Reinforcement Learning Based Energy Management Algorithm for Smart Energy Buildings

Authors:

Sunyong Kim, Hyuk Lim

Date Submitted: 2018-09-21

Keywords: Q-learning, reinforcement learning, Markov decision process, renewable energy sources, distributed energy resource, smart energy building, smart grid

Abstract:

A smart grid facilitates more effective energy management of an electrical grid system. Because both energy consumption and associated building operation costs are increasing rapidly around the world, the need for flexible and cost-effective management of the energy used by buildings in a smart grid environment is increasing. In this paper, we consider an energy management system for a smart energy building connected to an external grid (utility) as well as distributed energy resources including a renewable energy source, energy storage system, and vehicle-to-grid station. First, the energy management system is modeled using a Markov decision process that completely describes the state, action, transition probability, and rewards of the system. Subsequently, a reinforcement-learning-based energy management algorithm is proposed to reduce the operation energy costs of the target smart energy building under unknown future information. The results of numerical simulation based on the data measured in real environments show that the proposed energy management algorithm gradually reduces energy costs via learning processes compared to other random and non-learning-based algorithms.

DOI of Published Version: https://doi.org/10.3390/en11082010 energies

MDPI

Article

Reinforcement Learning Based Energy Management Algorithm for Smart Energy Buildings

Sunyong Kim and Hyuk Lim *

School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Korea; sunyongkim@gist.ac.kr

* Correspondence: hlim@gist.ac.kr; Tel.: +82-62-715-2229

Received: 16 July 2018; Accepted: 31 July 2018; Published: 2 August 2018

Abstract: A smart grid facilitates more effective energy management of an electrical grid system. Because both energy consumption and associated building operation costs are increasing rapidly around the world, the need for flexible and cost-effective management of the energy used by buildings in a smart grid environment is increasing. In this paper, we consider an energy management system for a smart energy building connected to an external grid (utility) as well as distributed energy resources including a renewable energy source, energy storage system, and vehicle-to-grid station. First, the energy management system is modeled using a Markov decision process that completely describes the state, action, transition probability, and rewards of the system. Subsequently, a reinforcement-learning-based energy management algorithm is proposed to reduce the operation energy costs of the target smart energy building under unknown future information. The results of numerical simulation based on the data measured in real environments show that the proposed energy management algorithm gradually reduces energy costs via learning processes compared to other random and non-learning-based algorithms.

Keywords: smart grid; smart energy building; distributed energy resource; renewable energy sources; Markov decision process; reinforcement learning; Q-learning

1. Introduction

The term “smart grid” refers to a method of operating within the electrical grid system that is associated with smart energy meters, Energy Storage Systems (ESSs), Renewable Energy Sources (RESs), and communication networks among others [1]. In a smart grid environment, where energy supply and demand are rapidly increasing in modern times, one of the most important issues is developing an effective Energy Management System (EMS) to achieve various goals such as reducing energy consumption, balancing energy supply and demand, increasing the utilization of RES, and minimizing energy costs [2]. However, energy management in a smart grid is a very challenging task given the informational unknowns about factors that change over time, such as load requirements, energy prices, and amount of energy generation. For example, the amount of energy generated by RES such as a Photovoltaic (PV) system is strongly influenced by weather conditions that change over time. The variability of weather conditions makes the amount of energy generation hard to be predicted, consequently leading to the difficulty in the energy management.

In recent years, many research groups working on the smart grid have focused on improving EMS in various environments. Many studies, including [3,4], have focused on energy management for Thermostatically Controlled Loads (TCLs). These works have tried to derive the optimal control policy of TCLs in order to reduce the energy consumption of the smart grid or to meet the requirement of ancillary services while guaranteeing users’ comfort. The energy management of ESS in microgrid environments has been researched in literature. The term “microgrid”, which refers to a localized small electricity grid, can be occasionally disconnected from the traditional centralized grid and operate in an island-mode by leveraging backup energy resources such as ESS, fuel generators, and RES. These research efforts in [5–8] focused on decreasing dependency on the external grid or balancing between energy supply and demand by increasing ESS utilization in the microgrids. There has also been a considerable increase in the interest in energy management of Vehicle-to-Grid (V2G) systems [9–12]. By intelligent management of Electric Vehicles (EVs) charging and discharging scheduling, the profits of V2G station or the costs charged to EV users are optimized.

As reported by the U.S. Energy Information Administration in [13], buildings accounted for the consumption of more than 20% of the total energy delivered worldwide in 2016. This percentage is expected to increase by an average of 1.5% per year through 2040. To prepare for a significant amount of energy consumption in buildings, many research groups have actively worked to develop an EMS for smart energy buildings. The smart energy building represents an Information Technology (IT) centric building that automates its energy operation for achieving the optimal energy consumption. These smart energy buildings can be considered as a type of microgrid systems since they can operate in island-mode with the help of ESS, RES, or V2G systems. The literatures of [14–21] aimed at optimizing the energy consumptions or minimizing operation costs by flexibly and intelligently managing the various kinds of controllable energy sources and loads such as ESS, RES, and V2G system in the smart energy building.

In basic terms, the Reinforcement Learning (RL) is the problem in an area of machine learning concerned with how a learning agent learns what to do (action) in a given situation (state) by interacting with an environment to maximize or minimize numerical returns (rewards) [22]. Because the actions of the agent can affect all subsequent rewards, including the instant reward and its later actions, RL is a closed-loop system. As one of popular methods in RL, Q-learning is widely used for its model-free characteristic, which requires no prior knowledge regarding rewards or transition probabilities in the system, which makes it suitable for operating a system dealing with real-time data without future information or any prediction process. For this reason, many researchers focusing on EMS in the smart grid, especially for TCLs, ESS, and V2G systems in [3,4,6,11,12], have adopted Q-learning in their algorithms to control energy by using real-time information. However, the extant literature contains little evidence of the use of Q-learning for managing energy in smart energy buildings. Much of the research regarding smart energy buildings is on calculating energy schedules by using given or predicted day-ahead information rather than real-time information.

In this study, an EMS for a smart energy building is considered. This study is motivated by a Research and Development (R&D) project that aims at developing an EMS for a smart energy building in the campus of Gwangju Institute of Science and Technology (GIST) in Republic of Korea. The main objective of this project is to reduce the energy costs of a campus building by applying Artificial Intelligence (AI) techniques to the EMS. The smart energy building investigated in this paper is associated with a utility, a PV system, an ESS, and a V2G station, and it can exchange energy with those systems in real-time. The EMS is modeled using a Markov Decision Process (MDP) that completely describes the state space, action space, transition probability, and reward function. Furthermore, the system accounts for all unknowns associated with future information on the load demand of the building, amount of PV generation, load demand of V2G station, and energy prices of utility and V2G. To reduce the energy operation costs associated with this unknown information, a Q-learning-based energy management algorithm is proposed that improves the recommended actions for energy dispatch at every moment by learning through experience without any prior knowledge. Through numerical simulation results, we verify that the proposed Q-learning-based energy management algorithm gradually reduces the daily energy cost of smart energy building as its learning process progresses. The main contributions of this study can be summarized as follows:

- The energy management of a smart energy building is modeled using MDP to completely describe the state space, action space, transition probability, and reward function.

- We propose a Q-learning-based energy management algorithm that provides an optimal action of energy dispatch in the smart energy building. The proposed algorithm can minimize the total energy cost that takes into consideration the future cost by using only current system information, while most existing work for energy management of smart energy buildings focuses on the optimization based on 24 h ahead given or predicted future information.
- To reduce the convergence time of the Q-learning-based algorithm, we propose a simple Q-table initialization procedure, in which each value of Q-table is set to an instantaneous reward directly obtained by the reward function with an initial system condition.
- From the simulations using real-life data sets of building energy demand, PV generation, and vehicles parking records, it is verified that the proposed algorithm significantly reduces the energy cost of smart energy building under both Time-of-Use (ToU) and real-time energy pricing approaches, compared to a conventional optimization-based approach as well as the greedy and random approaches.

The remainder of this paper is organized as follows: an overview of related work is presented in Section 2. In Section 3, the overall system structure of the smart energy building considered in this paper is described, and the system model is formulated using an MDP. In Section 4, a Q-learning-based energy management algorithm is proposed for the smart energy building. Section 5 presents a performance evaluation, and Section 6 concludes this paper.

2. Related Work

In this section, we briefly summarize the related work on energy management in the smart grid environment into three categories: energy management in microgrid, energy management in V2G, and energy management in smart energy building.

2.1. Energy Management in Microgrid

As previously defined in Section 1, the microgrid represents a small electricity grid that is able to operate in island-mode without the help of the external grid by utilizing its own backup energy resources. Most research work in this category focuses on the optimal ESS control for reducing the dependency on the external grid or balancing the energy supply and demand. Ju et al. [5] have proposed a two-layer EMS for a microgrid, where ESS is integrated to maintain energy balancing and minimize the operation cost. The lower layer is formulated as a quadratic mixed-integer problem to minimize power fluctuations induced by demand forecast errors, and the upper layer is formulated as a nonlinear mixed-integer problem to obtain the power dispatch schedules that minimize the total operation cost. In [6], Kuznetsova et al. have suggested an two steps-ahead Q-learning-based ESS scheduling algorithm that determines whether to charge or discharge the ESS under the unknown information about future load demands and wind power generation. The simulation results of a case study have verified that the algorithm increases the utilization of the ESS and wind turbine while reducing grid dependency. In [7], an optimal ESS control strategy in a grid-connected microgrid has been proposed. Especially, to improve the accuracy of State of Charge (SoC) calculation, the authors have applied an extended ESS model, which utilizes 2D efficiency maps of power and SoC, to the optimization problem. The simulation results have verified that the proposed strategy with the extended ESS model guarantees the cost robustness regardless of demand forecast errors. Farzin et al. [8] have focused on managing ESS under two possible scenarios in a microgrid: normal operation mode and unscheduled islanding event. To deal with the trade-off between minimizing the operation cost in the normal mode and increasing the load curtailment in the islanding event, the authors have proposed a multi-objective optimization-based ESS scheduling framework that can simultaneously optimize the cost and curtailment in each corresponding scenario.

2.2. Energy Management in V2G

The research efforts in this category focus on the charging and discharging scheduling of EVs at V2G station for optimizing the benefits of the V2G station or reducing the costs charged to EV users. He et al. [9] have developed a scheduling algorithm for charging and discharging multiple EVs that aims to minimize the costs charged to EV users. To optimize scheduling, they have proposed global and local optimization problems expressed as convex optimization problems and verified that the local optimization problem is more appropriate for a practical environment with a large population and dynamic arrivals of EVs while providing performance close to that of the global optimization problem. Yet another optimal scheduling model for EV charging and discharging has been proposed in [10], where the authors have considered not only the costs charged to EV users but also user preferences and battery lifetime of EVs. However, this model is not operated with real-time information, meaning that scheduling is based on given day-ahead energy information rather than real-time information. EV charging and discharging scheduling algorithms using real-time information have been suggested in [11,12]. Given the unknown future energy price information, these algorithms determine the optimal scheduling at each moment by using the RL technique based only on the current information regarding energy prices to increase the daily profit for an EV user. However, these algorithms are limited in that they can be applied only to a single EV.

2.3. Energy Management in Smart Energy Building

The research efforts in this category generally focus on the energy consumption optimization or the operation cost minimization in smart energy buildings by managing various kinds of controllable energy sources and loads. In [14], Zhao et al. have described the framework and algorithmic aspects of a Cyber Enabled Building EMS, called CEBEMS. The objective of CEBEMS is to minimize the energy cost of a building while satisfying the occupants' set lighting and cooling system points using decision-making control optimization. A case study involving a typical food service center as a test building has been executed to demonstrate the applicability of this framework to commercial buildings. Wang et al. [15] have proposed an intelligent multi-agent optimizer system to maximize occupant comfort and minimize building energy consumption. The multi-agent consists of a central coordinator-agent that coordinates the energy dispatch to local controller-agents and maximizes occupant comfort, and three local controller-agents that use multiple fuzzy logic controllers to satisfy different types of comfort demands. Within the range of set points for temperature, illumination level, and CO_2 concentration given by the occupants in advance, this optimizer system derives the optimal points of the three elements to balance power consumption and comfort demands. In [16], Wang et al. also have applied the intelligent multi-agent optimizer system proposed in [15] to a V2G-integrated building to evaluate the impact of the aggregation of EVs on the building energy consumption as a Distributed Energy Resource (DER). The results of simulation, carried out using data of a 24 h period, have verified that a larger number of EVs are beneficial for satisfying the comfort demands of occupants while reducing energy consumption. Missaoui et al. in [17] have proposed an EMS for smart homes to obtain two solutions: comfort-preferred and cost-preferred. They have conducted a case study with given energy prices and verified that the EMS significantly reduces cost with both solutions. In [18], Basit et al. have proposed a home EMS aiming at optimizing the operation cost by scheduling household appliances without violating the required operation duration of non-schedulable devices. A day-ahead multi-objective optimization model for building energy management has been proposed in [19]. Based on the forecasted information on PV generation, load demand, and temperature, the synergetic dispatch of source-load-storage is scheduled to minimize the operation cost under ToU energy pricing while maintaining the users' comfort. The EMSs operating in short-term to make real-time decisions have been considered in [20,21]. In [20], Yan et al. have focused on managing the EV charging station integrated with a commercial smart building. The chance-constrained optimization-based energy control algorithm that schedules power flows from and to the grid, EV charging and discharging, and ESS charging and discharging in real-time has been proposed to reduce the operational cost of the EV charging station. Piazza et al. [21] have proposed an EMS for smart energy buildings with a PV system and an ESS, where the system operates in two stages: planning stage that optimizes building energy cost by planning the grid-exchanged power profile and online replanning stage that aims at reducing building demand uncertainty.

Please note that our proposed Q-learning-based energy management algorithm falls into the third category. However, the proposed approach is distinct from existing work in that it determines real-time decisions of the optimal actions without any help of future information, while most existing work requires an additional prediction process or simply assume that a set of predicted data is available to solve optimization problems.

3. Energy Management System Model Using MDP

In this section, the overall system structure of a smart energy building is presented, and the system model is formulated using an MDP composed of a state space, action space, transition probabilities, and reward function.

3.1. Overall System Structure

The EMS of a smart energy building is considered with the aim of reducing building operation costs under unknown future information. Figure 1 describes the structure of the smart energy building considered in this paper. As shown, the smart energy building is connected to a utility and DERs including a PV system, ESS, and V2G station.

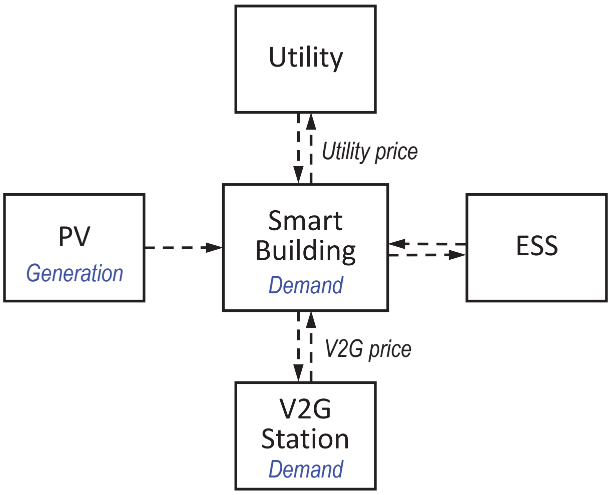


Figure 1. Illustration of smart energy building associated with utility, PV, ESS, and V2G station.

The four components associated with the smart energy building are characterized as follows:

- "Utility" represents a company that supplies energy in real-time. It is assumed that the smart energy building is able to trade (buy and sell) energy with the utility company at any time at the prices determined by the utility company.
- "PV system" is a power supply system that converts sunlight into energy by means of photovoltaic panels. The energy generated by the PV system can be consumed to help meet the load demands of the smart energy building and V2G station.
- "V2G station" describes a system where EVs request charging (energy flow from the grid to EVs) or discharging (energy flow from EVs to the grid) services. It is assumed that the smart energy building trades energy based on the net demand of the V2G station, which can be either positive (if the number of EVs that require charging service is greater than the number of EVs that require discharging service) or negative (vice versa).

- "ESS" represents an energy storage system, capable of storing and releasing energy as needed in a flexible way. As a simple example, the smart energy building can utilize the ESS to decrease its operation cost by charging the ESS when energy prices are low and discharging the ESS when energy prices are high. We assume that the ESS considered in this paper consists of a combination of battery and super-capacitor. This hybrid ESS can take both the advantages of battery (high energy density and low cost per kWh) and super-capacitor (quick charging/discharging and extended lifetime) [23].

Let t denote the index of the present time step, and τ denote the length of each time step (min), during which all system variables are considered to be constant. The EMS operates in a discrete-time manner based on this time step, and the time step is repeated infinitely. In this study, we assume that there are five unknowns in future information as follows:

- "*Building demand*" represents the energy demand required by the smart energy building itself due to its internal energy consumption. The amount of energy demand of the building at each time step t is denoted by e_t^{Bldg} (kWh).
- "*PV generation*" represents the energy generated by the PV system, and is denoted by e_t^{PV} (kWh).
- "*V2G demand*" represents the energy demand of the V2G station, denoted by e_t^{V2G} (kWh). Because the V2G station has bidirectional energy exchange capability with the building, there are two values of e_t^{V2G}, a positive value when the V2G station draws energy from the building and a negative value when it supplies energy to the building.
- "*Utility price*" represents the prices for energy transaction between the building and the utility company, and is denoted by p_t^{Util} ($/kWh).
- "*V2G price*" represents the energy prices for energy transaction when EVs are charged or discharged at V2G station, and is denoted by p_t^{V2G} ($/kWh).

The utility price is usually determined by the utility company in the form of a ToU pricing or a real-time pricing [24]. In the ToU pricing, the utility prices are just offered in a table with a few levels of prices according to time zone. In the real-time pricing, the prices are dynamically determined by wholesale energy market or energy supply and demand conditions. Likewise, V2G prices are available in several forms of pricing policy. The ToU pricing for V2G is generally determined by the utility company or V2G station itself in a way that the economic benefits are maximized. The real-time pricing policy for V2G is determined in accordance with the wholesale energy market or the real-time supply and demand condition by EVs charging and discharging [11]. Especially, there also exist some V2G pricing policies that provide discriminative incentives to EV users who permit the discharging in support of the grid operations [25]. Please note that no matter what policy the utility price and the V2G price follow, they are changing stochastically depending on the time of day, typically holding high values during the daytime and low values during the nighttime.

Future information on these five unknowns is not provided to the EMS at the present time step t, meaning that the information from time step $t + \tau$ is unknown, whereas past and current pieces of information are known to the system. The objective of the EMS is to gradually learn how to manage energy through experience gained over successive time steps under the unknowns of future information. In the following subsections, the state space, action space, transition probabilities, and reward function of the EMS are formulated using an MDP under this assumption of unknowns.

3.2. State Space

Let e_t^{ESS} denote the SoC of the ESS at time step t, and E^{ESS} represent the maximum capacity of the ESS. For safe use of the ESS, a guard ratio, denoted by η, at both ends of the ESS capacity is considered as follows:

$$\eta \cdot E^{\text{ESS}} \leq e_t^{\text{ESS}} \leq (1-\eta) \cdot E^{\text{ESS}}. \tag{1}$$

We define the net demand of the smart energy building at time step t, denoted by e_t^{Net}, as the sum of the energy demands from the building and the V2G station minus the energy generated by the PV system, as follows:

$$e_t^{\mathrm{Net}} = e_t^{\mathrm{Bldg}} + e_t^{\mathrm{V2G}} - e_t^{\mathrm{PV}}. \tag{2}$$

Because *Building demand* (e_t^{Bldg}), *PV generation* (e_t^{V2G}), and *V2G demand* (e_t^{PV}) are unknown variables, as mentioned in Section 3.1, the net demand of the smart energy building (e_t^{Net}) is an unknown variable. Note it is assumed that there is no energy/exergy loss during the charging, discharging, and idling of the ESS.

The time step t basically reaches infinity over time. However, because time is repeated with a period of one day, the state of the time step is considered to be repeated with a period of one day as well. Therefore, the state of the time step, denoted by $\widehat{t}$, can be defined as follows:

$$\widehat{t} = t \mod \frac{h \cdot m}{\tau}, \tag{3}$$

where h and m represent the number of hours per day (i.e., 24) and the number of minutes per hour (i.e., 60), respectively.

Taking all the above into consideration, the state of the EMS at time step t, denoted by s_t, is defined as follows:

$$s_t = [\, e_t^{\mathrm{ESS}} \; e_t^{\mathrm{Net}} \; \widehat{t}\,] \in \mathcal{S}, \tag{4}$$

where $\mathcal{S}$ is the state space of the EMS, and it is composed of five spaces: ESS SoC space $\mathcal{E}^{\mathrm{ESS}}$, energy demand space $\mathcal{E}^{\mathrm{demand}}$, and time space $\mathcal{T}$. Thus, $\mathcal{S} \equiv \mathcal{E}^{\mathrm{ESS}} \times \mathcal{E}^{\mathrm{demand}} \times \mathcal{T}$, where × represents Cartesian product. Please note that the values of utility price (p_t^{Util}) and V2G price (p_t^{V2G}), which are usually dependent on the state of the time step $\hat{t}$, are unknown, but they are not included in the state space. Instead, the stochastic price unknowns are included in the reward function to be used by the Q-learning.

3.3. Action Space

To satisfy e_t^{Net} in each time step t, the EMS of the smart energy building chooses one action from the action space $\mathcal{A}$, which is given by

$$\mathcal{A} = \{\mathit{Buying}, \mathit{Charging}, \mathit{Discharging}, \mathit{Selling}\}, \tag{5}$$

where *Buying* represents the action of buying energy from the utility company to satisfy e_t^{Net}, *Charging* represents charging the ESS for later use, *Discharging* represents discharging the ESS to satisfy e_t^{Net}, and *Selling* denotes the action of selling energy to the utility company. Please note that *Charging* and *Selling* obviously include the actions of *Buying* and *Discharging*, respectively. For example, if the action *Charging* is selected, the EMS buys more energy than the amount required to satisfy the net demand, e_t^{Net}, and the remaining energy is used for charging into the ESS.

We define $a_t \in \mathcal{A}_{s_t}$ as the action taken at time step t, where $\mathcal{A}_{s_t}$ denotes the possible action set in the action space $\mathcal{A}$ under state s_t. In each time step t, $\mathcal{A}_{s_t}$ is constrained by ESS capacity, meaning that at time step t, only actions that satisfy the SoC condition of the ESS in the next time step $t+\tau$, that is, $0 + \eta \cdot E^{\mathrm{ESS}} \leq e_{t+\tau}^{\mathrm{ESS}} \leq (1-\eta) \cdot E^{\mathrm{ESS}}$, can be included in $\mathcal{A}_{s_t}$. Therefore, $\mathcal{A}_{s_t}$ is determined as follows:

$$\mathcal{A}_{s_t} = \begin{cases} \{\mathit{Buying}, \mathit{Charging}\}, & \text{if } 0 \leq e_t^{\mathrm{ESS}} < e_t^{\mathrm{Net}}, \\ \{\mathit{Buying}, \mathit{Charging}, \mathit{Discharging}\}, & \text{if } e_t^{\mathrm{Net}} \leq e_t^{\mathrm{ESS}} < e_t^{\mathrm{Net}} + \Delta e, \\ \{\mathit{Buying}, \mathit{Charging}, \mathit{Discharging}, \mathit{Selling}\}, & \text{if } e_t^{\mathrm{Net}} + \Delta e \leq e_t^{\mathrm{ESS}} < E^{\mathrm{ESS}} - \Delta e, \\ \{\mathit{Buying}, \mathit{Discharging}, \mathit{Selling}\}, & \text{if } E^{\mathrm{ESS}} - \Delta e \leq e_t^{\mathrm{ESS}} \leq E^{\mathrm{ESS}}, \end{cases} \tag{6}$$

where Δe denotes the energy unit for charging the ESS and selling to the utility company. Once the possible action set $\mathcal{A}_{s_t}$ in each time step t is given by (6), the EMS selects one of the possible actions, a_t, from $\mathcal{A}_{s_t}$ according to a certain policy π, which denotes the decision-making rule for action selection. More information about π is covered in the next section.

Let $E(a_t)$ denote the function of the amount of energy charged into the ESS with the taken action a_t, represented by

$$E(a_t) = \begin{cases} 0, & \text{if } a_t = \textit{Buying}, \\ \Delta e, & \text{if } a_t = \textit{Charging}, \\ -e_t^{\text{Net}}, & \text{if } a_t = \textit{Discharging}, \\ -(e_t^{\text{Net}} + \Delta e), & \text{if } a_t = \textit{Selling}, \end{cases} \tag{7}$$

where negative values represent energy discharge from the ESS.

Please note that the derived state and action spaces can be easily extended if the EMS includes other energy components in the smart energy building. For instance, if a Combined Heat and Power (CHP) system is deployed, the state space may include more parameters such as the amount of energy generation by CHP and the indoor temperature, and the action space includes the consuming energy from CHP to meet the demand and the selling energy from CHP to the utility.

3.4. Transition Probability

The transition probability of the EMS from state s_t to state $s_{t+\tau}$ when action a_t is taken can be represented as follows:

$$\mathbb{P}(s_{t+\tau}|s_t, a_t) = \mathbb{P}(e_{t+\tau}^{\text{ESS}}|e_t^{\text{ESS}}) \cdot \mathbb{P}(e_{t+\tau}^{\text{Net}}|e_t^{\text{Net}}) \cdot \mathbb{P}(\widehat{t+\tau}|\hat{t}\,), \tag{8}$$

where $\mathbb{P}(e_{t+\tau}^{\text{ESS}}|e_t^{\text{ESS}})$ and $\mathbb{P}(\widehat{t+\tau}|\hat{t}\,)$ are given by

$$\mathbb{P}(e_{t+\tau}^{\text{ESS}}|e_t^{\text{ESS}}) = \begin{cases} 1, & \text{if } e_{t+\tau}^{\text{ESS}} = e_t^{\text{ESS}} + E(a_t), \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

and

$$\mathbb{P}(\widehat{t+\tau}|\hat{t}\,) = \begin{cases} 1, & \text{if } \hat{t} \leftarrow \widehat{t+\tau}, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

and $\mathbb{P}(e_{t+\tau}^{\text{Net}}|e_t^{\text{Net}})$ is represented by the product of $\mathbb{P}(e_{t+\tau}^{\text{Bldg}}|e_t^{\text{Bldg}})$, $\mathbb{P}(e_{t+\tau}^{\text{V2G}}|e_t^{\text{V2G}})$, and $\mathbb{P}(e_{t+\tau}^{\text{PV}}|e_t^{\text{PV}})$ as follows:

$$\mathbb{P}(e_{t+\tau}^{\text{Net}}|e_t^{\text{Net}}) = \mathbb{P}(e_{t+\tau}^{\text{Bldg}}|e_t^{\text{Bldg}}) \cdot \mathbb{P}(e_{t+\tau}^{\text{V2G}}|e_t^{\text{V2G}}) \cdot \mathbb{P}(e_{t+\tau}^{\text{PV}}|e_t^{\text{PV}}). \tag{11}$$

Because we assume that *Building demand* (e_t^{Bldg}), *V2G demand* (e_t^{V2G}), and *PV generation* (e_t^{PV}) are unknowns, the transition probability $\mathbb{P}(e_{t+\tau}^{\text{Net}}|e_t^{\text{Net}})$ is not known to the system in time step t in advance. However, because the present study applies the RL technique, it is not necessary to know these transition probabilities. This is especially true for Q-learning, a model-free algorithm, in which the transition probabilities are learned implicitly through the experience gained over successive time steps.

3.5. Reward Function

Let $r(s_t, a_t)$ denote the reward function that returns a cost value indicating how much money the smart energy building must pay for the energy used to operate the building. When action a_t is taken at state s_t, the reward function is defined by

$$r(s_t, a_t) = \begin{cases} e_t^{\text{Net}} \cdot p_t^{\text{Util}} - e_t^{\text{V2G}} \cdot p_t^{\text{V2G}}, & \text{if } a_t = \textit{Buying}, \\ (e_t^{\text{Net}} + \Delta e) \cdot p_t^{\text{Util}} - e_t^{\text{V2G}} \cdot p_t^{\text{V2G}}, & \text{if } a_t = \textit{Charging}, \\ -e_t^{\text{V2G}} \cdot p_t^{\text{V2G}}, & \text{if } a_t = \textit{Discharging}, \\ -\Delta e \cdot p_t^{\text{Util}} - e_t^{\text{V2G}} \cdot p_t^{\text{V2G}}, & \text{if } a_t = \textit{Selling}, \end{cases} \tag{12}$$

where the negative value of $r(s_t,\ a_t)$ implies that the smart energy building earns money, whereas the positive value is a cost that must be paid.

To account for the impact of the current action on future rewards, the total discounted reward at time step t under a given policy π, denoted by $R_t^{(\pi)}$, is defined as the sum of the instant reward at time step t and discounted rewards from the next time step, $t + \tau$, as follows:

$$R_t^{(\pi)} = r(s_t, a_t) + \sum_{i=1}^{\infty} \gamma^i \cdot r(s_{t+i\tau}, a_{t+i\tau}), \tag{13}$$

where $0 \le \gamma \le 1$ denotes the discount factor that determines the importance of future rewards from the next time step, $t + \tau$, to the infinity. For example, $\gamma = 0$ implies that the EMS will consider only the current reward, while $\gamma = 1$ implies that the system weighs both current reward and future long-term rewards equally. The objective of the EMS is to minimize the total discounted reward $R_t^{(\pi)}$ to reduce the operating cost of the smart energy building.

4. Energy Management Algorithm Using Q-Learning

In this section, we propose an RL-technique-based energy management algorithm that minimizes the total discounted reward defined by (13). Among the many types of algorithms included in the RL technique, the Q-learning algorithm was adopted owing to its model-free characteristic, in which transition probabilities can be learned implicitly through experience without any prior knowledge.

Let $Q(s_t, a_t)$, denoting the action-value function that returns the expected total discounted reward when action a_t is taken at state s_t by following a given policy π, be defined as follows:

$$\begin{aligned} Q(s_t, a_t) &= \mathbb{E}\Big\{ R_t^{(\pi)} \Big\} \\ &= \mathbb{E}\Big\{ r(s_t, a_t) + \sum_{i=1}^{\infty} \gamma^i \cdot r(s_{t+i\tau}, a_{t+i\tau}) \Big\}. \end{aligned} \tag{14}$$

The Q-learning algorithm tries to approximate the optimal action-value function Q^*, expressed as

$$Q^*(s, a) = \mathbb{E}\left\{ r(s_t, a_t) + \sum_{i=1}^{\infty} \min_{a_{t+i\tau} \in \mathcal{A}_{s_{t+i\tau}}} \gamma^i \cdot r(s_{t+i\tau}, a_{t+i\tau}) \middle| s_t = s, a_t = a \right\}, \tag{15}$$

by repeatedly updating the action-value function $Q(s_t, a_t)$ through experience of successive time steps.

To approximate the optimal action-value function $Q^*(S, A)$, we must to estimate the values of $Q(s_t, a_t)$ for all state-action pairs. Let $a_t^* (\in \mathcal{A}_{s_t})$ denote the greedy action minimizing the value of Q at state s_t, that is,

$$a_t^* = \arg \min_{a_t \in \mathcal{A}_{s_t}} Q(s_t, a_t), \tag{16}$$

and ε denote a positive small number between 0 and 1 ($0 \le \varepsilon \le 1$). To deal with the exploitation versus exploration tradeoff issue [26], we adopt a ε-greedy policy, where one of the actions from possible action set $\mathcal{A}_{s_t}$ is selected randomly with a probability of $\frac{\varepsilon}{|\mathcal{A}_{s_t}|}$ for exploration, whereas for the majority of the time, the greedy action a_t^* in $\mathcal{A}_{s_t}$ is selected with a probability of $1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}_{s_t}|}$ for exploitation. As a result, the probability of selecting action a_t at state s_t under the policy π, denoted by $\mathbb{P}_\pi(s_t, a_t)$, is represented as follows:

$$\mathbb{P}_{\pi}(s_t, a_t) = \begin{cases} 1 - \varepsilon + \dfrac{\varepsilon}{|\mathcal{A}_{s_t}|}, & \text{if } a_t = a_t^*, \\ \dfrac{\varepsilon}{|\mathcal{A}_{s_t}|}, & \text{if } a_t \neq a_t^*. \end{cases} \tag{17}$$

Once the action a_t is selected by following policy π, the reward function $r(s_t, a_t)$ is calculated using (12) and the state s_t evolves to the next state, $s_{t+\tau}$. Simultaneously, the action-value function $Q(s_t, a_t)$ is updated according to the following rule:

$$Q(s_t, a_t) \leftarrow (1 - \alpha) Q(s_t, a_t) + \alpha \Big[r(s_t, a_t) + \gamma \cdot \min_{a \in \mathcal{A}_{s_{t+\tau}}} Q(s_{t+\tau}, a) \Big], \tag{18}$$

where α is a learning rate that determines how much the newly obtained reward overrides the old value of $Q(s_t, a_t)$. For instance, $\alpha = 0$ implies that the newly obtained information is ignored, whereas $\alpha = 1$ implies that the system considers only the latest information.

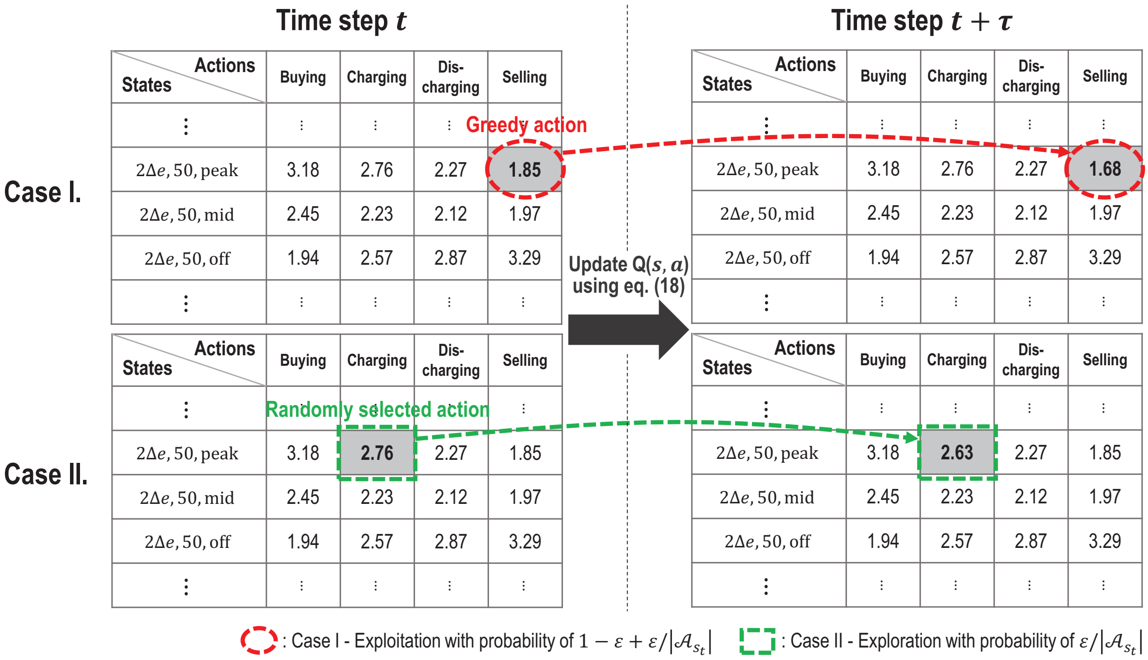
For the initialization problem, the typical Q-learning algorithm simply initializes the action-value function at time step 0, $Q(s_0, a_0)$, with the value of 0 or ∞. However, convergence of the action-value function requires significant time because a large number of time steps is required to explore and update the values of $Q(s_t, a_t)$ for all state-action pairs at least once. To reduce the convergence time of the proposed algorithm, here we suggest that each value of $Q(s_t, a_t)$ is initialized to the total discounted reward $R_0^{(\pi)}$ with $\gamma = 0$, which can be obtained easily as the instant reward at time step 0. That is, the values of $Q(s_0, a_0)$ for all state-action pairs can be explored preliminarily with instant rewards before the actual learning process begins. Through this simple additional initialization step, it is expected that the convergence time is significantly shortened.

Algorithm 1 shows the pseudocode of the main algorithm of the EMS using Q-learning. First, $Q(s, a)$ for all state-action pairs is initialized to the total discounted rewards with $\gamma = 0$ in line 1, and the learning parameters γ, α, and ϵ are initialized in line 2. Lines 3–11 show the loop for each time step t. The possible action set $\mathcal{A}_{s_t}$ satisfying the SoC condition of the ESS is determined in line 4, and the greedy action a_t^* is obtained in line 5. In lines 6–7, one action (a_t) is selected from $\mathcal{A}_{s_t}$, which now includes the greedy action a_t^* obtained in line 5, according to the probability of selecting an action under the policy π, and the potential reward $r(s_t, a_t)$ and the next state $s_{t+\tau}$ that will result from taking the selected action a_t are observed. Based on this observation, $Q(s_t, a_t)$ is updated according to the update rule (18) in line 8; finally, in lines 9–10, the time step t and the state s_t are transited to the next time step, $t + \tau$, and the next state, $s_{t+\tau}$, respectively.

Algorithm 1 Energy management algorithm using Q-learning

```
1:  Initialize $Q(s,a)$, $\forall s \in \mathcal{S}$, $\forall a \in \mathcal{A}$, to total discounted rewards with $\gamma = 0$
2:  Initialize learning parameters $\gamma$, $\alpha$, and $\varepsilon$
3:  for each time step $t$ do
4:      Determine possible action set $\mathcal{A}_{s_t}$ by (6)
5:      Obtain greedy action $a_t^*$ by (16)
6:      Select action $a_t$ from $\mathcal{A}_{s_t}$ by policy $\pi$
7:      Take action $a_t$ and observe $r(s_t, a_t)$, $s_{t+\tau}$
8:      Update $Q(s_t, a_t)$ according to (18)
9:      $t \leftarrow t + \tau$
10:     $s_t \leftarrow s_{t+\tau}$
11: end for
```

Figure 2 presents a simplified example of the Q-table updating procedure. The Q-table is given in the form of a matrix, with each row indicating the state and each column indicating the action. Suppose that the current state at time step t is $[2\Delta e,\ 50,\ \text{peak}]$. Then, *Selling* is supposed to be the greedy action because its Q-value, 1.85, is the minimum among all Q-values in the current state. According to the policy π, the final action is selected from either Case I for exploitation or Case II for exploration. In Case I, the greedy action, *Selling*, is selected with a probability of $1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}_{s_t}|}$. In Case II, an action is randomly chosen among all actions with a probability of $\frac{\varepsilon}{|\mathcal{A}_{s_t}|}$ regardless of Q-value. In both cases, the Q-value, $Q(s, a)$, for the selected action is updated according to (18).


Time step t
Time step $t + \tau$
Case I.
Case II.
Greedy action
Randomly selected action
Update Q(s, a) using eq. (18)
: Case I - Exploitation with probability of $1 - \varepsilon + \varepsilon/|\mathcal{A}_{s_t}|$
: Case II - Exploration with probability of $\varepsilon/|\mathcal{A}_{s_t}|$

States \ Actions	Buying	Charging	Dis-charging	Selling
⋮	⋮	⋮	⋮	
$2\Delta e, 50, \mathrm{peak}$	3.18	2.76	2.27	1.85
$2\Delta e, 50, \mathrm{mid}$	2.45	2.23	2.12	1.97
$2\Delta e, 50, \mathrm{off}$	1.94	2.57	2.87	3.29
⋮	⋮	⋮	⋮	⋮

States \ Actions	Buying	Charging	Dis-charging	Selling
⋮	⋮	⋮	⋮	⋮
$2\Delta e, 50, \mathrm{peak}$	3.18	2.76	2.27	1.68
$2\Delta e, 50, \mathrm{mid}$	2.45	2.23	2.12	1.97
$2\Delta e, 50, \mathrm{off}$	1.94	2.57	2.87	3.29
⋮	⋮	⋮	⋮	⋮

States \ Actions	Buying	Charging	Dis-charging	Selling
⋮				⋮
$2\Delta e, 50, \mathrm{peak}$	3.18	2.76	2.27	1.85
$2\Delta e, 50, \mathrm{mid}$	2.45	2.23	2.12	1.97
$2\Delta e, 50, \mathrm{off}$	1.94	2.57	2.87	3.29
⋮	⋮	⋮	⋮	⋮

States \ Actions	Buying	Charging	Dis-charging	Selling
⋮	⋮	⋮	⋮	⋮
$2\Delta e, 50, \mathrm{peak}$	3.18	2.63	2.27	1.85
$2\Delta e, 50, \mathrm{mid}$	2.45	2.23	2.12	1.97
$2\Delta e, 50, \mathrm{off}$	1.94	2.57	2.87	3.29
⋮	⋮	⋮	⋮	⋮

Figure 2. Simplified example of Q-table updating.

5. Performance Evaluation

To evaluate the performance of the proposed energy management algorithm using Q-learning, we consider a smart energy building in a smart grid environment, which is associated with a utility and DERs, including PV system, ESS system, and V2G station, and perform numerical simulations based on the data measured in real environments. As a simulation framework, MATLAB 2017b is used.

5.1. Simulation Setting

In the simulations, the length of each time step τ was set to 5 min, and the ESS capacity E^{ESS} was set to 500 kWh. The ESS guard ratio was set to $\eta = 0.02$, and the initial SoC of the ESS e_0^{ESS} was set to be 250 kWh. Here, the final SoC of each day is constrained to be e_0^{ESS} with a tolerance of ±10%, i.e., [225 275] kWh. The reason is that if the values of initial SoC vary every day, the initial condition of the learning process can differ from day to day. In this case, more explorations to learn the Q-values may be needed, resulting in longer convergence time. The energy unit for charging the ESS and selling to the utility company was set to $\Delta e = 25$ kWh. We set the learning rate α to 0.1. If α is too large, the values of $Q(s, a)$ may oscillate significantly. On the other hand, if it is too small, it may cause long convergence time of the Q-learning algorithm. The ε-greedy parameter ε was set to 0.2, and the discount factor γ was set to 0.95, as many studies dealing with long-term future rewards in RL typically take γ with the values slightly less than 1 [4,12]. To examine the validity of these learning parameters setting, the performance of the proposed algorithm will be analyzed with respect to ε and γ through simulations in the next subsection (see Figures 10 and 11). The simulation input parameters are summarized in Table 1.

Table 1. Simulation input parameters.

Name	Values
Length of time step	$\tau = 5$ (min)
ESS capacity	$E^{\text{ESS}} = 500$ (kWh)
ESS guard ratio	$\eta = 0.02$
Initial SoC of ESS	$e_0^{\text{ESS}} = 250$ (kWh)
Energy unit for *Charging* and *Selling*	$\Delta e = 25$ (kWh)
Learning rate	$\alpha = 0.1$
Discount factor	$\gamma = 0.95$
ε-greedy parameter	$\varepsilon = 0.2$

The building demand (e_t^{Bldg}) and PV generation (e_t^{PV}) follow the energy demand profile and PV generation profile, respectively, measured in a campus building in GIST during 100 weekdays (from 1 June 2016 to 18 October 2016) [27]. Examples of energy demand profile and PV generation profile for three days measured at intervals of 5 min are presented in Figures 3 and 4. Likewise, Figure 5 shows an example of vehicle parking records for three days in the district office in Gwangju [28]. Based on these records, we simply modeled V2G demand (e_t^{V2G}), in which the vehicles of commuters require charging and those of customers require discharging. Here, we assumed that every charger supports typical level 1 (dis)charging, where the (dis)charging rate is fixed to 7 kW/h (0.583 kW per 5 min). For the utility price (p_t^{Util}) and V2G price (p_t^{V2G}), the ToU energy price tables given by Korea Electric Power Corporation in 2017 were used as follows:

$$p_t^{\text{Util}} = \begin{cases} 0.14 \ (\$/\text{kWh}), & \text{if peak-load period}, \\ 0.08 \ (\$/\text{kWh}), & \text{if mid-load period}, \\ 0.04 \ (\$/\text{kWh}), & \text{if off-peak-load period}, \end{cases} \tag{19}$$

$$p_t^{\text{V2G}} = \begin{cases} 0.11 \ (\$/\text{kWh}), & \text{if peak-load period}, \\ 0.09 \ (\$/\text{kWh}), & \text{if mid-load period}, \\ 0.06 \ (\$/\text{kWh}), & \text{if off-peak-load period}, \end{cases} \tag{20}$$

where the peak-load periods are 10:00–12:00 and 13:00–17:00; mid-load periods are 09:00–10:00, 12:00–13:00, and 17:00–23:00; and off-peak-load period is 23:00–09:00 [29]. As we assumed that e_t^{Bldg}, e_t^{PV}, e_t^{V2G}, p_t^{Util}, and p_t^{V2G} are five unknown information in Section 3.1, the current values of them are measured at each time step t, but their future values are not available to the EMS.

Please note that as an effort to reduce the number of state-action pairs in Q-table, we discretized each element in the state space as follows: e_t^{ESS} into 20 levels, e_t^{Net} into 6 levels, and $\hat{t}$ into 3 levels (peak-load, mid-load, and off-peak-load periods). This discretization of the state space is expected to be effective in shortening the convergence time of the proposed algorithm, cooperating with the Q-table initialization procedure proposed in Section 4.

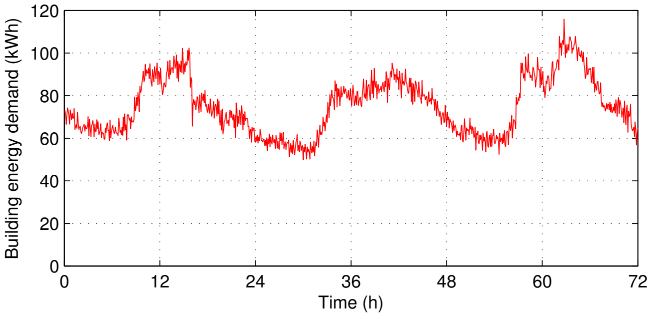


Figure 3. Energy demand profile example for 3 days of campus building in GIST.

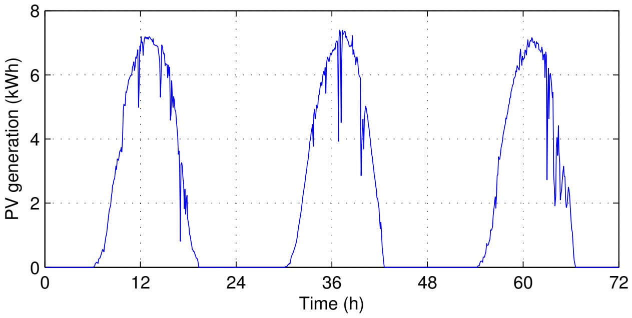


Figure 4. PV generation profile example for 3 days of campus building in GIST.

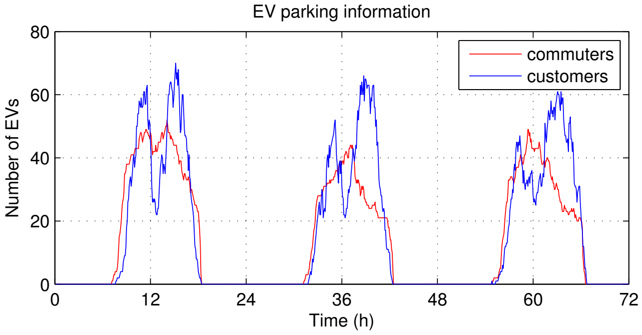


Figure 5. Vehicle parking records of district office in Gwangju.

5.2. Simulation Results

To verify the performance improvements achieved by applying the proposed algorithm, we compare the results obtained using the proposed algorithm to those obtained using three other algorithms, described as follows:

- Minimum instant reward—The system always chooses the action that gives the minimum instant reward $r(s_t, a_t)$ in the present time step t, without considering future rewards. This algorithm is expected to provide similar results as the proposed algorithm with $\gamma = 0$.
- Random action—The action is selected randomly from the possible action set $\mathcal{A}_{s_t}$, regardless of the value of $Q(s_t, a_t)$.
- Previous action maintain—This algorithm tries to always maintain its previous action while keeping the SoC of the ESS (e_t^{ESS}) between $(50 - \beta)\%$ and $(50 + \beta)\%$ of the maximum capacity of the ESS (E^{ESS}), regardless of any other information on the current state. Here, β is set to 20 so that e_t^{ESS} is kept between 30% and 70% of E^{ESS}. For example, if the previous action is *Buying* or *Charging* with e_t^{ESS} between 30% and 70%, the algorithm keeps selecting the *Buying* or *Charging* action until e_t^{ESS} reaches 70%; then, it changes the action to *Selling* or *Discharging*.

- Hourly optimization in [21]—At the beginning of every hour, this algorithm schedules the energy dispatch by using the hourly forecasted profiles. We assume that the hourly profiles of building demand (Figure 3) and PV generation (Figure 4) are given an hour in advance with Normalized Root-Mean-Square Error (NRMSE) of 11.7% and 9.41%, respectively, as in [21].

The primary evaluation is devoted to investigating how the proposed energy management algorithm using Q-learning improves performance as the learning process progresses. As a metric of performance evaluation, daily cost is calculated as the sum of rewards over $\frac{24\text{ h} \times 60\text{ min}}{\tau}$ time steps. The simulation results of daily cost variation versus increasing number of days experienced are shown in Figure 6. The results show that the daily cost obtained using the proposed algorithm is close to that obtained using the previous action maintain algorithm on the 1st day, but it quickly converged (within about 3–5 days) to around $400, which is even lower than the hourly optimization algorithm, with the help of the simple additional initialization step suggested in Section 4. This is because the proposed algorithm selects better actions by using the learning process as it experiences more state-action pairs. It is worth noting that the daily cost obtained using the minimum instant reward algorithm is higher than that calculated using the random action algorithm because the minimum instant reward algorithm tends to always sell energy from the ESS, disregarding expected future rewards, only to minimize the instant reward value. From the overall results, it can be inferred that the higher the utilization of ESS capacity, the higher is the reduction in daily cost.

In Figure 7, we compare the amount of energy bought daily from the utility by four algorithms. At first glance, it may seem strange that the order of magnitude of the amount of energy bought daily from the utility is almost opposite to that of daily cost shown in Figure 6. However, this result is valid because the daily cost can be reduced through the process of buying (selling) more energy from (to) the utility when the utility price is low (high) by utilizing the ESS. Especially, the minimum instant reward algorithm buys the least amount of energy from the utility because it tends to always sell energy for minimizing the instant reward value, while the proposed algorithm intelligently buys and sells the largest amount of energy to reduce the daily cost.

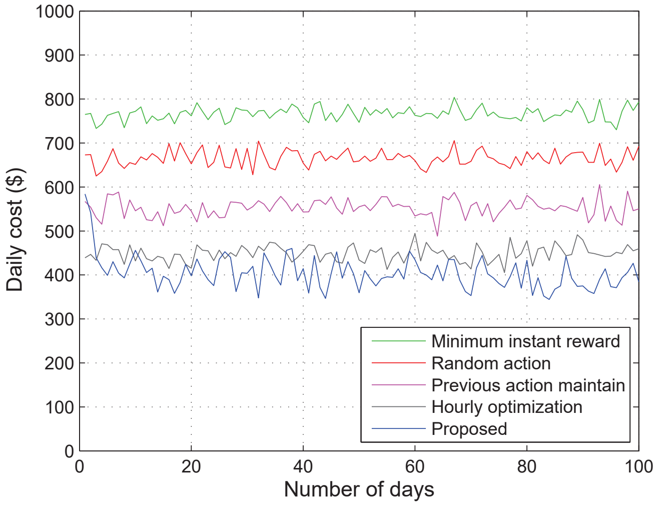


Figure 6. Daily cost comparison with respect to time in days.

To investigate the effect of ESS capacity on the average daily cost, we plotted the average daily cost for 100 days with varying ESS capacity, E^{ESS}, between 0 and 1000 kWh in Figure 8. Here it is assumed that e_0^{ESS} is set to the half of each E^{ESS}. Overall, the proposed algorithm gives the lowest average daily cost, the same as in Figure 6. It can be seen that when E^{ESS} is 0 kWh, all five algorithms provide the same, and highest, average daily costs because only the *Buying* action is possible for all five algorithms. However, as E^{ESS} increases, the average daily costs determined using the random action, previous action maintain, hourly optimization, and proposed algorithms decrease until E^{ESS} reaches 400 kWh because ESS capacity can be utilized to store and release energy by the *Charging*, *Discharging*, and *Selling* actions. For E^{ESS} values larger than 400 kWh, the previous action maintain, hourly optimization, and proposed algorithms show slightly decreasing average daily costs with respect to increasing E^{ESS}, whereas the random action algorithm is not affected by further changes in E^{ESS} at all. This is because the three algorithms can utilize a larger amount of ESS capacity as the value of E^{ESS} increases. However, increasing the ESS capacity generally requires a spike in purchasing and installation costs, with no remarkable performance improvements associated with increasing ESS capacity as shown in the Figure. Therefore, from the perspective of building operations, installing an ESS with a capacity between 400 and 600 kWh would be sufficient to ensure an average daily cost reduction.

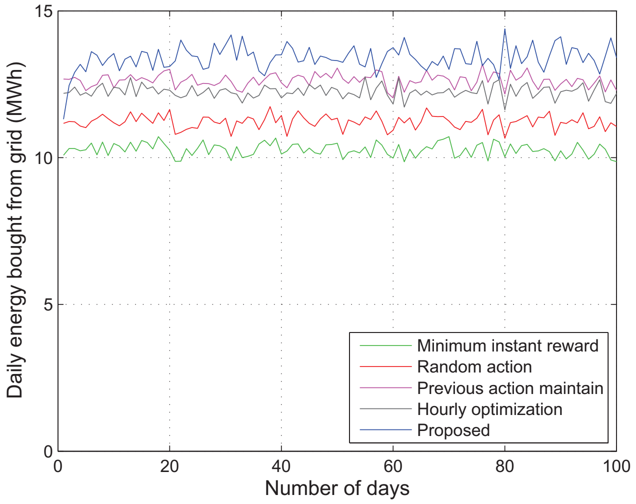


Figure 7. Amount of energy bought daily from utility with respect to time in days.

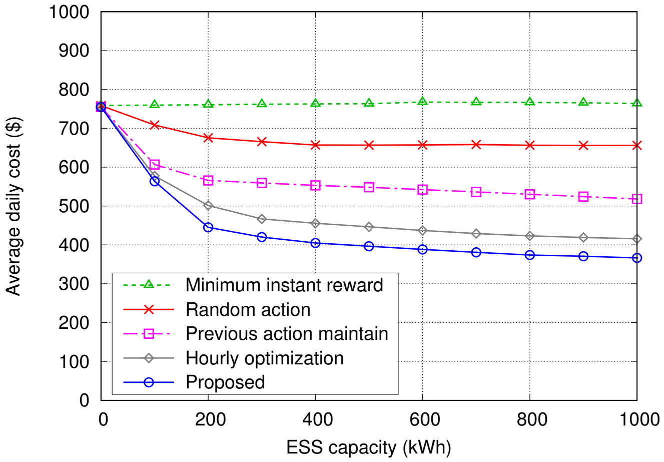


Figure 8. Average daily cost with respect to ESS capacity.

To study the impact of the scale of the PV system, we scaled up and down the PV system by multiplying PV generation (e_t^{PV}) with the scale factor ρ, where $0 \leq \rho \leq 2$. For example, $\rho = 2$ means that the scale of the PV system is doubled, whereas $\rho = 0$ means that the PV system is not associated with the smart energy building at all. Here, E^{ESS} is set to 500 kWh. Figure 9 shows the simulation results of the average daily cost with respect to ρ. Overall, the average daily costs decrease according to increasing ρ. Especially, the proposed algorithm exhibits the lowest daily cost for any value of

ρ, and its rate of decrease becomes slightly larger compared to those of the other four algorithms as ρ increases. This indicates that the proposed algorithm can reduce the daily cost by managing energy more intelligently with learning capability as the amount of energy generated by the PV system increases. Therefore, unless the installation cost of the PV system is taken into consideration, the larger the scale of the PV system, the more effective it would be for reducing the average daily cost of building operation.

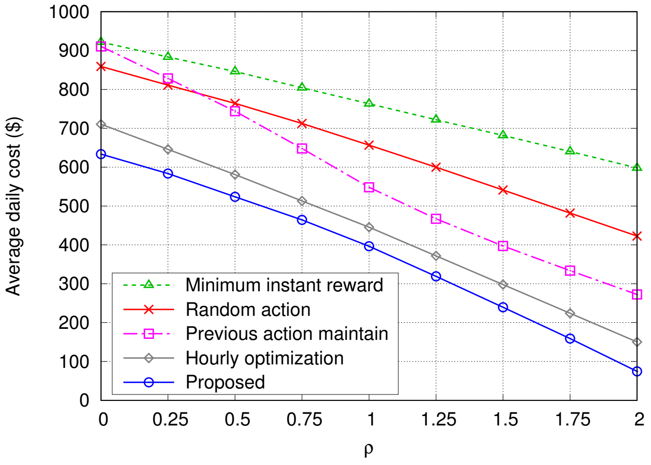


Figure 9. Average daily cost with respect to scale factor of PV system.

In Figures 10 and 11, we plotted the average daily cost variation of the proposed algorithm according to ε and γ, respectively, in order to examine the impact of these learning parameters on the performance. The costs of the four other algorithms are plotted for comparison. In Figure 10, the lowest average daily cost is achieved for $\varepsilon = 0.2$, and when $\varepsilon = 1$, almost the same cost as the random action algorithm is achieved because the action is always selected randomly. In Figure 11, the average daily cost gradually decreases as γ increases, and it becomes lower than the cost by the hourly optimization algorithm when $\gamma \geq 0.8$. Also please note that when $\gamma = 0$, the average daily cost of the proposed algorithm is the same as the minimum instant reward algorithm because the total discounted reward $R_t^{(\pi)}$ is composed of only the term of instant reward. These results imply that it was appropriate to set the learning parameters ε and γ to 0.2 and 0.95, respectively.

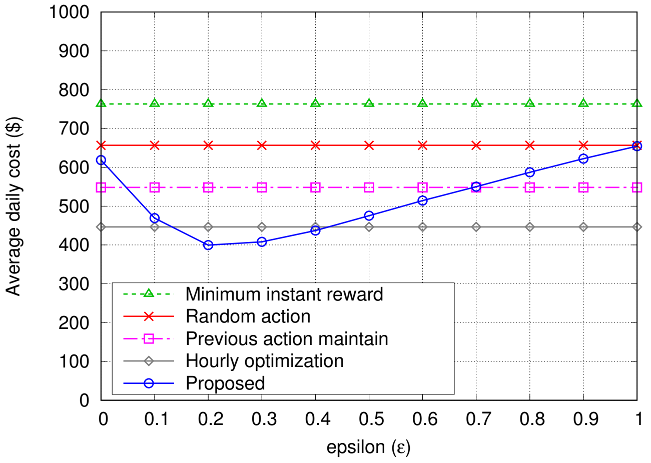


Figure 10. Average daily cost variation with respect to ε-greedy parameter.

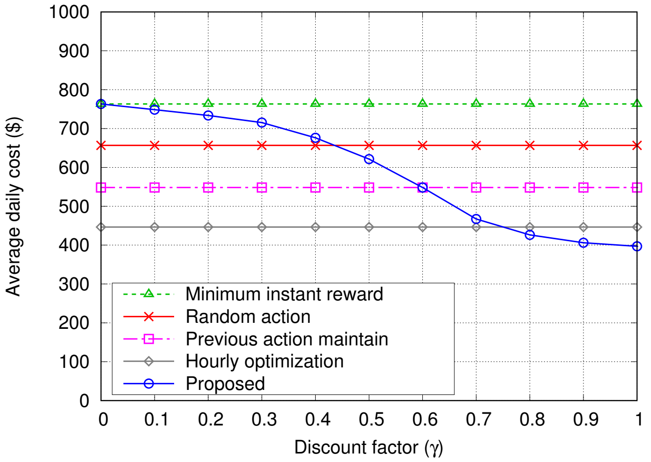


Figure 11. Average daily cost variation with respect to discount factor γ.

Finally, we investigated the performance in a case of the real-time energy pricing. Figure 12 shows an example of the real-time hourly energy price profiles for the same three days of Figures 3–5 in three U.S. states [30], and the average daily cost under the real-time pricing is presented in Figure 13. We simply assume that both p_t^{Util} and p_t^{V2G} are the same as shown in Figure 12. Figure 13 shows that the proposed algorithm provides the lowest cost among the five algorithms for all the three real-time energy prices. Therefore, it can be concluded that the proposed algorithm is significantly effective for energy cost reduction regardless of the type of energy pricing policy.

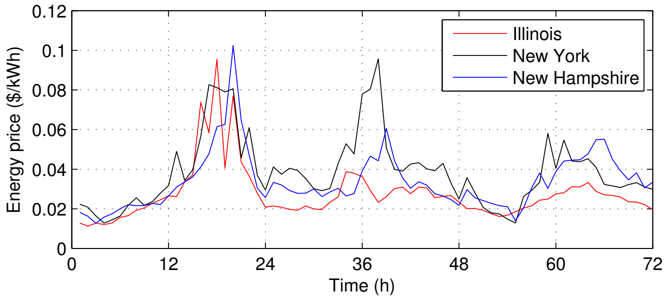


Figure 12. Real-time hourly energy price example for 3 days in 3 U.S. states.

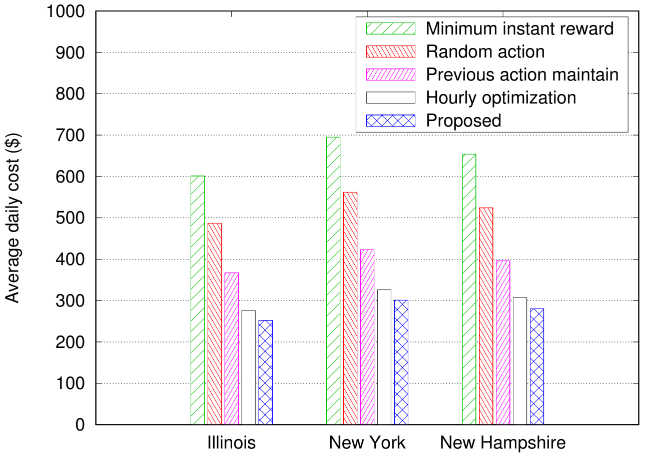


Figure 13. Average daily cost with real-time energy prices.

6. Conclusions and Discussion

In this study, we proposed an RL based energy management algorithm for smart energy buildings in a smart grid environment. Smart energy buildings are capable of exchanging energy with an external grid and DERs such as a PV, an ESS, and a V2G station in real-time. We first developed the energy management system model by using a Markov decision process that completely describes the state space, action space, transition probability, and reward function. To reduce the energy costs of a building given unknown future information about the amount of building load demand, V2G station load demand, and energy generation by PV system, a Q-learning-based energy management algorithm that identifies better energy dispatch actions by learning through experience without prior knowledge was proposed. Through numerical simulations based on data measured in the real environments, we verified that the proposed algorithm significantly reduces energy costs compared to the random and other existing algorithms. We showed that the proposed algorithm successfully reduces energy costs under widely-used energy pricing policies of ToU and real-time. It is expected that the proposed learning-based energy management algorithm is applicable in various smart grid environments such as residential microgrids and smart energy factories under different energy pricing policies. As future work, we will extend the proposed algorithm so that it can work in more complicated cases with additional energy components such as CHP, TCLs, or wind turbines, and empirically validate the proposed algorithm in real smart buildings.

Author Contributions: S.K. designed the algorithm, performed the simulations, and prepared the manuscript as the first author. H.L. led the project and research and advised on the whole process of manuscript preparation. All authors discussed the simulation results and approved the publication.

Acknowledgments: This work was partially supported by GIST Research Institute (GRI) grant funded by the GIST in 2018, by the KETEP grant funded by MOTIE (No. 20171210200810), and by IITP grant funded by MSIT (No. 2014-0-00065, Resilient CPS Research), Republic of Korea.

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

References

1. Hashmi, M.; Hänninen, S.; Mäki, K. Survey of smart grid concepts, architectures, and technological demonstrations worldwide. In Proceedings of the IEEE PES Conference on Innovative Smart Grid Technologies (ISGT), Medellin, Colombia, 19–21 October 2011.
2. Palensky, P.; Dietrich, D. Demand side management: Demand response, intelligent energy systems, and smart loads. *IEEE Trans. Ind. Inform.* **2011**, *7*, 381–388. [CrossRef]
3. Kara, E.C.; Berges, M.; Krogh, B.; Kar, S. Using smart devices for system-level management and control in the smart grid: A reinforcement learning framework. In Proceedings of the IEEE International Conference on Smart Grid Communications (SmartGridComm), Tainan, Taiwan, 5–8 November 2012.
4. Ruelens, F.; Claessens, B.J.; Vandael, S.; De Schutter, B.; Babuška, R.; Belmans, R. Residential demand response of thermostatically controlled loads using batch reinforcement learning. *IEEE Trans. Smart Grid* **2016**, *8*, 2149–2159. [CrossRef]
5. Ju, C.; Wang, P.; Goel, L.; Xu, Y. A two-layer energy management system for microgrids with hybrid energy storage considering degradation costs. *IEEE Trans. Smart Grid* **2017**. [CrossRef]
6. Kuznetsova, E.; Li, Y.F.; Ruiz, C.; Zio, E.; Ault, G.; Bell, K. Reinforcement learning for microgrid energy management. *Energy* **2013**, *59*, 133–146. [CrossRef]
7. Choi, J.; Shin, Y.; Choi, M.; Park, W.K.; Lee, I.W. Robust control of a microgrid energy storage system using various approaches. *IEEE Trans. Smart Grid* **2018**. [CrossRef]
8. Farzin, H.; Fotuhi-Firuzabad, M.; Moeini-Aghtaie, M. A stochastic multi-objective framework for optimal scheduling of energy storage systems in microgrids. *IEEE Trans. Smart Grid* **2017**, *8*, 117–127. [CrossRef]
9. He, Y.; Venkatesh, B.; Guan, L. Optimal scheduling for charging and discharging of electric vehicles. *IEEE Trans. Smart Grid* **2012**, *3*, 1095–1105. [CrossRef]
10. Honarmand, M.; Zakariazadeh, A.; Jadid, S. Optimal scheduling of electric vehicles in an intelligent parking lot considering vehicle-to-grid concept and battery condition. *Energy* **2014**, *65*, 572–579. [CrossRef]

11. Shi, W.; Wong, V.W. Real-time vehicle-to-grid control algorithm under price uncertainty. In Proceedings of the IEEE International Conference on Smart Grid Communications (SmartGridComm), Brussels, Belgium, 17–20 October 2011.
12. Di Giorgio, A.; Liberati, F.; Pietrabissa, A. On-board stochastic control of electric vehicle recharging. In Proceedings of the IEEE Conference on Decision and Control (CDC), Florence, Italy, 10–13 December 2013.
13. U.S. Energy Information Administration (EIA). International Energy Outlook 2016. Available online: https://www.eia.gov/outlooks/ieo/buildings.php (accessed on 3 March 2018).
14. Zhao, P.; Suryanarayanan, S.; Simões, M.G. An energy management system for building structures using a multi-agent decision-making control methodology. *IEEE Trans. Ind. Appl.* **2013**, *49*, 322–330. [CrossRef]
15. Wang, Z.; Yang, R.; Wang, L. Multi-agent control system with intelligent optimization for smart and energy-efficient buildings. In Proceedings of the IEEE Conference on Industrial Electronics Society (IECON), Glendale, AZ, USA, 7–10 November 2010.
16. Wang, Z.; Wang, L.; Dounis, A.I.; Yang, R. Integration of plug-in hybrid electric vehicles into energy and comfort management for smart building. *Energy Build.* **2012**, *47*, 260–266. [CrossRef]
17. Missaoui, R.; Joumaa, H.; Ploix, S.; Bacha, S. Managing energy smart homes according to energy prices: Analysis of a building energy management system. *Energy Build.* **2014**, *71*, 155–167. [CrossRef]
18. Basit, A.; Sidhu, G.A.S.; Mahmood, A.; Gao, F. Efficient and autonomous energy management techniques for the future smart homes. *IEEE Trans. Smart Grid* **2017**, *8*, 917–926. [CrossRef]
19. Wang, F.; Zhou, L.; Ren, H.; Liu, X.; Talari, S.; Shafie-khah, M.; Catalão, J.P. Multi-objective optimization model of source-load-storage synergetic dispatch for a building energy management system based on TOU price demand response. *IEEE Trans. Ind. Appl.* **2018**, *54*, 1017–1028. [CrossRef]
20. Yan, Q.; Zhang, B.; Kezunovic, M. Optimized operational cost reduction for an EV charging station integrated with battery energy storage and PV generation. *IEEE Trans. Smart Grid* **2018**. [CrossRef]
21. Di Piazza, M.; La Tona, G.; Luna, M.; Di Piazza, A. A two-stage energy management system for smart buildings reducing the impact of demand uncertainty. *Energy Build.* **2017**, *139*, 1–9. [CrossRef]
22. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
23. Manandhar, U.; Tummuru, N.R.; Kollimalla, S.K.; Ukil, A.; Beng, G.H.; Chaudhari, K. Validation of faster joint control strategy for battery-and supercapacitor-based energy storage system. *IEEE Trans. Ind. Electron.* **2018**, *65*, 3286–3295. [CrossRef]
24. Burger, S.P.; Luke, M. Business models for distributed energy resources: A review and empirical analysis. *Energy Policy* **2017**, *109*, 230–248. [CrossRef]
25. Mao, T.; Lau, W.H.; Shum, C.; Chung, H.S.H.; Tsang, K.F.; Tse, N.C.F. A regulation policy of EV discharging price for demand scheduling. *IEEE Trans. Power Syst.* **2018**, *33*, 1275–1288. [CrossRef]
26. Kearns, M.; Singh, S. Near-optimal reinforcement learning in polynomial time. *Mach. Learn.* **2002**, *49*, 209–232. [CrossRef]
27. Research Institute for Solar and Sustainable Energies (RISE). Available online: https://rise.gist.ac.kr/ (accessed on 15 February 2018).
28. Gwangju Buk-Gu Office. Available online: http://eng.bukgu.gwangju.kr/index.jsp (accessed on 12 March 2018).
29. Energy Price Table by Korea Electric Power Corporation (KEPCO). Available online: http://cyber.kepco.co.kr/ckepco/front/jsp/CY/E/E/CYEEHP00203.jsp (accessed on 17 March 2018).
30. ISO New England. Available online: https://www.iso-ne.com/ (accessed on 28 March 2018).