Energy Reports 10 (2023) 1135–1150

Contents lists available at ScienceDirect

Energy Reports

journal homepage: www.elsevier.com/locate/egyr

Research paper

Two-stage scheduling of smart electric vehicle charging stations and inverter-based Volt-VAR control using a prediction error-integrated deep reinforcement learning method

Sangyoon Lee, Dae-Hyun Choi *

School of Electrical and Electronics Engineering, Chung-ang University, Dongjak-gu, Seoul 156-756, Republic of Korea

ARTICLE INFO

Article history:
Received 22 April 2023
Received in revised form 3 July 2023
Accepted 26 July 2023
Available online xxxx

Keywords:
Electric vehicle charging station
Prediction error
Deep reinforcement learning
Power distribution grid
Profit maximization
Volt-VAR control

ABSTRACT

Smart electric vehicle charging stations (EVCSs) having distributed energy resources (DERs), including photovoltaic (PV) systems and energy storage systems (ESSs), are becoming vital devices for increasing their profit and maintaining stable distribution grid operations by scheduling the real/reactive power of DERs. However, prediction errors of PV generation outputs and electric vehicle (EV) loads from EVCSs may decrease their profit and destabilize the distribution grid owing to incorrect EV charging scheduling and voltage regulation. To address this issue, we propose a two-stage framework for smart EVCS scheduling and inverter-based Volt-VAR control (VVC) using prediction error-integrated deep reinforcement learning (DRL). In the first stage, the EVCS agents train their neural network model with a 30-min resolution to maximize their profit through a day-ahead charging/discharging scheduling of ESSs in the EVCSs while responding to various prediction errors of PV generation outputs and EV loads. The total real power consumption of each EVCS, including the charging/discharging schedules of the ESSs calculated in the first stage, is delivered to the second stage, in which the VVC agent trains its neural network model with a 5-min resolution to minimize the real power loss and voltage violations through real-time reactive power scheduling of the ESSs in the EVCSs via their inverters. The proposed approach was tested in the IEEE 33-node and IEEE 123-node distribution systems. The results show that the proposed approach outperforms DRL methods that do not consider the prediction errors in terms of profitability of the EVCS and reduction of real power loss.

1. Introduction

As more electric vehicles (EVs) are connected to power distribution systems, thereby reducing environment pollution (e.g., greenhouse gas emissions and carbon pollution) and making the most of their flexible load capability (e.g., vehicle-to-grid (V2G) technology for peak shaving and voltage regulation), electric vehicle charging stations (EVCSs) are becoming vital entities for economically managing the charging schedules of EVs and reliably maintaining stable power distribution grid operations (Farzin et al., 2016). Recently, conventional EVCSs have been transformed into smart EVCSs equipped with distributed energy resources (DERs), including solar photovoltaic (PV) systems and energy storage systems (ESSs). Smart EVCSs can reduce the power consumption from the distribution grid using power generated by the PV systems and stored in the ESSs, thereby maximizing their charging profit while avoiding transformer overloading and power equipment degradation (Datta et al., 2020).

Evidently, scheduling the charging of EVs in smart EVCS without considering power distribution grid operations may result in abnormal grid operations, such as an increase in power losses and voltage violations. To resolve this issue, smart EVCSs must cooperate with the Volt-VAR control (VVC) module, which is one of the key functions in distribution management systems to maintain stable distribution grid operations. Recently, VVC has leveraged on smart inverters of DERs as new voltage regulating devices through which reactive power absorption and injection of DERs from and to the grid are controlled to determine optimal nodal voltage magnitudes along with reduction of the power loss (Wang et al., 2020b). In this context, smart inverters of DERs connected to smart EVCSs can be exploited as voltage-regulating devices to perform VVC. Therefore, the smart inverter-based reactive power capability of DERs in smart EVCSs suggests the need for a system-wide coordinated framework in which smart EVCSs and inverter-based VVC cooperate to manage stable distribution grid operations while ensuring economical EVCS operation.

However, the development of a framework for smart EVCS–VVC coordination presents several challenges. First, the prediction errors of PV generation output at smart EVCSs may lead to

* Corresponding author.
E-mail address: dhchoi@cau.ac.kr (D.-H. Choi).

https://doi.org/10.1016/j.egyr.2023.07.054

abnormal EV charging scheduling, thereby yielding a distorted calculation of the EVCS profit. Second, incorrectly aggregated EV load consumption schedules due to such prediction errors are fed into the VVC as wrong input data, which in turn gives rise to miscalculations of the power loss and voltage profile along the distribution feeder via the malfunction of VVC. Third, because conventional model-based VVC optimization methods rely on inaccurate large distribution system models with uncertain line parameters, their solutions may not be optimal and may not scale well for real-time applications. To resolve the aforementioned challenges, we present a model-free deep reinforcement learning (DRL)-based coordination framework in which multiple EVCS DRL agents interact with a VVC DRL agent to calculate economic charging schedules of EVs under various prediction-error scenarios of PV generation output and aggregated EV load from a smart EVCS while maintaining robust distribution grid operations against uncertainty of the distribution grid model.

Many recent studies have proposed model-based optimization approaches for the scheduling of smart EVCSs and VVC in power distribution systems. These approaches can be categorized as follows:

(L1) **Optimization-based smart EVCS scheduling:** A stochastic optimization model was presented in Katrin et al. (2019) in which the operation cost of PV-integrated EVCSs is minimized along with various prediction options for PV generation outputs. In Yan et al. (2021b), a two-stage optimization method was proposed in which the power allocation of a smart EVCS integrated with a PV system and an ESS is conducted in the first stage, and the charging schedules of EVs are coordinated in the second stage. In Yan et al. (2019), a multi-stage optimization algorithm comprising day-ahead energy management and real-time control was developed to reduce the total operating cost of a PV-ESS-integrated EVCS while considering the satisfaction of EV users under uncertain EVCS operation. An optimization framework that coordinates the exchange of energy between the distribution grid and smart EVCSs was presented in Li et al. (2020); in this framework, a chance-constrained optimization method is adopted to handle the uncertainty in the prediction of PV generation output from smart EVCSs. In Yang et al. (2021), a time-of-use (TOU) price tariff-based energy management framework for smart EVCSs was designed to maximize the profit of smart EVCSs using the peak–valley price difference. A robust optimization problem was formulated in Li et al. (2022) in which the optimal location of smart EVCSs combined with wind and PV systems and ESSs was determined based on the power distribution and transportation networks under uncertainties in wind/PV generation output and EV charging demand. In Kriekinge et al. (2021), the cost and peak-minimizing EV charging strategy was developed using model predictive control method in which the EV charging schedule is calculated based on the forecasted PV generation and EV load demand based on the deep neural network method. In Sierra et al. (2020), a simulation model was developed to quantify the feasibility of PV-ESS integrated EVCSs located in the United States and China from a technical, financial, and environmental perspective.

(L2) **Optimization-based VVC:** Many studies have addressed the development of VVC model using the smart inverters of PV systems and/or EVs based on V2G technology. In Jabr (2019), two decentralized optimization schemes considering the uncertainty in PV generation output were formulated using robust and distributionally robust optimization methods to reduce voltage violations by dispatching the reactive power of PV systems via their smart inverters. A fully distributed VVC framework using aggregated PV inverters with two timescales was presented in Wang et al. (2020c). In this framework, the reactive power of PV aggregators is scheduled to minimize the power loss in a 15-min time resolution using the alternating direction multiplier method while the fast fluctuations of PV generation outputs are handled using the droop control of PV systems in real-time. A two-layer VVC method was proposed in Hu et al. (2020) in which the optimal rates of Volt-VAR droop control curves for PV systems are calculated to reduce the network power loss at the global layer, and the reactive power of PV systems is tuned using the optimized droop control curves to suppress voltage violations at the local layer. Demand response-integrated VVC method using legacy voltage regulators (e.g., on-load tap changers (OLTCs) and capacitor banks (CBs)) and smart inverters of PV systems was developed in Vineeth et al. (2021) to minimize both the real power loss and peak load in unbalanced active distribution systems. In Wenjie et al. (2018), a multi-agent-based VVC model integrated with V2G technology of EVs was presented. In this model, reactive power dispatch of EVs via V2G is conducted along with EV charging coordination while ensuring the performance of VVC under uncertain EV charging scenarios. A two-stage VVC model was presented in Sun et al. (2021) in which OLTCs and CBs are scheduled using an optimal power flow method with 1-h resolution in the first stage; in the second stage, the reactive power of PV systems and EVs is controlled in real-time to mitigate voltage violations.

Although the aforementioned model-based approaches in the literature (L1) and (L2) yield the desired performance for EVCS scheduling and VVC, they heavily rely on accurate knowledge of the EV charging behavior (e.g., arrival/departure time of EVs and initial/desired state of charge (SOC) of the EV battery) and power distribution system model; however, such knowledge varies dynamically and is difficult to obtain in real-world scenarios. In addition, under a large number of scenarios for EVCS operation and VVC, the model-based approach may not calculate the optimal solution efficiently and rapidly owing to the high complexity of heterogeneous EV charging behaviors and power distribution systems.

To resolve these challenges encountered by the aforementioned model-based approaches, DRL, which is a reinforcement learning (RL) integrated with artificial neural networks (ANNs), has recently attracted attention as a model-free methodology for efficient scheduling of EVCSs and VVC. We next present a literature review related to our study divided into two categories:

(L3) **DRL-based EV/EVCS scheduling:** From the planning perspective of EVCSs, a hybrid approach that ensures their long-term revenue was proposed in Tao et al. (2022). In this approach, DRL and mixed-integer linear programming methods are jointly used to calculate the best match between the EVs and available charging/swapping infrastructure. In Dorokhova et al. (2021), a DRL method using double deep Q-networks learning and deep deterministic policy gradient approaches was applied to the EV charging scheduling problem to maximize both the PV self-consumption and SOCs of EVs when they depart from a smart EVCS. In Felix et al. (2021), the deep Q-network method was employed to schedule the charging of EVs without knowledge of future information, including arrival/departure time and energy consumption of EVs. In Wang et al. (2022), a novel cluster-based EV charging scheduling algorithm was developed using the DRL method

to response the real-time price signals from the distribution system operators. A soft actor critic (SAC)-based DRL framework was developed in Yan et al. (2021a) in which the performance of an individual EV charging problem with dynamic EV-user behaviors is improved using supervised learning and RL methods. In Zhao and Lee (2022), a new dynamic pricing framework was presented to maximize the quality of service with a differentiated service requirement for EVs. A novel multi-agent DRL (MADRL) algorithm for PV-ESS-integrated EVCSs was developed in Shin et al. (2020). In this algorithm, EVCSs minimize their charging cost by sharing the surplus energy stored at their ESSs. In Yan et al. (2022), a MADRL-based decentralized and cooperative charging strategy was proposed to control the charging of multiple EVs simultaneously considering varying environmental factors such as electricity price and EV driver's characteristics. A MADRL method was adopted to build an optimal energy purchasing strategy for EVCSs in Zhang et al. (2023) in which a long short-term memory neural network is used to predict the EV charging demand. In Zhang et al. (2021a), a joint charging and traveling route scheduling problem of EVs was formulated using the DRL method. A federated DRL model using the SAC method was proposed in Lee and Choi (2021). In this model, multiple smart EVCSs maximize their profit using the calculated profitable electricity price while preserving their private data. A DRL-based EV charging scheduling algorithm was proposed in Jin and Xu (2021) in which the power distribution system operation conditions are considered along with uncertainties of renewable energy generation and electricity prices. There were many DRL-based algorithms including Lee and Choi (2021), Jin and Xu (2021) for the scheduling of EV and EVCS. However, no studies proposed the DRL framework for the coordination of smart EVCSs and VVC while considering the various prediction errors of PV generation outputs and aggregated EV loads in smart EVCSs.

(L4) **DRL-based VVC:** A two-timescale hybrid voltage regulation scheme was presented in Sun and Qiu (2021) in which OLTC and CBs are dispatched using a mixed-integer second-order cone programming method in a slow timescale and PV inverters are controlled using a DRL method in a fast timescale to mitigate fast voltage violations. A distributed SAC method was applied to the voltage regulation problem in Cao et al. (2022). In this method, an entire distribution network is first decomposed into several sub-networks using voltage-reactive power sensitivity to achieve fast control of PV inverters. In each sub-network, PV inverters cooperate with OLTCs and CBs to minimize the total voltage deviations and long-term switching numbers of OLTCs and CBs. In Wang et al. (2020a), a safe off-policy DRL problem using the constrained SAC method was formulated in a constrained Markov decision process (MDP) problem to better satisfy the operation constraints in power distribution systems. Following a similar approach to that reported in Wang et al. (2020a), a novel safety layer was added to the DRL framework in Yuanqi and Nanpeng (2022). This layer enables the VVC agents to perform a safe exploration during the training process by satisfying the physical constraints of the distribution system while improving the training convergence performance. A novel MADRL-based VVC algorithm was presented in Zhang et al. (2021b) in which DRL agents for OLTC, CBs, and smart inverters of PV systems interact to perform VVC in unbalanced distribution systems with voltage-dependent loads. More recently, a consensus MADRL-based fully distributed VVC model without a central controller was developed in Gao et al. (2021). In this model, the DRL agents of heterogeneous legacy voltage regulators minimize the real power loss, voltage violations, and switching costs of the voltage regulators by efficiently communicating their local information with their neighbors while maintaining resilience against failures of individual controllers and communication links. In Liu et al. (2021), a MADRL-based robust VVC algorithm was adopted to minimize bus voltage deviations and network power losses in which uncertainties of PV generation outputs and loads are modeled by stochastic programming. In Nguyen and Choi (2022), a novel stand-alone safety module was proposed and integrated with the DRL-based VVC to remove voltage violations during the training process.

Previous studies related to categories (L3) and (L4) have two limitations. First, these studies on the scheduling of smart EVCSs were formulated as DRL problems in which the prediction values of PV generation outputs and EV loads are used as crucial input data. However, these predicted values are not very accurate and do not reflect all variations in PV generation outputs and EV loads. Evidently, this would distort the charging scheduling of EVs, thereby yielding incorrect profits for smart EVCSs. Second, DRL approaches for the scheduling of smart EVCSs and VVC were implemented separately without considering their interdependence. Note that VVC performs its task based on the aggregated EV loads scheduled by smart EVCSs. Therefore, incorrect aggregated EV charging schedules due to prediction errors in the PV generation outputs and EV loads of smart EVCSs would degrade VVC performance.

In short, the limitations of the aforementioned literature (L1) ∼(L4) are summarized as follows. In the literature (L1) and (L2), the algorithms for the smart EVCS scheduling and VVC were formulated using the model-based optimization problem, respectively. However, they assumed the unrealistic situation in which the knowledge of the system model for smart EVCS and VVC operation is very accurate. Furthermore, the model-based optimization approach may increase the computation complexity significantly with larger system models. To tackle these limitations, the model-free DRL algorithms for the EV/EVCS scheduling and VVC were developed in the literature (L3) and (L4), respectively. However, the studies in (L3) have the strict assumption that the prediction values of PV generation outputs and EV loads are accurate (i.e., no prediction errors exist), thereby calculating an incorrect EV/EVCS schedule. Furthermore, DRL approaches for the scheduling of EVCS and VVC were implemented separately without considering their interdependence. Given the interdependence between the EVCS and VVC, the distorted EV/EVCS schedule due to such prediction errors has an detrimental impact on the performance of the VVC algorithm in the literature (L4).

To address all limitations from the literature (L1)∼(L4), we propose a two-stage DRL framework for the efficient coordination of smart EVCSs and VVC in power distribution systems while quickly responding to various prediction error scenarios for the PV generation outputs and aggregated EV loads in smart EVCSs. In this framework, smart EVCSs and VVC cooperate to maintain stable power distribution system operations while ensuring the profit of the smart EVCSs under various prediction error scenarios for the PV generation outputs and aggregated EV loads. The key point of the proposed two-stage DRL framework is the regulation of charging/discharging real/reactive powers of the ESSs of smart EVCSs to maximize the profits of the smart EVCSs and minimize the real power loss and voltage violations in power distribution systems. Thus, the main contributions of this study can be summarized as follows:

- Given a coupling between the smart EVCS operation and VVC under uncertain environment, we present a two-stage coordinated DRL model that schedules smart EVCSs (Stage 1) and VVC (Stage 2) to achieve profitable smart EVCS and stable power distribution grid operations simultaneously.
- We formulate the DRL algorithm for Stage 1. In comparison with existing DRL methods for the smart EVCS scheduling, our DRL method explicitly incorporates the prediction errors of PV generation output and aggregated EV load of smart EVCSs into the state space of the EVCS agents, thereby quickly responding to these various prediction errors. In Stage 1, each EVCS agent maximizes its profit by regulating the real power charging/discharging of its ESS while taking into account various prediction errors.
- We formulate the DRL algorithm for Stage 2. The error-induced real power consumption schedules of all EVCSs calculated by Stage 1 are embedded and updated in the state space of the VVC agent in Stage 2. Based on the updated state, the VVC agent minimizes the real power loss and voltage violation in the distribution grid by adjusting the reactive power of the ESSs of all EVCSs.
- Simulation results show that, compared to DRL approaches excluding the prediction errors of EVCSs, the proposed approach yields a greater increase of the profits of EVCSs and a greater reduction of the real power loss in the distribution grid.

The remainder of this paper is organized as follows. Section 2 introduces the power distribution system model and the SAC method based on an RL framework. The system model, with its mathematical notation and training procedure in the proposed two-stage DRL framework, is described in Section 3. Section 4 presents a two-stage DRL algorithm using the SAC method along with the formulation of the state/action spaces and reward functions for each stage. The simulation results for the IEEE 33-node and 123-node distribution systems are presented and analyzed in Section 5. The limitation of the presented DRL algorithm is discussed in Section 6. Finally, concluding remarks are presented in Section 7.

2. Backgrounds

2.1. Power distribution system model

Let us consider a power distribution system that includes a set $\mathcal{N}$ of nodes and a set $\mathcal{L}$ of distribution lines connecting nodes. The power flow model (Baran and Wu, 1989) of a radial distribution system is expressed as follows:

$$P_{ij,t} = P^c_{j,t} - P^g_{j,t} + \sum_{jk\in\mathcal{L}} P_{jk,t} + r_{ij} I^2_{ij,t} \tag{1}$$

$$Q_{ij,t} = Q^c_{j,t} - Q^g_{j,t} + \sum_{jk\in\mathcal{L}} Q_{jk,t} + x_{ij} I^2_{ij,t} \tag{2}$$

$$V^2_{j,t} = V^2_{i,t} - 2(r_{ij}P_{ij,t} + x_{ij}Q_{ij,t}) + (r^2_{ij} + x^2_{ij}) I^2_{ij,t} \tag{3}$$

$$I^2_{ij,t} V^2_{i,t} = P^2_{ij,t} + Q^2_{ij,t} \tag{4}$$

where $P_{ij,t}$ and $Q_{ij,t}$ are the real and reactive power flows from node i to j at time t, respectively; $P^c_{j,t}$ and $P^g_{j,t}$ are the real power consumption and generation at node j and time t, respectively; $Q^c_{j,t}$ and $Q^g_{j,t}$ are the reactive power consumption and generation at node j and time t, respectively; r_{ij} and x_{ij} are resistance and reactance of the line between nodes i and j; $I_{ij,t}$ is the current that flows from node i to j at time t; and $V_{i,t}$ denotes the voltage magnitude at node i and time t.

2.2. RL framework

An MDP is a fundamental environment in which RL approaches can be mathematically formulated. An MDP is defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{S}', \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{S}'$ denote the sets of states s_t and actions a_t at the current time t and states s_{t+1} at the next time $t+1$ for an RL agent in the given environment, respectively. $\mathcal{P}$ is a function that calculates the state transition probability of the agent; it can be expressed as $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to P(\mathcal{S})$, where $P(\mathcal{S})$ denotes the probability of transition $s_t \in \mathcal{S}$ into $s_{t+1} \in \mathcal{S}'$ by action $a_t \in \mathcal{A}$. $\mathcal{R}$ is a function that calculates a numerical reward for the agent when the agent moves from state s_t to state s_{t+1} according to the selected action a_t. The reward function is $\mathcal{R} : \mathcal{S}\times\mathcal{A}\times\mathcal{S}' \to \mathbb{R}$, where R_{t+1} is formulated as $R_{t+1} = \mathcal{R}(s_t, a_t, s_{t+1})$. The primary goal of the agent is to find the optimal policy $\boldsymbol{\pi}^*_{\boldsymbol{\theta}}$ parameterized with the weights $\boldsymbol{\theta}$ of the agent's neural network, which generates the largest numerical reward through the best selection of the action in a certain state. To evaluate the numerical value of the policy, we define the Q-value as $Q(s_t, a_t)$, which is expressed as $Q(s_t, a_t) = E[\sum_{k=0}^{T} \gamma^k R_{k+t+1} \mid s = s_t, a = a_t]$. In the Q-value, γ is the discount rate that represents the relative importance of the future reward to the current reward, and T is the terminal time of the agent's learning process. In summary, the RL agent aims to obtain the optimal policy $\boldsymbol{\pi}^*_{\boldsymbol{\theta}}$ that maximizes the Q-value by updating its parameter $\boldsymbol{\theta}$ given s_t and a_t as follows:

$$\boldsymbol{\pi}^*_{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} E\left[\sum_{k=0}^{T} \gamma^k R_{k+t+1} \mid s = s_t, a = a_t\right]. \tag{5}$$

2.3. SAC method

SAC (Haarnoja et al., 2018) is a state-of-the-art DRL method that determines continuous actions given continuous states. Compared to conventional DRL methods, the SAC method significantly improves its performance in terms of stability and sample efficiency. To this end, the agent using SAC maximizes the augmented Q-function with an entropy term $\mathcal{H}(\pi(a_t|s_t))$ to calculate the optimal policy $\boldsymbol{\pi}^{*,\mathrm{SAC}}_{\boldsymbol{\theta}}$ as follows:

$$\boldsymbol{\pi}^{*,\mathrm{SAC}}_{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} E\Big[\sum_{k=0}^{T} \gamma^k \{R_{k+t+1} + \zeta\,\mathcal{H}(\pi(a_t \mid s_t))\} \mid s = s_t, a = a_t\Big] \tag{6}$$

where $\mathcal{H}(\pi(a_t|s_t)) = -\sum_{a_t} \pi(a_t|s_t) \log \pi(a_t|s_t)$ represents the entropy value given the probability $\pi(a_t|s_t)$ of selecting action a_t in state s_t and ζ is a temperature coefficient that represents the relative importance between reward and entropy. Compared to conventional DRL-methods, the SAC method improves its performance in terms of exploration and sample efficiency by adding the entropy term (6). In addition, a replay buffer $\mathcal{B}$ is employed to further obtain stable convergence of the training curves of the SAC agent. The experience at each time is stored in the replay buffer $\mathcal{B}$ as follows: $\mathcal{B} \leftarrow (s, a, r, s') \cup \mathcal{B}$. To prevent correlation between samples, the agent randomly samples a tuple (s, a, r, s') from the replay buffer $\mathcal{B}$ and renews the weights $\boldsymbol{\theta}$.

SAC comprises four neural networks: a value network ($\boldsymbol{\theta}_v$), a target value network ($\widehat{\boldsymbol{\theta}_v}$), a critic network ($\boldsymbol{\theta}_c$), and an actor network ($\boldsymbol{\theta}_a$). The value, critic, and actor networks are updated by minimizing the following three loss functions, respectively:

$$L_{\mathrm{v}}(\boldsymbol{\theta}_v) = E\left[\frac{1}{2}(V_{\boldsymbol{\theta}_v}(s) - E[Q_{\boldsymbol{\theta}_c}(s,a) - \zeta \log \pi_{\boldsymbol{\theta}_a}(a|s)])^2\right] \tag{7}$$

$$L_{\mathrm{c}}(\boldsymbol{\theta}_c) = \frac{1}{2} E\left[Q_{\boldsymbol{\theta}_c}(s,a) - \widehat{\mathrm{Q}}(s,a)\right]^2 \tag{8}$$

$$L_{\mathrm{a}}(\boldsymbol{\theta}_a) = E[\log \pi_{\boldsymbol{\theta}_a}(h_{\boldsymbol{\theta}_a}(\kappa; s)|s) - Q_{\boldsymbol{\theta}_c}(s, h_{\boldsymbol{\theta}_a}(\kappa; s))]. \tag{9}$$

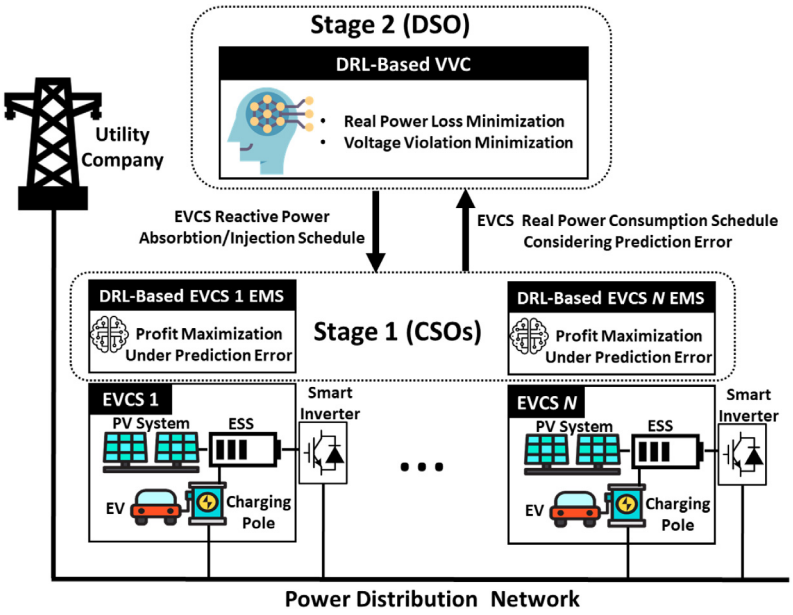

Fig. 1. Architecture of the proposed two-stage DRL framework.

For notation convenience, the time index t of the subscript of the variables in (7)–(9) is omitted. In (7), $V_{\boldsymbol{\theta}_v}(s)$ is state-value function which identifies the value of state by the policy $\boldsymbol{\theta}_v$, $Q_{\boldsymbol{\theta}_c}(s,a)$ is the Q-value of state s and action a by the policy $\boldsymbol{\theta}_c$, $\pi_{\boldsymbol{\theta}_a}(a|s)$ is the probability of state–action pair calculated from the policy $\boldsymbol{\theta}_a$. In (8), $\widehat{Q}(s,a)$ is the target value of $Q_{\boldsymbol{\theta}_c}(s,a)$. In (9), $h_{\boldsymbol{\theta}_a}(\boldsymbol{\kappa};s)$ is the actor network re-parameterized with the random noise vector $\boldsymbol{\kappa}$. In each training episode i, the weights $\boldsymbol{\theta}_x(i)$ of each network with its gradient $\nabla L_x(\boldsymbol{\theta}_x)$ and learning rate α_x are updated as follows: $\boldsymbol{\theta}_x(i+1) \leftarrow \boldsymbol{\theta}_x(i) - \alpha_x \nabla L_x(\boldsymbol{\theta}_x)$, where x represents v, c and a for the value, critic, and actor networks, respectively.

To further improve the training stability, the target value network is iteratively updated along with a periodic copy of the weights of the value network as follows:

$$\widehat{\boldsymbol{\theta}}_v(i+1) \leftarrow \delta\boldsymbol{\theta}_v(i+1) + (1-\delta)\widehat{\boldsymbol{\theta}}_v(i), \tag{10}$$

where $\widehat{\boldsymbol{\theta}}_v(i)$ is the weights for the target value network during training episode i, and δ is a smoothing parameter.

3. System model for the proposed two-stage DRL approach

3.1. Architecture of the proposed two-stage framework

Let us consider a system in which multiple smart EVCSs integrated with PV systems and ESSs are located in a power distribution grid. Smart EVCSs are connected to the power distribution grid via smart inverters of PV systems and ESSs. Using these inverters, smart EVCSs can absorb or inject their reactive power from or to the grid. As shown in Fig. 1, the proposed DRL-based framework consists of two stages that correspond to the operations of (i) DRL-based energy management systems (EMSs) in smart EVCSs in Stage 1; and (ii) DRL-based VVC in Stage 2. We consider the situation in which the charging station operators (CSOs) and distribution system operator (DSO) execute the DRL-based EMSs for smart EVCSs and DRL-based VVC, respectively. In the two-stage DRL framework, the CSOs and DSO interact with each other to achieve stable power distribution grid operation while ensuring economical smart EVCS operation. In Stage 1, each EVCS DRL agent of the EMS schedules charging and discharging operations of the ESS in the EVCS to maximize the profit of the EVCS while considering various prediction errors of the PV generation outputs and aggregated EV loads. Here, charging and discharging of the ESS correspond to energy bought from the grid and energy sold to EV users under TOU pricing, respectively. When the ESS discharging power is insufficient to fully support the EV loads, the EVCS agent can buy additional power directly from the grid. In Stage 2, using the grid operation data along with the real power consumption schedule of each EVCS transmitted from Stage 1, the VVC DRL agent schedules the reactive power absorption and injection of each EVCS from and to the grid via its smart inverter to minimize the total real power loss and voltage violations in the power distribution grid.

3.2. Notation

We denote the set of nodes as $\mathcal{N} = \mathcal{N}^{\text{EVCS}} \cup \mathcal{N}^{\text{non-EVCS}}$. $\mathcal{N}^{\text{EVCS}}$ and $\mathcal{N}^{\text{non-EVCS}}$ represent sets of nodes with and without EVCSs, respectively. In Stage 1, a day-ahead real power charging/discharging scheduling of the ESS in the EVCS is performed with a scheduling period $t \in \mathcal{T}^{(1)} = \{1,\ldots,T^{(1)}\}$ based on a 30-min resolution. In Stage 2, a real-time VVC is executed with a scheduling period $t \in \mathcal{T}^{(2)} = \{1,\ldots,T^{(2)}\}$ based on a 5-min resolution.

In Stage 1, $\forall\, n \in \mathcal{N}^{\text{EVCS}}$, $\forall\, t \in \mathcal{T}^{(1)}$, $P_{n,t}^{\text{ch/dch}}$ indicates the charging/discharging power schedule of the EVCS at node n and time t. $\overline{P}_n^{\text{ch/dch}}$ and $\underline{P}_n^{\text{ch/dch}}$ are the maximum and minimum capacities for charging/discharging of the EVCS at node n, respectively. The SOC of the EVCS at node n and time t is denoted by $SOC_{n,t}$, and its maximum and minimum capacities are $\overline{SOC}_n$ and $\underline{SOC}_n$, respectively. The SOC dynamics of the EVCS at node n and time t is expressed as $SOC_{n,t} = SOC_{n,t-1} + \frac{\eta_n^{\text{ch}} P_{n,t}^{\text{pch}} + \widehat{P}_{n,t}^{\text{PV}} + \Delta P_{n,t}^{\text{PV}}}{E_n^{\text{cap}}} - \frac{P_{n,t}^{\text{dch}}}{\eta_n^{\text{dch}} E_n^{\text{cap}}}$. In this SOC dynamics, $\eta_n^{\text{ch(dch)}}$ is the charging (discharging) efficiency of the EVCS, E_n^{cap} is the energy capacity of the EVCS, and $\widehat{P}_{n,t}^{\text{PV}}$ and $\Delta P_{n,t}^{\text{PV}}$ are the predicted real power output of the PV system in the EVCS and its corresponding prediction error, respectively. $\widehat{P}_{n,t}^{\text{EV}}$ and $\Delta P_{n,t}^{\text{EV}}$ are the predicted aggregated EV load and its corresponding prediction error for the EVCS at node n and time t, respectively. $\pi_{n,t}^{\text{sell(buy)}}$ denotes the selling (buying) price of the EVCS at node n and time t. The deficient power of the EVCS at node n and time t is defined as $P_{n,t}^{\text{EV,de}}$, which is the positive difference between

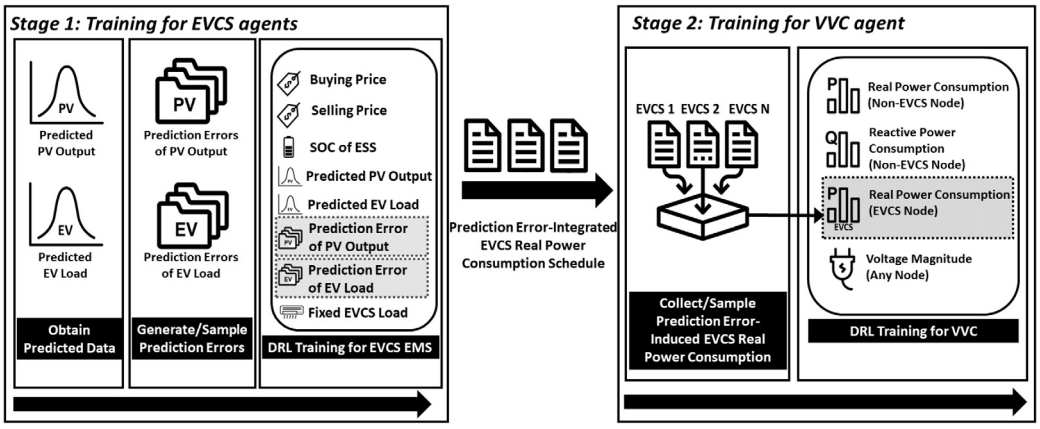

Fig. 2. Illustration of the training procedures of the proposed two-stage DRL algorithm.

the actual aggregated EV load $(\widehat{P}^{\text{EV}}_{n,t} + \Delta P^{\text{EV}}_{n,t})$ and the discharging power $(P^{\text{dch}}_{n,t})$ defined as follows: $P^{\text{EV,de}}_{n,t} = \widehat{P}^{\text{EV}}_{n,t} + \Delta P^{\text{EV}}_{n,t} - P^{\text{dch}}_{n,t}$. Note that each EVCS n can buy its deficient energy $P^{\text{EV,de}}_{n,t}$ from the grid directly and sell it back to EVs without discharging of ESS. The predicted fixed load (e.g., heating, ventilation, and air conditioning appliances) of the EVCS at node n and time t is denoted by $\widehat{P}^{\text{Fixed}}_{n,t}$.

In Stage 2, $\forall\ t \in \mathcal{T}^{(2)}$, the vectors of the real and reactive power consumptions at non-EVCS nodes are denoted by $\mathbf{P}^c_t = [P^c_{n,t}]$ and $\mathbf{Q}^c_t = [Q^c_{n,t}]$, respectively, where $P^c_{n,t}$ and $Q^c_{n,t}$ are the real and reactive power consumptions at node $n \in \mathcal{N}^{\text{non-EVCS}}$ and time t, respectively. The vector of voltage magnitudes at any node is defined as $\mathbf{V}_t = \left[V_{n,t}\right]$, where $V_{n,t}$ is the voltage magnitude at any node $n \in \mathcal{N}$ and time t. The vector of real power consumption schedules for all EVCSs at time t is denoted by $\mathbf{P}^{\text{EVCS}}_t = \left[P^{\text{EVCS}}_{n,t}\right]$, where $P^{\text{EVCS}}_{n,t}$ is the real power consumption of the EVCS at node $n \in \mathcal{N}^{\text{EVCS}}$ and time t; it is expressed as $P^{\text{EVCS}}_{n,t} = P^{\text{ch}}_{n,t} + P^{\text{EV,de}}_{n,t} + \widehat{P}^{\text{Fixed}}_{n,t}$. Note that $P^{\text{EVCS}}_{n,t}$ is used as the training data for the VVC agent in Stage 2 and is obtained by randomly sampling the prediction error-integrated EVCS real power consumption schedules that are generated using the trained model of the EVCS agents under various prediction error scenarios in Stage 1. The vector of reactive power generation or consumption schedules for all the EVCSs at time t is denoted by $\mathbf{Q}^{\text{EVCS}}_t = \left[Q^{\text{EVCS}}_{n,t}\right]$, where $Q^{\text{EVCS}}_{n,t}$ is the reactive power generation or consumption of the EVCS at node $n \in \mathcal{N}^{\text{EVCS}}$ and time t.

3.3. Training process description

As depicted in Fig. 2, the training procedure of the proposed two-stage DRL algorithm can be summarized as follows:

- Stage 1: As the preliminary step of Stage 1, a day-ahead predicted PV generation output and aggregated EV load data are obtained and their corresponding error data are generated and sampled. Then, for $t \in \mathcal{T}^{(1)}$, each EVCS DRL agent n trains its own neural network model and finds its optimal policy (i.e., charging/discharging schedule $P^{\text{ch/dch}}_{n,t}$ of the EVCS) using the following data: (i) buying/selling price $(\pi^{\text{buy}}_{n,t}/\pi^{\text{sell}}_{n,t})$, (ii) SOC $(SOC_{n,t-1})$ of ESS in the EVCS, (iii) prediction values $(\widehat{P}^{\text{PV}}_{n,t}, \widehat{P}^{\text{EV}}_{n,t})$ and prediction errors $(\Delta P^{\text{PV}}_{n,t}, \Delta P^{\text{EV}}_{n,t})$ of PV generation outputs and aggregated EV loads, and (iv) fixed EVCS load $(\widehat{P}^{\text{Fixed}}_{n,t})$. After completing the training of each EVCS agent in Stage 1, all EVCS agents send a set of their real power consumption schedules generated with various prediction errors to the VVC agent in Stage 2.
- Stage 2: Prediction error-integrated real power consumption schedules of EVCSs from Stage 1 are gathered and randomly sampled by the VVC agent. Then, for $t \in \mathcal{T}^{(2)}$, the VVC DRL agent trains its own neural network model and find its optimal policy (i.e., reactive power schedule $Q^{\text{EVCS}}_{n,t}$ of the EVCS) using the sampled real power consumption schedules of the EVCSs $(\mathbf{P}^{\text{EVCS}}_t)$ along with the real/reactive power consumption $(\mathbf{P}^c_t/\mathbf{Q}^c_t)$ at the non-EVCS nodes and voltage magnitude $(\mathbf{V}_{t-1})$ at any node.

4. Mathematical formulation of the proposed two-stage DRL approach

In this section, we formulate the state/action spaces and reward functions for the EVCS agents in Stage 1 and VVC agent in Stage 2.

4.1. DRL formulation of EVCS agent in stage 1

State space: For $t \in \mathcal{T}^{(1)} = \{1, \ldots, T^{(1)}\}$, state space $\mathcal{S}^{(1)}_{n,t}$ of EVCS agent at node n and time t is defined as follows:

$$\mathcal{S}^{(1)}_{n,t} = \{\pi^{\text{buy}}_{n,t}, \pi^{\text{sell}}_{n,t}, SOC_{n,t-1}, \widehat{P}^{\text{PV}}_{n,t}, \widehat{P}^{\text{EV}}_{n,t}, \Delta P^{\text{PV}}_{n,t}, \Delta P^{\text{EV}}_{n,t}, \widehat{P}^{\text{Fixed}}_{n,t}\} \tag{11}$$

where $\pi^{\text{buy(sell)}}_{n,t}$ is the buying (selling) energy price of EVCS n at time t; $SOC_{n,t-1}$ is the SOC of the ESS in the EVCS n at time $t-1$; $\widehat{P}^{\text{PV(EV)}}_{n,t}$ is the predicted PV generation output (EV aggregated load) of EVCS n at time t; $\Delta P^{\text{PV(EV)}}_{n,t}$ is the prediction error of PV generation output (EV aggregated load) of EVCS n at time t; and $\widehat{P}^{\text{Fixed}}_{n,t}$ is the predicted fixed load of EVCS n at time t.

Action space: Action space $\mathcal{A}^{(1)}_{n,t}$ of EVCS agent at node n and time t is expressed as follows:

$$\mathcal{A}^{(1)}_{n,t} = \{P^{\text{ch/dch}}_{n,t}\} \tag{12}$$

where $P^{\text{ch/dch}}_{n,t}$ represents the charging/discharging power of the ESS for EVCS n at time t.

Reward function: Reward function $R^{(1)}_{n,t}$ of EVCS agent at node n and time t is formulated as follows:

$$R^{(1)}_{n,t} = \pi^{\text{sell}}_{n,t} P^{\text{sell}}_{n,t} - \pi^{\text{buy}}_{n,t} P^{\text{buy}}_{n,t} - SOC^{\text{pen}}_{n,t} \tag{13}$$

where

$$P^{\text{sell}}_{n,t} = \widehat{P}^{\text{EV}}_{n,t} + \Delta P^{\text{EV}}_{n,t} = P^{\text{dch}}_{n,t} + P^{\text{EV,de}}_{n,t} \tag{14}$$

$$P^{\text{buy}}_{n,t} = P^{\text{ch}}_{n,t} + P^{\text{EV,de}}_{n,t} + \widehat{P}^{\text{Fixed}}_{n,t} \tag{15}$$

$$SOC^{\text{pen}}_{n,t} = \begin{cases} \mu_1(SOC^{\text{reg}}_n - SOC_{n,t}), & \text{if } SOC_{n,t} < SOC^{\text{reg}}_n \\ \mu_2(SOC_{n,t} - \overline{SOC}_n), & \text{if } SOC_{n,t} > \overline{SOC}_n \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The reward function in (13) consists of two parts: (i) the profit of the EVCS (i.e., the difference between the revenue $\pi_{n,t}^{\text{sell}} P_{n,t}^{\text{sell}}$ from selling power to EVs and the cost $\pi_{n,t}^{\text{buy}} P_{n,t}^{\text{buy}}$ from buying power from the grid) and (ii) the penalty cost $SOC_{n,t}^{\text{pen}}$ resulting from undercharging and overcharging the ESS. In (14), the selling power $P_{n,t}^{\text{sell}}$ is defined as the actual aggregated EV load $(\widehat{P}_{n,t}^{\text{EV}} + \Delta P_{n,t}^{\text{EV}})$, which is equal to the sum of the discharging power $(P_{n,t}^{\text{dch}})$ of the ESS and the deficient EV load $(P_{n,t}^{\text{EV,de}})$ that cannot be supported by the ESS discharging. In this study, if the discharging power $P_{n,t}^{\text{dch}}$ is less than the actual aggregated EV load $(\widehat{P}_{n,t}^{\text{EV}} + \Delta P_{n,t}^{\text{EV}})$, the EVCS directly buys the deficient EV load power $P_{n,t}^{\text{EV,de}}$ from the grid and sells it to EVs. In (15), the buying power $P_{n,t}^{\text{buy}}$ is expressed as the sum of the charging power $(P_{n,t}^{\text{ch}})$ of the ESS, deficient EV load $(P_{n,t}^{\text{EV,de}})$, and predicted fixed load of the EVCS $(\widehat{P}_{n,t}^{\text{Fixed}})$. The penalty $SOC_{n,t}^{\text{pen}}$ for the SOC of the ESS, defined in (16), avoids undercharging and overcharging the ESS with positive parameters μ_1 and μ_2. SOC_n^{reg} in (16) is denoted by $SOC_n^{\text{reg}} = \max\left\{SOC_n^{\text{reg}}, \underline{SOC}_n\right\}$, which is a predefined SOE regulation parameter of the EVCS at node n to constrain the current discharging capability of the EVCS to store sufficient energy in response to an unexpected power shortage.

4.2. DRL formulation of VVC agent in stage 2

State space: For $t \in \mathcal{T}^{(2)} = \{1, \ldots, T^{(2)}\}$, state space $\mathcal{S}_t^{(2)}$ of VVC agent at time t is defined as follows:

$$\mathcal{S}_t^{(2)} = \{\mathbf{P}_t^c, \mathbf{Q}_t^c, \mathbf{P}_t^{\text{EVCS}}, \mathbf{V}_{t-1}\} \tag{17}$$

where $\mathbf{P}_t^c$ and $\mathbf{Q}_t^c$ are the vectors of the real and reactive power consumptions at non-EVCS nodes $n \in \mathcal{N}^{\text{non-EVCS}}$ and time t, respectively; $\mathbf{P}_t^{\text{EVCS}}$ is the vector of the real power consumptions at EVCS nodes $n \in \mathcal{N}^{\text{EVCS}}$ and time t; and $\mathbf{V}_{t-1}$ is the vector of voltage magnitudes at every node $n \in \mathcal{N}$ and time $t-1$. Note that $\mathbf{P}_t^{\text{EVCS}}$ are the randomly sampled data of the prediction error-integrated real power consumption schedules of the EVCSs calculated in Stage 1.

Action space: Action space $\mathcal{A}_t^{(2)}$ of VVC agent at time t is expressed as follows:

$$\mathcal{A}_t^{(2)} = \{\mathbf{Q}_t^{\text{EVCS}}\}, \tag{18}$$

where $\mathbf{Q}_t^{\text{EVCS}} = \left[Q_{n,t}^{\text{EVCS}}\right]$ is the vector of reactive powers for all the EVCSs, where $Q_{n,t}^{\text{EVCS}}$ is the reactive power generation or consumption of the EVCS at node $n \in \mathcal{N}^{\text{EVCS}}$ and time t. The allowable range of $Q_{n,t}^{\text{EVCS}}$ is defined as $\left|Q_{n,t}^{\text{EVCS}}\right| \le \sqrt{(S_n)^2 - (P_{n,t}^{\text{EVCS}})^2}$, where S_n denotes the apparent power of the EVCS at node n.

Reward function: Reward function $R_t^{(2)}$ of VVC agent at time t is formulated as the sum of two negative cost functions:

$$R_t^{(2)} = -\beta_1 \sum_{n=1}^{|\mathcal{N}|} \Delta V_{n,t} - \beta_2 P_t^{\text{loss}}. \tag{19}$$

In (19), the first cost function $\sum_{n=1}^{|\mathcal{N}|} \Delta V_{n,t}$ represents the total voltage violation for all nodes, where $\Delta V_{n,t}$ is the deviation of the voltage magnitude from its admissible range $[\underline{V}, \overline{V}]$. The second cost function P_t^{loss} represents the total real power loss, which is expressed as $P_t^{\text{loss}} = \sum_{nm \in \mathcal{L}} r_{nm} \left[\frac{(P_{nm,t})^2 + (Q_{nm,t})^2}{(V_{n,t})^2}\right]$. β_1 and β_2 are the penalties for both negative cost functions. The penalty-based multi-reward function in (19) shows a trade-off relationship between the reduction of the total voltage violation and real power loss in terms of β_1 and β_2. On this trade-off relationship, the DSOs may adaptively adjust these penalties to situations in which they aim to reduce the total voltage violation or total real power loss further in distribution systems. For example, the DSO can adaptively adjust the penalty β_1 (or β_2) to reduce the total voltage violation (or real power loss) further with higher β_1 (or β_2).

Algorithm 1: Two-stage DRL algorithm using SAC method

1 Initialize $[\boldsymbol{\theta}_v^{(1)}, \widehat{\boldsymbol{\theta}_v^{(1)}}, \boldsymbol{\theta}_c^{(1)}, \boldsymbol{\theta}_a^{(1)}]$ and $[\boldsymbol{\theta}_v^{(2)}, \widehat{\boldsymbol{\theta}_v^{(2)}}, \boldsymbol{\theta}_c^{(2)}, \boldsymbol{\theta}_a^{(2)}]$
2 Each EVCS agent n receives $\{\widehat{P}_{n,t}^{\text{PV}}, \widehat{P}_{n,t}^{\text{EV}}\}$ and generates prediction errors
3 %%Training for an optimal charging/discharging scheduling of each EVCS agent
4 **for** *training episode i = 1, maximum training episode* **do**
5 **for** *time step t = 1,$T^{(1)}$* **do**
6 ▷ Randomly sample $\Delta P_{n,t}^{\text{PV}}$ and $\Delta P_{n,t}^{\text{EV}}$ from prediction error sets
7 ▷ Extract $P_{n,t}^{\text{ch/dch}}$ from the actor network based on $\pi_{\boldsymbol{\theta}_a^{(1)}}(a_{n,t}^{(1)}|s_{n,t}^{(1)})$
8 ▷ Compute the action and receive $R_{n,t+1}^{(1)}$ and $s_{n,t+1}^{(1)}$
9 ▷ Store the tuple $[s_{n,t}^{(1)}, a_{n,t}^{(1)}, R_{n,t+1}^{(1)}, s_{n,t+1}^{(1)}]$ in $\mathcal{B}^{(1)}$
10 **end**
11 ▷ Randomly sample batches of $[s^{(1)}, a^{(1)}, r^{(1)}, s'^{(1)}]$ from $\mathcal{B}^{(1)}$
12 ▷ Calculate $L_v(\boldsymbol{\theta}_v^{(1)})$, $L_a(\boldsymbol{\theta}_a^{(1)})$, and $L_c(\boldsymbol{\theta}_c^{(1)})$
13 ▷ Update the networks with the weight optimizer $\Upsilon_x^{(1)}$: $\nabla\boldsymbol{\theta}_x^{(1)} = \Upsilon_x^{(1)}\left(L_x(\boldsymbol{\theta}_x^{(1)})\right)$, $\boldsymbol{\theta}_x^{(1),\text{new}} \leftarrow \boldsymbol{\theta}_x^{(1),\text{old}} - \alpha_x^{(1)}\nabla\boldsymbol{\theta}_x^{(1)}$, $x = v$, a, and c
14 ▷ Update the target value network using an updated value network as follows: $\widehat{\boldsymbol{\theta}_v^{(1)}}(i+1) \leftarrow \delta^{(1)}\boldsymbol{\theta}_v^{(1)}(i+1) + (1-\delta^{(1)})\widehat{\boldsymbol{\theta}_v^{(1)}}(i)$
15 **end**
16 Each EVCS agent n transmits the set of prediction error-integrated EVCS real power consumption schedules to the VVC agent
17 %%Training for optimal EVCS reactive power scheduling of VVC agent
18 **for** *training episode i = 1, maximum training episode* **do**
19 ▷ Randomly sample $\mathbf{P}^{\text{EVCS}} = [P_{n,t}^{\text{EVCS}}]$ from the transmitted sets
20 **for** *time step t = 1,$T^{(2)}$* **do**
21 ▷ Extract $\mathbf{Q}_t^{\text{EVCS}}$ from the actor network based on $\pi_{\boldsymbol{\theta}_a^{(2)}}(a_t^{(2)}|s_t^{(2)})$
22 ▷ Compute the action and receive $R_{t+1}^{(2)}$ and $s_{t+1}^{(2)}$
23 ▷ Store tuple $[s_t^{(2)}, a_t^{(2)}, R_{t+1}^{(2)}, s_{t+1}^{(2)}]$ in $\mathcal{B}^{(2)}$
24 **end**
25 ▷ Conduct the same procedure in line 11∼14 with $[s_t^{(2)}, a_t^{(2)}, R_{t+1}^{(2)}, s_{t+1}^{(2)}]$, $\mathcal{B}^{(2)}$, $L_x(\boldsymbol{\theta}_x^{(2)})$, $\nabla\boldsymbol{\theta}_x^{(2)}$, $\Upsilon_x^{(2)}$, $\alpha_x^{(2)}$, $\widehat{\boldsymbol{\theta}_v^{(2)}}$, and $\delta^{(2)}$, $x = v$, a, and c
26 **end**

4.3. Algorithm description of two-stage SAC framework

The proposed two-stage DRL approach (Algorithm 1) is implemented using the SAC method, as explained in Section 2.3. In Algorithm 1, the superscripts (1) and (2) of both variables and parameters represent Stages 1 and 2, respectively. Algorithm 1 consists of three parts: (i) initialization of the weights of neural networks for SAC-based EVCS and VVC agents along with the generation of prediction errors of the PV generation outputs and aggregated EV loads (lines 1∼2); (ii) training procedure of each EVCS agent in Stage 1 (lines 4∼15); and (iii) training procedure of VVC agent in Stage 2 (lines 16∼26). In Stage 1, all SAC-based EVCS agents independently train their neural network model using their own data, including the prediction errors of PV generation outputs and aggregated EV loads, and find their optimal policy of charging and discharging schedule of the ESSs in the EVCSs. Various prediction error-integrated EVCS real power consumption schedules generated by the EVCS agents are randomly sampled and embedded into the elements of the state space for the SAC-based VVC agent in Stage 2. Then, using the sampled EVCS consumption schedule and distribution grid operation data, the VVC agent trains its neural network model and find its optimal policy of reactive power schedule for all the EVCSs. Fig. 3 depicts the SAC structure of the proposed two-stage DRL framework.

5. Simulation results

5.1. Simulation setup

The proposed two-stage DRL algorithm was applied to the IEEE 33-node and 123-node distribution systems (Kersting, 1991),

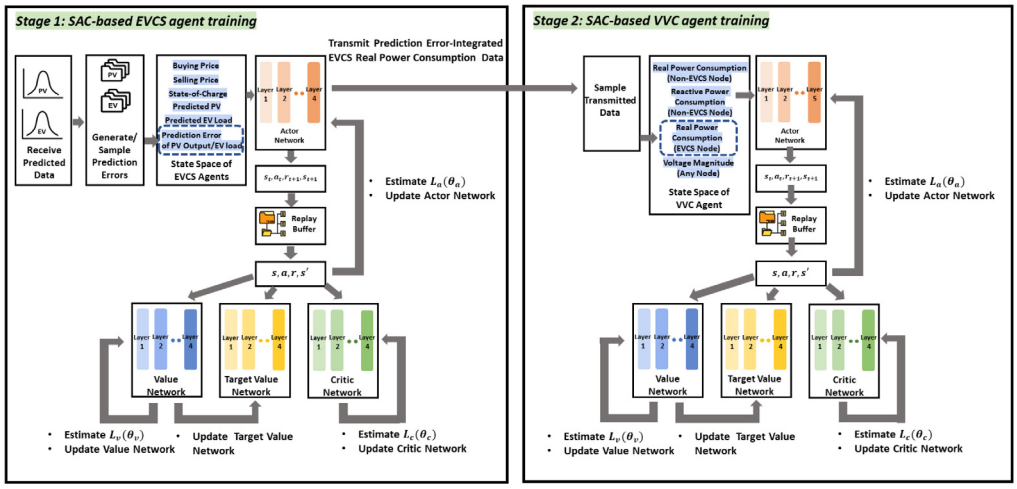

Fig. 3. Structure of the SAC method employed for the proposed two-stage DRL framework.

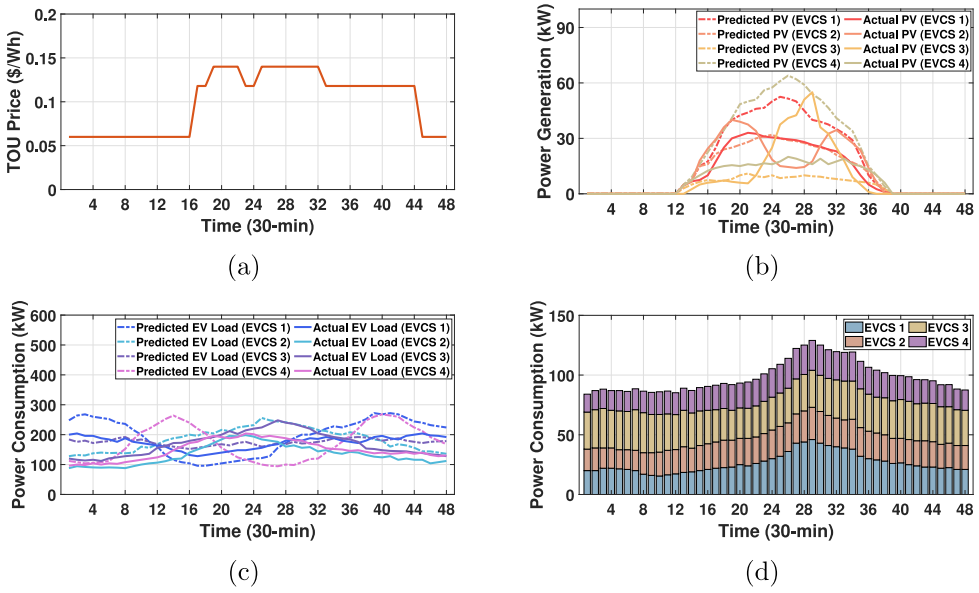

Fig. 4. Profiles of electricity price, PV generation output and EVCS load for four smart EVCSs in the IEEE 33-node distribution system: (a) TOU price, (b) predicted/actual PV generation outputs, (c) predicted/actual aggregated EV loads, and (d) predicted fixed loads.

which present four and eight smart EVCSs integrated with the PV systems and ESSs, respectively. EVCSs $1 \sim 4$ were connected to nodes 2, 10, 15, and 22 in the former system whereas EVCSs $1 \sim 8$ were connected to nodes 3, 8, 21, 34, 40, 48, 55, and 74 in the latter system. In this study, EVCS n represents an EVCS connected to node n. Concerning the ESS of each EVCS n, its capacity was set to $E_n^{\text{cap}} = 1$ MWh; the maximum and minimum charging/discharging powers were set to $\overline{P}_n^{\text{ch/dch}} = 300$ kW and $\underline{P}_n^{\text{ch/dch}} = 0$ kW, respectively; the initial SOC, maximum and minimum SOCs, and regulation SOC of the ESS were set to $SOC_{n,0} = 0.5$, $\overline{SOC}_n = 1.0$, $\underline{SOC}_n = 0.1$, and $SOC_n^{\text{reg}} = 0.1$, respectively; the charging and discharging efficiencies for the ESS were set to $\eta_n^{\text{ch}} = \eta_n^{\text{dch}} = 1.0$; and the size of the ESS was set to $S_n = 0.885$ MVA. Under a TOU pricing tariff as depicted in Fig. 4(a), each EVCS buys power from the grid and sells it to EVs with a gain factor of 1.5 with respect to the TOU price. Figs. 4(b) and (c) show the predicted/actual PV generation outputs and the predicted/actual aggregated EV loads for each EVCS n and time t in the IEEE 33-node distribution system, respectively. The data in these figures are referred to and modified from Zhang et al. (2020), Wood et al. (2018). Fifteen prediction error scenarios of the PV generation outputs and aggregated EV loads were generated by sampling the predicted data from the continuous uniform distributions $\mathcal{U}(0, 2\times \widehat{P}_{n,t}^{\text{PV}})$ and $\mathcal{U}(0, 2 \times \widehat{P}_{n,t}^{\text{EV}})$, respectively. The actual PV generation outputs and aggregated EV loads were used as test data for the EVCS agents. Twelve real power consumption scenarios generated by each EVCS agent in Stage 1 are transmitted to VVC agent for its training process in Stage 2. The predicted fixed load $\widehat{P}_{n,t}^{\text{Fixed}}$ for each EVCS n in the IEEE 33-node distribution system is shown in Fig. 4(d). The predicted/actual data for the PV generation outputs and aggregated EV loads of the eight EVCSs in the IEEE 123-node distribution system were also modified from Zhang et al. (2020), Wood et al. (2018), respectively. The prediction errors of the eight EVCSs were sampled from the aforementioned uniform distributions of the four EVCSs in the IEEE 33-node distribution system. In addition, the predicted fixed loads of the eight EVCSs were generated by inserting numerical noise into the data shown in Fig. 4(d). The maximum and minimum limits of the allowed voltage range for any node were set to $\overline{V} = 1.05$ and $\underline{V} = 0.95$ p.u., respectively.

Table 1
Specification of neural networks for SAC-based EVCS agent (Stage 1) and SAC-based VVC agent (Stage 2).

Stage	Neural network	Number of layers	Number of Neurons per Layer					Transfer function	Optimization method	Learning rate
			Layer 1	Layer 2	Layer 3	Layer 4	Layer 5			
Stage 1	Actor Network	4	256	256	256	128	–	Hyperbolic Tangent Function	ADAM Optimization Method	5×10^{-7}
	Critic Network	4	256	256	128	128	–	Sigmoid Function	ADAM Optimization Method	10^{-6}
	Value Network	4	256	256	128	128	–	Sigmoid Function	ADAM Optimization Method	10^{-6}
	Target Value Network	4	256	256	128	128	–	Sigmoid Function	ADAM Optimization Method	–
Stage 2	Actor Network	5	512	256	256	128	128	Hyperbolic Tangent Function	ADAM Optimization Method	10^{-6}
	Critic Network	4	512	256	256	128	–	Sigmoid Function	ADAM Optimization Method	5×10^{-5}
	Value Network	4	512	256	256	128	–	Sigmoid Function	ADAM Optimization Method	5×10^{-5}
	Target Value Network	4	512	256	256	128	–	Sigmoid Function	ADAM Optimization Method	–

Table 2
Classification for case studies.

Case	Stage 1		Stage 2
	Neural network	State space	Neural network
Case 1	ANN	$\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \notin \mathcal{S}^{(1)}_{n,t}$	ANN
Case 2	RNN	$\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \notin \mathcal{S}^{(1)}_{n,t}$	RNN
Case 3	GRU	$\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \notin \mathcal{S}^{(1)}_{n,t}$	GRU
Case 4	ANN	$\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \notin \mathcal{S}^{(1)}_{n,t}$, $\widetilde{P}^{\mathrm{PV}}_{n,t}, \widetilde{P}^{\mathrm{EV}}_{n,t} \in \mathcal{S}^{(1)}_{n,t}$	ANN
Case 5	RNN	$\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \in \mathcal{S}^{(1)}_{n,t}$	RNN
Case 6	GRU	$\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \in \mathcal{S}^{(1)}_{n,t}$	GRU
Proposed	ANN	$\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \in \mathcal{S}^{(1)}_{n,t}$	ANN

Table 1 provides the specifications of four SAC-based neural networks for EVCS and VVC agents in Stages 1 and 2, respectively. Since the complexity of the problem in Stage 2 is higher than that in Stage 1, an extra hidden layer is added to actor network of Stage 2 for ensuring stable training performance. For both stages, the ADAM optimization method (Kingma and Ba, 2014) was adopted to optimally update the weights of the neural networks. For the EVCS and VVC agents, the maximum sizes of the replay buffer and sampling sizes of the batch in the training were set to $\{40000, 50000\}$ and $\{48, 72\}$, respectively. The smoothing parameter in the target value network and the temperature coefficient in the augmented Q-function were set to $\delta = 0.2$ and $\zeta = 1$, respectively. The penalties in the reward functions of the EVCS and VVC agents were set to $\mu_1 = \mu_2 = 20$ and $\{\beta_1 = 1000, \beta_2 = 500\}$, respectively. The SAC algorithms for Stages 1 and 2 were run for 24 h with $T^{(1)} = 48$ (30-min scheduling resolution) and $T^{(2)} = 288$ (5-min scheduling resolution), respectively. The values of the maximum training episode for Stages 1 and 2 in Algorithm 1 were set to 4500 and 800, respectively. Simulations for both stages were conducted using pytorch 1.6.0 in Python 3.7.0.

Table 2 shows a classification of the results of our simulation study into six cases. Cases 1∼3 exclude the prediction errors of PV generation outputs and aggregated EV loads in the state space for each EVCS agent (i.e., $\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \notin \mathcal{S}^{(1)}_{n,t}$) having as neural networks an ANN, a recurrent neural network (RNN) (Williams et al., 1986), and a gated recurrent unit (GRU) (Cho et al., 2014), respectively. RNNs constitute an advanced approach with respect to ANNs to efficiently manage sequential data by employing a hidden state that implies iterative memory during the training of the agent. GRUs in turn constitute an advanced approach with respect to RNNs to solve the long-term dependencies during the training of the agent. Case 4 denotes a benchmarking method in which the actual PV generation outputs ($\widetilde{P}^{\mathrm{PV}}_{n,t} = \widehat{P}^{\mathrm{PV}}_{n,t} + \Delta P^{\mathrm{PV}}_{n,t}$) and aggregated EV loads ($\widetilde{P}^{\mathrm{EV}}_{n,t} = \widehat{P}^{\mathrm{EV}}_{n,t} + \Delta P^{\mathrm{EV}}_{n,t}$) belong to the elements of the state space for each EVCS agent (i.e., $\widetilde{P}^{\mathrm{PV}}_{n,t}, \widetilde{P}^{\mathrm{EV}}_{n,t} \in \mathcal{S}^{(1)}_{n,t}$) without their prediction errors (i.e., $\Delta P^{\mathrm{PV}}_{n,t}, \Delta P^{\mathrm{EV}}_{n,t} \notin \mathcal{S}^{(1)}_{n,t}$). Cases 5 and 6 correspond to the proposed approach, which explicitly embeds the prediction errors of PV generation outputs and aggregated EV loads into the state space of the EVCS agent; however, the ANN in the proposed approach was replaced by the RNN and GRU for Cases 5 and 6, respectively.

5.2. Performance evaluation

The performance results of the proposed approach are presented in the following four subsections.

- Section 5.2.1: The performance of four EVCS agents in Stage 1 was quantified in the IEEE 33-node distribution system in terms of charging and discharging power schedules and sensitivities of their SOCs and profits with respect to changes in the SOC regulation parameter.

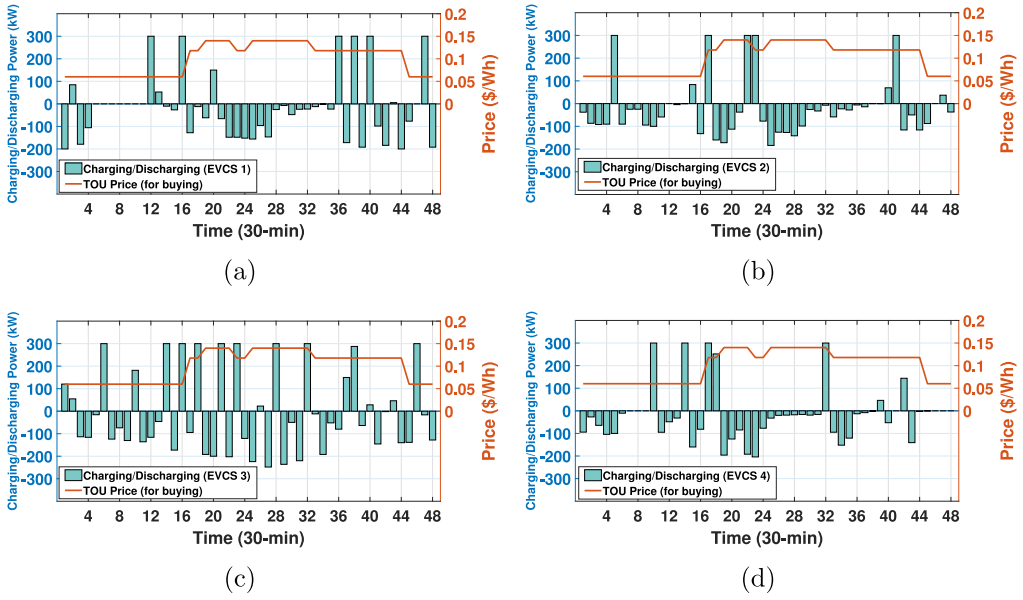

Fig. 5. Charging/discharging power ($P_{n,t}^{\text{ch/dch}}$) of four EVCSs in Stage 1 during a 24-h period: (a) EVCS 1, (b) EVCS 2, (c) EVCS 3, and (d) EVCS 4.

- Section 5.2.2: Through the comparison of Cases 1∼4 with the proposed approach, the performance of four EVCS agents in Stage 1 and a VVC agent in Stage 2 was assessed in the IEEE 33-node distribution system in terms of EVCS profit (Stage 1) and total real power loss and voltage magnitude (Stage 2).
- Section 5.2.3: Through the comparison of Cases 5 and 6 with the proposed approach, the performance improvement of the proposed approach resulting from the utilization of advanced neural networks was verified in the IEEE 33-node distribution system in terms of EVCS profit, total real power loss, and voltage magnitude.
- Section 5.2.4: The scalability of the proposed approach was tested in the IEEE 123-node distribution system. In addition, the convergence of the training curves of the agents along with their execution times in Stages 1 and 2, was validated in both the IEEE 33-node and 123-node distribution systems.

5.2.1. Performance results of EVCS agents

The performance of the EVCS agents considering various prediction errors were validated using actual data of PV generation output and aggregated EV load as shown in Figs. 4(b) and (c). Figs. 5 show the charging schedule (positive power consumption) and discharging schedule (negative power consumption) for the four EVCSs. The charging/discharging schedule ($P_{n,t}^{\text{ch/dch}}$) represents the action of each EVCS agent n. From these figures, we first observe that the discharging schedules of the EVCSs rely on actual data of the requested aggregated EV loads. For example, by comparing Figs. 4 and Figs. 5, it can be concluded that EVCS 1 discharges a larger amount of power than the other three EVCSs in the time period [00:30 a.m., 02:00 a.m.] owing to the request of a larger amount of aggregated EV loads for EVCS 1. In addition, as shown in Figs. 5(b) and (d), in general, EVCSs 2 and 4 discharge more power in the time period [08:00 a.m., 2:00 p.m.] than in the other time slots. This is because the agents of these EVCSs aim to maximize their profit by selling power to EVs in a high TOU pricing period.

Table 3 shows the sensitivities of the average SOC ($SOC_n^{\text{avg}}(x)$) and average profit deviation ($\Delta Pr_n(x)$) of EVCS n with respect to the varying SOC regulation parameter $x = SOC_n^{\text{reg}}$ during a total of $T^{(1)}$ scheduling periods, calculated using the following indices:

Table 3
Average SOCs ($SOC_n^{\text{avg}}(x)$) and average profit deviation ($\Delta Pr_n(x)$) of four EVCSs during a 24-h period.

Index	EVCS 1 (%)	EVCS 2 (%)	EVCS 3 (%)	EVCS 4 (%)
$SOC_n^{\text{avg}}(0.1)$	31.2	24.2	68.4	24.5
$SOC_n^{\text{avg}}(0.25)$	39.0	30.7	72.1	40.1
$SOC_n^{\text{avg}}(0.35)$	47.3	47.1	78.3	49.8
$\Delta Pr_n(0.25)$	−7.3	−4.9	−8.2	−8.3
$\Delta Pr_n(0.35)$	−14.5	−10.4	−12.8	−12.3

$$SOC_n^{\text{avg}}(x) = \frac{1}{T^{(1)}} \sum_{t=1}^{T^{(1)}} SOC_{n,t}(x) \times 100\% \tag{20}$$

$$\Delta Pr_n(x) = \frac{Pr_n(x) - Pr_n(0.1)}{Pr_n(0.1)} \times 100\% \tag{21}$$

where, for a given $x = SOC_n^{\text{reg}}$, $SOC_{n,t}(x)$ is the value of the SOC of EVCS n at time t, and $Pr_n(x) = \frac{1}{T^{(1)}} \sum_{t=1}^{T^{(1)}} (\pi_{n,t}^{\text{sell}} P_{n,t}^{\text{sell}} - \pi_{n,t}^{\text{buy}} P_{n,t}^{\text{buy}})$ is the average profit of EVCS n with respect to a total of $T^{(1)}$ scheduling periods when $x \neq 0.1$. Note from Table 3 that, as SOC_n^{reg} increases from $x = 0.1$ to $x = 0.35$, SOC_n^{avg} increases, whereas ΔPr_n decreases. This is because an increase in SOC_n^{reg} raises the minimum limit of the SOC according to $SOC_n^{\text{reg}} = \max\{SOC_n^{\text{reg}}, \underline{SOC}_n\}$ and curtails the EVCS selling power to EVs through the discharging process, thereby decreasing the profit of the EVCS. However, the limited discharging capability of the EVCS enables it to store more energy in the ESS to respond to unexpected power consumption in the future.

5.2.2. Performance comparison between cases 1∼4 and the proposed approach

In this subsection, the performance of the proposed two-stage DRL approach is evaluated and compared with that of Cases 1∼4 in Table 2 for the IEEE 33-node distribution system.

Table 4 shows three performance results: (i) the relative average profit (Rel. Avg. Profit) of the four EVCSs in the four cases for the proposed approach (Stage 1); (ii) the relative real power loss (Rel. Real Power Loss) in the four cases for the proposed approach (Stage 2); and (iii) the minimum ($V^{\min}$) and maximum

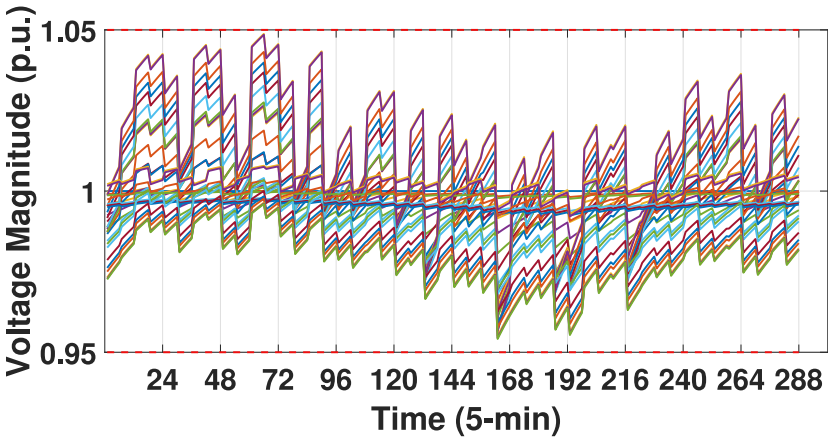

Fig. 6. Voltage profile for 33 nodes during a 24-h period in the proposed approach.

($V^{\max}$) voltage magnitudes in the four cases (Stage 2). In this table, the positive (negative) relative average profit and real power loss represent their increase (decrease) in each case with respect to the average profit and real power loss, respectively, for the proposed approach. The main observations from Table 4 can be summarized as follows:

(O1) No voltage violations occurred during an entire scheduling period for any of the four cases. In addition, Fig. 6 demonstrates that the proposed approach maintains a normal voltage profile within its acceptable range of $[\underline{V}, \overline{V}] = [0.95 \text{ p.u.}, 1.05 \text{ p.u.}]$.

(O2) Cases 1∼3 present negative relative average profits, which implies that the average profits in Cases 1∼3 are less than those in the proposed approach. This is due to the fact that the EVCS agents in the three cases for which the prediction errors of both the PV generation output and aggregated EV load are not considered do not respond to various prediction error scenarios adequately, and hence present low discharging efficiency. Here, the discharging efficiency of the EVCSs requires the utilization of discharging power of the ESSs to support EV loads, which is defined as $\frac{1}{T^{(1)}} \sum_{t=1}^{T^{(1)}} \frac{P_{t,n}^{\text{dch}}}{P_{t,n}^{\text{sell}}} \times 100\%$ where $P_{t,n}^{\text{dch}}$ and $P_{t,n}^{\text{sell}}$ represent the discharging and selling powers of EVCS n at time t. According to the defined discharging efficiency and (14) (i.e., $P_{n,t}^{\text{sell}} = P_{n,t}^{\text{dch}} + P_{n,t}^{\text{EV,de}}$), an increasing discharging efficiency implies that the EVCS can sell more ESS discharging power ($P_{n,t}^{\text{dch}}$) to EVs while buying less power ($P_{n,t}^{\text{EV,de}}$) from the grid, consequently leading to an increase of the EVCS profit. As expected, Fig. 7 shows that the discharging efficiency of each EVCS in the proposed approach is higher than that in Cases 1∼3.

(O3) Cases 1∼3 are listed in the following decreasing order of relative average profit: Case 3 (with GRU) > Case 2 (with RNN) > Case 1 (with ANN). This observation justifies that EVCS agents with a more advanced neural network can achieve greater profit. However, Cases 2 and 3 still have a negative relative profit, even though the neural networks in both cases (RNN and GRU, respectively) are more advanced than the neural network (ANN) of the proposed approach. This demonstrates that the prediction error embedded in the state space of the EVCS agent is more influential in terms of profit than having an advanced neural network structure for the EVCS agent.

(O4) The relative real power losses for Cases 1∼3 are positive (i.e., the real power loss in the proposed approach is less than that of Cases 1∼3). The three cases can be listed in increasing order of the real power loss as follows: Case 3 < Case 2 < Case 1. Note from (O2) that each EVCS agent with higher discharging efficiency makes more profit by discharging ESS power to EVs instead of purchasing power from the grid and selling it to EVs. This decreasing grid power request of the EVCS agent enables the VVC agent to further reduce real power loss. Furthermore, similar to the results in (O3), as the EVCS and VVC agent are implemented with more advanced neural networks, the real power loss can be further reduced.

(O5) The EVCS agent in Case 4 was trained using actual (test) data of the PV generation output and aggregated EV load. Therefore, Case 4 outperforms the proposed approach in terms of profit and real power loss. However, Case 4 cannot be deployed in an actual power distribution system because the prediction of the PV generation output and aggregated EV load without error is extremely difficult in real-world scenarios. The proposed approach performs more practical scheduling of EVCS operations and VVC under realistic situations with various prediction errors.

Table 4
Performance comparison between four cases (Cases 1∼4) and the proposed approach for the IEEE 33-node distribution system.

Case	Stage 1	Stage 2		
	Rel. Avg. Profit (%)	Rel. Real Power Loss (%)	$V^{\min}$ (p.u.)	$V^{\max}$ (p.u.)
Case 1	−24.25	4.1	0.9502	1.0451
Case 2	−18.98	3.1	0.9515	1.0463
Case 3	−17.98	2.9	0.9523	1.0470
Case 4	10.78	−2.4	0.9547	1.0470

5.2.3. Performance comparison between cases 5, 6 and the proposed approach

In this subsection, we investigate the impact of different neural networks for the SAC-based EVCS and VVC agents on the performance of the proposed approach for the IEEE 33-node distribution system.

Table 5 compares Cases 5 and 6 in terms of relative average profit/real power loss and voltage magnitude. According to Table 2, the EVCS and VVC agents in Cases 5 and 6 are implemented using advanced neural networks, namely RNN and GRU, which outperform the ANN used for the proposed approach. Note from Table 5 that in both cases the average profit and real power loss are improved with respect to the proposed approach while maintaining an acceptable voltage profile. Furthermore, a greater profit and less real power loss are achieved in Case 6 than in Case

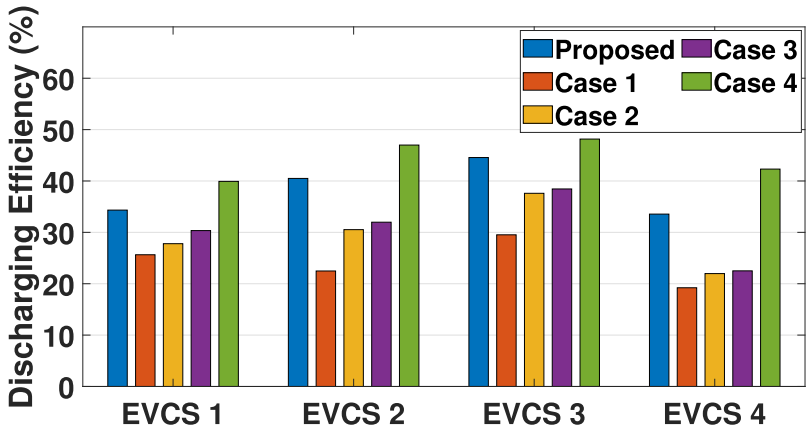

Fig. 7. Discharging efficiencies of four EVCSs for Cases 1∼4 and the proposed approach.

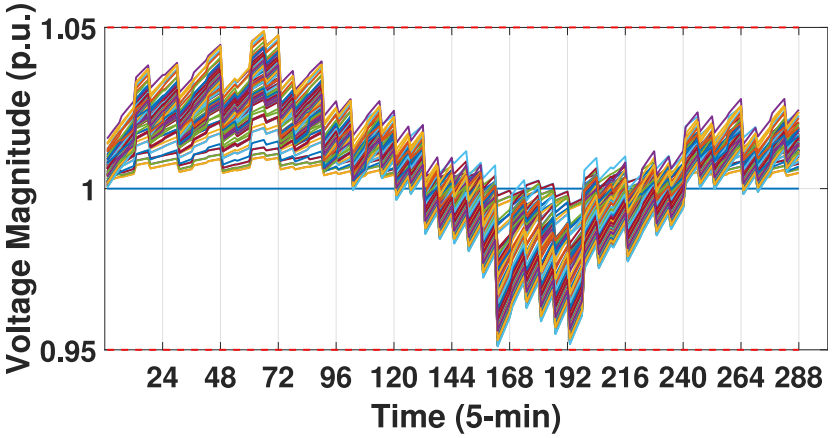

Fig. 8. Voltage profile for 123 nodes during a 24-h period in the proposed approach.

5. Therefore, we conclude from the aforementioned observations that SAC-based EVCS and VVC agents built with more advanced neural networks contribute to the performance improvement of the proposed approach.

Recently, many studies have proposed new DRL methods combined with advanced neural network structures to improve the training performance of DRL agents. In this context, the proposed two-stage DRL approach provides a universal framework to ensure both the profitability of EVCSs and the stability of power distribution grids in an environment with various prediction errors. It can be extended to a framework with better performance by employing DRL methods with upgraded neural networks.

5.2.4. Analysis of scalability, training convergence, and computation time

In this subsection, the performance of the proposed approach is analyzed in terms of scalability, training convergence, and computation time. Concerning scalability, the proposed approach was tested using an IEEE 123-node distribution system. Fig. 8 shows voltage magnitudes at any node and time during a 24-h period when the proposed approach is executed. Note from this figure that the proposed approach maintains a normal voltage profile within its acceptable range, $[\underline{V}, \overline{V}] = [0.95 \text{ p.u.}, 1.05 \text{ p.u.}]$, even in a larger power distribution system. Fig. 9 compares the relative average profit of eight EVCSs and real power loss in Cases 1∼6 with respect to the proposed approach for the IEEE 123-node distribution system. All observations for the IEEE 33-node distribution system pointed out in Sections 5.2.2 and 5.2.3 were also verified for the larger IEEE 123-node distribution system: (i) less profitability and less real power loss reduction (Cases 1∼3) and greater profitability and real power loss reduction (Cases 4∼6); (ii) the impact of different neural networks on the profit and real power loss can be expressed in decreasing order of performance as follows: Case 3 > Case 2 > Case 1 and Case 6 > Case 5; and (iii) the greatest profitability with highest real power loss reduction is achieved in Case 4.

Table 5
Performance comparison between two cases (Cases 5 and 6) and the proposed approach for the IEEE 33-node distribution system.

Case	Stage 1	Stage 2		
	Rel. Avg. Profit (%)	Rel. Real Power Loss (%)	$V^{\min}$ (p.u.)	$V^{\max}$ (p.u.)
Case 5	3.1	−1.1	0.9508	1.0488
Case 6	3.5	−1.2	0.9504	1.0491

In addition, the performance of the proposed SAC-based DRL method is compared with that of the actor-critic (AC) and advantage actor-critic (A2C) methods (Silver et al., 2014) that are also well known as the DRL approaches that determine continuous actions given continuous states. In this performance comparison, the AC, A2C, and SAC methods were implemented using GRU. Figs. 10 compare the training curves of the EVCS agents (Stage 1) and VVC agent (Stage 2) between the AC, A2C, and SAC methods in the IEEE 33-node and 123-node distribution systems. Note from these figures that the AC and A2C methods show a poor performance of the convergence. By contrast, the proposed SAC method shows that the training curves increase and converge well to optimal policy of the EVCS and VVC agents during the

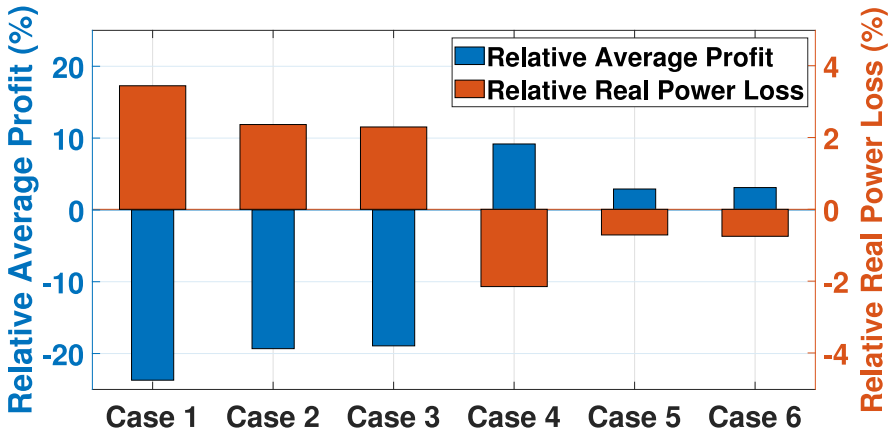

Fig. 9. Relative average profit and real power loss among six cases (Cases 1∼6) for the IEEE 123-node distribution system.

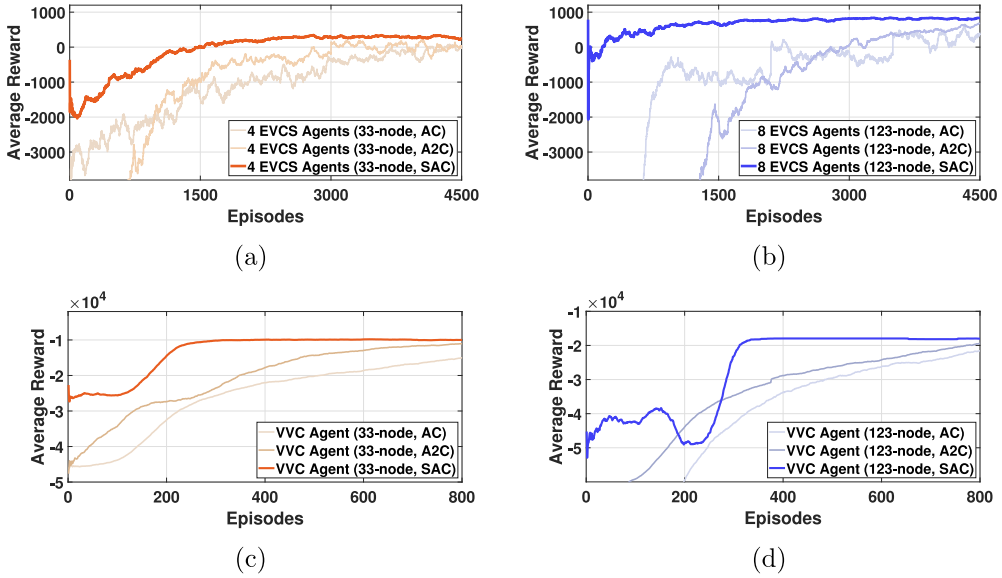

Fig. 10. Comparison of average reward convergence between the AC, A2C, and SAC methods for the IEEE 33-node and 123-node distribution systems: (a) EVCS agents in Stage 1 (33-node), (b) EVCS agents in Stage 1 (123-node), (c) VVC agent in Stage 2 (33-node), and (d) VVC agent in Stage 2 (123-node).

training period in both distribution systems. After completing their training process, the EVCS and VVC agents calculate their optimal actions (i.e., charging/discharging power and reactive power schedules of EVCSs) based on the trained models during the execution process. The execution times of Stages 1 and 2 for the IEEE 33-node and 123-node distribution systems can be summarized as follows: (i) 0.039 s in Stage 1 and 3.182 s in Stage 2 for the IEEE 33-node distribution system and (ii) 0.041 s in Stage 1 and 3.891 s in Stage 2 for the IEEE 123-node distribution system. Note that Stages 1 and 2 in our system model are assumed to be carried out with 30-min and 5-min scheduling resolutions, respectively. Therefore, the proposed approach is computationally efficient and applicable to the system model.

In addition, the result show that, in contrast with the AC and A2C methods, the proposed SAC method increases the EVCS profit (Stage 1) by 21.7% and 19.4% in the IEEE 33-node distribution system, respectively, and increases 20.3% and 19.9% in the IEEE 123-node distribution system, respectively. The proposed SAC method decreases the real power loss (Stage 2) by 4.3% and 4.1% in the IEEE 33-node distribution system, respectively, and decreases 2.1% and 2.0% in the IEEE 123-node distribution system, respectively. Furthermore, the performance of the proposed SAC-based DRL method is compared with that of the model-based optimization method. The model-based optimization methods for Stages 1 and 2 are formulated as the mixed-integer linear programming and mixed-integer second-order cone programming problems, respectively. These two model-based optimization methods are selected as the benchmarking algorithms that yield optimal solution of Stages 1 and 2 when no prediction errors of PV generation output and EV load occur. The simulation results show that, in comparison with the optimization methods, the proposed DRL method in Stage 1 decreases the average profit of EVCSs by 12.3% and 13.7% in the IEEE 33-node and 123-node distribution systems, respectively, and increases the real power loss by 3.5% and 2.3% in the IEEE 33-node and 123-node distribution systems, respectively. These degraded performance results of the proposed DRL method over the optimization methods are natural because the optimization methods have an assumption that the predictions of PV generation output and EV load are accurate. However, such assumption is too strict and unrealistic because it is impossible to predict PV generation output and EV load with 100% accuracy. In addition, the solving times for the optimization method in Stage 2 are 211 s and 314 s in the IEEE 33-node and 123-node distribution systems; however, the execution times for the proposed DRL method in Stage 2 are 3.182 s and 3.891 s that are much faster than the solving

Table 6
Performance comparison between the baseline approaches (AC, A2C, and optimization) and the proposed DRL approach for the IEEE 33-node and 123-node distribution systems.

Baseline method			AC	A2C	Optimization
Stage 1	Rel. Avg. Profit Increase (%)	33-node	21.7	19.4	−12.3
		123-node	20.3	19.9	−13.7
Stage 2	Rel. Real Power Loss Decrease (%)	33-node	4.3	4.1	−3.5
		123-node	2.1	2.0	−2.3

times of the optimization method. In summary, compared to the optimization method without considering the prediction error, the proposed DRL method calculates the solution more rapidly while responding to various real-time prediction errors. Table 6 summarizes the relative average profit increase and real power loss decrease of the proposed approach in Stages 1 and 2 with respect to the baseline methods (AC, A2C, and optimization) in the IEEE 33-node and 123-node distribution systems.

Lastly, the advantages and main observations of the proposed approach can be summarized as follows:

- The proposed two-stage DRL framework can further increase the profits of smart EVCSs and decrease the real power losses of power distribution systems by incorporating the prediction errors of the PV generation outputs and aggregated EV loads into the state space of EVCS agents.
- Compared to Cases 1∼3, which do not consider the prediction errors, the proposed approach leads to further increase of the total profit of EVCSs and reduction of the real power loss by 24.25% and 4.1% (Case 1), 18.98% and 3.1% (Case 2), and 17.98% and 2.9% (Case 3), respectively, for the IEEE 33-node distribution system (See Table 4).
- The proposed approach integrating more advanced neural networks (RNN and GRU) results in further increase of the total profit of EVCSs and reduction of the real power loss by 3.1% and 1.1% (Case 5 with RNN) and 3.5% and 1.2% (Case 6 with GRU), respectively, for the IEEE 33-node distribution system (See Table 5).
- From the perspective of scalability, the merits of the proposed approach observed in the IEEE 33-node distribution system were also verified in the IEEE 123-node distribution system (See Fig. 9).

6. Discussions

This study is motivated by a desire to develop a CSO–DSO coordinated DRL framework that achieves: (i) profitable smart EVCS operation from the perspective of CSO and (ii) stable power distribution grid operation from the perspective of DSO. To satisfy the goals of CSO and DSO simultaneously, ESS in smart EVCS is a crucial device, which serves as a buffer between the power distribution grid and EVs to maximize the EVCS profit given the aggregated EV load and perform voltage regulation by exploiting real/reactive charging/discharging capability of ESS. In this context, this study (the DRL algorithm in Stage 1) is limited to the charging/discharging scheduling of ESS in smart EVCS without considering the charging/discharging scheduling of individual EVs. Previous studies in Chaudhari et al. (2018), Datta et al. (2020) similar to this study focused on the charging/discharging role of ESS in smart EVCS to minimize the electricity purchase cost and the degradation of ESS and avoid transformer overloading while fulfilling the aggregated EV load completely; in these previous studies, the charging/discharging scheduling of individual EVs was not considered. Nonetheless, an important extension of this study would be to incorporate the scheduling of individual EVs into our DRL framework in Stage 1. A key part of this task would be to expand state/action space and reward function for EVCS agent in terms of individual EVs scheduling. However, the performance analysis of EVCS agent with the expanded state/action space and reward function is beyond the scope of this study and it is referred to as our future work.

7. Conclusions

This paper presents a two-stage DRL framework in which multiple PV-ESS integrated smart EVCSs and VVC are coordinated to simultaneously maximize the profit of EVCSs and minimize the real power loss and voltage violations in power distribution grids under uncertain operating environments of EVCSs with various prediction errors. In the first stage, each EVCS agent trains its robust neural network model against the prediction errors, and calculates the profitable charging/discharging schedule of the ESS in the EVCS. The profitable charging/discharging schedules calculated by all EVCS agents are used by the VVC agent in the second stage, which trains its neural network model and calculates the reactive power schedules of the ESSs in all EVCSs to minimize the real power loss and voltage violations. Numerical examples demonstrate that, in comparison with DRL methods that do not consider prediction errors, the proposed DRL method increases the EVCS profit by 17.9∼24.2% and 18.9∼23.7% and decreases the real power loss by 2.9∼4.1% and 2.3∼3.4% in the IEEE 33-node and 123-node distribution systems, respectively. The impact of different neural networks of the agents on the performance of the proposed framework is also analyzed. The results show that the RNN-and GRU-based agents increases the EVCS profit by 3.1∼3.5% and 2.9∼3.1% and decreases the real power loss by 1.1∼1.2% and 0.71∼0.75% in the IEEE 33-node and 123-node distribution systems, respectively. Furthermore, the performance of the proposed SAC method is compared with that of the other DRL methods (the AC and A2C methods). The results show that, in comparison with the AC and A2C methods, the proposed SAC method increases the EVCS profit by 19.4∼21.7%, and decreases the real power loss by 2.0∼4.3% in both distribution systems.

In future studies, the proposed method will be extended to a more practical DRL framework in which conventional voltage regulators, including OLTCs and CBs, will be coordinated with the smart inverters of EVCSs to ensure profitable operation of the EVCSs and efficient voltage regulation in realistic unbalanced distribution systems. In the future study, the DRL algorithm that schedules the optimal charging and discharging of individual EVs will be incorporated into the proposed two-stage DRL framework to minimize their charging cost while satisfying the preferences of EV users.

CRediT authorship contribution statement

Sangyoon Lee: Conceptualization, Methodology, Investigation, Writing – original draft. **Dae-Hyun Choi:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article

Acknowledgments

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2022R1F1A1062888, and in part by the NRF funded by the Korea government (MSIT) unded Grant 2021R1A4A1031019.

References

Baran, M., Wu, F.F., 1989. Optimal sizing of capacitors placed on a radial distribution system. IEEE Trans. Power Deliv. 4 (1), 735–743. http://dx.doi.org/10.1109/61.19266.

Cao, D., et al., 2022. Deep reinforcement learning enabled physical-model-free two-timescale voltage control method for active distribution systems. IEEE Trans. Smart Grid 13 (1), 149–165. http://dx.doi.org/10.1109/TSG.2021.3113085.

Chaudhari, K., Ukil, A., Kumar, K.N., Manandhar, U., Kollimalla, S.K., 2018. Hybrid optimization for economic deployment of ESS in PV-integrated EV charging stations. IEEE Trans. Ind. Inform. 14 (1), 106–116. http://dx.doi.org/10.1109/TII.2017.2713481.

Cho, K., et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches. pp. 1–9, arXiv preprint arXiv:1409.1259.

Datta, U., Kalam, A., Shi, J., 2020. Smart control of BESS in PV integrated EV charging station for reducing transformer overloading and providing battery-to-grid service. J. Energy Storage 28 (101224), 1–10. http://dx.doi.org/10.1016/j.est.2020.101224.

Dorokhova, M., Martinson, Y., Ballif, C., Wyrsch, N., 2021. Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation. Appl. Energy 301, 117504. http://dx.doi.org/10.1016/j.apenergy.2021.117504.

Farzin, H., Moeini-Aghtaie, M., Fotuhi-Firuzabad, M., 2016. Reliability studies of distribution systems integrated with electric vehicles under battery-exchange mode. IEEE Trans. Power Deliv. 31 (6), 2473–2482. http://dx.doi.org/10.1109/TPWRD.2015.2497219.

Felix, T., Niklas, E., Jonas, S., Marco, P., 2021. Development and evaluation of a smart charging strategy for an electric vehicle fleet based on reinforcement learning. Appl. Energy 285, 116382. http://dx.doi.org/10.1016/j.apenergy.2020.116382.

Gao, Y., Wang, W., Yu, N., 2021. Consensus multi-agent reinforcement learning for Volt-VAR control in power distribution networks. IEEE Trans. Smart Grid 12 (4), 3594–3604. http://dx.doi.org/10.1109/TSG.2021.3058996.

Haarnoja, T., Zhou, A., Abbeel, P., Levine, S., 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning. ICML, Stockholm, Sweden, pp. 1861–1870.

Hu, Y., Liu, W., Wang, W., 2020. A two-layer Volt-VAR control method in rural distribution networks considering utilization of photovoltaic power. IEEE Access 8, 118417–118425. http://dx.doi.org/10.1109/ACCESS.2020.3003426.

Jabr, R.A., 2019. Robust Volt/VAR control with photovoltaics. IEEE Trans. Power Syst. 34 (3), 2401–2408. http://dx.doi.org/10.1109/TPWRS.2018.2890767.

Jin, J., Xu, Y., 2021. Optimal policy characterization enhanced actor-critic approach for electric vehicle charging scheduling in a power distribution network. IEEE Trans. Smart Grid 12 (2), 1416–1428. http://dx.doi.org/10.1109/TSG.2020.3028470.

Katrin, S., Patrick, J., Wolf, F., 2019. Two-stage stochastic optimization for cost-minimal charging of electric vehicles at public charging stations with photovoltaics. Appl. Energy 242, 769–781. http://dx.doi.org/10.1016/j.apenergy.2019.03.036.

Kersting, W., 1991. Radial distribution test feeder. IEEE Trans. Power Syst. 6 (3), 975–985. http://dx.doi.org/10.1109/59.119237.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. pp. 1–15, arXiv preprint arXiv:1412.6980.

Kriekinge, G.V., Cauwer, C.D., Sapountzoglou, N., Coosemans, T., Messagie, M., 2021. Peak shaving and cost minimization using model predictive control for uni- and bi-directional charging of electric vehicles. Energy Rep. 7, 8760–8771. http://dx.doi.org/10.1016/j.egyr.2021.11.207.

Lee, S., Choi, D.-H., 2021. Dynamic pricing and energy management for profit maximization in multiple smart electric vehicle charging stations: A privacy-preserving deep reinforcement learning approach. Appl. Energy 304, 117754. http://dx.doi.org/10.1016/j.apenergy.2021.117754.

Li, C., Zhang, L., Ou, Z., Wang, Q., Zhou, D., Ma, J., 2022. Robust model of electric vehicle charging station location considering renewable energy and storage equipment. Energy 238, 121713. http://dx.doi.org/10.1016/j.energy.2021.121713.

Li, D., Zouma, A., Liao, J.-T., Yang, H.-T., 2020. An energy management strategy with renewable energy and energy storage system for a large electric vehicle charging station. eTransportation 6, 100076. http://dx.doi.org/10.1016/j.etran.2020.100076.

Liu, H., Zhang, C., Chai, Q., Meng, K., Guo, Q., Dong, Z.Y., 2021. Robust regional coordination of inverter-based Volt/Var control via multi-agent deep reinforcement learning. IEEE Trans. Smart Grid 12 (6), 5420–5433. http://dx.doi.org/10.1109/TSG.2021.3104139.

Nguyen, H.T., Choi, D.H., 2022. Three-stage inverter-based peak shaving and Volt-VAR control in active distribution networks using online safe deep reinforcement learning. IEEE Trans. Smart Grid 13 (4), 3266–3277. http://dx.doi.org/10.1109/TSG.2022.3166192.

Shin, M., Choi, D.H., Kim, J., 2020. Cooperative management for PV/ESS-Enabled electric vehicle charging stations: A multi-agent deep reinforcement learning approach. IEEE Trans. Ind. Inform. 16 (5), 3493–3503. http://dx.doi.org/10.1109/TII.2019.2944183.

Sierra, A., Gercek, C., Geurs, K., Reinders, A., 2020. Technical, financial, and environmental feasibility analysis of photovoltaic EV charging stations with energy storage in China and the United States. IEEE J. Photovolt. 10 (6), 1892–1899. http://dx.doi.org/10.1109/JPHOTOV.2020.3019955.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M., 2014. Deterministic policy gradient algorithms. In: International Conference on Machine Learning. PMLR, pp. 387–395.

Sun, X., Qiu, J., 2021. Two-stage Volt/VAR control in active distribution networks with multi-agent deep reinforcement learning method. IEEE Trans. Smart Grid 12 (4), 2903–2912. http://dx.doi.org/10.1109/TSG.2021.3052998.

Sun, X., Qiu, J., Zhao, J., 2021. Real-time Volt/VAR control in active distribution networks with data-driven partition method. IEEE Trans. Power Syst. 36 (3), 2248–2461. http://dx.doi.org/10.1109/TPWRS.2020.3037294.

Tao, Y., Qiu, J., Lai, S., Sun, X., Zhao, J., Zhou, B., Cheng, L., 2022. Data-driven on-demand energy supplement planning for electric vehicles considering multi-charging/swapping services. Appl. Energy 311, 118632. http://dx.doi.org/10.1016/j.apenergy.2022.118632.

Vineeth, V., Abheejeet, M., S.N., S., 2021. Demand response with Volt/VAR optimization for unbalanced active distribution systems. Appl. Energy 300, 117361. http://dx.doi.org/10.1016/j.apenergy.2021.117361.

Wang, K., Wang, H., Yang, J., Feng, J., Li, Y., Zhang, S., Okoye, M.O., 2022. Electric vehicle clusters scheduling strategy considering real-time electricity prices based on deep reinforcement learning. Energy Rep. 8, 695–703. http://dx.doi.org/10.1016/j.egyr.2022.01.233.

Wang, W., Yu, N., Gao, Y., Shi, J., 2020a. Safe off-policy deep reinforcement learning algorithm for Volt-VAR control in power distribution systems. IEEE Trans. Smart Grid 11 (4), 3008–3018. http://dx.doi.org/10.1109/TSG.2019.2962625.

Wang, Y., Zhao, T., Ju, C., Xu, Y., Wang, P., 2020b. Two-level distributed Volt/Var control using aggregated PV inverters in distribution networks. IEEE Trans. Power Deliv. 35 (4), 1844–1855. http://dx.doi.org/10.1109/TPWRD.2019.2955506.

Wang, Y., Zhao, T., Ju, C., Xu, Y., Wang, P., 2020c. Two-level distributed Volt/VAR control using aggregated PV inverters in distribution networks. IEEE Trans. Power Deliv. 35 (4), 1844–1855. http://dx.doi.org/10.1109/TPWRD.2019.2955506.

Wenjie, Z., Oktoviano, G., Hao, Q., Carlos, R., Dipti, S., 2018. A multi-agent based integrated Volt-VAR optimization engine for fast vehicle-to-grid reactive power dispatch and electric vehicle coordination. Appl. Energy 229, 96–110. http://dx.doi.org/10.1016/j.apenergy.2018.07.092.

Williams, R.J., Hinton, G.E., David, D.E., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536. http://dx.doi.org/10.1038/323533a0.

Wood, E.W., et al., 2018. California plug-in electric vehicle infrastructure projections: 2017–2025-future infrastructure needs for reaching the state's zero emission-vehicle deployment goals. no. nrel/tp-5400-70893.

Yan, L., Chen, X., Chen, Y., Wen, J., 2022. A cooperative charging control strategy for electric vehicles based on multi-agent deep reinforcement learning. IEEE Trans. Ind. Inform. 18 (12), 8765–8775. http://dx.doi.org/10.1109/TII.2022.3152218.

Yan, L., Chen, X., Zhou, J., Chen, Y., Wen, J., 2021a. Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors. IEEE Trans. Smart Grid 12 (6), 5124–5134. http://dx.doi.org/10.1109/TSG.2021.3098298.

Yan, D., Yin, H., Li, T., Ma, C., 2021b. A two-stage scheme for both power allocation and EV charging coordination in a grid-tied PV–Battery charging station. IEEE Trans. Ind. Inform. 17 (10), 6994–7004. http://dx.doi.org/10.1109/TII.2021.3054417.

Yan, Q., Zhang, B., Kezunovic, M., 2019. Optimized operational cost reduction for an EV charging station integrated with battery energy storage and PV generation. IEEE Trans. Smart Grid 10 (2), 2096–2106. http://dx.doi.org/10.1109/TSG.2017.2788440.

Yang, M., Zhang, L., Zhao, Z., Wang, L., 2021. Comprehensive benefits analysis of electric vehicle charging station integrated photovoltaic and energy storage. J. Clean. Prod. 302, 126967. http://dx.doi.org/10.1016/j.jclepro.2021.126967.

Yuanqi, G., Nanpeng, Y., 2022. Model-augmented safe reinforcement learning for Volt-VAR control in power distribution networks. Appl. Energy 313, 118762. http://dx.doi.org/10.1016/j.apenergy.2022.118762.

Zhang, Z., Li, R., Li, F., 2020. A novel peer-to-peer local electricity market for joint trading of energy and uncertainty. IEEE Trans. Smart Grid 11 (2), 1205–1215. http://dx.doi.org/10.1109/TSG.2019.2933574.

Zhang, C., Liu, Y., Wu, F., Tang, B., Fan, W., 2021a. Effective charging planning based on deep reinforcement learning for electric vehicles. IEEE Trans. Intell. Transp. Syst. 22 (1), 542–554. http://dx.doi.org/10.1109/TITS.2020.3002271.

Zhang, Y., Wang, X., Wang, J., Zhang, Y., 2021b. Deep reinforcement learning based Volt-VAR optimization in smart distribution systems. IEEE Trans. Smart Grid 12 (1), 361–371. http://dx.doi.org/10.1109/TSG.2020.3010130.

Zhang, Y., Yang, Q., An, D., Li, D., Wu, Z., 2023. Multistep multiagent reinforcement learning for optimal energy schedule strategy of charging stations in smart grid. IEEE Trans. Cybern. 53 (7), 4292–4305. http://dx.doi.org/10.1109/TCYB.2022.3165074.

Zhao, Z., Lee, C.K.M., 2022. Dynamic pricing for EV charging stations: A deep reinforcement learning approach. IEEE Trans. Transp. Electrif. 8 (2), 2456–2468. http://dx.doi.org/10.1109/TTE.2021.3139674.