{A}^{(1)} = \{E_t^{\text{WM}}, E_t^{\text{AC}}\} \tag{18}$$

where E_t^{WM} and E_t^{AC} represent the energy consumption of the WM and AC at time t, respectively. In this study, E_t^{AC} has a continuous value, whereas E_t^{WM} has a discrete value; $E_t^{\text{WM}} = E^{\text{WM,max}}$ when the WM turns on, otherwise, $E_t^{\text{WM}} = 0$.

3.1.3. Reward

The reward function for the first level is formulated as the sum of the negative cost of electricity and dissatisfaction of WM and AC related to the consumer's preferred comfort and appliance's operation characteristics. The total reward at the first level is expressed as:

$$\mathcal{R}_t^{(1)} = -(c_t^{\text{WM}} + c_t^{\text{AC}}) \tag{19}$$

where c_t^{WM} and c_t^{AC} are the cost functions for the WM and AC, respectively. Each cost function includes the cost of electricity for the appliance along with the cost of the consumer's dissatisfaction for the undesired operation of the WM and the indoor thermal discomfort.

First, the cost function of the WM is defined as

$$c_t^{\text{WM}} = \begin{cases} \pi_t E_t^{\text{WM}} + \overline{\delta}(\omega_s^{\text{pref}} - t), & \text{if } t < \omega_s^{\text{pref}} \\ \pi_t E_t^{\text{WM}} + \underline{\delta}(t - \omega_f^{\text{pref}}), & \text{if } t > \omega_f^{\text{pref}} \\ \pi_t E_t^{\text{WM}}, & \text{otherwise} \end{cases} \tag{20}$$

where ω_s^{pref} and ω_f^{pref} are the consumer's preferred starting and finishing times of the WM, respectively, while $\overline{\delta}$ and $\underline{\delta}$ are the penalties for early and late operations, respectively, compared to the consumer's preferred operation interval. The cost of dissatisfaction is added to the cost function if the WM schedules the WM energy consumption earlier than ω_s^{pref} or later than ω_f^{pref}; otherwise, the cost function includes only the cost of electricity.

The cost function of the AC is expressed as:

$$c_t^{\text{AC}} = \begin{cases} \pi_t E_t^{\text{AC}} + \overline{\kappa}(T^{\min} - T_t^{\text{in}}), & \text{if } T_t^{\text{in}} < T^{\min} \\ \pi_t E_t^{\text{AC}} + \underline{\kappa}(T_t^{\text{in}} - T^{\max}), & \text{if } T_t^{\text{in}} > T^{\max} \\ \pi_t E_t^{\text{AC}}, & \text{otherwise} \end{cases} \tag{21}$$

where $\overline{\kappa}$ and $\underline{\kappa}$ are the penalties for the consumer's thermal discomfort. The cost of dissatisfaction is defined as the deviation of the consumer's preferred temperature T_t^{in} from $T^{\min}$ and $T^{\max}$.

It is noted that two terms in (20) and (21) have a trade-off relationship between the saving of the electricity cost and the reduction of the consumer's dissatisfaction cost in terms of the penalties $\{\overline{\delta}, \underline{\delta}\}$ and $\{\overline{\kappa}, \underline{\kappa}\}$, respectively. On the trade-off relationship, HEMS operators using our proposed DRL algorithm can adaptively adjust and tune the penalty to the situations where the consumer aims to save the electricity cost more or maintain the consumer's desired comfort and preference. The selection of the values of these penalties would become different depending on the consumer's desired comfort level and environment.

3.2. Energy Management Model for ESS and EV: Level 2

The optimal schedules of the energy consumption of the WM and AC from the first level along with the fixed load of the uncontrollable appliances are embedded into the actor–critic module at the second level. In the second level, the agent for the ESS and EV initiates the learning process to determine the optimal charging and discharging schedules of the ESS and EV to minimize the cost of electricity. During the learning process in this second level, the energy generated by the PV system is assumed to be charged first to the ESS; then, the ESS will select an appropriate action.

3.2.1. State Space

The state space of the agent at the second level, which manages the operations of the ESS and EV, is defined as

$$\mathcal{S}^{(2)} = \{t, \pi_t, SOE_t^{\text{ESS}}, SOE_t^{\text{EV}}, \widehat{E}_t^{\text{PV}}, E_t^{(1)}\} \tag{22}$$

where the states t, π_t, SOE_t^{ESS}, SOE_t^{EV}, $\widehat{E}_t^{\text{PV}}$, and $E_t^{(1)}$ are the scheduling time of the ESS and EV, the TOU price, SOE of the ESS and EV, the predicted PV generation output, and the aggregated energy consumption schedule calculated at the first level, respectively, at time t.

3.2.2. Action Space

Similar to the action space of the WM and AC in Section 3.1.2, the action space of the ESS and EV at the second level is expressed as

$$\mathcal{A}^{(2)} = \{E_t^{\text{ESS}}, E_t^{\text{EV}}\} \tag{23}$$

where E_t^{ESS} and E_t^{EV} represent the continuous energy charging and discharging of the ESS and EV, respectively, at time t.

Note that, in the proposed two-level DRL architecture, the agent for the ESS and EV selects their optimal charging and discharging action using $E_t^{(1)}$ (22) that includes the action of the agent for the WM and AC along with the fixed load of the uncontrollable appliances at the first level. If the DRL-based HEMS algorithm is modelled as a single-level framework (i.e., the state spaces (17), (22) and the action spaces (18), (23) at the first and second levels are combined, respectively), the agent for the ESS and EV may not find its optimal policy because the ESS and EV have no consumption data of other appliances in their state space. This is verified in Section 4.2.

3.2.3. Reward

The reward for the second level is formulated as the sum of the negative cost of electricity and dissatisfaction of ESS and EV associated with the consumer's preferred comfort and appliance's operation characteristics. The total reward at the second level is defined as

$$\mathcal{R}_t^{(2)} = -(c_t^{\text{ESS}} + c_t^{\text{EV}}). \tag{24}$$

In (24), c_t^{ESS} and c_t^{EV} represent the cost functions for the ESS and EV, respectively. Each cost function includes the cost of electricity of the appliance along with the cost of dissatisfaction for

underdischarging and overcharging of the ESS and EV. Notably, these cost functions include the discharging energy from the ESS and EV, which supports the uncovered energy consumption of aggregated load for WM, AC, and uncontrollable appliances.

First, the cost function of the ESS is expressed as follows:

$$c_t^{\mathrm{ESS}} = \begin{cases} \pi_t E_t^{\mathrm{ESS}} + \overline{\tau}(SOE_t^{\mathrm{ESS}} - SOE^{\mathrm{ESS,max}}), \\ \qquad\qquad \text{if } SOE_t^{\mathrm{ESS}} > SOE^{\mathrm{ESS,max}} \\ \pi_t E_t^{\mathrm{ESS}} + \underline{\tau}(SOE^{\mathrm{ESS,min}} - SOE_t^{\mathrm{ESS}}), \\ \qquad\qquad \text{if } SOE_t^{\mathrm{ESS}} < SOE^{\mathrm{ESS,min}} \\ \pi_t E_t^{\mathrm{ESS}}, \qquad\qquad \text{otherwise,} \end{cases} \tag{25}$$

where $\overline{\tau}$ and $\underline{\tau}$ are the penalties for ESS overcharging and undercharging, respectively. In this case, energy underutilization and dissipation of the ESS occur if the SOE becomes lower than $SOE^{\min}$ (undercharging) or greater than $SOE^{\max}$ (overcharging).

Next, the cost function of the EV is expressed as

$$c_t^{\mathrm{EV}} = \begin{cases} \pi_t E_t^{\mathrm{EV}} + \overline{\nu}(SOE_t^{\mathrm{EV}} - SOE^{\mathrm{EV,max}}), \\ \qquad \text{if } SOE_t^{\mathrm{EV}} > SOE^{\mathrm{EV,max}}, \ \ t \in [\omega^{\mathrm{arr}}, \omega^{\mathrm{dep}}] \\ \pi_t E_t^{\mathrm{EV}} + \underline{\nu}(SOE^{\mathrm{EV,min}} - SOE_t^{\mathrm{EV}}), \\ \qquad \text{if } SOE_t^{\mathrm{EV}} < SOE^{\mathrm{EV,min}}, \ \ t \in [\omega^{\mathrm{arr}}, \omega^{\mathrm{dep}}] \\ \pi_t E_t^{\mathrm{EV}} + \underline{\eta}(SOE^{\mathrm{Pref}} - SOE_t^{\mathrm{EV}}), \\ \qquad \text{if } t = \omega^{\mathrm{dep}} \text{ and } SOE_t^{\mathrm{EV}} < SOE^{\mathrm{Pref}} \\ \pi_t E_t^{\mathrm{EV}}, \qquad \text{otherwise,} \end{cases} \tag{26}$$

where $\overline{\nu}$ and $\underline{\nu}$ are the penalties for overcharging and undercharging of the EV, respectively. Similar to the operation of the ESS, energy underutilization and dissipation occur if the SOE of the EV becomes lower than $SOE^{\mathrm{EV,min}}$ or higher than $SOE^{\mathrm{EV,max}}$. Unlike the reward function of the ESS, the reward function of the EV includes the parameter $\underline{\eta}$, which denote the consumer's preference penalty of the EV, corresponding to the deviation of the SOE of the EV from the consumer's preferred SOE when the EV departs. If SOE_t^{EV} is lower than SOE^{Pref} at departure time ω^{dep}, the cost of dissatisfaction increases owing to insufficient SOE.

3.3. Proposed Actor–Critic-Based HEMS Algorithm

In this subsection, we illustrate the proposed DRL method based on the actor–critic method that determines the optimal policy to minimize the electricity bill within the consumer's preferred comfort level and the appliance operation characteristics. Compared to value-based RL methods, the policy gradient approach is appropriate for engineering problems with continuous action spaces. In general, the continuous policy gradient network obtains state information from the agent and returns the appropriate action using a normal distribution. The network yields the mean and variance to achieve a normal distribution, and the agent samples the action randomly based on the resulting distribution. In the actor–critic approach, the additory method of criticizing the Q-value for its efficiency and convergence is added. Therefore, the network provides the mean, variance, and Q-values to find the optimal actions. As shown in Figure 2, the proposed actor–critic network model for each level consists of one input layer for state elements, four hidden layers for a common body network with 512 neurons, one hidden layer for each actor and critic networks with 256 neurons, and one output layer with means and variances of the operation schedules of appliances and Q-values. In this study, a hyperbolic tangent function was used as a transfer function. In addition, the adaptive moment estimation (ADAM) optimization algorithm [35] was used for training the proposed DRL model with a learning rate of 0.00004. Finally, Algorithm 3 illustrates the procedure of the actor–critic-based HEMS algorithm that

learns the energy management policies, which optimize the cost of electricity and consumer's comfort level for level 1 (WM and AC) and level 2 (ESS and EV).

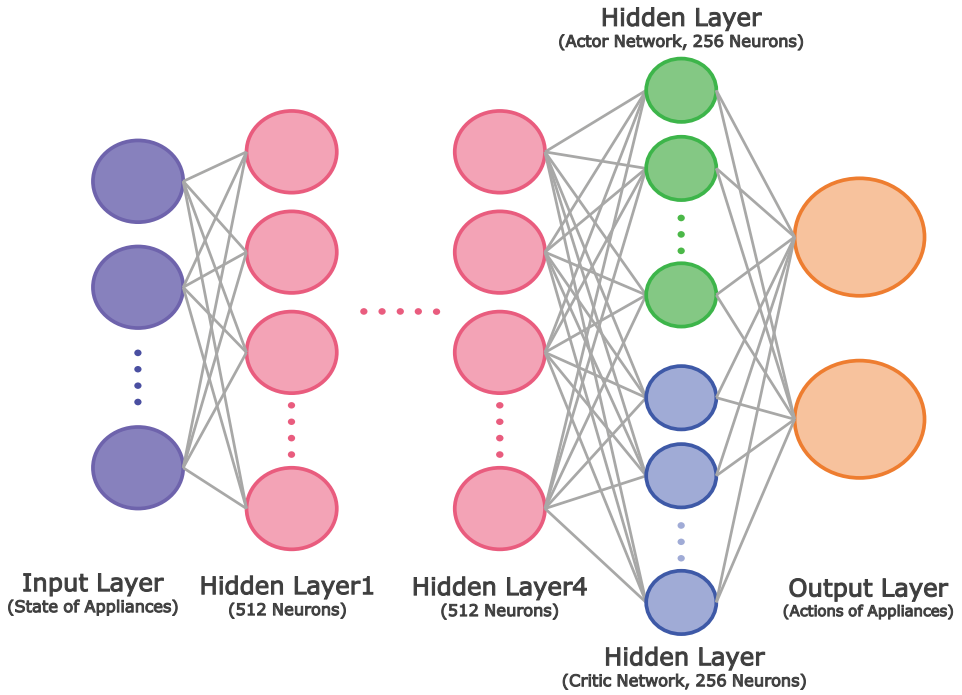

Figure 2. Architecture of the neural network model for the proposed actor–critic method.

Algorithm 3: Proposed actor–critic-based energy management of smart home at level 1 (or level 2).

```
Initialize each appliance's energy demand, dissatisfaction parameters, and actor critic method parameters
%%Learning the optimal energy demand scheduling of the agent for level 1 (or level 2)
Initialize the policy gradient and Q-value of the agent for level 1 (or level 2)
for episode = 1, MaxEpisode do
    ▷ Initialize state, action and time period
    for time step = 1, 24 do
        ▷ With probability of action in state, select a set of actions for each state
        ▷ Execute the action chosen from the set of actions and calculate the Q-value of present state and next state
        ▷ Calculate the loss function of each actor network and critic network
        ▷ Store each loss of actor and critic network in buffer
        ▷ Update the parameters of the full network by applying the SG descent method to the loss in the buffer
    end
    ▷ Find the optimal policy with the largest Q-value
end
```

4. Numerical Examples

4.1. Simulation Setup

Under the TOU pricing shown in Figure 3a, we considered a household where the proposed DRL algorithm schedules the operation of two tasks: (i) two major controllable home appliances (WM and AC) at the first level and (ii) controllable DERs (PV-integrated ESS and EV) at the second level. The simulations were executed for 24 h with a 1-h scheduling resolution.

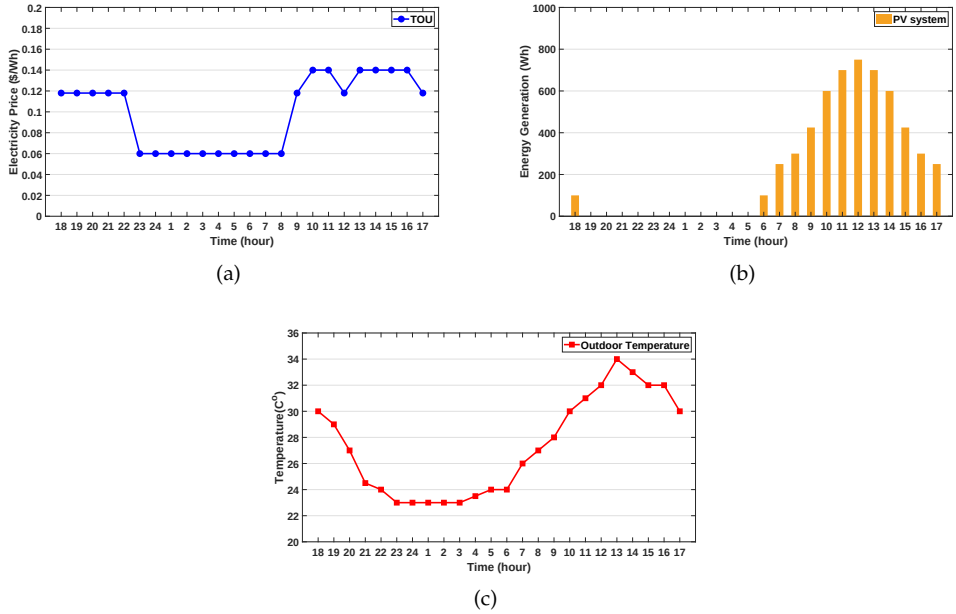

Figure 3. Profiles of electricity price and weather. (**a**) time-of-use (TOU) price; (**b**) photovoltaic (PV) generation; (**c**) outdoor temperature.

We assumed that the predicted PV generation energy $\widehat{E}_t^{\mathrm{PV}}$ in Figure 3b and predicted outdoor temperature $\widehat{T}_t^{\mathrm{out}}$ in Figure 3c could be calculated accurately. The maximum energy consumptions of the WM and AC were set to 300 and 1000 Wh, respectively. For the ESS, the battery capacity is 4000 Wh, and the minimum ($SOE^{\mathrm{ESS,min}}$), maximum ($SOE^{\mathrm{ESS,max}}$) and initial SOE ($SOE_{\mathrm{ini}}^{\mathrm{ESS}}$) of the ESS were set to 400 Wh (10% $SOE^{\mathrm{ESS,max}}$), 4000 Wh (100% $SOE^{\mathrm{ESS,max}}$), and 2000 Wh (50% $SOE^{\mathrm{ESS,max}}$), respectively. The maximum charging and discharging energy of the ESS were both 1200 Wh. For the EV, the battery capacity was 17,000 Wh, and the minimum ($SOE^{\mathrm{EV,min}}$), maximum ($SOE^{\mathrm{EV,max}}$) and initial SOE ($SOE_{\mathrm{ini}}^{\mathrm{EV}}$) of the EV were set to 1700 Wh (10% $SOE^{\mathrm{EV,max}}$), 17,000 Wh (100% $SOE^{\mathrm{EV,max}}$), and 9350 Wh (55% $SOE^{\mathrm{EV,max}}$), respectively. The maximum charging and discharging energy of the EV were both 10,000 Wh. For the reward function, the consumer's preferred operating period $[\omega_s^{\mathrm{pref}}, \omega_f^{\mathrm{pref}}]$ for the WM was set to [9:00 a.m., 10:00 p.m.] along with 2 h of consecutive operation time. The range of consumer's comfortable indoor temperature $[T^{\min}, T^{\max}]$ controlled by the AC was set to [22.5°C, 25.5°C]. For the EV, the preferred SOE (SOE^{pref}) and the departure time (ω^{dep}) of the EV were set to 12,750 Wh (75% $SOE^{\mathrm{EV,max}}$) and 8:00 a.m., respectively. The pairs of penalties for the cost of dissatisfaction of the WM, AC, ESS, and EV were $\{(\overline{\delta} = 50, \underline{\delta} = 50)$, $(\overline{\kappa} = 200, \underline{\kappa} = 200)$, $(\overline{\tau} = 100, \underline{\tau} = 100)$, $(\overline{\nu} = 100, \underline{\nu} = 100)$, and $(\underline{\eta} = 50)\}$, respectively.

The performance of the proposed approach was tested for the following four cases according to different weather, weekday/weekend, and initial SOE $SOE_{\mathrm{ini}}^{\mathrm{EV}}$ of the EV:

- Case 1: Sunny, weekday, $SOE_{\mathrm{ini}}^{\mathrm{EV}} = 0.55 \times SOE^{\mathrm{EV,max}}$,
- Case 2: Rainy, weekday, $SOE_{\mathrm{ini}}^{\mathrm{EV}} = 0.55 \times SOE^{\mathrm{EV,max}}$,
- Case 3: Sunny, weekend, $SOE_{\mathrm{ini}}^{\mathrm{EV}} = 0.55 \times SOE^{\mathrm{EV,max}}$,
- Case 4: Sunny, weekday, $SOE_{\mathrm{ini}}^{\mathrm{EV}} = 0.15 \times SOE^{\mathrm{EV,max}}$.

On a sunny day, the predicted PV generation output at time t ($\widehat{E}_t^{\mathrm{PV}}$) follows the profile in Figure 3b, and on a rainy day, it is set to zero. On a weekday, the EV is assumed to arrive at the household at 6:00 p.m. and then the charging and discharging processes are conducted until the EV departs from the household at 8:00 a.m. During the weekend, the EV charges or discharges energy during 24 h. $SOE_{\mathrm{ini}}^{\mathrm{EV}}$ denotes the SOE when the EV arrives at home, and different values of $SOE_{\mathrm{ini}}^{\mathrm{EV}}$ represent different

driving distances of the EV. All the cases were tested using Python 3.7.0 with the machine learning package pytorch 1.1.0.

4.2. Simulation Results at Level 1

In this subsection, we report the simulation results of the proposed approach associated with level 1 and verify the optimal energy consumption schedule of the WM and AC along with the consumer's comfort level. Figure 4a shows the energy consumption schedule of the WM. We observe from Figure 4a that the operation period is selected as [7:00 p.m., 8:00 p.m.] with two consecutive operation hours. This scheduling policy is optimal because the WM operates at the lowest TOU pricing during the consumer's preferred operation period [9:00 a.m., 10:00 p.m.], which in turn reduces the electricity bill while satisfying the consumer's preference and operation characteristics of the WM. Figure 4b illustrates the energy consumption schedule of the AC. Unlike the observation in Figure 4a, the AC energy consumption is scheduled at an even higher TOU pricing during the period [12:00 p.m., 4:00 p.m.]. This was expected because the agent at the first level considers the consumer's thermal comfort as well as saving on electricity bills in the reward function. As shown in Figure 4b, the AC turns off at midnight when the outdoor temperature is within the range of consumer's preferred temperatures. When an indoor temperature violation occurs at 7:00 a.m. owing to a sharp increase of the outdoor temperature, the AC turns on and its energy consumption increases to maintain the consumer's preferred indoor temperature. The maximum energy consumption schedule is verified during the interval [12:00 p.m., 4:00 p.m.], which presents the highest TOU pricing. In this interval, the AC aims to satisfy the consumer's comfort at the expense of increased electricity bills. Figure 4c shows the total energy consumption schedule of the WM, AC, and uncontrollable appliances with fixed loads at level 1. Note that the sum of energy consumption schedules of WM, AC, and uncontrollable appliances at each period is used by the actor–critic module at the second level, which in turn determines the optimal policy of charging and discharging for the ESS and EV.

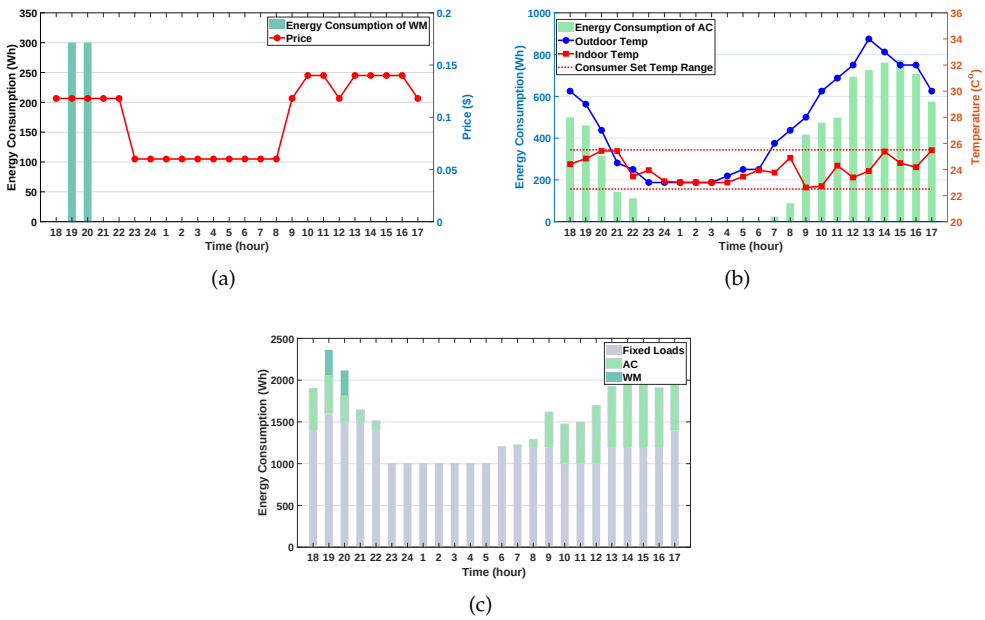

Figure 4. DRL-based energy consumption schedule. (**a**) washing machine (WM); (**b**) air conditioner (AC); (**c**) WM+AC+Uncontrollable appliances at Level 1.

4.3. Simulation Results at Level 2

In this subsection, we present the simulation results of the proposed approach at Level 2. These results are divided into three comparison tests using four cases described in Section 4.1: {Case 1,

Case 2}, {Case 1, Case 3}, and {Case 1, Case 4}. The discharging ratio of the EV and ESS to cover the energy consumption at the first level simultaneously was set to 0.8 and 0.2, respectively. This ratio was determined as the battery capacity of the EV divided by the battery capacity of the ESS.

4.3.1. Case 1 vs. Case 2

In this simulation, we investigate and compare the performance between Case 1 (with the PV generation output) and Case 2 (without the PV generation output). Figure 5a,b show the charging/discharging and SOE schedules of the ESS for Case 1 and Case 2, respectively. We observe from Figure 5a that, in general, more discharging (negative energy consumption) of the ESS for both cases occurs at high TOU pricing to support the household energy demand, thereby leading to consumer's energy savings. In addition, it can be observed through the comparison of Figure 5a,b that the SOE of the ESS increases (or decreases) as the ESS charges (or discharges) energy. However, Case 1 shows an unexpected phenomenon where the SOE of the ESS in the scheduling period between 10:00 a.m and 4:00 p.m is higher than in other periods even though the ESS conducts a significant energy discharging in this period. This phenomenon occurs because the PV generation energy is injected into the ESS further than the ESS discharging energy in this period.

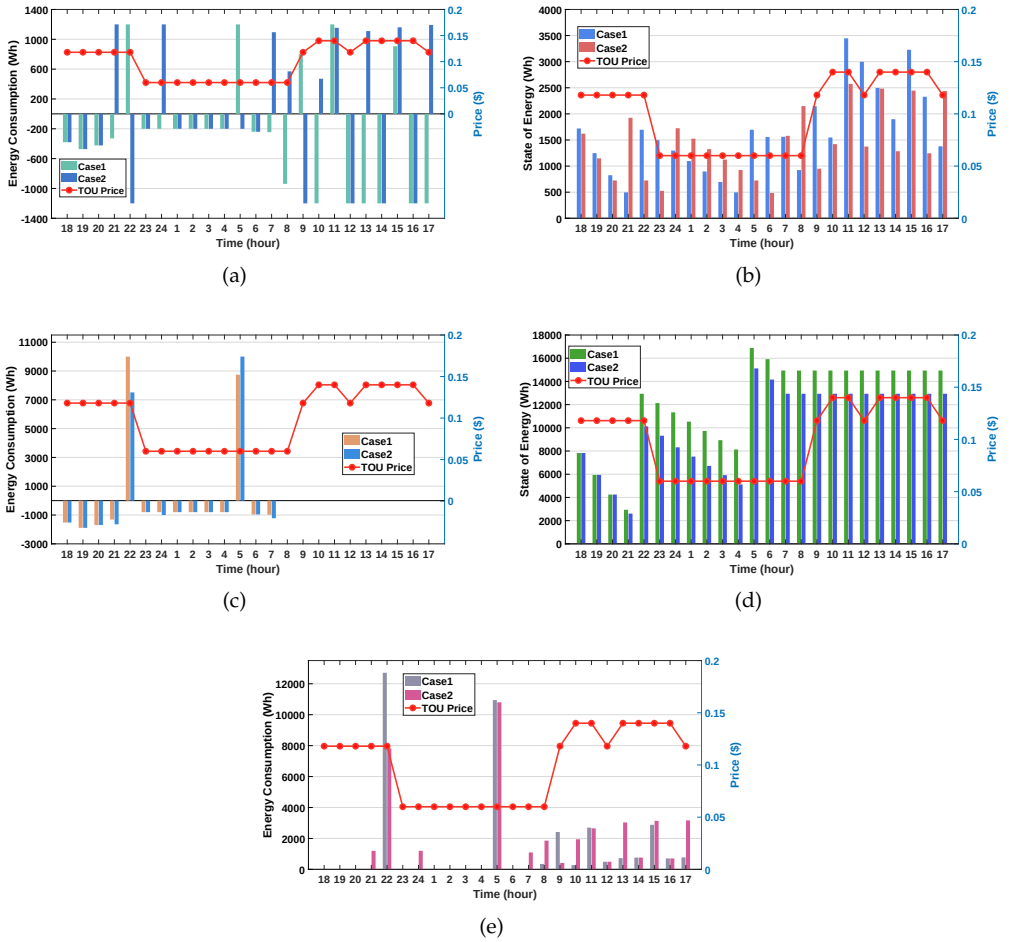

Figure 5. Performance comparison between Case 1 and Case 2. (**a**) energy consumption of the energy storage system (ESS); (**b**) state of energy (SOE) of the ESS; (**c**) energy consumption of the electric vehicle (EV); (**d**) SOE of the EV; (**e**) net consumption of household.

However, we observe from Figure 5a,b that the ESS charges more power at high TOU pricing for Case 2 than for Case 1. This observation justifies that the different weather conditions influence the charing and discharging schedules of the ESS significantly.

Figure 5c,d show the charging/discharging and SOE schedules for the EV. We observe from these figures that the EVs for Cases 1 and 2 does not perform neither charging nor discharging processes after 7:00 a.m. and that the SOE level remains unchanged, respectively. This observation is consistent with the EVs departure time setting (ω^{dep}=8:00 a.m.). We also observe from Figure 5c that the EVs charge significantly large amounts of energy from the grid at 10:00 p.m. and 5:00 a.m., whereas it discharges energy to support the household energy demand in the other time periods. This charging process derives from the fact that the EVs charge sufficient energy in advance to satisfy the consumer's preferred SOE condition (SOE^{pref}=12,750 Wh). It can be verified from Figure 5d that the SOE at 8:00 a.m. exceeds 12,750 Wh for both cases. However, no unexpected phenomenon observed in Figure 5a,b is identified in Figure 5c,d. This is because the PV generation output affects only the ESS charging and discharging.

Figure 5e shows the net energy consumptions of the household for Cases 1 and 2, which is the difference between the energy consumption of the controllable/uncontrollable appliances and the predicted PV generation output. As shown in this figure, the values of the net energy consumption of both cases are much larger at 10:00 p.m. and 5:00 a.m. than at the other time slots owing to the large EV charging at these two time slots. However, in the period between 8:00 a.m. and 5:00 p.m., we can identify the large amount of energy consumption in Case 2. This is because the ESS needs to charge more power in Case 2 owing to no PV generation than in Case 1.

4.3.2. Case 1 vs. Case 3

In this case study, we investigate and compare the performance between Case 1 (on a weekday) and Case 3 (on a weekend). We considered that the EV stays at home during 24 h in the weekend. Given that the EV does not depart, the consumer's preferred SOE of the EV is not considered in the proposed algorithm. First, we can observe from Figure 6a that the SOE of the ESS in Case 3 is generally lower than that in Case1 even if the weather is sunny. We can interpret this observation as follows.

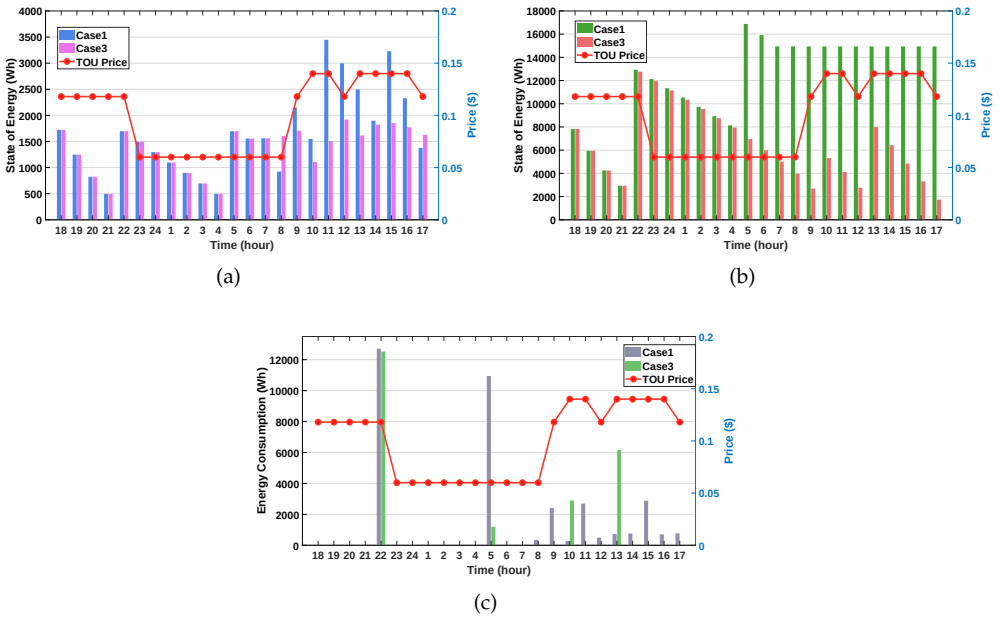

Figure 6. Performance comparison between Case 1 and Case 3. (**a**) SOE of the ESS; (**b**) SOE of the EV; (**c**) net consumption of household.

In Case 3, the ESS does not need to allocate much stored energy to satisfy the consumer's preferred SOE of the EV because the EV stays at home all day. Thus, the ESS increasingly supports the household energy demand through its charging process along with the EV charging. Unlike the results for Cases 1 and 2, we can verify from Case 3 in Figure 6b that the consumer's preferred SOE of EV condition is ignored during the scheduling process of EV energy consumption, in which the SOE of the EV at 8:00 a.m. is much lower than SOE^{pref} = 12,750 Wh. After 8:00 a.m., the EV keeps charging and discharging the energy in the same way as the ESS. We also observe from Figure 6c that in Case 3 a high net energy consumption occurs at 10:00 p.m. and 1:00 p.m. owing to EV charging for discharging plan in future scheduling. Moreover, the net energy consumption at 5:00 a.m. in Case 3 is much smaller than in Case 1 because the consumer's preferred SOE of EV is ignored. Note from Case 3 in Figure 6c that zero-energy consumption is verified in the period between 8:00 a.m. and 5:00 p.m. except 10:00 a.m. and 1:00p.m. This result shows that more energy saving can be obtained during the EV charging and/or discharging during the whole day.

4.3.3. Case 1 vs. Case 4

In this simulation, we investigate and compare the performance between Case 1 (with a high initial SOE) and Case 4 (with a low initial SOE). Figure 7a,b illustrate the SOE schedules of the ESS and EV for Case 1 and Case 4, respectively. First, we observe from Figure 7b that the SOE of the EV at 6:00 p.m. is larger in Case 4 than in Case 1. This is because the low initial SOE of the EV enable the EV to require more charging power to support the upcoming household load demands. This observation is also verified for the ESS as shown Figure 7a where the ESS charges more power in Case 4 than in Case 1 at 6:00 p.m. with the same reason. We also observe from Case 4 in Figure 7c that the highest net energy consumption occurs at 6:00 p.m., owing to a significant charging of the ESS and EV. It is noted that this large amount of the net energy consumption is not observed in previous cases. This is because in these cases the EV has a high initial SOE so that it does not have to charge energy from grid in advance. Except the time slot at 6:00 p.m., the schedules of the net energy consumption at other time slots are similar to the schedules in Case 1.

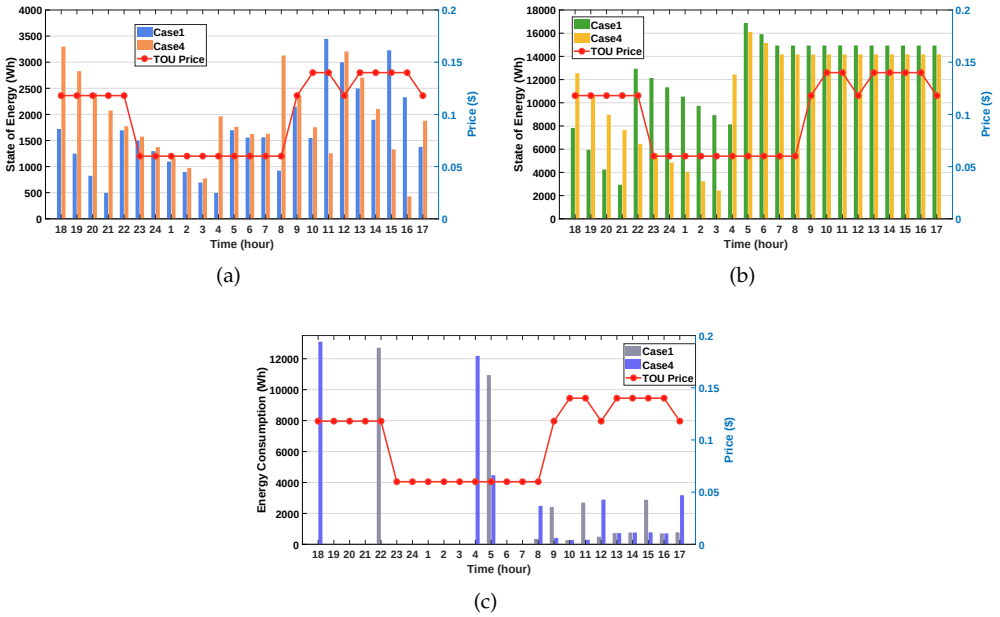

Figure 7. Performance comparison between Case 1 and Case 4. (**a**) SOE of the ESS; (**b**) SOE of the EV; (**c**) net consumption of household.

Figure 8 shows a relative increase of the total electricity bill for the aforementioned Cases 1, 2, and 4 with respect to Case 3 with the minimum total electricity bill using the following metric:

$$\frac{X_n^{\text{bill}} - X_3^{\text{bill}}}{X_3^{\text{bill}}} \times 100(\%), \tag{27}$$

where X_3^{bill} is the total electricity bill for Case 3 and X_n^{bill} is the total electricity bill for Case n where n =1, 2, and 4.

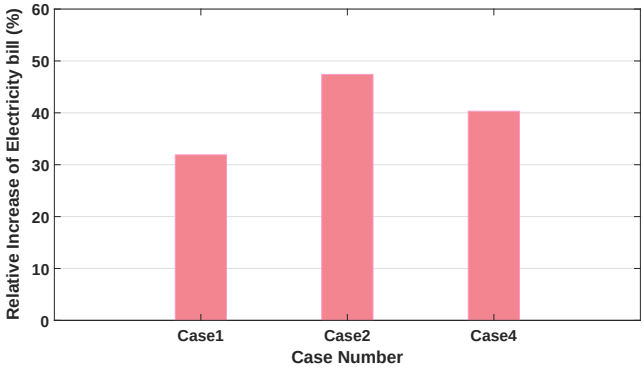

Figure 8. Comparison of a relative increase of the total electricity bill in Case 3 among the considered three cases.

Note in this figure that the relative total electricity bills for the three cases are listed in decreasing order of their bills as follows: Case 2 > Case 4 > Case 1. Note in turn from this list that the relative increase of the electricity bill in Case 1 (a sunny weekday with high $SOE_{\text{ini}}^{\text{EV}}$) is smaller than the other two cases (a sunny or rainy weekday with high or low $SOE_{\text{ini}}^{\text{EV}}$). Thus, a fraction of the charging and discharging periods of the EV constitutes the most influential aspect for saving on electricity bills.

Figure 9a,b show training curves that present the convergence of the total cost for the first and second levels, respectively. Each figure compares the three training curves using policy gradient method, actor–critic method with separate actor and critic neural networks (NNs), and proposed actor–critic method with a common NN. We observe from Figure 9a,b that policy gradient and actor–critic with separate NNs show a poor performance of the convergence. By contrast, the proposed actor–critic shows that the training curves steadily decrease and then converge to an optimal policy within a moderate training period.

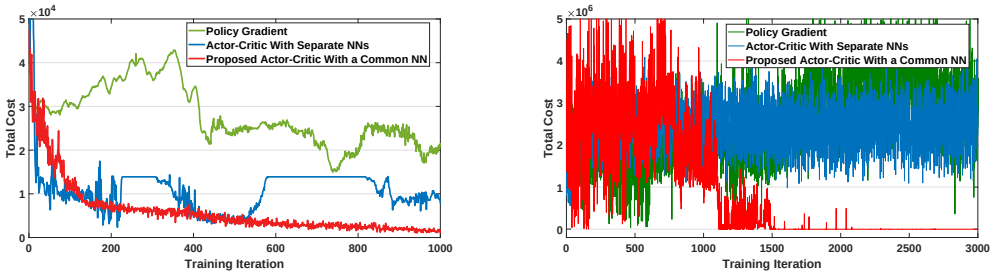

Figure 9. Convergence of the total cost. (**a**) Level 1; (**b**) Level 2.

Figure 10 compares the training curves for the level 1 and level 2 in the proposed DRL method and the single-level DRL method. We can observe from this figure that each training curve for the level 1 and level 2 in the proposed approach steadily decreases and converges to an optimal policy in the training periods of [1, 1,000] and [1,001, 4,000], respectively. By contrast, the training curve for

the single-level approach shows a large fluctuation during the training process and a poor value of the result compared to the proposed two-level DRL method. The poor convergence performance of the single-level approach derives from the fact that the complexity of the HEMS problem increases dramatically as the state-action space dimension becomes larger in the single level. Furthermore, since the agent for the EV and ESS is not able to obtain energy consumption data of other appliances in the single level, the single-level approach may not calculate optimal charging and discharging actions of the ESS and EV.

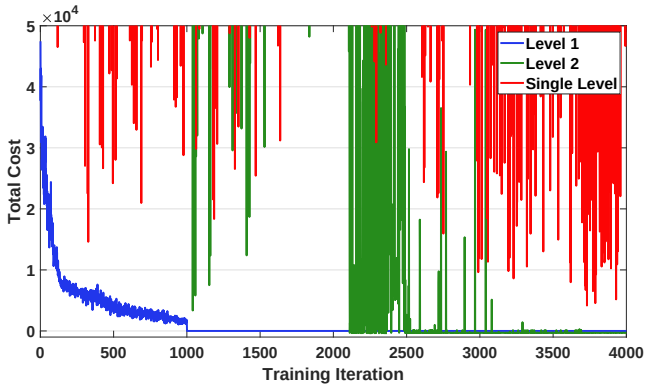

Figure 10. Comparison of the total cost convergence between the proposed two-level DRL approach and the single-level DRL approach.

Figure 11 shows a relative increase of the total electricity bill in Cases 1–4 using a Building Energy Optimization Tool (BEopt) [36] and MILP optimization method with respect to the proposed approach. BEopt is widely used as an energy simulation program for the residential building. To fairly compare the performance of the proposed method to that of the MILP method, the MILP method was executed for 24 h with a 1-h scheduling resolution. We can verify from this figure that our proposed method is the most economical for all four cases compared to two existing approaches using BEopt and MILP methods. In particular, it is observed that Case 3 (a sunny weekend with high $SOE_{\text{ini}}^{\text{EV}}$) shows the largest cost increase. This observation implies that the proposed DRL approach schedules the charging and discharging energy of EV in a much more cost-effective way.

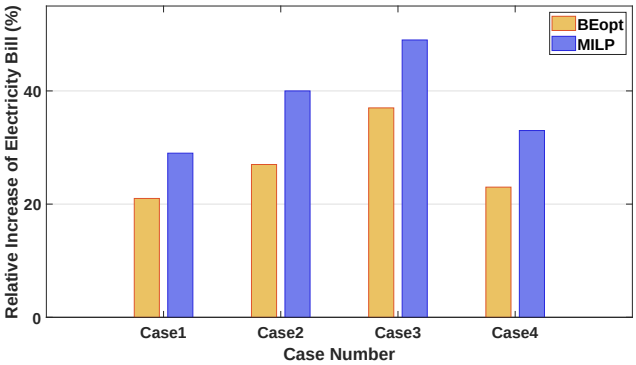

Figure 11. Comparison of a relative increase of the total electricity bill in the proposed DRL method using building energy optimization tool (BEopt) and mixed-integer linear programming (MILP) programs for four cases.

The meaningful observations of the proposed HEMS approach in the numerical examples can be summarized as follows:

- Through a comparison between {Case 1, Case 3, Case 4} (with PV system) and Case 2 (without PV system), we conclude that PV generation has a significant impact on the reduction of the total cost of electricity. For example, the total cost of electricity in Case 2 is 11% higher than in Case 1.
- Given that the battery capacity of the EV is approximately four times larger than that of the ESS, the EV can discharge more power than the ESS to cover the total cost of electricity. This can be verified through a comparison between {Case 1, Case 2} (in weekday) and Case 3 on weekends. In contrast, different driving patterns associated with the initial SOE of the EV significantly influence the total cost of electricity. We conclude from a comparison between Case 1 (with high SOE) and Case 4 (with low SOE) that the total cost of electricity in Case 4 is 7% higher than in Case 1. This is because the EV with low SOE needs to charge more power than with high SOE to satisfy the consumer's preferred SOE at departure time.

5. Conclusions

In this study, we propose a two-level distributed deep reinforcement learning algorithm to minimize the cost of electricity through the energy consumption scheduling of two controllable home appliances (an air conditioner and a washing machine) and the charging and discharging of an energy storage system and an electric vehicle while maintaining the consumer's comfort level and appliance operation characteristics. In the proposed deep reinforcement learning method, two agents interact with each other to schedule the optimal home energy consumption efficiently. One agent for a washing machine and an air conditioner determines their continuous actions in the first level to schedule optimal energy consumption within the consumer's preferred indoor temperature and operation period, respectively. Based on the optimal energy consumption schedules from the first level, the other agent for an energy storage system and an electric vehicle conducts their continuous charging and discharging actions in the second level to support the aggregated load for controllable and uncontrollable appliances. The comparative case studies under different weather and driving patterns of the electric vehicle with different initial state of energy confirm that the proposed approach can successfully minimize the cost of electricity within the consumer's preference.

In future work, we plan to develop a multi-agent reinforcement learning algorithm based on a continuous action space that schedules the energy consumption of multiple smart homes with home appliances and distributed energy resources. A key challenge is to design an information exchange scheme between households to minimize the cost of electricity and maintain each consumer's comfort level. This future work can be implemented using advanced deep reinforcement learning methods such as deep deterministic policy gradient and asynchronous advantage actor–critic methods.

Author Contributions: S.L. proposed the DRL-based home energy system model and conducted the simulation study. D.-H.C. coordinated the approach that is proposed in this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Research Foundation of Korea (NRF) Grant through the Korea Government (MSIP) under Grant 2018R1C1B6000965, and in part by the Competency Development Program for Industry Specialists of the Korean Ministry of Trade, Industry and Energy (MOTIE), operated by Korea Institute for Advancement of Technology (KIAT) under Grant P0002397 HRD program for Industrial Convergence of Wearable Smart Devices.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. U.S. Energy Information Administration. Electricity: Detailed State Data. 2018. Available online: https://www.eia.gov/electricity/data/state/ (accessed on 26 March 2020).
2. Althaher, S.; Mancarella, P.; Mutale, J. Automated demand response from home energy management system under dynamic pricing and power and comfort constraints. *IEEE Trans. Smart Grid.* **2015**, *6*, 1874–1883. [CrossRef]

3. Paterakis, N.G.; Erdinc, O.; Pappi, I.N.; Bakirtzis, A.G.; Catalão, J.P.S. Coordinated operation of a neighborhood of smart households comprising electric vehicles, energy storage and distributed generation. *IEEE Trans. Smart Grid.* **2016**, *7*, 2736–2747. [CrossRef]
4. Liu, Y.; Xiao, L.; Yao, G. Pricing-based demand response for a smart home with various types of household appliances considering customer satisfaction. *IEEE Access* **2019**, *7*, 86463–86472. [CrossRef]
5. Wang, C.; Zhou, Y.; Wu, J.; Wang, J.; Zhang, Y.; Wang, D. Robust-Index Method for Household Load Scheduling Considering Uncertainties of Customer Behavior. *IEEE Trans. Smart Grid.* **2015**, *6*, 1806–1818. [CrossRef]
6. Joo, I.-Y.; Choi, D.-H. Distributed optimization framework for energy management of multiple smart homes with distributed energy resources. *IEEE Access* **2017**, *5*, 2169–3536. [CrossRef]
7. Mhanna, S.; Chapman, A.C.; Verbic, G. A Fast Distributed Algorithm for Large-Scale Demand Response Aggregation. *IEEE Trans. Smart Grid.* **2016**, *7*, 2094–2107. [CrossRef]
8. Rajasekharan, J.; Koivunen, V. Optimal Energy Consumption Model for Smart Grid Households With Energy Storage. *IEEE J. Sel. Topics Signal Process.* **2014**, *8*, 1154–1166. [CrossRef]
9. Ogata, Y.; Namerikawa, T. Energy Management of Smart Home by Model Predictive Control Based on EV state Prediction. In Proceedings of the 12th Asian Control Conference (ASCC), Kitakyushu-shi, Japan, 9–12 June 2019; pp. 410–415.
10. Hou, X.; Wang, J.; Huang, T.; Wang, T.; Wang, P. Smart Home Energy Management Optimization Method Considering ESS and PEV. *IEEE Access* **2019**, *5*, 144010–144020. [CrossRef]
11. Erdinc, O.; Paterakis, N.G.; Mendes, T.D.P.; Bakirtzis, A.G.; Catalão, J.P.S. Smart household operation considering bi-directional EV and ESS utilization by real-time pricing-based DR. *IEEE Trans. Smart Grid.* **2015**, *6*, 1281–1291. [CrossRef]
12. Tushar, M.H.K.; Zeineddine, A.W.; Assi, C. Demand-side Management by Regulating Charging and Discharging of the EV, ESS, and Utilizing Renewable Energy. *IEEE Trans. Ind. Informat.* **2018**, *14*, 117–126. [CrossRef]
13. Floris, A.; Meloni, A.; Pilloni, V.; Atzori, L. A QoE-aware approach for smart home energy management. In Proceedings of the IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA, 6–10 December 2015; pp. 1–6. [CrossRef]
14. Chen, Y.; Lin, R.P.; Wang, C.; Groot, M.D.; Zeng, Z. Consumer operational comfort level based power demand management in the smart grid. In Proceedings of the 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), Berlin, Germany, 14–17 October 2012; pp. 14–17. [CrossRef]
15. Kuzlu, M. Score-based intelligent home energy management (HEM) algorithm for demand response applications and impact of HEM operation on customer comfort *IET Gener. Transm. Distrib.* **2015**, *9*, 627–635. [CrossRef]
16. Li, M.; Jiang, C.-W. QoE-aware and cost-efficient home energy management under dynamic electricity prices. In Proceedings of the IEEE ninth International Conference on Ubiquitous and Future Networks (ICUFN 2017), Milan, Italy, 4–7 July 2017; pp. 498–501. [CrossRef]
17. Li, M.; Li, G.-Y.; Chen, H.-R.; Jiang, C.-W. QoE-Aware Smart Home Energy Management Considering Renewables and Electric Vehicles. *Energies* **2018**, *11*, 2304. [CrossRef]
18. Leitao, J.; Gil, P.; Ribeiro, B.; Cardoso, A. A Survey on Home Energy Management. *IEEE Access* **2020**, *8*, 5699–5722. [CrossRef]
19. Cononnioni, M.; D'Andrea, E.; Lazzerini, B. 24-Hour-Ahead Forecasting of Energy Production in Solar PV Systems. In Proceedings of the 2011 International Conference on Intelligent Systems Design and Application, Cordoba, Spain, 22–24 November 2011; pp. 410–415. [CrossRef]
20. Chow, S. Lee, E.; Li, D. Short-Term Prediction of Photovoltaic Energy Generation by Intelligent Approach. *Energy Build.* **2012**, *55*, 660–667. [CrossRef]
21. Zaouali, K.; Rekik, R.; Bouallegue, R. Deep Learning Forecasting Based on Auto-LSTM Model for Home Solar Power Systems. In Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications, Exeter, UK, 28–30 June 2018; pp. 235–242. [CrossRef]
22. Megahed, T.F.; Abdelkader, S.M.; Zakaria, A. Energy Management in Zero-Energy Building Using Neural Network Predictive Control. *IEEE Internet Things J.* **2019**, *6*, 5336–5344. [CrossRef]
23. Ryu, S.; Noh, J.; Kim, H. Deep Neural Network Based Demand Side Short Term Load Forecasting. *Energies* **2016**, *10*. [CrossRef]

24. Baghaee, S.; Ulusoy, I. User comfort and energy efficiency in HVAC systems by Q-learning. In Proceedings of the 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4. [CrossRef]
25. Lee, S.; Choi, D.-H. Reinforcement Learning-Based Energy Management of Smart Home with Rooftop PV, Energy Storage System, and Home Appliances. *Sensors.* **2019**, *19*. [CrossRef]
26. Lu, R.; Hong, S.H.; Yu, M. Demand Response for Home Energy Management using Reinforcement Learning and Artificial Neural Network. *IEEE Trans. Smart Grid.* **2019**, *10*, 6629–6639. [CrossRef]
27. Alfaverh, F.; Denai, M.; Sun, Y. Demand Response Strategy Based on Reinforcement Learning and Fuzzy Reasoning for Home Energy Management. *IEEE Access* **2020**, *8*, 39310–39321. [CrossRef]
28. Wan, Z.; Li, H.; He, H. Residential Energy Management with Deep Reinforcement Learning. In Proceedings of International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–7. [CrossRef]
29. Wei, T.; Wang, Y.; Zhu, Q. Deep Reinforcement Learning for Building HVAC Control. In Proceedings of the 54th ACM/EDAC/IEEE Design Automation Conference (DAC), Austin, TX, USA, 18–22 June 2017; pp. 1–6. [CrossRef]
30. Ding, X.; Du, W.; Cerpa, A. OCTOPUS: Deep Reinforcement Learning for Holistic Smart Building Control. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, New York, NY, USA, 13–14 November 2019; pp. 326–335. [CrossRef]
31. Zhang, C.; Kuppannagari, S.R.; Kannan, R.; Prasanna, V.K. Building HVAC Scheduling Using Reinforcement Learning via Neural Network Based Model Approximation. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, New York, NY, USA, 13–14 November 2019; pp. 287–296. [CrossRef]
32. Le, D.V.; Liu, Y.; Wang, R.; Tan, R.; Wong, Y.-W.; Wen, Y. Control of Air Free-Cooled Data Centers in Tropics via Deep Reinforcement Learning. In Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, New York, NY, USA, 13–14 November 2019; pp. 306–315. [CrossRef]
33. Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; Riedmiller, M. Deterministic Policy Gradient Algorithms. In Proceedings of the 2014 International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014; pp. 1–9.
34. Tesau, C.; Tesau, G. Temporal Difference Learning and TD-Gammon. *Commun. ACM* **1995**, *38*, 58-68. [CrossRef]
35. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
36. National Renewable Energy Laboratory. Building Energy Optimization Tool (BEopt). Available online: https://beopt.nrel.gov/home (accessed on 26 March 2020).