Cooperative Management for PV/ESS-Enabled Electric Vehicle Charging Stations: A Multiagent Deep Reinforcement Learning Approach

MyungJae Shin, Dae-Hyun Choi, *Member, IEEE*, and Joongheon Kim, *Senior Member, IEEE*

***Abstract*—This article proposes a novel multiagent deep reinforcement learning method for the energy management of distributed electric vehicle charging stations with a solar photovoltaic system and energy storage system. In the literature, the conventional method is to calculate the optimal electric vehicle charging schedule in a centralized manner. However, in general, the centralized approach is not realistic under certain environments where the system operators for multiple electric vehicle charging stations handle dynamically varying data, such as the status of the energy storage system and electric vehicle-related information. Therefore, this article proposes a method that can compute the scheduling solutions of multiple electric vehicle charging stations in a distributed manner while handling run-time time-varying dynamic data. As shown in the data-intensive performance evaluation, it can be observed that the proposed method achieves a desirable performance in terms of reducing the operation costs of electric vehicle charging stations.**

***Index Terms*—Electric vehicles, multi-agent systems, neural networks, scheduling algorithms.**

I. Introduction

RECENTLY, energy-efficient technologies for enabling smart city applications have been receiving considerable attention, especially electric vehicles (EVs) and their related researches. In order to facilitate EV deployment in the real world, it is important to establish and distribute corresponding infrastructures such as EV charging stations (EVCSs). Based on this trend, it is obvious that guaranteeing the sustainability of EVCSs energy/power supply is essential. In order to ensure the sustainability, it is important to develop technologies that supply energy/power consistently, e.g., photovoltaic (PV) power generation and utilization of energy storage systems (ESS).

It is essential to conduct joint optimization of the energy consumption and the operation cost of EVCS based on the knowledge regarding the behavior of suppliers/consumers. A conventional technology in EVCS management is the optimization method for optimal operation and EVCS infrastructure planning. Recently, a considerable amount of literature has been published on the optimization of charging station operation in EVCS infrastructure [1]–[6], including the chance-constrained optimization-based operation cost reduction for a single CS with PV/ESS considering the uncertainties (e.g., PV generation and charging demand from CS) [1], optimal scheduling for multiple CSs without PV/ESS based on distributed model predictive control [2] and convex relaxation technique [3], the minimization of operational cost and grid power for multiple CSs with PV/ESS using hybrid optimization with a rule-based deterministic approach [4], hierarchical framework considering EV and CS status information [5], and calculation of the adequate charging price using stochastic dynamic programming and greedy algorithms [6]. For the efficient planning of EVCSs, the optimal placement policy for EVCSs was proposed, in which the competition among CSs is formulated as a Bayesian game, which determines the locations of EVCSs in a distributed optimization manner [7]. A new coordinated planning method was presented considering both the transportation network and the electric distribution network [8]. More recently, a multistage stochastic expansion planning framework for the distribution system and EVCSs was presented to determine the optimal locations of EVCSs, substation, and feeders along with the PV and capacitor banks [9].

Even though these previous research results show good performance in terms of their own objectives, the solution approaches are all centralized, i.e., the solutions are computed in a centralized optimization problem solving. In this case, it is obvious that online real-time solution computation is not possible for massive dynamic time-varying data processing. As a result, the centralized approaches are hard to be used in real world. To solve the given problem in a distributed manner, machine learning based approaches are widely used. To handle the data (e.g., ESS related data, EV load, PV generation, etc.) with high uncertainty and dynamic updates, a new multiagent

Manuscript received May 31, 2019; revised September 1, 2019; accepted September 24, 2019. Date of publication September 27, 2019; date of current version February 6, 2020. This work was supported in part by the National Research Foundation of Korea under Grant 2019R1A2C4070663 and in part by the Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning Grant funded by the Korea Government Ministry of Trade, Industry and Energy under Grant 20184030202070. Paper no. TII-19-2218. *(Corresponding authors: Dae-Hyun Choi; Joongheon Kim.)*

M. Shin is with the School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea (e-mail: mjshin.cau@gmail.com).

D.-H. Choi is with the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, South Korea (e-mail: dhchoi@cau.ac.kr).

J. Kim is with the School of Electrical Engineering, Korea University, Seoul 02841, South Korea (e-mail: joongheon@korea.ac.kr).

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

Digital Object Identifier 10.1109/TII.2019.2944183

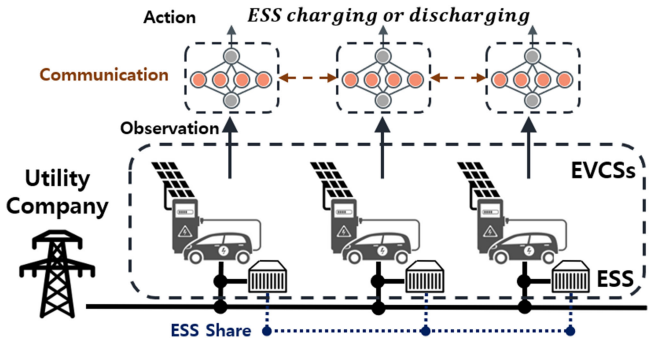

Fig. 1. Communication based multicharging station systems.

(for distributed computation over EVCSs) deep reinforcement learning algorithm is designed in this article, considering the ESS and PV power generation for EVCSs.

Fig. 1 shows an overall architecture for multicharging station systems. A single private enterprise manages multiple EVCSs equipped with renewable energy resources (e.g., solar power). Multiple EVCSs are independently connected to the ESSs. Each EVCS manages the charging and discharging of the connected ESS. Then, EVCSs preferentially consume the energy of the connected ESS, and EVCSs can use the energy stored in the ESSs of the other connected EVCSs when needed. At this time, EVCSs only share the surplus energy remaining after meeting the own net demand. EVCSs cannot only meet their net demand, but also manage the surplus energy to reduce overall operating costs. Therefore, it is important to cooperatively manage the charging and discharging of the energy stored in the ESSs.

In recent years, compared to the aforementioned model-based EVCS optimization approaches, data-driven approaches using machine learning methods have gained popularity owing to their more efficient energy management for CSs. This is because the existing model-based approach is limited to deterministic decision-making under an uncertain environment and approximated system models, thereby leading to undesired EV charging scheduling. Deep reinforcement learning for residential EV charging management was proposed where the randomness in the commuting behavior and the electricity price was considered [10], [11]. Using state-action-reward-state-action RL algorithm, the optimal pricing and scheduling strategy that maximizes the long-term profit of CS was investigated [12]. In [13], a group of EVCSs were coordinated to learn an optimal charging policy using fitted Q-iteration-based batch RL algorithm while considering individual EV charging characteristics. Multiple agent framework based on the multistep Q-learning was proposed to model the charging loads of plug-in electric taxis [14]. The simulation results showed that the proposed approach outperforms the conventional Q-learning method in terms of the convergence speed and an amount of reward. The control strategy for EVs using Q-learning was developed to quickly respond the frequency regulation signals from the grid operators, which enables the aggregator to optimally allocate the regulation power [15]. In [16], batch reinforcement learning integrated with the prediction of the electricity prices using a Bayesian neural network was presented to reduce the EV user charging cost.

TABLE I
SYSTEM DESCRIPTION PARAMETERS

Parameter	Description
N	Total number of EVCSs
Z	Total number of utility company
$\mathcal{C}$	The set of EVCSs
$\mathcal{P}$	The set of PV system
$\mathcal{E}$	The set of ESS
$\mathcal{L}$	The set of internal load of EVCS
$\mathcal{V}$	The set of PV generation
$\mathcal{H}$	The set of EV demands
$\mathcal{O}$	The set of the amount of energy charged in ESS
$\mathcal{J}$	The set of the amount of energy required to be purchased
$\mathcal{U}$	The set of residual energy in the ESS
$\mathcal{W}$	The set of the amount of energy required to be supplied
$\mathcal{Q}^n$	The set of shared energy at nth EVCS
$\mathcal{D}$	The set of total demands
$\mathcal{F}$	The set of net demand
$E^{\max}$	The set of the maximum capacity of ESS
$E_g^{\min}$	The set of the safe minimum capacity of ESS
$E_g^{\max}$	The set of the safe maximum capacity of ESS
η	A guard ratio of ESS
σ	A coefficient for the importance of $\mathcal{O}$
$\mathcal{S}$	The set of observation states
$\mathcal{A}$	The set of actions
$\mathcal{R}_{total}$	The set of total reward of the EVCSs
$\mathcal{R}_p$	The set of pay rewards
$\mathcal{R}_b$	The set of benefit rewards
$\mathcal{R}_s$	The set of benefit rewards by sharing
$\mathcal{R}_o$	The set of overcharge rewards

The main contributions of this article are as follows.

1) We propose a new multiagent deep reinforcement learning (MADRL) algorithms to achieve cooperation in a distributed charging station context for conceiving an algorithm to learn the electrical patterns and to optimize the energy consumption and operation cost of the CSs.
2) To the best of our knowledge, the energy management optimization of multiple charging stations (under the consideration of PV generation and ESS related dynamic time-varying data) via MADRL has not been studied yet; therefore the proposed scheme will be a new guideline for dynamic PV/ESS related studies in future.
3) Through the performance evaluation of the proposed *CommNet* based energy management scheme, we show that the proposed method is able to successfully perform cooperative energy management of distributed multiple charging stations with PV/ESS related features.

II. Deep Reinforcement Learning

This section presents conventional and modern reinforcement learning algorithms and their extensions to multiagent systems. Then, the limitations of the predecessors in terms of learning in distributed multiagent cooperation are presented.

A. Preliminaries

The Markov Decision Process (MDP) can be defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R, T)$,[1] where $\mathcal{S}$ is the finite set of all valid states, and $\mathcal{A}$ represents the finite set of all valid actions. The

[1]The main notations used in this article are summarized in Table I.

function $P : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$ is the transition probability function, with $P(s' \mid s, a)$ being the probability of transitioning into state s' if an agent starts executing action a in state s. The function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function, with $r_t = R(s_t, a_t, s_{t+1})$. The MDP has a finite time horizon T. Solving an MDP means finding a policy $\pi_\theta \in \Pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, where π is parameterized with θ (e.g., the weights and biases of a neural network). The policy π_θ specifies which action $a \in \mathcal{A}$ must be executed in each s for maximizing the discounted cumulative rewards received during finite time T.

When the environment transitions and the policy are stochastic, the probability of a T-step trajectory is defined as $P(\tau \mid \pi_\theta) = \rho(s_0) \prod_{t=0}^{T-1} P(s_{t+1} \mid s_t, a_t)\pi_\theta(a_t \mid s_t)$ where ρ is the starting state distribution. Then, the expected return $\mathcal{J}(\pi_\theta)$ is defined as $\mathcal{J}(\pi_\theta) = \int_\tau P(\tau \mid \pi_\theta)R(\tau) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$ where trajectory τ is a sequence of states and actions in the environment. The goal in reinforcement learning is to learn a policy that maximizes the expected return $\mathcal{J}(\pi_\theta)$ when the agent acts according to the policy π_θ. Therefore, the optimization objective can then be expressed by

$$\pi_\theta^* = \arg\max_\theta \mathcal{J}(\pi_\theta) \tag{1}$$

with π_θ^* being the optimal policy. The multiagent MDP (MMDP) [17] generalizes the MDP to the multiagent system, where the state space is defined by taking the Cartesian product of the state spaces of the individual agents, and actions represent the joint actions that can be executed by the agents. Because the MMDP is a regular MDP, the optimization problem can be solved using the same solutions.

B. Deep Reinforcement Learning

1) Deep Q-Network (DQN) (see [18]): A DQN is a model free reinforcement learning method, designed to learn the optimal policy with a high-dimensional state space. The DQN is inspired by Q-learning[19]. A neural network is used to approximate the Q-function. Experience replay $\mathcal{D}$ and target network are two key features used to stabilize the optimization. Experiences $e_t = (s_t, a_t, r_{t+1}, s_{t+1})$ of the agent are stored in the experience buffer $\mathcal{D} = (e_1, e_2, \ldots, e_T)$, and are periodically resampled to train the Q-networks. A mini-batch resampled experience is used to update the parameters θ_i of the policy with the loss function at the ith training iteration where the loss function is defined as

$$L(\theta_i) = \mathbb{E}\left[(r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a'; \theta_i^-) - Q(s_t, a_t; \theta_i))^2\right]$$

where θ_i^- are the target network parameters. The target network parameters θ_i^- are updated using the Q-network parameters θ in every predefined step. The stochastic gradient descent method is used to optimize the loss function.

2) Proximal Policy Optimization (PPO) (see [20]): PPO is one of the breakthroughs of deep reinforcement learning algorithms, which adopts the trust region concept[21] to improve training stability by ensuring that π_θ updates at every iteration are small by clipping the probability ratio $r_\pi(\theta) = \pi_\theta(a \mid s)/\pi_{\theta_{\text{old}}}(a \mid s)$. A clipped surrogate objective to prevent the new policy from straying away from the old one is used to train the policy π_θ. The clipped objective function is as follows:

$$L_t^{\text{CLIP}}(\theta) = \min(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t) \tag{2}$$

where θ is the parameter of the new policy while θ_{old} is that of the old policy, A_t is the estimated advantage value under hyperparameter $\epsilon < 1$, which means how far away the new policy is allowed to update from the old policy. PPO uses the stochastic gradient descent to maximize the objective (2).

3) Limitations of the Predecessors: The predecessors are designed with a single agent system. In the system, the agent considers only the changes that are the outcomes of its own actions. However, in a multiagent system, the agent needs to concurrently observe the effect of its own actions as well as the behavior of other agents. This characteristic of the multiagent system constantly reshapes the environment and leads to nonstationarity (i.e., it becomes a nonstationary problem). As a result, the convergence theory of the predecessors is generally not guaranteed in a multiagent system [22]. Therefore, the information collecting and processing method should not affect the convergence stability of the agents in multiagent systems.

III. Cooperative Energy Management

A. System Description

Suppose there exists a single private enterprise that manages N EVCSs equipped with renewable energy resources (e.g., solar power). The set of EVCSs is denoted as $\mathcal{C} = \{c_0, c_1, \ldots, c_n, \ldots, c_{N-1}\}$. Here, c_n represents the nth EVCS, where $\forall c_n \in \mathcal{C}, n \in [0, N)$. We assume that every c_n is associated with only one utility company. However, the utility company can be related to multiple EVCSs. A set of utility companies is denoted as $\mathcal{Y} = \{y_0, y_1, \ldots, y_z, \ldots, y_{Z-1}\}$. Here, y_z represents the zth utility company, where Z is the number of the utility company. The utility company trades energy following the requests by EVCSs in real-time at the prices determined by the utility company. Each EVCS has an energy management system (EMS) to schedule energy flows from and to the utility company, EV charging, and ESS charging and discharging in real-time. The EMS of EVCS aims to reduce the operational costs under unknown future information. In this article, we assume that the EMS can only manage charging and discharging of directly connected ESS. In addition, it is assumed that the EVCSs managed by a single private enterprise can share the energy stored in the ESS.

There exists a set of PV systems $\mathcal{P} = \{p_0, p_1, \ldots, p_n, \ldots, p_{N-1}\}$, and each p_n is equipped with EVCS c_n such that $n \in [0, N)$. Each PV system p_n of c_n converts sunlight into energy by means of PV panels. The PV generation is used to help meet the demands (e.g., load demands and EV demands) at the beginning of every time step. Furthermore, $\mathcal{E} = \{e_0, e_1, \ldots, e_n, \ldots, e_{N-1}\}$ represents the set of ESSs with which each EVCS is connected. Each ESS has a spatial maximum capacity of $E_n^{\max}$. The ESS is capable of storing and releasing energy as needed. As mentioned before, the energy stored in the ESS can be shared with the other connected EVCSs. We assumed that the shared energy would not be sold.

We also assume that the utility price is determined by the utility company in the form of a time-of-use (ToU); and therefore the prices are predetermined and constant.

In the charging management problem, the time step of EMS reaches infinity over time. However, the time step is considered to be repeated with a period of one day as $t = t \mod h \times m$, where h and m represent the number of hours per day (i.e., 24) and the number of minutes per hour (i.e., 60). The problem is studied in a finite time horizon T with equal length time slots $t = 0, 1, 2, \ldots, h \cdot m$. In time slot t, the energy demand required by the EVCS itself due to its internal energy consumption at each time is denoted by $\mathcal{L} = \{l_t^0, l_t^1, \ldots, l_t^n, \ldots, l_t^{N-1}\}$. Here, l_t^n represents the internal load of the nth EVCS at time step t. A set of PV generation is denoted as $\mathcal{V} = \{v_t^0, v_t^1, \ldots, v_t^n, \ldots, v_t^{N-1}\}$. Here, v_t^n denotes the amount of energy generated from the PV system p_n at time step t. In addition, there exists a set of energy demand of the vehicles, denoted as $\mathcal{H} = \{h_t^0, h_t^1, \ldots, h_t^n, \ldots, h_t^{N-1}\}$, where h_t^n is the energy demand of the vehicles in the nth EVCS at t.

1) State Space: The state space of the energy management scheme consists of the ESS state, the price state, and the load state. The set of the amount of energy charged in the ESS at time step t is denoted as $\mathcal{O} = \{o_t^0, o_t^1, \ldots, o_t^n, \ldots, o_t^{N-1}\}$. Here, o_t^n represents the amount of energy charged in e_n connected to the nth EVCS c_n at t. For safe use of the ESS, a guard ratio η is considered as $\eta \cdot E^{\max} \leq o_t^n \leq (1-\eta) \cdot E^{\max} \forall o_t^n \in \mathcal{O}$ where $\eta \in [0, 0.5)$ [23].

We define the set of total demand of the EVCSs at time step t as $\mathcal{D} = \{d_t^0, d_t^1, \ldots, d_t^n, \ldots, d_t^{N-1}\}$, where d_t^n represents the sum of energy demands from the internal load l_t^n and vehicle demand h_t^n at the EVCS c_n as $d_t^n = l_t^n + h_t^n$.

Taking all the above into consideration, the set of states at time step t is denoted as $\mathcal{S} = \{s_t^0, s_t^1, \ldots, s_t^n, \ldots, s_t^{N-1}\}$. The state of the EVCS c_n at time step t, denoted by s_t^n, is as $s_t^n = [o_t^n, E_{g,n}^{\min}, E_{g,n}^{\max}, price_t, price_t^{\text{avg}}, v_t^n, d_t^n]$ where $E_{g,n}^{\min} \in E_g^{\min}$ represents the safeguard minimum capacity of the ESS and $E_{g,n}^{\max} \in E_g^{\max}$ is the safeguard maximum capacity of the ESS at the nth EVCS. The set of safeguard maximum capacities is denoted as $E_g^{\max} = \{E_{g,0}^{\max}, E_{g,1}^{\max}, \ldots, E_{g,n}^{\max}, \ldots, E_{g,N-1}^{\max}\}$, and the set of safeguard minimum capacities is denoted as $E_g^{\min} = \{E_{g,0}^{\min}, E_{g,1}^{\min}, \ldots, E_{g,n}^{\min}, \ldots, E_{g,N-1}^{\min}\}$. Here, $price_t$ represents the price of energy at time t, and $price_t^{\text{avg}}$ represents the average price of energy over the last 24 h. This lets EVCSs know the current price is cheap enough to buy energy.

It should be noted that, the EVCS can observe only the past and current pieces of information, whereas future information is not observable. In addition, it is assumed that the observation from other EVCSs is not needed. This makes the EMS learn how to manage energy using the partially observable information and past experiences under unknown future states.

2) Action Space: To satisfy net demand in each time step t, the EVCS c_n buys the amount of net demand from the utility company. The set of net demand energies at time step t is denoted as $\mathcal{F} = \{f_t^0, f_t^1, \ldots, f_t^n, \ldots, f_t^{N-1}\}$. We define the net demand f_t^n as the total demand d_t^n minus PV generation v_t^n as $f_t^n = \max(d_t^n - v_t^n, 0)$ where f_t^n represents the net demand of the nth EVCS at time step t.

The EVCS c_n first strives to meet the net demand f_t^n through the energy stored in the ESS e_n it manages. Then, the amount of energy that cannot be met by the stored energy o_t^n is supplied by cooperation. It is defined as the residual energy w_t^n which can be calculated by $w_t^n = \max(f_t^n - (o_t^n - E_{g,n}^{\min}), 0)$ where $w_t^n \in \mathcal{W} = \{w_t^0, w_t^1, \ldots, w_t^n, \ldots, w_t^{N-1}\}$.

When the residual energy w_t^n is not 0, the EVCS c_n requests the cooperation to meet the residual energy. The required energy w_t^n is supplied by the ESSs connected to the other EVCSs. Then, the amount of energy of the ESS that can be supplied to the other EVCSs is defined as follows:

$$u_t^n = \max(v_t^n - d_t^n, 0) + \max((o_t^n - E_{g,n}^{\min}) - f_t^n, 0). \quad (3)$$

The set of residual energy in the ESS at time step t is defined as $\mathcal{U} = \{u_t^0, u_t^1, \ldots, u_t^n, \ldots, u_t^{N-1}\}$.

We assume that cooperative EVCSs use the maximum available energy in its own ESS to meet the net demand of the other EVCSs. Therefore, the amount of scarce energy, denoted by j_t^n, has to be purchased from the utility company. Then, the amount of energy that need to be purchased is defined as $j_t^n = \max(w_t^n - \sum_{n' \neq n} u_t^{n'}), 0)$ where $j_t^n \in \mathcal{J}$ when the set of the amount of energy that need to be purchased is denoted as $\mathcal{J} = \{j_0, j_1, \ldots, j_n, \ldots, j_{N-1}\}$.

This article assumes that the nearest EVCS first shares the energy, e.g., c_2 is the nearest EVCS of EVCS c_1. Then, j_t^1 is 100 and u_t^2 is 40. EVCS c_2 shares all the energy 40 for cooperation.

The amount of energy shared for cooperation is denoted by $q_t^{n,m} \forall n, m \in \mathcal{N}$ where $q_t^{n,m}$ is the amount of energy shared to the mth EVCS from nth EVCS at t. The set of the amount of shared energy at the nth EVCS is denoted as $\mathcal{Q}^n = \{q_t^{n,0}, \ldots, q_t^{n,m}, \ldots, q_t^{n,N-1}\}, m \in [0, N), m \neq n$.

After the sharing step, the amount of energy stored in the ESS, denoted by b_t^n, is calculated as $b_t^n = u_t^n - \sum_{m \neq n} q_t^{n,m}$.

The result of the actions is based on the status of the ESS b_t^n. Discretized actions can be categorized into discharging action and charging action. The energy unit $\Delta\mathsf{u}$ denotes the amount of energy used for charging the ESS and selling to the utility company. The discrete set of actions is defined as $\mathcal{A} = \{-\Delta\mathsf{u}, \ldots, -K\Delta\mathsf{u}, +\Delta\mathsf{u}, \ldots, +K\Delta\mathsf{u}\}$ where $K\Delta\mathsf{u}$ is the maximum amount of energy that can be charged/discharged from the ESS in each discrete instant. We define $a_t^n \in \mathcal{A}_{s_t}$ as the action taken at time step t by EVCS c_n, where $A_{s_t^n}$ denotes the possible action set in the action space $\mathcal{A}$ under state s_t^n. In each time step t, $A_{s_t^n}$ is constrained by the amount of energy charged in the ESS o_t^n as follows:

$$\mathcal{A}_{s_t} = \begin{cases} \{Charging\}, & \text{if } E_g^{\min} \leq o_t^n < \Delta\mathsf{u} \\ \{Charging, Discharging\}, & \text{else} \end{cases}. \quad (4)$$

The function $\text{ESS}(a_t^n)$ denotes the amount of energy charged into the ESS e_n. The function $U(a_t^n)$ denotes the amount of energy trade determined by the selected action a_t^n. For example, if the EVCS c_n takes action $a_t^n = -K\Delta\mathsf{u}$ of *Discharging*, the value of $U(a_t^n)$ is $K\Delta\mathsf{u}$. Therefore, the function $\text{ESS}(a_t^n)$ is

defined as

$$\mathrm{ESS}(a_t^n) = \begin{cases} -U(a_t^n), & \text{if } a_t^n = Discharging \\ +U(a_t^n), & \text{if } a_t^n = Charging \end{cases}. \quad (5)$$

We assume that when the taken action *charging* increases $o_t^n + U(a_t^n)$ to more the maximum guard capacity $E_g^{\max}$, only the amount of chargeable energy $E_g^{\max} - o_t^n$ is charged and the overcharged energy is discarded.

3) Reward: The comprehensive reward $r(s_t, a_t)$ of the energy management scheme is configured to assess three aspects of the system management: first, how much money the EVCS pays for the energy used to operate the EVCS, second, benefits provided by the precharged energy, and third, loss caused by overcharging during the charging process. These subrewards are denoted as $\mathcal{R}_p = \{r_p^0, \ldots, r_p^n, \ldots, r_p^{N-1}\}$, $\mathcal{R}_b = \{r_b^0, \ldots, r_b^n, \ldots, r_b^{N-1}\}$, and $\mathcal{R}_o = \{r_o^0, \ldots, r_o^n, \ldots, r_o^{N-1}\}$, respectively. Here, r_p^n and r_b^n represent the pay reward and benefit reward of the nth EVCS, respectively, and r_o^n represents the overcharged reward of the nth EVCS. Note that these subrewards are added together. First, the pay reward r_p^n is determined by the action taken by the EVCS c_n. As a result of the taken action a_t^n, the EVCS c_n is rewarded if the amount of energy sold is larger than the amount of energy purchased. Otherwise, the EVCS receives a negative reward of r_p^n as penalty, i.e., $r_p^n = -(j_t^n + ESS(a_t^n)) * p_t$.

Next, r_b^n is calculated by the amount of energy to meet the net demand f_t^n from the energy charged in the ESS o_t^n. As the current price $price_t$ increases, the benefit reward becomes high. The reward indicates the reduced payment resulting from the use of the charged energy in the ESS at time t

$$r_b^n = \begin{cases} f_t^n \cdot price_t \cdot \sigma, & \text{if } f_t^n \le o_t^n - E_{g,n}^{\min} \\ (o_t^n - E_{g,n}^{\min}) * price_t, & \text{else} \end{cases}. \quad (6)$$

where σ is a coefficient that influences the importance of the amount of energy that is charged in the ESS to meet the net demand. Then, r_s^n is calculated by the amount of energy shared for cooperation. The share reward indicates the benefit of reduced payment obtained by sharing the charged energy in the ESS at time step t, i.e., $r_s^n = \sum_{m \neq n} q_t^{n,m} \cdot price_t$.

Finally, r_o^n is calculated by the amount of overcharged energy when the EVCS c_n charges the ESS e_n by the taken action a_t^n. As the current transaction price $price_t$ becomes high, the overcharged reward becomes low. The overcharged reward indicates the unnecessary payment when the EVCS charges the ESS at t, i.e., $r_o^n = -((b_t^n + U(a_t^n)) - E_{g,n}^{\max}) * price_t$.

Therefore, we define r_{total}^n for each agent as follows:

$$r_{\text{total}}^n = r_p^n + r_b^n + r_s^n + r_o^n. \quad (7)$$

B. Algorithm for Learning Energy Management

The EMSs of the EVCSs connected to the PV system and ESS learn the energy management policies that optimize the operational cost by utilizing the *CommNet* algorithm, which is shown as Algorithm 1. *CommNet* is a representative communication based MADRL algorithm. In *CommNet*, each agent uses only its observable state as shown in Fig. 2. Note that the

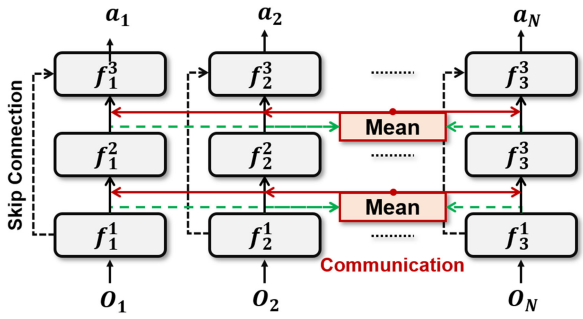

Fig. 2. Structure of CommNet.

Algorithm 1: CommNet-based Energy Management Deep Reinforcement Learning Method for PV/ESS-based Multiple EVCSs.

Initialize the *critic* and *actor* networks with weights θ^Q and θ^μ
Initialize the target networks as: $\theta^{\hat{Q}} \leftarrow \theta^Q, \theta^{\hat{\mu}} \leftarrow \theta^\mu$
for *episode = 1, MaxEpisode* **do**
  ▷ Initialize **EVCS Environments**
  for *time step = 1, T* **do**
    ▷ With probability $\mu(\mathcal{S}|\theta^\mu)$ select a set of actions $\mathcal{A}$ for each $s_t^n \in \mathcal{S}$
    ▷ Execute actions at in **Simulation Environments** and observe reward $\mathcal{R}_{total}$ and the next set of states $\mathcal{S}'$
    ▷ Store the transition pairs $\xi = (\mathcal{S}, \mathcal{A}, \mathcal{R}_{total}, \mathcal{S}')$ in replay buffer Φ
    If *time step* **is update period, do followings:**
    ▷ Sample a random minibatch from Φ
    ▷ Set $y_i = r_i + \delta \hat{Q}(s_i^{'}, \mu(s_i^{'}|\theta^{\hat{\mu}})|\theta^{\hat{Q}})$
    ▷ Update the θ^Q by applying stochastic gradient descent to the loss function of *critic* network: $L = \frac{1}{\varphi} \sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$
    ▷ Update the θ^μ by applying stochastic gradient ascent with respect to the gradient of *actor* network: $\nabla_{\theta^\mathcal{A}} J(\theta^\mu) \approx \frac{1}{\varphi} \sum_i \nabla_a Q(s, a|\theta^Q) \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s=s_i, a=\mu(s_i|\theta^\mu)}$ ▷ *Target Update* $\theta^{\hat{Q}}$ and $\theta^{\hat{\mu}}$
  end
end

communication structure of *CommNet* makes the convergence stable, even in a multiagent system. *CommNet* maps the states of all agents to their actions. During the mapping procedure, each agent has access to a broadcasting communication structure to share information. At a communication step, each agent sends its embedded state information as the communication message to the channel. The averaged message from the other agents is taken as the input of the next layer. The output of the final layer determines the action of the agent at time step t. Here, $m \in \{0, .., M\}$, where M is the number of communication steps in the network. Each f^m takes two input vectors for each EVCS $c_n \in \mathcal{C}$: the hidden state h_n^m and the communication $comm_n^m$, and outputs a vector h_n^{m+1}. In *CommNet*, the communication and hidden state are calculated as follows:

$$h_n^{m+1} = g^m(h_n^m, comm_n^m) \quad (8)$$

$$comm_n^{m+1} = \frac{1}{N-1} \sum_{n' \neq n} h_{n'}^{m+1}. \quad (9)$$

The softmax activation function is placed at the output layer $output = softmax(h_n^M)$. We can then interpret the output of the softmax as the probabilities that action a_t^n is taken when the EVCS c_n observes state s_t^n at t. We use the *actor* and *critic* reinforcement learning model for configuring *CommNet*.

Finally, the overall learning procedures are as follows.

1) First, the parameters of the *actor* and *critic* networks (μ and Q), which activate and evaluate the action of the EVCSs $\mathcal{C}$, are initialized (line 1).
2) Then, the target networks of both the *actor* and *critic* networks, $\hat{\mu}$ and $\hat{Q}$, are initialized (line 2).
3) The set of EVCSs $\mathcal{C}$ repeats the following procedures to the learn the energy management policies.
   i) For every episode, the transition pairs, which are obtained from the environments, consist of four types of information: first, a set of states $\mathcal{S}$; second, a set of actions $\mathcal{A}$ generated by the *actor* network μ; third the reward $\mathcal{R}_{\text{total}}$; and fourth, the observed set of next state spaces $\mathcal{S}'$. The pairs are stored in *replay buffer* Φ (lines 4–8).
   ii) After Φ is fully stored, a minibatch is randomly sampled from Φ. Then, the ith transition pair of the minibatch is utilized for calculating the mean squared Bellman error [24] between the target value y_i and $Q(s_i, a_i|\theta^Q)$ to update the *critic* network (line 10–12). In addition, to update the parameters of the *actor* network, the gradient of θ^μ is computed (line 13–14).
   iii) After the above procedures, the updated parameters of the *critic* and *actor* networks are utilized to update the target parameters $\theta^{\hat{Q}}$ and $\theta^{\hat{\mu}}$.
4) The neural network parameters are shared for EVCSs. That is, the EVCSs have the same policy for the EMS; and thus the private enterprises can easily extend the number of EVCSs.

Throughout the interaction between the EVCSs and the environments, the energy management operations of the EVCSs are sufficient to optimize the operational cost. The performance evaluation in Section IV shows that the proposed scheme achieves reasonable performance in the energy management of PV/ESS-enabled multiple EVCSs.

IV. Experiments

A. Simulation Setup

In this section, we elaborate the implementation details of the proposed *CommNet* learning-based cooperative energy management scheme for charging PV- and ESS-enabled electric vehicles. A Xavier initializer is used to initialize the weights of the neural networks for stabilizing the learning phase. The neural network was constructed with a dense layer, and the number of nodes in the hidden layer was 512. Furthermore, the number of nodes for communication channel was 512. Note that we implemented the *CommNet*-based optimal energy management algorithm and customized the cooperative energy management for PV- and ESS-enabled electric vehicle charging scenario as described in Section III. The agents in the *CommNet*-based method continuously interact with the dynamic energy management environment and obtain the pairs of state transition. Based on the pairs, the policies of the agents are optimized to manage the EVCSs. The operational energy management policies can be optimized using the policy gradient method after convergence of the learning phase. We evaluated the performance of the proposed cooperative scheme for charging PV/ESS-enabled electric vehicles.

We consider three EVCSs with one utility company. Fig. 3 shows the newly generated load demand of EVs, denoted as $\mathcal{H}$, and the internal load demand of EVCSs, denoted as $\mathcal{L}$. Furthermore, the PV generation and price are external variables, following a typical pattern that is related to the time of the day. This can also be viewed in Fig. 3. During the training procedure, we added Gaussian noise to the training data set for preventing the overfitting problem and obtaining various experiences. For all EVCSs, the maximum capacity of the ESS is identically set to $E^{\max} = 620$ kWh with $\eta = 0.02$. Therefore, the safeguard maximum and minimum capacities of the ESS for EVCS n are set to $E_{g,n}^{\max} = 607.6$ kWh and $E_{g,n}^{\min} = 12.4$ kWh, respectively. The energy unit $\Delta\mathsf{u}$ is set to 25. Then, K is 5. Therefore, the discrete set of actions is defined as follows: $\mathcal{A} = \{-25, -50, \ldots, -150, +25, +50, \ldots, +150\}$. We assumed the experimental results are executed for one day. One day consists of 1440 timeslots each of which lasts for 60 s.

B. Performance Evaluation

This section provides the numerical results to evaluate the performance of our cooperative energy management method.

1) Reward Convergence: Fig. 4 shows the training curve, which shows the reward convergence tendency of three different deep reinforcement learning algorithms in the proposed cooperative energy management scheme. Note that the results in Fig. 4 were obtained by using the same setting, i.e., $N = 3$. Typically, the training curve increases and then converges, as DRL learns from the cumulated experience and finally achieves an optimized policy. In Fig. 4, DQN and PPO fail to learn an optimized policy. The convergence of DQN and PPO usually relies on an underlying transition model that is stationary. When we use DQN and PPO, as the EVCS is not aware of the actions of the other stations, it does not know if the other EVCSs are changing their policies or taking exploratory actions. In the proposed energy management scheme with multiple EVCSs, the transition probabilities associated with the action of a single EVCS from one state to another are not stationary and change over time as the action choices of the other EVCSs change. On the other hand, *CommNet* shows that the training curve steadily increases and then becomes stable. *CommNet* uses the communication between the EVCSs before making a decision. Then, the communication informs the EVCS that the other EVCSs have changed their policies or have taken exploratory actions. As a result, in the proposed energy management scheme with multiple EVCSs, *CommNet* is able to obtain the optimized policies.

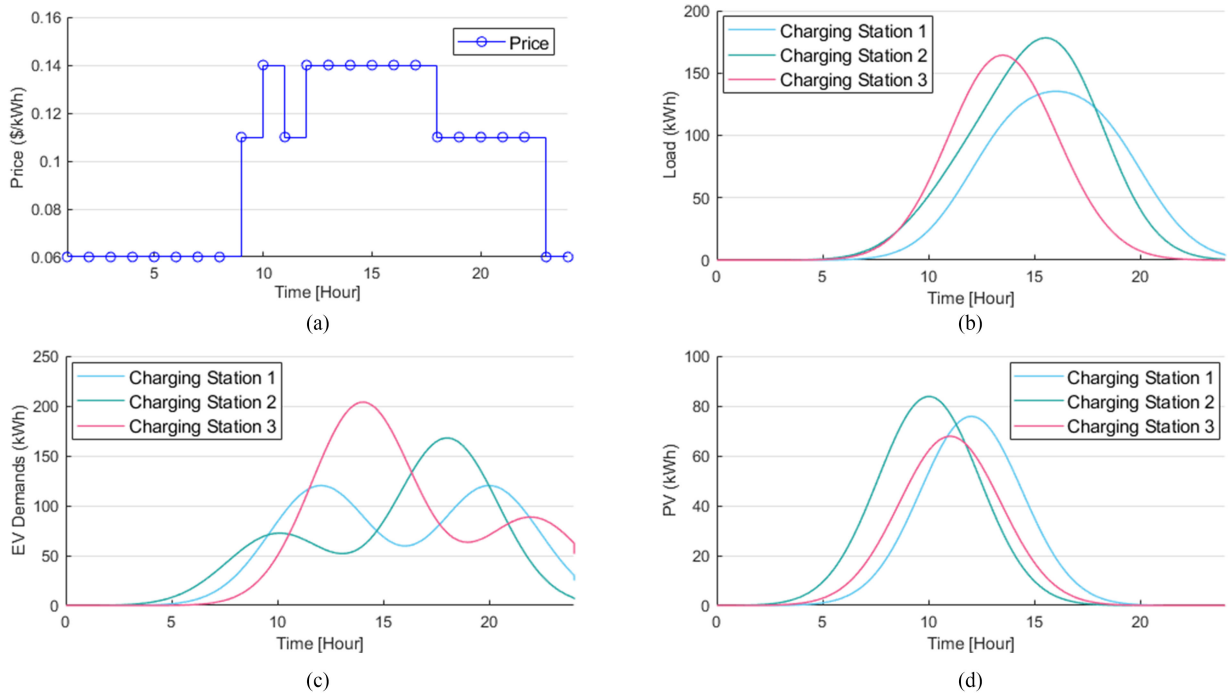

Fig. 3. Profile of EVCS for 1 day ($N = 3$). (a) ToU price profile. (b) Energy demands profile. (c) EV demands profile. (d) PV generation profile.

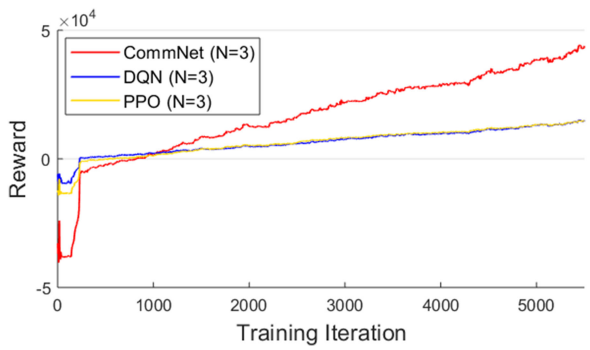

Fig. 4. Reward convergence ($N = 3$).

TABLE II
AVERAGE SOC OF THE ESS [SEE FIG. 5(A)–(C)]

σ	EVCS 1 c_0	EVCS 2 c_1	EVCS 3 c_2
σ=1	0.1884	0.2096	0.1881
σ=2	0.2527	0.2444	0.2554
σ=3	0.2608	0.2914	0.3281

2) Energy Management: We now study the impact of the reward coefficient σ on the performance of the EVCSs. Fig. 5(a)–(c) shows the changes in the SoC of the ESS with varying σ during the day. The coefficient σ affects the evaluation criteria for the action that charges the ESS to maximize the usage of energy stored in the ESS when the EVCS tries to meet the net demand. In terms of the operation cost, it is important to charge the ESS when the utility price is low and to discharge the ESS when the price is high. Furthermore, it is important to use the energy stored in the ESS when the price is high. As mentioned before, because the energy usage is the unknown information, an efficient ESS energy sharing strategy among the EVCSs can meet the unexpected high net demand effectively, consequently reducing the operation cost of the EVCSs. In Fig. 5(a), EVCS 1 gradually increases the SoC of the ESS to meet the highest net demand. At the same time, EVCS 1 charges the surplus energy to help the other EVCSs. Since the maximum net demand of EVCS 1 is about 210 kWh at around 21:00, the SoC of the ESS is charged until about 0.55 at around 20:00. Then, the utility price and net demand are high as 0.11 ($/kWh), and 210 kWh, respectively. After that, EVCS 1 obtains revenue by discharging the ESS before the utility price is lowered by 0.06 ($/kWh). Similarly, the maximum net demand of EVCS 2 is about 307 kWh at around 17:00. Therefore, EVCS 2 also gradually increases the SoC of the ESS for the highest net demand. After the peak net demand time of the EVCS 2, the SoC decreases for a while, and then increases again to meet the increasing net demand at around 19:00. After that, EVCS 2 obtains revenue by discharging the ESS until around 24:00. EVCS 3 has the maximum net demand about 334 kWh at around 17:00. Similar to EVCS 1 and EVCS 2, the SoC of the ESS gradually increases. After the SoC of the ESS decreases for a while, the SoC of the ESS increases to meet the increasing net demand again.

Fig. 5(b)–(c) shows the similar pattern. Table II gives the average SoC of the EVCSs as σ changes. As the coefficient σ increases, the average SoC of all the EVCSs also increases. As the coefficient σ increases by 1, the average SoC increases

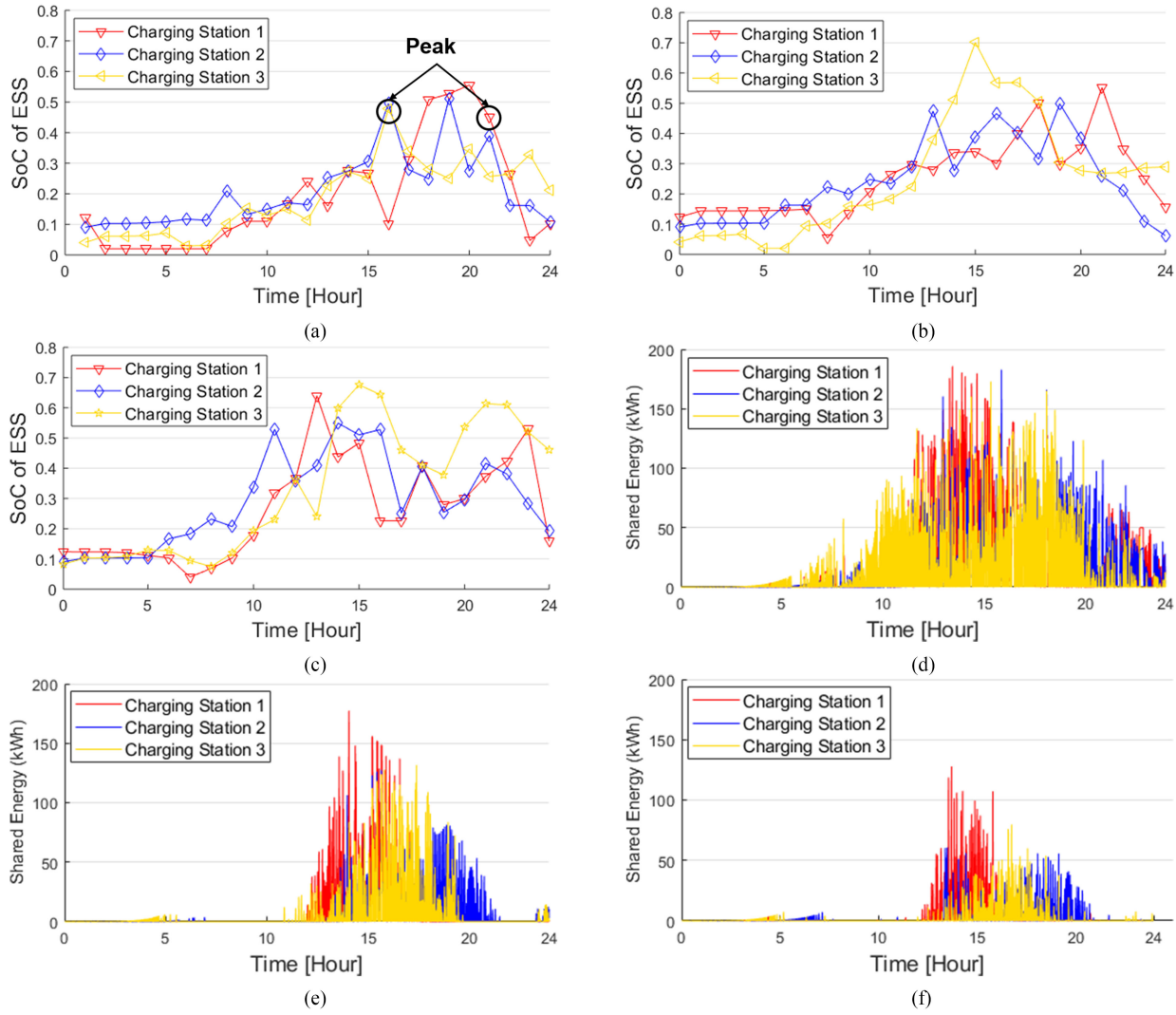

Fig. 5. Tendency of the SoC of the ESS and the amount of shared energy in the proposed scheme in each σ ($N = 3$). (a) SoC of the ESS ($\sigma = 1$). (b) SoC of the ESS ($\sigma = 2$). (c) SoC of the ESS ($\sigma = 3$). (d) Amount of shared energy ($\sigma = 1$). (e) Amount of shared energy ($\sigma = 2$). (f) Amount of shared energy ($\sigma = 3$).

by about 25%. This is because the policy does not consider the energy sharing from the other EVCSs when the coefficient σ increases. Then, EVCSs charge more energy to prepare the unknown net demand. Thus, our experiments were able to confirm that, with the reward coefficient σ, the proposed energy management scheme for cooperation is capable of learning the optimal energy management policy with respect to the ability to meet net demands with energy sharing. As the coefficient σ increases, the EVCSs can learn policies to increase the SoC of the ESS to meet the net demand independently with less energy sharing.

Fig. 5(d)–(f) shows the amount of energy shared by the EVCSs during the day. In Fig. 5(d), EVCS 1 shares the most energy at around 13:00. EVCS 1 shares a large amount of the energy stored in the ESS at around 15:00 to help the other EVCSs meet their net demands. Because the trained policy of EVCS 1 has chosen charging action, EVCS 1 has surplus energy for the sharing. Therefore, EVCS 1 can share its own stored energy with the other EVCSs. This is because the EVCSs are rewarded when they share energy to help the other EVCSs. As shown in Fig. 5(d), EVCS 1 gradually charges the ESS to meet own maximum net demand and to share energy for helping the other EVCSs. This charging action can lower the overall operating cost by purchasing energy when the price of energy is low. When the other EVCSs, through the communication, request EVCS 1 for energy sharing, EVCS 1 initiates the sharing process. This cooperation reduces the total operation cost of the entire EVCSs. EVCS 1 sells its surplus energy when it is considered beneficial to sell the stored energy and later purchase it again. Therefore, the SoC of the ESS rapidly decreases after around 16:00 when the peak demands of other EVCSs have ended. In Table III, EVCS 2 shows a low amount of energy sharing, compared to the other EVCSs. However, EVCS 2 shares more energy between around 20:00 and 24:00 compared to the EVCSs.

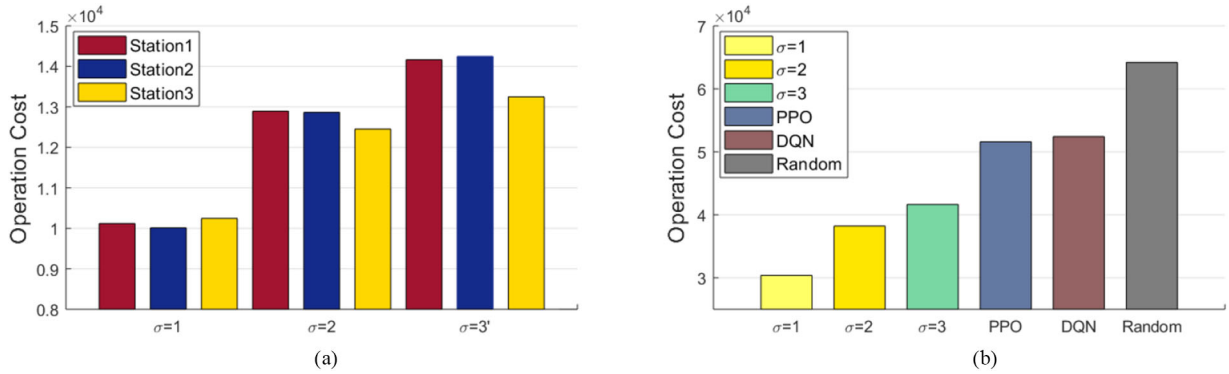

Fig. 6. Operation cost comparison for each coefficient σ and the baseline methods ($N = 3$). (a) Operation cost for each coefficient σ. (b) Performance comparison with baseline algorithms.

TABLE III
AVERAGE AMOUNT OF SHARED ENERGY [SEE FIG. 5(D)–(F)]

σ	EVCS 1 c_0	EVCS 2 c_1	EVCS 3 c_2
σ=1	14.8855	12.0142	17.4110
σ=2	6.2965	5.3240	5.7170
σ=3	2.4711	1.8716	1.6117

EVCS 3 shares considerable energy between around 13:00 and 17:00. This result demonstrates the benefit of a communication-based energy management scheme. Without the communication, it is possible that all EVCSs purchase the surplus energy for sharing. However, in the proposed approach, when one EVCS charges ESS to share a large amount of energy, the other EVCSs reduce the amount of surplus energy charged to the ESS. The communication of *CommNet* allows EVCSs to schedule the charging times of the ESSs with each other. When $\sigma = 2$, the average amount of energy shared is reduced to about 30% because EVCSs try to utilize the energy stored in their ESSs to meet the net demand. As a result, the amount of energy sharing required to meet the net demand is reduced, and the average amount of shared energy is reduced.

3) Operation Cost: In Fig. 6(a), the experiment shows that the average operation cost of EVCSs changes as the value of the coefficient σ increases. As mentioned before, the net demand is an unknown external information for the EVCS. Therefore, as the coefficient σ increases, the EVCSs try to charge as much energy as possible for meeting their own net demands while less considering the surplus energy from the other EVCSs. As a result, a lot of surplus energy is charged to the ESSs, which in turn increases the overall operation cost. In view of the total operation cost reduction, the proposed algorithm shows the best performance with $\sigma = 1$. However, given the unpredictable demands such as internal loads and EV demands, it is important that the surplus energy stored in the ESS can cope with unexpected situations. In Fig. 6(b), we compare the operation cost of the EVCSs based on our proposed cooperative method with that of the conventional deep reinforcement learning algorithms (e.g., DQN and PPO). PPO and DQN were trained in the energy management environment with the coefficient σ is set to 1. As shown in Fig. 4, the conventional DRL algorithms fail to converge. As a result, the operation cost of the EMS of the EVCSs trained using PPO or DQN is found to be about two times that of the proposed *CommNet* based scheme when the coefficient σ is 1. The proposed *CommNet* based scheme requires only about 20% of the operation cost compared to that when the EVCS energy management scheme charges or discharges the ESS without considering the system state.

4) System Scalability: Fig. 7 shows the scalability of the proposed model with $N = 4$. In this experiment, we set the reward coefficient σ as 1. In Fig. 7(e), EVCS 1 gradually increases the SoC of the ESS and prepares the highest net demand. Since the maximum net demand of the EVCS 1 is about 250 kWh at around 22:00, EVCS 1 purchases the energy because the energy price is relatively cheap from 18:00 to 20:00. Similarly, the maximum net demand of EVCS 2 is about 330 kWh at around 17:00. Therefore, EVCS 2 also gradually increase the SoC of the ESS for preparing the highest net demand. After the peak net demand time of the EVCS 2, the SoC decreases for a while, then increases again to meet the increasing net demand at around 19:00. After that, EVCS 2 obtains the revenue by discharging the ESS until around 24:00. As shown in Fig. 7(f), EVCS 2 shares large amount of the remaining energy from 15:00 to 21:00. The results show that the EVCSs that maintain high SoC values after the energy consumption peak time share energy to meet the energy requirements of other EVCSs. EVCS 3 has the maximum net demand, with about 350 kWh at around 13:00. Similar to EVCS 1 and EVCS 2, the SoC of the ESS gradually increases for preparing the peak net demand. EVCS 3 also helps to reduce the overall operating costs by sharing the remaining energy with other EVCSs after the peak time. Unlike the previous experiment, the added EVCS 4 shows the sharing of the remaining energy after the peak time 13:00, which requires energy about 250 kWh. Thus, our experiments can confirm that, with the reward coefficient $\sigma = 1$, the proposed energy management scheme is capable of being extended.

Table IV gives the inference time required for the proposed algorithm to make observations and determine the actions that need to be performed in order to operate in EVCS environment in real-time. The experiment is an average of 10 000 experiments. The result shows that the inference time is only 0.0004 s slower

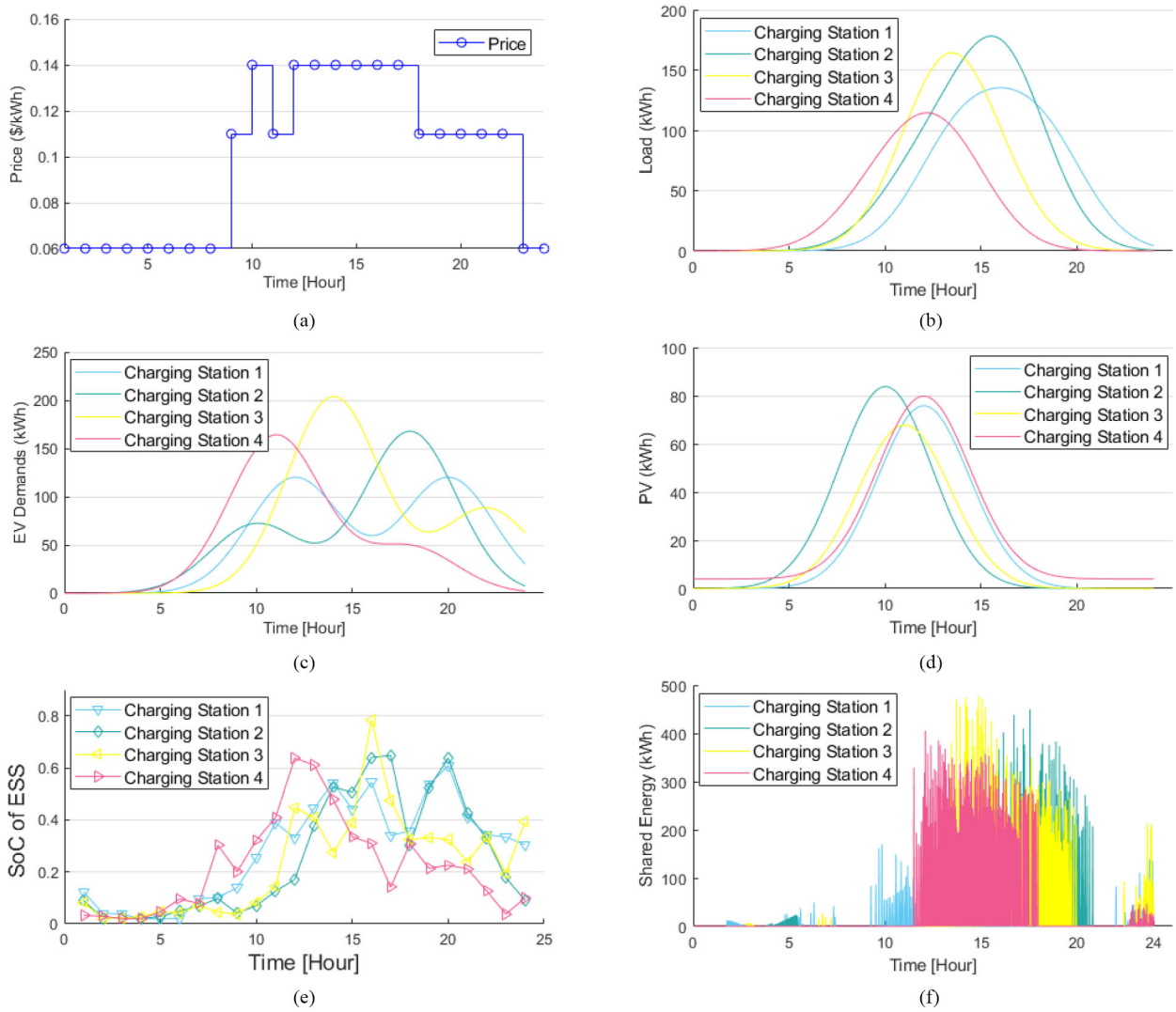

Fig. 7. Tendency of SoC of the ESS and the amount of shared energy of the proposed management scheme for $\sigma = 1$ ($N = 4$). (a) ToU price profile. (b) Energy demands profile. (c) EV demands profile. (d) PV generation profile. (e) SoC of the ESS. (f) SoC of the ESS.

TABLE IV
AVERAGE INFERENCE TIME

Model	Inference time (seconds)
CommNet (proposed)	0.002714
DQN (no communication)	0.002413
PPO (no communication)	0.002315

than the general DRL model. As a result, the proposed algorithm can be used efficiently in a real-time EVCS system.

5) Summarization: In this article, we consider the situation in which a single private enterprise manages multiple EVCSs equipped with renewable energy resources. The primary goal of the proposed algorithm is to reduce the entire operation cost of multiple EVCSs by sharing the ESS resource based on the communication among EVCSs. Without the communication, each EVCS determines the charging and discharging action independently so that the effective surplus energy management for energy sharing can be limited. However, the proposed method based on *CommNet* enables multiple EVCSs to communicate with each other for sharing their stored ESS energy, consequently leading to the reduction of the overall operation cost. The experimental results in Fig. 5 show that when $\sigma = 3$, the EVCSs try to share about $1/7$ of the energy compared to that than when $\sigma = 1$. As a result, the operating costs increase by about 30%. Thus, it is shown that communication based multiple EVCSs energy management can efficiently reduce the overall operation cost.

V. CONCLUSION

This article proposed a new distributed MADRL method for the energy management of PV/ESS-enabled electric-vehicle charging stations. The main novelty of this article was the computation of scheduling solutions in a distributed manner while handling run-time time-varying dynamically changing charging related data of an electronic vehicles. As shown in the data-intensive performance evaluation with large-scale data,

it is confirmed that the proposed method achieves a desirable performance.

REFERENCES

[1] Q. Yan, B. Zhang, and M. Kezunovic, "Optimized operational cost reduction for an EV charging station integrated with battery energy storage and PV generation," *IEEE Trans. Smart Grid*, vol. 10, no. 2, pp. 2096–2106, Mar. 2019.

[2] Y. Zheng, Y. Song, D. J. Hill, and K. Meng, "Online distributed MPC-based optimal scheduling for EV charging stations in distribution systems," *IEEE Trans. Ind. Informat.*, vol. 15, no. 2, pp. 638–649, Feb. 2019.

[3] Y. Song, Y. Zheng, and D. J. Hill, "Optimal scheduling for EV charging stations in distribution networks: A convexified model," *IEEE Trans. Power Syst.*, vol. 32, no. 2, pp. 1574–1575, Mar. 2017.

[4] K. Chaudhari, A. Ukil, K. N. Kumar, U. Manandhar, and S. K. Kollimalla, "Hybrid optimization for economic deployment of ESS in PV-integrated EV charging stations," *IEEE Trans. Ind. Informat.*, vol. 14, no. 1, pp. 106–116, Jan. 2018.

[5] J. Zhang *et al.*, "A hierarchical distributed energy management for multiple PV-based EV charging stations," in *Proc. 44th Annu. Conf. IEEE Ind. Elect. Soc.*, Oct. 2018, pp. 1603–1608.

[6] C. Luo, Y.-F. Huang, and V. Gupta, "Stochastic dynamic pricing for EV charging stations with renewable integration and energy storage," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1494–1505, Mar. 2018.

[7] C. Luo, Y. F. Huang, and V. Gupta, "Placement of EV charging stations-balancing benefits among multiple entities," *IEEE Trans. Smart Grid*, vol. 8, no. 2, pp. 759–768, Mar. 2017.

[8] X. Wang, M. Shahidehpour, C. Jiang, and Z. Li, "Coordinated planning strategy for electric vehicle charging stations and coupled traffic-electric networks," *IEEE Trans. Power Syst.*, vol. 34, no. 1, pp. 268–279, Jan. 2019.

[9] A. Ehsan and Q. Yang, "Active distribution system reinforcement planning with EV charging stations—Part I: Uncertainty modelling and problem formulation," *IEEE Trans. Sustain. Energy*, to be published, doi: 10.1109/TSTE.2019.2915338.

[10] Z. Wan, H. Li, H. He, and D. Prokhorov, "A data-driven approach for real-time residential EV charging management," in *Proc. IEEE Power Energy Soc. General Meeting*, Aug. 2018, pp. 1–5.

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

[12] S. Wang, S. Bi, and Y.-J. A. Zhang, "A reinforcement learning approach for EV charging station dynamic pricing and scheduling control," in *Proc. IEEE Power Energy Soc. General Meeting*, Aug. 2018, pp. 1–5.

[13] N. Sadeghianpourhamami, J. Deleu, and C. Develder, "Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2019.2920320.

[14] C. Jiang, Z. Jing, X. Cui, T. Ji, and Q. Wu, "Multiple agents and reinforcement learning for modelling charging loads of electric taxis," *Appl. Energy*, vol. 222, pp. 158–168, 2018.

[15] X. Z. Ye, T. Y. Ji, M. S. Li, and Q. H. Wu, "Optimal control strategy for plug-in electric vehicles based on reinforcement learning in distribution networks," in *Proc. Int. Conf. Power Syst. Technol.*, Nov. 2018, pp. 1706–1711.

[16] A. Chiş, J. Lundén, and V. Koivunen, "Reinforcement learning-based plug-in electric vehicle charging with forecasted price," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3674–3684, May 2017.

[17] C. Boutilier, "Planning, learning and coordination in multiagent decision processes," in *Proc. 6th Conf. Theor. Aspects Rationality Knowl.*, Mar. 1996, pp. 195–210.

[18] V. Mnih, K Kavukcuoglu, D Silver, A Graves, I Antonoglou, D Wierstra, and M Riedmiller, "Playing atari with deep reinforcement learning," 2013, *arXiv:1312.5602*.

[19] C. J. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, no. 3/4, pp. 279–292, May 1992.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[21] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," 2017, *arXiv:1502.05477*.

[22] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multi-agent systems: A review of challenges, solutions and applications," 2018, *arXiv:1812.11794*.

[23] S. Kim and H. Lim, "Reinforcement learning based energy management algorithm for smart energy buildings," *Energies*, vol. 11, no. 8, pp. 1–19, Aug. 2018.

[24] R. S. Sutton, "On the significance of Markov decision processes," in *Proc. Int. Conf. Artif. Neural Netw.*, Oct. 1997, pp. 273–282.

[25] H. Berlink and A. H. Costa, "Batch reinforcement learning for smart home energy management," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2015, pp. 2561–2567.

MyungJae Shin received the B.S. degree computer science and engineering from Chung-Ang University (CAU), Seoul, South Korea, in 2018, where he is currently working toward the M.S. degree in computer science and engineering.

His research interests include various econometric theories and their deep-learning-based computational solutions.

Mr. Shin was the recipient of the Second Highest Honor from the CAU College of Engineering. He was also the recipient of the National Science and Technology Scholarship (2016–2017).

Dae-Hyun Choi (S'10–M'19) received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2002, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from Texas A&M University, College Station, TX, USA, in 2008 and 2014, respectively.

He is currently an Associate Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul. From 2002 to 2006, he was a Researcher with Korea Telecom, Seoul, where he worked on designing and implementing home network systems. From 2014 to 2015, he was a Senior Researcher with LG Electronics, Seoul, where he developed home energy management systems. His research interests include power system state estimation, electricity markets, the cyber-physical security of smart grids, and the theory and application of cyber-physical energy systems.

Dr. Choi was the recipient of the Best Paper Award at the 2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm) in Tainan City, Taiwan.

Joongheon Kim (M'06–SM'18) received the B.S. and M.S. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 2004 and 2006, respectively, and the Ph.D. degree in computer science from the University of Southern California (USC), Los Angeles, CA, USA, in 2014.

Before joining Korea University as an Assistant Professor, he was with LG Electronics Seocho R&D Campus as a Research Engineer (Seoul, 2006–2009), InterDigital as an Intern (San Diego, CA, USA, 2012), Intel Corporation as a Systems Engineer (Santa Clara, CA, 2013–2016), and Chung-Ang University as an Assistant Professor of Computer Science and Engineering (Seoul, 2016–2019). He is currently an Assistant Professor of Electrical Engineering with Korea University.

Dr. Kim was a recipient of the Annenberg Graduate Fellowship with his Ph.D. admission from USC (2009) and the Haedong Young Scholar Award (2018), which is for recognizing a young Korean Researcher under the age of 40 who has made outstanding scholarly contributions to communications and information sciences research.