Physics-Informed Reinforcement Learning Framework for Real-Time Coordination of EV Charging with Renewable Energy Sources

Qiuchen Zhang

School of Mechanical and Power Engineering, Cangzhou Jiaotong College, Cangzhou, Hebei province, 061199, China
E-mail: jssrgstwc@163.com

Received 25 July 2025; Accepted 11 June 2026

Abstract

We made a Physics-Informed Reinforcement Learning (PI-RL) framework to coordinate the charging stations for electric vehicles (EVs) in real time. These stations are powered by different renewable energy sources (RES), like wind and photovoltaic (PV). Our methodology explicitly integrates energy conservation laws, state-of-charge (SOC) dynamics, and inverter limitations into the training process, unlike previous reinforcement learning (RL)-based methodologies that are confined to single-source renewable energy systems (RES) and do not incorporate physical system constraints. We changed the Soft Actor-Critic (SAC) algorithm by adding domain-informed reward shaping and adaptive Lagrangian multipliers in order to make sure that constraints were met. We subsequently structured the EV-RES coordination issue as a physics-constrained Markov Decision Process (MDP). We evaluated the proposed PI-RL approach using actual datasets of solar irradiance, synthetic wind generation profiles, and electric vehicle arrival patterns. In terms of

Distributed Generation & Alternative Energy Journal, Vol. 41_4, 1153–1180.
doi: 10.13052/dgaej2156-3306.41411

operational profit, safety (constraint violation rate), and use of renewable energy, our approach worked better than traditional SAC and model-based rolling optimization. Also, our model made it much less common for charge-discharge switching to happen, which led to control strategies that are easier to understand and that help the battery last longer. These results show that the PI-RL framework is a reliable and widely applicable way to manage energy in real time in EV charging infrastructures that are getting more and more complicated as they use renewable energy.

Keywords: Electric vehicle (EV), EV charging management, physics-informed reinforcement learning (PI-RL), renewable energy sources (RES).

1 Introduction

The world is moving toward energy systems that don't emit carbon, and more and more people are driving electric cars (EVs). This means that we need smarter, more flexible, and longer-lasting charging solutions than ever before [1]. When there is a lot of demand for electric vehicles (EVs) and the output of renewable energy isn't always predictable, standard grid-connected EV charging stations typically have problems [2]. Recently, RL-based charging planning approaches haven't taken into account system-level physical restrictions such battery dynamics and power balancing. This makes them less practical in the actual world [3]. Most approaches just look at one type of RES, thus they don't take advantage of the benefits of employing both PV and wind systems together. That's why it's so vital to have a control framework that knows about physics and makes sure that EV-RES coordination in real time is safe, efficient, and easy to understand. Scientists are growing more interested in frameworks that integrate deep learning with knowledge of a specific sector as more electric vehicle charging stations that use both renewable energy and fossil fuels are created. We aim to create a PI-RL system that enables individuals control energy in real time, under certain constraints, and with the goal of making as much money as possible in regions where electric cars can charge from more than one source.

There are various papers addressing EV charging coordination, particularly concerning RES. Huang [4] presents a robust scheduling approach that accounts for wind power uncertainty via worst-case generation scenario optimization. Frequency regulation and system stability through EV charging coordination are investigated in [5] and [6] when renewable generation and electric vehicles support the grid. Zhang [7] proposes a distributed scheme founded upon peer-to-peer transactions and Nash bargaining between distribution system operators and charging facilities for the other's profitability. Zheng [8] studies an online optimal scheduling scheme via distributed predictive control in protecting data privacy and accounting for EV behavior uncertainty. Traditional model-based optimization methodologies for EV-PV coordination are well developed and require accurate deterministic or stochastic models of all variables involved [9]. Nevertheless, due to the small size of EV stations, these models typically have no exposure to high-quality long-term forecasts and must instead use forecast methods to learn renewable generation, EV arrivals, and price patterns from history in order to optimize all-day profit.

Such doubts are controlled increasingly with the help of data-oriented approaches. These methods obtain variable relations from past observations and tackle the complexity of coupled systems [10, 11]. RL, for example, is particularly adept at dealing with sequential, high-dimensional decision-making in uncertain environments [12]. Fitted Q-iteration is employed in [13] to determine aggregate charging policies, which are then communicated to individual EVs through binning algorithms. An RL integrated framework combining centralized planning and decentralized agent-based execution is discussed in [14], optimizing charging policy and vehicle placement through learned value functions. Charging and pricing are simultaneously scheduled in [15] where RL agents consider EV stochastic arrival and departure. Residential EV scenarios are considered through multi-agent RL frameworks in [16, 17] where every agent manages a vehicle separately. Cooperative reward mechanisms are used to achieve global optimality. Human factors such as driver stress and behavioral tendencies are modeled in [18, 19] to minimize operational cost and user satisfaction. Dynamic electricity prices are used in the RL input space in [20] in order to adapt policies under fluctuating pricing schemes.

Recent research studies using RL reinforce the need for operational constraints even more. References [21, 22] address power limitations due to EV arrival and departure instances. [21] specifically addresses a safe deep RL method that guarantees charging limitations in the absence of penalty-based training. PV-EV coordination is addressed expressly in [23, 24], where [23] addresses fuel cell electric vehicles and [24] focuses on algorithmic performance under varying RL methods. Despite the huge literature on EV charging coordination, two critical gaps remain that this work addresses:

1. *Lack of Physical Constraint Integration in RL-Based Approaches*: Existing RL techniques tend to overlook system-level physical constraints such as power balance, battery SOC dynamics, and inverter capacity. Without incorporating these physics-based constraints into the learning process explicitly, learned policies prove unsafe or inefficient when deployed in the real world.
2. *Limited Use of Predictive Information in RL Policies*: Most existing work relies solely on current observations and immediate or reactive control based on these observations without making use of multi-time-step predictions (e.g., PV generation, electricity prices, EV arrivals) in their decision-making process. This approach tends to produce short-sighted decisions that do not take into account future demand or supply fluctuations. The development of forecast-aware RL frameworks that incorporate predictive models into the state and reward structure is an important development area for enabling proactive and economically efficient control.

To address these limitations, our work offers the following contributions:

1. We formulated the EV-RES coordination problem as a physics-constrained MDP and solved it using an adapted SAC algorithm. Such formulation integrates domain-specific information such as energy conservation rules and operation constraints into training the agent so that it learns safer and more stable results. PI-RL improves convergence stability and constraint satisfaction without requiring external penalty tuning.
2. We tested the new approach on real EV usage and PV power generation, with synthesized results for simulated wind generation profiles. Performance comparison against elementary RL baselines (Proximal Policy Optimization (PPO)) and model-based rolling optimization indicated improved profit, safety (constraint violation rate), and energy efficiency.

The structure of the paper is: Section 2 presents problem background, including EV charging limitations, PV integration, and constrains of existing RL approaches. Section 3 presents the proposed PI-RL framework and describes the MDP formulation, policy architecture, and learning process. Section 4 demonstrates simulation results and performance assessment for different types of EV stations compared to alternative methods. Section 5 is a summary of main findings, emphasizes the strengths of the use of physics-based constraints in learning and states directions for future research.

2 Problem Background

2.1 System Architecture

It is suggested that the electric vehicle charging device work with either a Microgrid (MG) or a Virtual Power Plant (VPP). In this case, electricity for electric vehicles (EVs) comes from both utility power and renewable energy sources (RES) in the area, such as solar panels and wind turbines. The utility grid, spread photovoltaics, and wind power are the three main sources of power for the system. These sources are shown in Figure 1. It also has EV chargers that can work in both directions, so cars can charge and release power to and from the grid (V2G). This combination design makes the station more adaptable and self-sufficient, and it also lets energy be managed in real time. In our system, renewable energy is given the most weight. During different times of the day, the company buys energy from the infrastructure and then sells it to people who drive electric cars for a set price. But the company can only get more wind and solar power for less money. More cost-effective than sending energy to the grid, it is better to use renewable energy sources that are close to electric cars (EVs). This makes sure that the times for charging and sending power are as close to perfect as they can be.

To make the coordination tractable, we apply several assumptions. First, EV charging demands are not constant at maximum power; instead, their power levels can be adaptively controlled. Second, EV users may depart early

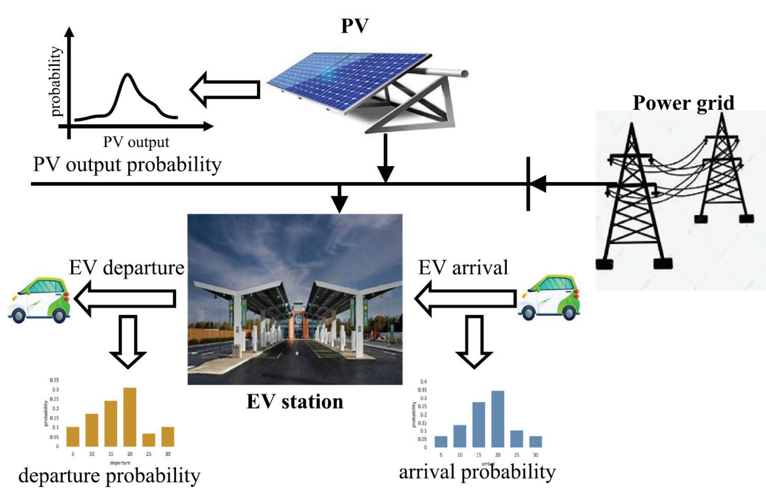

Figure 1 System architecture of the EV charging station with integrated RES, and grid sources.

if their charging completes before the estimated departure time, promoting efficient use of limited charging slots. Third, PV energy sales to the grid are financially less attractive, incentivizing on-site consumption. Finally, wind and PV output, EV arrivals, and price fluctuations are treated as stochastic but learnable through historical data. A PI-RL framework is applied to govern real-time decision-making. Unlike standard RL, the PI-RL approach embeds system dynamics—including battery SOC evolution, inverter power constraints, and power flow equations—into the learning process. This improves the agent's ability to make safe, interpretable, and energy-efficient decisions under uncertainty.

2.2 EV Station Model

The EV station consists of several smart charging poles, each with the capability of bidirectional power flow. These chargers are able to adjust the charging/discharging power of connected EVs, constrained by user-defined parameters such as desired energy demand and time of planned departure. The charging of each EV *i* within its stay duration which runs from arrival time t_i^{ar} to departure t_i^{de} is represented with the following equations. The total energy delivered by the EV station, must satisfy, or exceed, the specified energy demand D_i:

$$\sum_{t=t_i^{\mathrm{ar}}}^{t_i^{\mathrm{de}}} p_{t,i} \geq D_i \tag{1}$$

The state of charge (SOC) evolves over time as:

$$SOC_{t+1,i} = SOC_{t,i} + p_{t,i} \tag{2}$$

SOC and power bounds enforce the physical limits of battery and hardware:

$$-SOC_{t,i} \leq p_{t,i} \leq (SOC_i^{\max} - SOC_{t,i}) \tag{3}$$

$$-p_i^{\mathrm{dis,max}} \leq p_{t,i} \leq p_i^{\mathrm{cha,max}} \tag{4}$$

The EV sets for scheduling are defined as follows. The set $\mathcal{E}_t^c$ includes EVs already under charging at time t, and $\mathcal{E}_t^a$ contains new arrivals at time t:

$$\mathcal{E}_t^c = \left\{ i | t_i^{\mathrm{ar}} < t < t_i^{\mathrm{de}}, \sum_{t=t_i^{\mathrm{ar}}}^{t} p_{t,i} < D_i \right\} \tag{5}$$

$$\mathcal{E}_t^a = \{ i | t_i^{\mathrm{ar}} = t \} \tag{6}$$

Similarly, the set $\mathcal{E}_t^o$ includes EVs whose charging is complete but are still parked:

$$\mathcal{E}_t^o = \left\{ i | t_i^{\text{ar}} < t < t_i^{\text{de}}, \sum_{t=t_i^{\text{ar}}}^{t} p_{t,i} \geq D_i \right\} \tag{7}$$

To increase charger turnover, we modeled early departures probabilistically. Each EV in $\mathcal{E}_t^o$ departs early with a small probability ε:

$$t_i^{\text{de}} = \begin{cases} t_i^{\text{de}}, & \text{with probability } 1-\varepsilon \\ 0, & \text{with probability } \varepsilon \end{cases} \quad i \in \mathcal{E}_t^o \tag{8}$$

A problem with scheduling based on reinforcement learning comes from the fact that the action space $p_{t,i}$ is state-dependent and limited. We used a normalized adjustment rate $x_t \in [0, 1]$ as the control variable so that it would work with standard RL algorithms.

$$x_t = \frac{\sum_{i \in \mathcal{E}_t^c \cup \mathcal{E}_t^a} (p_{t,i} - p_{t,i}^{\text{lower}})}{\sum_{i \in \mathcal{E}_t^c \cup \mathcal{E}_t^a} (p_{t,i}^{\text{upper}} - p_{t,i}^{\text{lower}})} \tag{9}$$

Where upper and lower bounds of charging power are defined as:

$$p_{t,i}^{\text{upper}} = \min(p_i^{\text{cha,max}}, d_{t,i}) \tag{10}$$

$$p_{t,i}^{\text{lower}} = \max(-p_i^{\text{dis,max}}, -SOC_{t,i}, d_{t,i} - p_i^{\text{cha,max}}(t_i^{\text{de}} - t - 1)) \tag{11}$$

$$d_{t,i} = D_i - \sum_{t=t_i^{\text{ar}}}^{t} p_{t,i} \tag{12}$$

This formulation ensures that the control variable x_t always lies within a fixed interval, enabling tractable optimization via PI-RL. The physics-informed agent learns how to choose x_t values that take into account short-term costs, long-term rewards, and physical charging limits.

2.3 Coordinated Charging Control in EV Station

After the PI-RL agent determines the station-level adjustment rate x_t, the next step is to convert this normalized control action into the unique charging and discharging powers of the individual EVs. However, normally allocating the net charging/discharging energy of the station evenly among all connected

EVs can yield rapid and oscillatory switching between charging to discharging. This can impact battery health, create challenges in energy management, and in general be cumbersome. To improve this situation, we will use a coordinated control scheme leveraging virtual discharging. In the first stage of this scheme, all the currently connected EVs within the set $\mathcal{E}_t^c \cup \mathcal{E}_t^a$ are virtually discharged at their lower power limit levels, a lower bound $p_{t,i}$ as defined by Equations (10)–(12). Since this is only a computational action and therefore doesn't involve real energy transfer, the virtual discharging recalibrates each EV's available upper charging bound by subtracting the lower bound from the original maximum charging limit so that a charging capacity is effectively reserved based on the EV's flexibility and physical constraints. By having all EVs virtually discharged, the total allocatable charging power is calculated based on the station-level adjustment rate x_t. This is achieved by adding together the differences between the upper limits and lower limits of the adjusted limits of every EV and multiplying by x_t. The allocatable power represents the portion of total energy that the station plans to use for actual charging.

In the second stage, the system provides this power sequentially to individual EVs. The EVs are sorted in advance by expected remaining time where the EVs with the shortest remaining time get the power first. For each EV in sorted order, the EV gets a charge if the remaining energy is above its readjusted upper bound, and the power is allocated at that cap level. The remaining allocable power is then reduced by that amount. If the available charging power is not enough to fully charge the next EV, then it gets to charge with whatever power remains and the allocation stops. This two-phase allocation ensures that the EVs that need charging the most, charge ahead of the others, while other EVs can wait or discharge virtually. This ensures that charging and discharging do not unnecessarily switch back and forth. The end product of this cooperative scheduling is that most EVs are charged exactly to their capacity or discharged to the limits set by their temporal limits and residual energy demand d_t. After this control step, the subsequent sets of EVs are set as $\mathcal{E}_t^c, \mathcal{E}_t^a$, and $\mathcal{E}_t^o$ for the next time increment, by the means as given in Equations (5), (6), and (7) respectively.

2.4 MDP Formulation for Prediction-Based EV-PV Coordination

In the context of PI-RL, a station's charging of electric vehicles (EVs) and use of PV energy can be formalized as an MDP. The PI-RL framework functions in a dynamic environment that mimics real-world EV stations that operate with renewable energy sources, while maintaining both physical and stochastic characteristics. As the agent works within the PI-RL framework, it will always be in a closed-loop with the environment, observing a given system state s_t, taking an action a_t, obtaining an immediate reward rt, and taking the system to a next state s_{t+1}. This closed-loop interaction is an iterative process for each time step t, allowing the agent to learn the optimal policy π that maximizes its expected future cumulative rewards, typically over the duration of a full operation cycle or "episode" period which might be a day. A system state s_t can consist of either instantaneous or predicted parameters. For example, an instantaneous system state might include the practically upper and lower bounds of the aggregate charging/discharging power of the station p_t^{upper} and p_t^{lower}, the aggregate residual energy demand d_t^{agg}, the number of vacant charging slots n_t^{vacant}, and current time index t. While predicted state parameters useful for decision-making based on forecasts, might include the predicted vector of the PV power output at $\hat{p}_t^{\text{PV}}$, number of predicted EV arrivals $\hat{n}_t^{\text{arrival}}$, or predicted electricity purchasing prices ρ_t^{buy}. These forecasted vectors span n future time steps, allowing for prediction of system behavior and evolution. The result is the overall state vector:

$$s_t = (p_t^{\text{upper}}, p_t^{\text{lower}}, d_t^{\text{agg}}, \hat{p}_t^{\text{PV}}, \hat{n}_t^{\text{arrival}}, \rho_t^{\text{buy}}, n_t^{\text{vacant}}, t) \tag{13}$$

The action a_t performed by the agent at each time step corresponds to the station-level action rate $x_t \in [0, 1]$. The continuous control variable x_t specifies how much of the available charging capacity (between the station level's upper and lower limits) will be allocated. Thus, the overall charging power can be summarized as $p_t^{\text{lower}} + x_t(p_t^{\text{upper}} - p_t^{\text{lower}})$, and then distributed to the individual EVs using the coordinated allocation process presented in Section Analytical Model First. Depending on the action taken (i.e., what x_t is) and the stochastic nature of arrival and departure (of EVs) as well as stochastic variations in PV output, the environment creates a next state s_{t+1}. The transition begins by updating the active EV set $\mathcal{E}_{t+1}^c$ to include new EV's that have arrived and to remove EV's that have completed their charge or departed early. The updated residual demand is found in $d_{t+1,i}$, which corresponds to the remaining demand of the $\mathrm{i_{th}}$ active EV and the entire state of charge (SOC) for each EV from the allocated charging/discharging power $p_{t,i}$. The predictive vectors $\hat{p}_{t+1}^{\text{PV}}, \hat{n}_{t+1}^{\text{arrival}}$, and ρ_{t+1}^{buy} are produced by the embedded forecasting models ϕ_{PV}, ϕ_{space}, and ϕ_{price}, which accept historical or contextual features as inputs. In this analysis, it's assumed that the predictors create unbiased estimates to differentiate decision performance from forecasting error.

The reward function r_t describes the economic and operational performance of the station, at each time step. It includes profits from selling PV power to the grid, costs from buying power when PV power underdelivers, income from charging EVs, and a penalty term to enforce charging efficiency and discourage unmet EV demand. The reward is formally expressed as:

$$r_t = \rho^{\text{sell}}\, p_t^{\text{sell}} - \rho_t^{\text{buy}} p_t^{\text{buy}} + \rho^{\text{charge}}\, p_t^{\text{charge}} - \lambda d_t^{\text{agg}} \tag{14}$$

where ρ^{sell} and ρ^{charge} are exogenous unit prices corresponding to selling power to the grid and charging EVs, p_t^{sell} and p_t^{buy} represent the energy traded with the grid, where either the station has more PV output than required to meet the total EV charging, or cannot supply enough power to charge EVs. The final term also has an unmet demand penalty factor, λ, scaled by the station's total unmet demand, d_t^{agg}. This MDP structure supports the PI-RL framework to optimize the coordination strategy in direct consideration of physical constraints, renewable variability, forecast horizons, and battery degradation. By including physics-based components into the state transitions and reward definitions, the learning agent can develop strategies that are both economically optimal and feasible and capable of dealing with uncertainty in a robust way.

3 Proposed Framework

3.1 SAC-Based Policy for EV Station Charging

When considering EV charging coordination amid uncertainty, typical methods such as SAC have significant optimization capabilities for RL approaches. However, typical RL approaches such as SAC do not include the constraints through power system physical and operational knowledge. For example, energy balance, charging limits and pricing mechanisms. Therefore, the proposed framework introduces a PI-RL approach where the physical laws of the particular domain are embedded directly into the RL model's learning and policy inference cycle. This section will provide a rewriting of the SAC-based policy for an EV station charging coordinated under the use of physical constraints on the EV-PV-grid system, and embedding those physically-based constraints into the policy and value functions of the SAC agent. The objective does not change and continues aim to maximize the cumulative reward over the full operational horizon (normally one day), and now represented as:

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} r_t\right] \text{s.t. dynamics and physical constraints} \tag{15}$$

Unlike normal SAC resting entirely on data to develop the policy $\pi(a_t|s_t)$, the PI-RL policy has been changed to proxy in energy-aware Q-functions that consider physical constraints. In this approach, the policy is still sampling an action $a_t = x_t \in [0, 1]$, which is the rate of adjustment of the action, however, in this case, the selection is biased towards actions that meet power system constraints by nature. The fomerly stated policy of PI-RL is stated as:

$$\pi_{\theta}(a_t|s_t) \propto exp\left(-\frac{1}{\alpha}Q^{\pi}_{\text{phys}}(s_t, a_t)\right) \tag{16}$$

The term $Q^{\pi}_{\text{phys}}(s_t, a_t)$ is a physics-informed soft Q-value that implements operational knowledge of the EV station. This ensures that the policy will not only learn the rewards to maximize, but dictates respect for state-action constraints dictated by physics laws, ie, SoC dynamics, grid power limits, and PV generation limits. The energy function in PI-RL is similar to SAC's energy function, however, PI-RL extends the equation to learn constraints of the domain. It is laid out as follows:

$$E(s_t, a_t) = -\frac{1}{\alpha}Q^{\pi}_{\text{phys}}(s_t, a_t) \tag{17}$$

The energy function prioritizes actions involving higher, long-term utility while ensuring power balance and battery safety considerations are satisfied. The constraints mentioned above – namely the SoC dynamics $SOC_{t+1,i} = SOC_{t,i} + p_{t,i}$ and the upper and lower bounds of power – are hard-coded into the architecture of Q^{π}_{phys} through differentiable surrogate losses either during training or via constrained sampling schemes. In addition to the Q-value, the PI-RL framework introduces a physics-informed soft value function $V^{\pi}_{\text{phys}}(s_t)$, which is calculated using:

$$V^{\pi}_{\text{phys}}(s_t) = \mathbb{E}_{a_t \sim \pi}[Q^{\pi}_{\text{phys}}(s_t, a_t) - \alpha \log \pi(a_t|s_t)] \tag{18}$$

The entropy term $-\alpha \log \pi(a_t|s_t)$ maintains the randomness of SAC in the PI-RL framework by using the physical feasibility of highly infeasible actions to assign very low probabilities to infeasible actions, which have been penalized in the Q-function. The entropy coefficient α is learned through dual gradient descent to weight exploration relative to not violating physical laws. The math presented is similarly structured to a SAC-based framework, but only includes physical feasible actions in the expectation. Implementing domain knowledge directly into a value estimate and policy optimization allows the PI-RL method to significantly improve efficiency of training and demonstrate practicality in real world operations, particularly in the case of data sparsity or high variability in system state. Building on the policy structure we've laid out, the PI-RL architecture is built upon new value function updates that honor both the stochastic domain of SAC and the deterministic domain of EV-PV-grid systems. The quality of the policy is measured via a modified Bellman backup for the physics-informed soft Q-value:

$$Q^{\pi}_{\text{phys}}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \mathcal{T}, a_{t+1} \sim \pi}[V^{\pi}_{\text{phys}}(s_{t+1})] \tag{19}$$

In the above equation, the transition function $\mathcal{T}$ is bound by the real-world physics associated with EV arrival/departure, state-of-charge dynamics, and expected renewable outputs. These systems dynamics, along with the exploration and exploitation of actions, are integrated either by constructing simulation sites where transition is limited to the feasible ones from that transition function, or by incorporating the dynamics into the learning loss function using Lagrangian relaxation or penalties. In doing so, we give assurance that the learned Q-function will not overestimate the potential to realize long-term utility accompanied by an infeasible action. To further reinforce the physical interpretability of the learned policy, all actions sampled from the policy are projected back onto the feasible action space defined the relation:

$$\mathcal{A}_t = \left[\frac{\sum_i p^{\text{lower}}_{t,i}}{P_{\max}}, \frac{\sum_i p^{\text{upper}}_{t,i}}{P_{\max}} \right] \subseteq [0, 1] \tag{20}$$

Whereby the adjustment rate x_t has to be within operationally acceptable transitions that are based on the model equation schemas for the EV characteristics (Equations (2)–(4)). This is especially important when determining the adjustment rate is what actually controls the aggregate charging power is constrained to the EV stay time and urgency to charge.

3.2 Training Process of SAC

In the context of the Physics-Informed Soft Actor-Critic (PI-SAC) framework, the aim for training. is to learn an optimal policy $\pi(a_t|s_t)$ that conforms with physics principles by also incorporating other pre-existing knowledge about the system into the Q- and V-values approximators. To achieve this, the training is built upon three deep neural networks (DNNs): an actor network, a Q-critic networks, and the V-critic networks, which are each respectively parameterized by by ϕ, θ, and ψ. A slowly-updated target V-network, parameterized by $\widehat{\psi}$, is also maintained to stabilize training.

The actor network is responsible for modeling the policy $\pi_\phi(a_t|s_t)$. It receives as input, the current system state s_t and outputs the mean and covariance of a Gaussian distribution over the continuous action space $a_t \in [0, 1]$ that represents the station's adjustment rate. In PI-RL, action sampling must also abide known physical feasibility constraint, thus action sampling is followed by a projection onto the physically admissible action set $\mathcal{A}_t$. The projection of the action is performed to meet the charging power selection is compliant with the EV upper/lower power bounds and total energy needs over time.

For critic learning, two independent Q-networks estimate the soft Q-values $Q_{\theta_1}(s_t, a_t)$ and $Q_{\theta_2}(s_t, a_t)$. This mitigates the common overestimation bias of Q-learning by using the lowest value from the Q-networks when evaluating the policy. Each Q-network takes the state-action pair (s_t, a_t) as input, and they are trained to minimize the squared temporal difference error with respect to the soft target value $\hat{Q}(s_t, a_t)$:

$$\hat{Q}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V_{\widehat{\psi}}(s_{t+1})] \tag{21}$$

where $r(s_t, a_t)$ is the instantaneous reward, and $V_{\widehat{\psi}}$ is the target value output from the target V-network. The target Q-value contains a physics informed reward function where violations of unsatisfied EV charging demands or disallowed battery oscillations yield a penalty, allowing us to use knowledge from the domain directly in the training objective. The V-network, with parameters ψ, approximates the soft value function $V^\pi(s_t)$, which is the expected Q-value minus the entropy of the policy:

$$J_V(\psi) = \mathbb{E}_{s_t \sim D}\left[\frac{1}{2}(V_\psi(s_t) - \mathbb{E}_{a_t \sim \pi_\phi}[Q_{\min}(s_t, a_t) - \log \pi_\phi(a_t|s_t)])^2\right] \tag{22}$$

Loss pushes V_ψ to accurately predict expected return when the action is sampled from the policy, and is necessary for consistency of updates to the Q-functions. The experience buffer D will collect experience of the form (s_t, a_t, r_t, s_{t+1}) from interacting with the environment. The Q-networks will be updated by minimizing the loss function:

$$J_Q(\theta) = \mathbb{E}_{(s_t,a_t)\sim D}\left[\frac{1}{2}(Q_\theta(s_t,a_t) - \hat{Q}(s_t,a_t))^2\right] \tag{23}$$

At the same time, the actor network will be updated to minimize the KL div ergence from its distribution to the Boltzmann distribution defined by the soft Q-function:

$$J_\pi = \mathbb{E}_{s_t\sim D, a_t\sim\pi_\phi}[\log \pi_\phi(a_t|s_t) - Q_{\min}(s_t,a_t)] \tag{24}$$

Thus, the objective function defines the actor will try to assign a greater probability to actions with better expected return, while still exploring other actions due to the entropy regularization. To support stable training, the target V-network will be softly updated after each gradient step for each target V update:

$$\widehat{\psi} \leftarrow \tau\psi + (1-\tau)\widehat{\psi} \tag{25}$$

Where $\tau \in (0,1)$ is the soft update rate and is usually a small number such as 0.005. In practical implementation the full training schedule will be simply collected data and gradient updates. For every interaction loop:

1. The actor samples an action $a_t \sim \pi_\phi(\cdot|s_t)$.
2. The environment (EV charging station) executes a_t, transitioning to state s_{t+1}, yielding reward r_t.
3. The tuple (s_t, a_t, r_t, s_{t+1}) is stored in the buffer D.
4. Mini-batches are sampled from D for critic and actor updates.

Every iteration of training modifies the policy away from the optimal actions while respecting the physical constraints, and allows for exploration by keeping some policy entropy while relying on the critic networks' feedback. This physically-based and data-driven learning loop is the essence of PI-RL training, and uniquely fits the high-dimensional, uncertainty- laden world of EV-PV coordination. The overall structure of the proposed EV-RES coordination framework based on PI-RL is shown in Figure 2. At every time t, the actor network generates some parameters (mean and variance) of a Gaussian distribution conditioned on the current state s_t that comprises physical features of the system and multi-step observations of PV output, electricity price, and EV arrivals over time. An adjustment rate x_t (i.e., action a_t) is sampled from the Gaussian distribution and inputs into the charging control module to coordinate charging schedule. The charging control module uses the physical constraints of the system along with their own scheduling

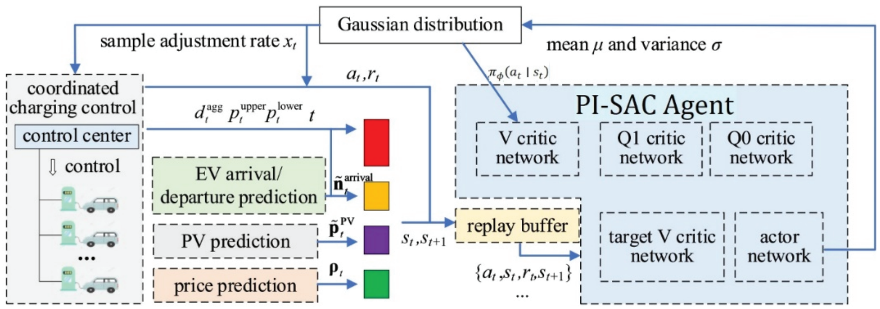

Figure 2 Overall architecture of the proposed physics-informed reinforcement learning framework for EV-RES coordination.

Table 1 Parameters of different types of EVs

EV Type	Parking Time (h)	Power Demand (kWh)	Initial SOC (%)
Emergent	<2	8–12	20
Normal	2–6	18–24	20
Residential	6–12	26–32	10

logic to determine how to distribute the resultant aggregate charging power across a set of EVs while respecting their station capacity, deadlines, and SOC limits.

4 Case Study

4.1 Environment Setup

To analyze the performance of the proposed PI-RL framework in optimizing the coordination of EV charging with RES generation, a complete simulation environment is developed. This environment is designed to provide characteristics representative of real-life distributed EV stations with renewable sources available, especially PV systems. The electric vehicles (EVs) are intuitively defined in three broad groups (emergent, normal, and residential), reflecting some typical usage, dwell time, and flexibility in discretionary energy management. These groups reflect a range of service expectations and adjustments. The defining will be the combination of parking duration, power demand for charging, and initial state of charge (SOC) presented in Table 1.

Emergent EVs represent vehicles that are used for a short, high-priority task with relatively high initial SOCs, but short parking durations. Thus, there is little flexibility in charging these vehicles. The normal and residential EVs will have longer dwell times and wider energy demands to allow for considerably more potential load adjustment, and more connections with EVs to cater to the renewable's variability. The simulation made decisions at a 15-minute time resolution which matched both the finer input data and decision-making steps for this study. Vehicle arrival data will be based and altered from real data sources [15]. PV output profiles have been taken from an industrial park in order to capture seasonal and temporal solar irradiance processes. Three representative sampling profiles of EV stations were modeled:

1. Type 1 – Residential-Dominant: High EV inflow in the morning, largely composed of residential users.
2. Type 2 – Evening-Dominant: Lower EV volume, concentrated in the evening hours.
3. Type 3 – Emergent-Dominant: High-volume station with a majority of emergent EVs arriving throughout the day.

Electricity prices are structured by a tier-based system based on time-of-use (TOU) and peak pricing. The TOU price contains Low rate (0.25 CNY/kWh), and High rate (0.5 CNY/kWh). Peak pricing is to represent grid stresses during high air-conditioning loads (e.g., during the summer in midday) as high as 2.25 CNY/kWh. There is also random timing for both the high and peak prices within practical bounds (High price: between 08:00–10:00 and 17:00–20:00, Peak price: from 12:00–13:00, and 14:00–16:00). Each episode starts with sampling one day of EV arrivals and coinciding PV output data, and their relevant stochastic electricity price profiles, from the historical datasets. The EV station has 200 charging piles and there is an exit probability $\varepsilon = 0.1$, indicating whether EVs would leave the EV station after they complete their charge ahead of the scheduled departure time. If an EV arrives and the EV station is at max capacity, (i.e., there is no longer charging piles available) the new arriving EV is denied entry. However, there are those long dwell times, not considering too low an initial SOC, queuing a recent arrival waiting, would not be modeled. The cost/reward parameters in the environment are fixed as:

- Selling price to grid (ρ_sell): 0.1 CNY/kWh
- Charging price to EVs (ρ_charge): 0.75 CNY/kWh
- Penalty coefficient (λ) for unmet energy demand: 0.001

A single EV's maximum charging power is 10 kW, while discharging is inactive during this circumstance (i.e., lower power bound = 0). Each episode is 96 steps long (15 minutes intervals for 24 hours) and within these episodes, the selected PI-RL agent will have the opportunity to act in the environment and learn optimal policies for energy allocation while considering physical and operational constraints.

4.2 SAC Training Process

Within the proposed PI-RL framework, the SAC agent's training goal was to learn suitable policies for optimal control, which considers the charging demand and EV charging, PV generation, and time-dependent electricity price. The actor network was a key component of the models and is responsible for outputting the mean and variance (also known as exploration rate) of a Gaussian distribution that models probabilistic policy denoted $\pi(a_t|s_t)$. The action was the sampled action $x_t \in [0,1]$, which was seen in this case to be the adjusted rate of EV charging station total power. The conversion to an actual EV charging operation was made by the resource coordinator as previously referenced. The state vector's dimension which included the system constraints, the forecasts, and availability is depicted in Table 2. The neural network design for all the components (actor, critic Q, critic V) were the same and were two hidden layers with 128 neurons and 64 neurons and used ReLU activation functions. This architecture was chosen to promote convergence stability and expressiveness of nonlinear reward dynamics resulting from constraints about physical systems.

The training process entails training three different agents based on the types of EV stations (i.e., residential-dominant, evening-dominant, and emergent-dominant agents) on their respective EV arrival profiles. Each

Table 2 Parameters of the SAC algorithm

Parameter	Value
Actor Network Hidden Layers	[128, 64]
Q-Critic Network Hidden Layers	[128, 64]
V-Critic Network Hidden Layers	[128, 64]
Activation Function	ReLU
Replay Buffer Size	1,000,000
Batch Size	256
Learning Rate (Actor/Critic)	3e-4
Discount Factor (γ)	0.99
Soft Update Rate (τ)	0.005
Entropy Coefficient (α)	Automatic Tuning
Time Steps per Episode	96
Environment Time Step	15 minutes
Total Training Steps	$>$500,000

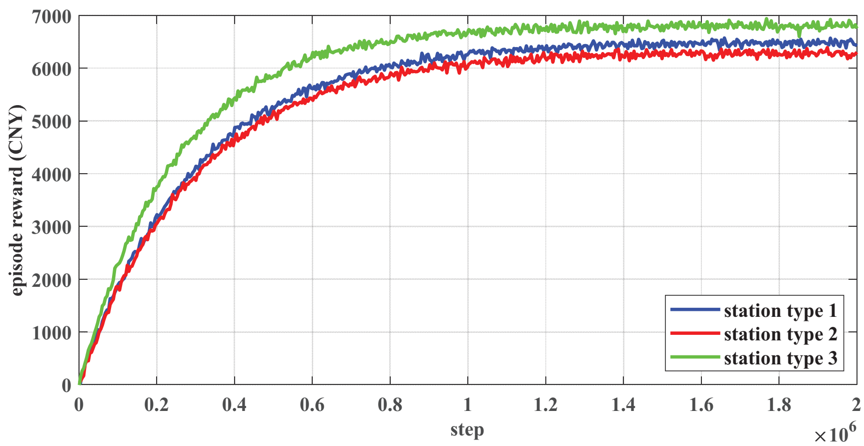

Figure 3 Convergence curve during the PI-SAC agent's training process.

episode is comprised of 24 hours (i.e., 96-time steps), at every time step, the agent observes the state s_t, samples its action x_t, obtains its reward r_t, and evolves to s_{t+1}. Each interaction is recorded in the replay buffer from which mini-batches of interactions are used for gradient-based updates, and over time, the agent improves its policy by maximizing expected long term cumulated rewards, while at the same time implicitly satisfying physical constraints of the charging process as well as uncertainties due to PV generation and variable electricity pricing. At the beginning of the training, when actions are nearly random, the total reward of an episode (i.e., daily profit) is typically less than 2000 CNY. However, as the agent gradually learns over more than 500,000 total timesteps (or approximately 5000 episodes), the agent converges towards a stable policy that nets substantially higher profits.

The accumulated rewards increase continuously in the learning graph (Figure 3) demonstrating improved quality of decision making by the PI-SAC agent. We observe variations for the different types of EV charging stations based on differences in arrival patterns, and energy flexibility in the operational profiles of the departure patterns. This demonstrates the versatility of the learning framework PI-RL for different operational situations.

4.3 Performance Analysis

The effectiveness of the proposed PI-RL framework within each type of EV station in real-time coordination of EV charging and renewables is evaluated

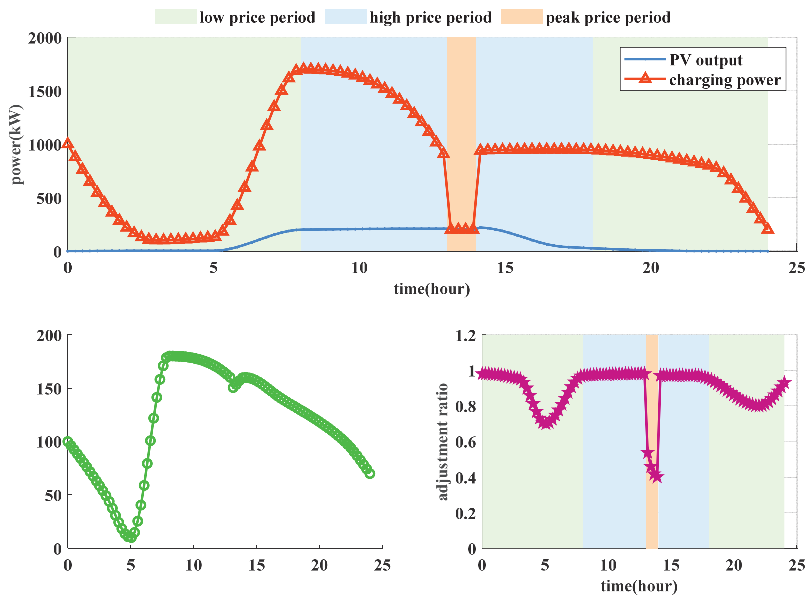

Figure 4 Performance of the trained SAC agent in the first testing phase.

by testing the trained agent in ten unseen episodes for every type of EV station. The investigation is focused mainly on EV station Type 1, where residential EVs account for most station arrivals (morning hours) and with average daily profit of 6523 CNY shows the agent's ability to manage charging loads amid changing prices and consumer demand.

A representative performance from one episode is presented in Figure 4, which shows the charging power, PV output, adjustment rate x_t, and number of charging piles occupied throughout the day. Electricity pricing periods are colored for low-price hours (green), high-price intervals (blue), and peak-price periods (orange). The agent is capable of successfully maximising profit by adjusting charging power in harmony with PV generation while the electricity prices are high. For example, in the peak price windows, the charging power remains tightly coupled with the PV output, avoiding purchasing power at expensive prices. Notably, in the first high-price window, the charging power exceeds the PV output, suggesting the agent chose to meet expected demand in the early window of higher grid costs instead. In the second high-price window, the gap notably reduces, suggesting that the agent responds to expected lower prices in the subsequent periods up ahead, shifting its behavior.

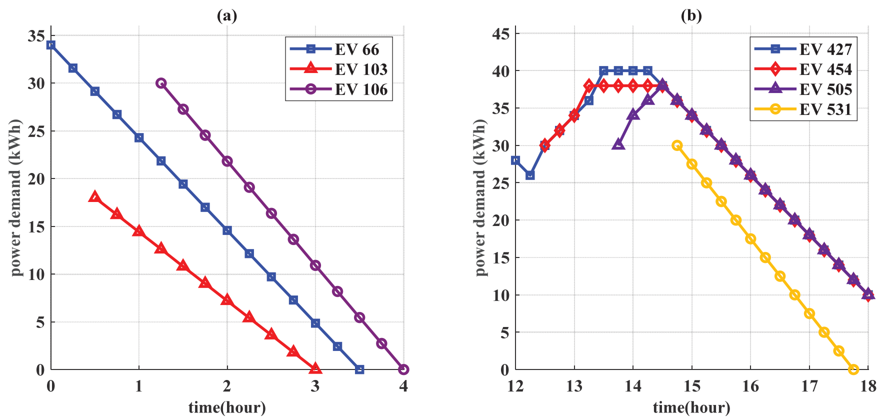

Figure 5 Charging and discharging profiles of 7 individual EVs during a sample episode.

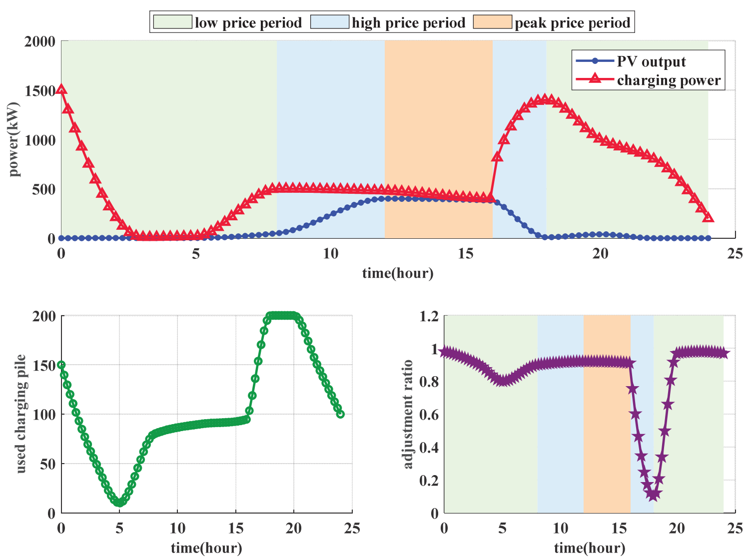

Figure 6 Agent's performance in Case 2 testing.

Figure 5 shows charging and discharging profiles for seven individual EVs that also show personalized energy management for the EVs. For instance, EVs 66, 103, 106, and 531 arrive to the terminal during high adjustment periods and avail themselves of charging the second they arrive. Conversely, EVs 454, and 505, enter their adjustment phase during lower adjustment periods and they are discharged. EV 427 is unique in that it

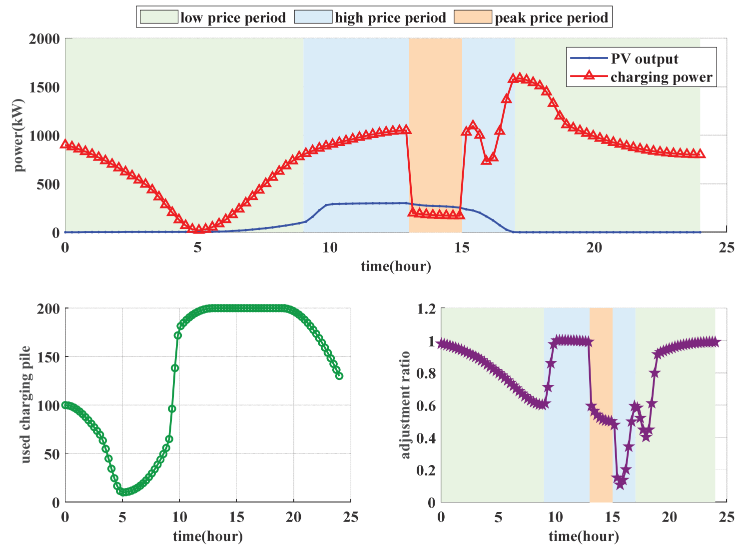

Figure 7 Agent's performance in Case 3 testing.

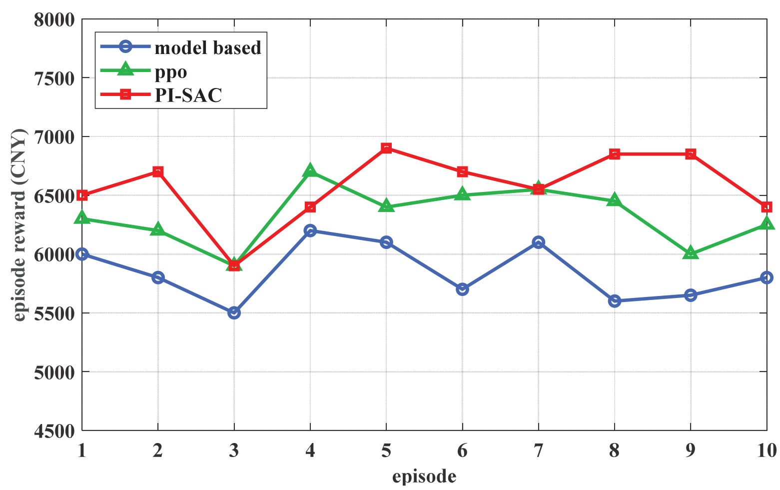

Figure 8 Episode reward comparison of PI-RL (SAC), model-based optimization, and PPO-based reinforcement learning.

charges, and then as the adjustment rate drops, it discharges. These profiles demonstrate the responsiveness of the policy, without forcing the system to transition frequently between action states, and as a result, protect EV battery health.

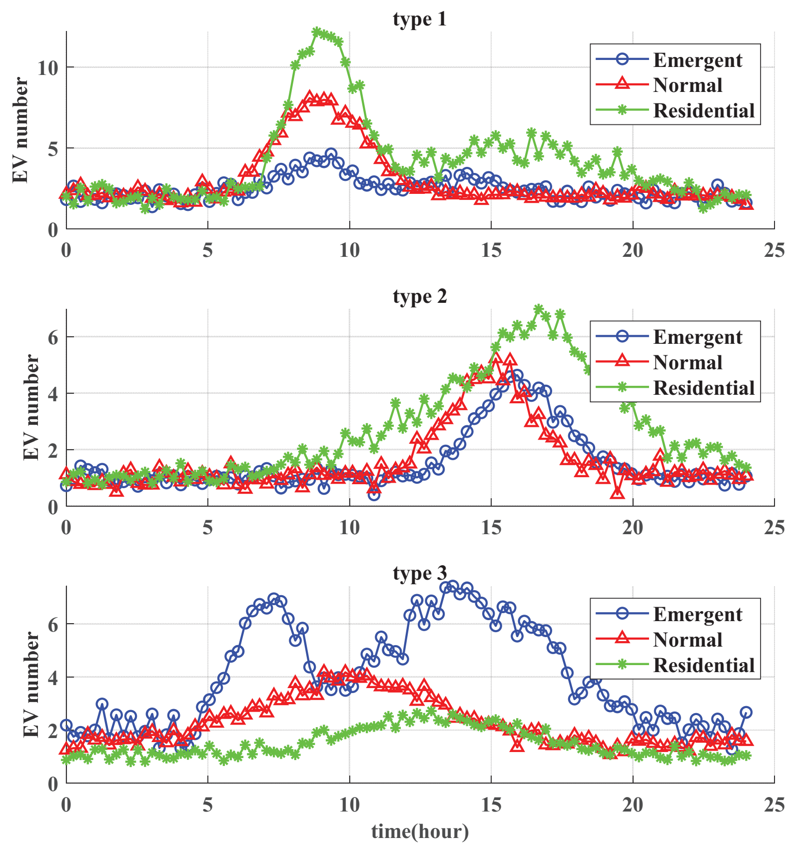

Figure 9 EV arrival pattern across the three EV station types.

We see also nuanced responses when we consider agent performance by EV station type. As shown in Figure 6 concerning Type 2 stations, where most of the arrivals occur in the evening, the agent is able to increase charging power significantly during the second high-price window of the evening to increase the turnover, prior to future arrivals, and as a result, we see more successful weight on responses evidential as after the initial period of adjustment (vs baseline). We see a high volume of emergent types of EVs at Type 3 station (Figure 7), and with this we often see all spaces full, thus the adjustment strategy is very sensitive to electricity price in the adjustment strategy, and the agent essentially behaves as a Type 1 station.

The efficacy of the PI-RL agent is further corroborated by Figure 8 which considers episode rewards across three learning methods, including the proposed SAC PI-RL agent, a model-based optimization mechanism using rolling horizon, and Proximal Policy Optimization (PPO). The model-based learning method constraints the optimization over short prediction horizons, applying only the first-step and thereby leading to limited long-term benefits. The SAC agent performs well because it is long-duration, recognizing patterns that are in the best interest of the agent, and therefore the economic performance. The PI-RL averaged improvement to episode reward was 13.7% from the model-based approach and 3.0% from PPO.

To offer additional context, Figure 9 shows arrival patterns for the three EV stations included in the study. It is clear from the figures that daily distributions of residential, normal, and emergent EVs are significantly different, because these arrival patterns are critical to agent behavior the framework is designed to incorporate this into agent behavior to control a system optimally given expected arrival profiles, electricity price variation, and renewable energy supply.

5 Conclusion

This study investigated the development of a new PI-RL framework to coordinate the EV charging with the PV generation under asymmetric time-varying electricity prices. The framework integrates physical constraints, forecasts of future states, and learning via deep policy learning to mediate controlled flexibility with greater economic efficiency than standard rule-based and model-less approaches. It begins by re-formulating the coordinated EV charging as a constrained MDP, where the level of control is further defined through the rate of adjustment of charging power at each EV station. The proposed PI-RL approach embedded operational constraints directly into the learning process rather than utilizing a standard SAC approach. In addition to the EV states and rewards, the proposed method included forecasts of future PV generation, electricity prices, and distributions of EV arrivals for the specific EV station. As a remedy to battery degradation caused by many transitions between charge and discharge states, this framework assumes a coordinated charging control strategy. It utilized a virtual discharging mechanism in the model, by allowing an EV with a longer expected stay at the EV station to be more responsible for the flexibility of grid-interaction while safeguarding the operational stability of the site. The proposed framework received extensively simulated demonstrations using three distinctive types of EV stations and showed the best learning performance outcomes. According to the results obtained from the study, the PI-RL based SAC agent was able to increase the profit for each station on a daily basis while successfully lowering electricity procurement costs and maximizing the utilization of PV energy generation. In summary, the PI-SAC agent was able to achieve an average profit improvement of 13.7% over a model-based rolling optimization method and 3.0% over PPO. These levels of profit improvement demonstrate a value in the robustness and ability of the agent for suitable and generalization associates with different types of station and different pricing scenarios as well. what the agent did to adaptively charge to the way electricity is priced and the time the EV were staying was evidenced as a desirable capability to consider economic benefits with station operations and constraints. Furthermore, the learned policy was structured to the extent that the charging behaviors stayed within battery health guidelines, while mitigating any extreme power fluctuations. Future development of this framework may consider bidirectional energy exchanges (i.e., Vehicle-to-Grid operations), uncertainty modelling around PV and EV arrival predictions, and multi-agent coordination around a shared or distributed charging network system.

References

[1] Q. Peng, W. Liu, Y. Zhang, S. Zeng, and B. Graham, "Generation planning for power companies with hybrid production technologies under multiple renewable energy policies," *Renewable and Sustainable Energy Reviews*, vol. 176, p. 113209, 2023.

[2] O. Almughram, S. Abdullah ben Slama, and B. A. Zafar, "A reinforcement learning approach for integrating an intelligent home energy management system with a vehicle-to-home unit," *Applied Sciences*, vol. 13, no. 9, p. 5539, 2023.

[3] F. Sun et al., "Prediction-based EV-PV coordination strategy for charging stations using reinforcement learning," *IEEE Trans Ind Appl*, vol. 60, no. 1, pp. 910–919, 2023.

[4] Q. Huang, Q.-S. Jia, and X. Guan, "Robust scheduling of EV charging load with uncertain wind power integration," *IEEE Trans Smart Grid*, vol. 9, no. 2, pp. 1043–1054, 2016.

[5] S. F. Aliabadi, S. A. Taher, and M. Shahidehpour, "Smart deregulated grid frequency control in presence of renewable energy resources by EVs charging control," *IEEE Trans Smart Grid*, vol. 9, no. 2, pp. 1073–1085, 2016.

[6] M. F. M. Arani and Y. A.-R. I. Mohamed, "Cooperative control of wind power generator and electric vehicles for microgrid primary frequency

regulation," *IEEE Trans Smart Grid*, vol. 9, no. 6, pp. 5677–5686, 2017.

[7] J. Zhang, L. Che, X. Wan, and M. Shahidehpour, "Distributed hierarchical coordination of networked charging stations based on peer-to-peer trading and EV charging flexibility quantification," *IEEE Transactions on Power Systems*, vol. 37, no. 4, pp. 2961–2975, 2021.

[8] Y. Zheng, Y. Song, D. J. Hill, and K. Meng, "Online distributed MPC-based optimal scheduling for EV charging stations in distribution systems," *IEEE Trans Industr Inform*, vol. 15, no. 2, pp. 638–649, 2018.

[9] X. Chen, G. Qu, Y. Tang, S. Low, and N. Li, "Reinforcement learning for selective key applications in power systems: Recent advances and future challenges," *IEEE Trans Smart Grid*, vol. 13, no. 4, pp. 2935–2958, 2022.

[10] F. Sun, X. Kong, J. Wu, B. Gao, K. Chen, and N. Lu, "DSM pricing method based on A3C and LSTM under cloud-edge environment," *Appl Energy*, vol. 315, p. 118853, 2022.

[11] Y. Zheng, Y. Song, D. J. Hill, and K. Meng, "Online distributed MPC-based optimal scheduling for EV charging stations in distribution systems," *IEEE Trans Industr Inform*, vol. 15, no. 2, pp. 638–649, 2018.

[12] F. Sun et al., "Prediction-based EV-PV coordination strategy for charging stations using reinforcement learning," *IEEE Trans Ind Appl*, vol. 60, no. 1, pp. 910–919, 2023.

[13] N. Sadeghianpourhamami, J. Deleu, and C. Develder, "Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning," *IEEE Trans Smart Grid*, vol. 11, no. 1, pp. 203–214, 2019.

[14] Z. Ye, Y. Gao, and N. Yu, "Learning to operate an electric vehicle charging station considering vehicle-grid integration," *IEEE Trans Smart Grid*, vol. 13, no. 4, pp. 3038–3048, 2022.

[15] S. Wang, S. Bi, and Y. A. Zhang, "Reinforcement learning for real-time pricing and scheduling control in EV charging stations," *IEEE Trans Industr Inform*, vol. 17, no. 2, pp. 849–859, 2019.

[16] F. L. Da Silva, C. E. H. Nishida, D. M. Roijers, and A. H. R. Costa, "Coordination of electric vehicle charging through multiagent reinforcement learning," *IEEE Trans Smart Grid*, vol. 11, no. 3, pp. 2347–2356, 2019.

[17] L. Yan, X. Chen, Y. Chen, and J. Wen, "A cooperative charging control strategy for electric vehicles based on multiagent deep reinforcement

learning," *IEEE Trans Industr Inform*, vol. 18, no. 12, pp. 8765–8775, 2022.

[18] T. Qian, C. Shao, X. Li, X. Wang, Z. Chen, and M. Shahidehpour, "Multi-agent deep reinforcement learning method for EV charging station game," *IEEE Transactions on Power Systems*, vol. 37, no. 3, pp. 1682–1694, 2021.

[19] L. Yan, X. Chen, J. Zhou, Y. Chen, and J. Wen, "Deep reinforcement learning for continuous electric vehicles charging control with dynamic user behaviors," *IEEE Trans Smart Grid*, vol. 12, no. 6, pp. 5124–5134, 2021.

[20] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans Smart Grid*, vol. 10, no. 5, pp. 5246–5257, 2018.

[21] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Trans Smart Grid*, vol. 11, no. 3, pp. 2427–2439, 2019.

[22] H. Li et al., "Constrained large-scale real-time EV scheduling based on recurrent deep reinforcement learning," *International Journal of Electrical Power & Energy Systems*, vol. 144, p. 108603, 2023.

[23] G. Guo and Y. Gong, "Energy management of intelligent solar parking lot with EV charging and FCEV refueling based on deep reinforcement learning," *International Journal of Electrical Power & Energy Systems*, vol. 140, p. 108061, 2022.

[24] M. Dorokhova, Y. Martinson, C. Ballif, and N. Wyrsch, "Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation," *Appl Energy*, vol. 301, p. 117504, 2021.

Biography

Qiuchen Zhang graduated from City University of London in 2015 with a Master's degree in vehicle engineering, currently, he works in Cangzhou Jiaotong College. He has completed a education reform project. His research interest includes vehicle construction, and vehicle semantics.