Model-free safe deep reinforcement learning for grid-to-vehicle management considering grid constraints and transformer thermal stress

Zhewei Zhang, Rémy Rigo-Mariani, Nouredine Hadjsaid, Yan Xu

Engineering Applications of Artificial Intelligence 162 (2025) 112529

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Model-free safe deep reinforcement learning for grid-to-vehicle management considering grid constraints and transformer thermal stress

Zhewei Zhang [a,b,d,*], Rémy Rigo-Mariani [a], Nouredine Hadjsaid [a], Yan Xu [c]

[a] Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab, 38000, Grenoble, France
[b] CNRS@CREATE, 1 Create Way, #08-01 Create Tower, 138602, Singapore
[c] Centre for Power Engineering (CPE), School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore
[d] Energy Research Institute @ NTU, Interdisciplinary Graduate Programme, Nanyang Technological University, 639798, Singapore

ARTICLE INFO

Keywords:
Electric vehicle
Distribution grid
Safe deep reinforcement learning
Thermal loading

ABSTRACT

The increasing penetration of Electric Vehicles (EVs) presents challenges to the distribution grid, due to more volatile power profiles and higher peak demand. One key research question is how to accommodate EVs with limited-capacity grid equipment, such as transformers and lines. However, uncertainties from the EV side and the complexity of grid equipment models challenge the performance of the control strategies implemented. Moreover, the thermal loading of the transformer is often neglected. In this work, we propose a fully model-free, safe Deep Reinforcement Learning (DRL)- based grid-to-vehicle management strategy to avoid electric and thermal overloading of the transformer and power grid constraint violation. The management strategy is based on Projection-based Constraint Policy Optimization (PCPO) and takes only the observable information from the grid and vehicles. The target is to maximize energy delivery to the EV fleet while considering safe constraints, such as transformer thermal loading, voltage magnitude limits, and line loading limits. We compared the proposed strategy with conventional DRL and other safe DRL methods and investigated its robustness against higher ambient temperatures. The results show that the proposed strategy can deliver 92 % energy and reduce violations of the grid and transformers, while the other benchmarks deliver less than 80 %. The robustness test demonstrates that the proposed strategy is effective in various temperature. Moreover, the proposed strategy can effectively reduce at most 90 % of the transformer aging incurred by the thermal stress, compared with the uncontrolled charging.

* Corresponding author. Univ. Grenoble Alpes, CNRS, Grenoble INP, G2Elab, 38000, Grenoble, France.
E-mail addresses: zhewei.zhang@g2elab.grenoble-inp.fr (Z. Zhang), remy.rigo-mariani@g2elab.grenoble-inp.fr (R. Rigo-Mariani), nourdine.hadjsaid@g2elab.grenoble-inp.fr (N. Hadjsaid), xuyan@ntu.edu.sg (Y. Xu).

https://doi.org/10.1016/j.engappai.2025.112529
Received 4 June 2025; Received in revised form 24 September 2025; Accepted 24 September 2025
Available online 29 September 2025

Table of Notations

Category	Notation	Description	Category	Notation	Description
Set	$\mathscr{C}$	Charging Stations	Parameters	$G_{bb'}$	Real Part of Admittance Matrix
	$\mathscr{N}$	Charging Ports		$B_{bb'}$	Imaginary Part of Admittance Matrix
	$\mathscr{B}$	Buses		$Y_{bb'}$	Admittance between b and b'
	$\mathscr{L}$	Lines		Δt	Timestep
	$\mathscr{T}$	Time Range		$\overline{p}_n$	Maximum Charging Power of Port n
Index	c	Charging Station		$\overline{E}_n$	Energy Demand at Port n
	n	Charging Port		$\underline{v}$	Voltage Lower Limit
	b	Bus		$\overline{v}$	Voltage Upper Limit
	l	Line		$\overline{i}$	Maximum Line Loading
	t	Timestamp		$\overline{\theta}^o$	Maximum Oil Temperature
Variables	$v_{b,t}$	Voltage Magnitude of Bus b		$\overline{\theta}^h$	Maximum Hotspot Temperature
	$i_{l(b\to b'),t}$	Current Magnitude of Line l		δt	Calculation Resolution of Transformer Thermal Model
	$p^l_{b,t}$	Active Power Injection of Bus b		T^{arr}_n	Arrival Time at Port n
	$q^l_{b,t}$	Reactive Power Injection of Bus b		T^{dep}_n	Departure Time at Port n
	$p^d_{b,t}$	Active Power Demand of Bus b	Safe Deep Reinforcement Learning Formulation	s_t	State Vector
	$q^d_{b,t}$	Reactive Power Demand of Bus b		o_t	Observation Vector
	$p^g_{b,t}$	Active Power Generation of Bus b		r_t	Reward of the Agent
	$q^g_{b,t}$	Reactive Power Generation of Bus b		c_t	Cost of the Agent
	$\phi_{bb',t}$	Voltage Angle Difference between b and b'		π	Policy of the Agent
	$p_{n,t}$	Charging Power at Port n		$J^R(\pi)$	Expected Rewards
	$\Omega_{n,t}$	Connection of Port n		$J^C(\pi)$	Expected Costs
	$\Delta E_{n,t}$	Energy Delivered of Port n		$Q^R(s,a)$	Action Value Function of Rewards
	$\theta^o_{c,t}$	Oil Temperature of Station c at time t		$Q^C(s,a)$	Action Value Function of Costs
	$\theta^h_{c,t}$	Hotspot Temperature of Station c at time t		$V^R(s)$	State Value Function of Rewards
	$\Delta\theta^1_{c,t}$	Auxiliary Variable of Transformer Thermal Model		$V^C(s)$	State Value Function of Costs
	$\Delta\theta^1_{c,t}$	Auxiliary Variable of Transformer Thermal Model		$A^R(\pi)$	Advantage of Rewards
Abbreviations	EV	Electrical Vehicle		$A^C(\pi)$	Advantage of Costs
	G2V	Grid to Vehicle		γ	Discount Factor
	DRL	Deep Reinforcement Learning		Π^C	Feasible Space
	SDRL	Safe Deep Reinforcement Learning		D_{KL}	KL Divergence Function
	IEA	International Energy Agency		g	Gradient of Rewards Advantage Expectation
	PPO	Proximal Policy Optimization		H	Hessian of KL Divergence
	DDPG	Deep Deterministic Policy Gradient		a	Gradient of Costs Advantage Expectation
	TD3	Twin Delayed DDPG		b	Constraints Violation
	PDO	Primal-Dual Optimization		d	Tolerance of Violations
	CPO	Constrained Policy Optimization		δ	Tolerance of Kullback-Leibler Divergence
	PCPO	Projection-based Policy Optimization		ϑ_k	Policy Parameter at Episode k
	SOTA	State of the Art		λ_k	Lagrangian Multiplier at Episode k
	KL	Kullback-Leibler		ε	Clip Range
				η	Learning Rate of Lagrangian Multiplier

1. ıntroduction

As the transportation sector contributes to most of the greenhouse gas emissions, its electrification becomes vital in order to meet the EU 2050 carbon neutrality target (Martins et al., 2023). According to the IEA, registrations of new Electric Vehicles (EVs) grew by 37 % in 2023 (Trends in Electric Cars). By 2050, EVs are expected to account for up to 80 % of total vehicles (World Energy Outlook 2024). However, this increasing EV penetration brings considerable challenges to the power grid. According to the ENTSO-E, significant challenges can be expected for power transformers and lines due to increasing power flows and charging demand (Electric Vehicle Integration into Power Grids, 2021), while installing new equipment is time-consuming and expensive. This scarce capacity can then hinder the transportation electrification process (Gnanavendan et al., 2024). Furthermore, Rezaee et al. conducted a quantitative analysis of the impact of EV charging demand on the power grid, and they found that more than 8 % EV penetration can lead to voltage out-of-bounds and heavy power losses (Rezaee et al., 2013). Therefore, to promote EV utilization, effective Grid-to-Vehicle (G2V) control strategies based on available grid capacity are necessary.

However, effective G2V control suffers from two significant difficulties. The first is the uncertain future EV arrivals and their energy demand (time-series-related uncertainty), and the second is the complexity of physical models (model nonconvexity). Both make it hard for conventional model-based controllers to achieve good performance. Deep Reinforcement Learning (DRL) has recently gained popularity due to its ability to handle time series-related uncertainty and non-explicit models (DRL is a model-free approach). DRL agents learn a control policy by interacting with the environment, maximizing the rewards, and incorporating Deep Learning to approximate the policy and value function (Görges, 2017). There are several DRL applications in the field of EV fleet charging (Qiu et al., 2023), considering load curve smoothing in a power grid (Tuchnitz et al., 2021), valley-filling in an active distribution grid (Qi et al., 2023), congestion management with cost reduction (Yang et al., 2024), charging station revenue maximization (Zheng et al., 2025), charging cost minimization with PV and storage integration (Zhang et al., 2023a), (Hussain et al., 2022), (Zhang et al., 2024a) and multi-agent formulation (Li et al., 2025). In these works, the reward function treats the constraints as penalties. However, because the DRL agent generates action usually based on a stochastic policy, one key research problem is to ensure safety.

The safe DRL (SDRL) algorithms can be classified into two major

Table 1
Summary of safe deep reinforcement learning for grid-to-vehicle control.

Ref	Objectives	Alg.	Embedded	Constraints	Grid	EV Info
Zhang et al. (2024b)	Energy	PPO	No	EV SOC Capacity	No	Full
Liu et al. (2023)	Cost, Voltage Violation	DDPG	No	EV SOC Capacity, Generator Capacity, Voltage Level	Yes	Full
Wu et al. (2023)	Energy, Cost	TD3	No	Charging Power, EV SOC Capacity	Yes	Full
Chen et al. (2025)	Cost	Lagrangian	Yes	EV Demand	No	Full
Ding et al. (2022)	Cost, Degradation	Lagrangian	Yes	Energy Balance, Equipment Power Capacities	No	Full
Zhang et al. (2023b)	Profit	PDO	Yes	EV Demand, Energy Storage Capacity	Yes	Full
Li et al. (2020)	Cost	CPO	Yes	EV Capacity	No	Full
Our Work	Energy	PCPO	Yes	EV Demand, Transformer Thermal & Power Load, Voltage Level, Current Load	Yes	Partial

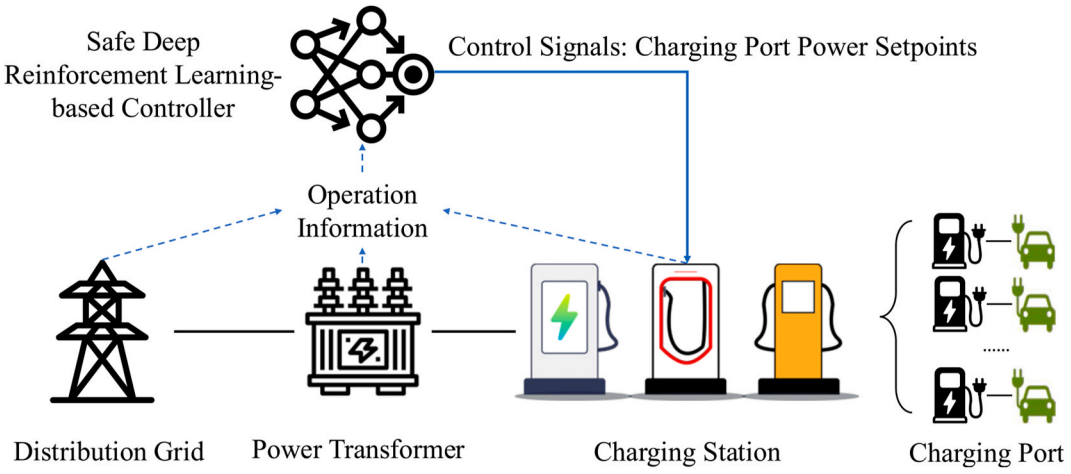

Fig. 1. Schematic of G2V problem.

categories, depending on whether safety awareness is embedded with the DRL scheme or not. One common approach for the non-embedded option consists of correcting the DRL actions through an additional safe module, either an optimization problem or a heuristic. Zhang et al. implemented a rule-based module with multi-agent Proximal Policy Optimization (PPO) to ensure the state-of-charge (SOC) is not out of bounds (Zhang et al., 2024b). That work considered the balance between EV charging and grid dispatching command, yet the grid infrastructure was neglected. Liu et al. proposed a safe layer based on an optimal power flow to avoid voltage fluctuations during EV charging (Liu et al., 2023). Wu et al. employed an optimization-based safe module to ensure that the charging power is not exceeded and that the charging demand is met (Wu et al., 2023). However, the safe module-based SDRL depend on the convexity of the model and the case-by-case physical rules. To facilitate more general usage, embedded approaches are currently popular. The DRL approaches that embed safety can also be classified into two categories: combining the reward/cost function or limiting the agent's search space (Ji et al., 2024). Usually, the combination methods employ convex optimization methodologies, such as Lagrangian decomposition and primal-dual decomposition, utilizing the Lagrangian multiplier. This multiplier increases when the cost is severe and the agent's reward is reduced. Therefore, the agent can be aware of unsafe actions during the training. For example, Chen et al. proposed an augmented Lagrangian-based SDRL to ensure the battery charging limits (Chen et al., 2025). However, that work did not consider any safety requirements for grid assets. Ding et al. proposed a Lagrangian-based DDPG for smart-home multi-energy management with EV (Ding et al., 2022). Similarly, Zhang et al. proposed a primal-dual optimization-based SDRL with a safe layer to ensure constraints for storage systems and EV energy capacity (Zhang et al., 2023b). As an example of algorithms that directly limit the agent search space, Li et al. proposed a constrained policy optimization (CPO) approach to ensure charging power constraint (Li et al., 2020). The researchers focused on energy delivery and used CPO to ensure delivery. However, the impact of EV charging on the grid has not been studied. A summary of SDRL approaches reviewed in the field of EV charging is presented in Table 1. Especially, we add a comparison metric, "EV Info", denoting the available information from the EV side in the management. A "Full Info" refer to cases where the actual departure time and energy demand of connected EV are perfectly known. Usually, the existing research assumes that such information is fully available, and includes EV info in the SDRL formulation, as seen from (Zhang et al., 2024b) to (Li et al., 2020). As discussed in previous works from the same authors (Zhang et al., 2025), this is not realistic, and any uncertainty in that information can have a significant impact on control performance.

The first research gap, namely the partial availability of EV information, is identified in the literature reviewed. A second gap relates to the thermal limitations of critical grid infrastructures that are not systematically considered. From the existing works above (Zhang et al., 2023a), considered the power limit of the transformer. However, none have considered the thermal loading constraints, which are an essential requirement for the safe operation of the distribution grid. What's more, the models of those infrastructures, such as power transformers, are always complex (nonlinear and nonconvex), which hinders the usage of the model-based safe module.

Existing works often overlook critical physical constraints, such as transformer thermal loading, and lack consideration of unknown EV information. Thus, there is an urgent need for safe and intelligent charging strategies that can manage EV fleets effectively based on available EV information while protecting grid assets. Therefore, we proposed a model-free SDRL-based G2V strategy by projecting the policy into a safe space. The target is to deliver as much energy as possible to the EV fleet while considering the safe constraints of transformer temperature, bus voltage level, and line loading. The contributions of this work are:

- A G2V management strategy, considering unknown energy demand, departure time and non-observability of physical models: this strategy is more similar to the real-world case, thus is more practical for implementation.
- A model-free SDRL-based approach that considers transformer thermal loading and grid constraints. This method addresses the complex physical model and safety requirements of the system.
- Comparison of the proposed algorithm with other SDRL methods implemented in previous works, as shown in Tables 1 and in terms of sensitivity to observability and hyperparameters. This provides justification for the effectiveness of the proposed methods against the State-of-the-Art (SOTA).

This paper is organized as follows: in section II, we present case study and its model equation to introduce the control problem; in section III, we introduce the model-free SDRL proposed; in section IV, we introduce the baseline algorithms used to benchmark the proposed method's performance and analyze the results; and in section V, we summarize this paper and propose further research potentials.

2. Problem formulation

2.1. Case study

The schematic of investigated G2V case study is shown in Fig. 1. In this work, we consider a power distribution grid with charging stations ($c \in \mathscr{C}$) based on IEEE 33 test case (Baran and Wu, 1989) with each charging station consisting in several charging ports ($n \in \mathscr{N}$). Each charging station is connected to the grid with a power transformer. At each timestep Δt, the controller takes operation information from the grid, transformer and charging station, then computes the control signals, which are charging port power setpoints, and sends signals to the charging station.

2.2. Network model

Giving a power grid with $\mathscr{B}$ buses ($b \in \mathscr{B}$) and $\mathscr{L}$ lines ($l \in \mathscr{L}$), the AC power flow model can be formulated in (1), where $p^l_{b,t}$ and $q^l_{b,t}$ are active and reactive power injection to the bus b, v_b are voltage magnitude of bus b, $G_{bb'}$ and $B_{bb'}$ are real and imaginary part of admittance matrix, $\phi_{bb',t}$ are voltage angle difference between bus b and b'. For buses $b \in \mathscr{B}$, according to Kirchhoff's Law, the power balance is formulated as (2), where $p^g_{b,t}, q^g_{b,t}$ are active/reactive power generation at bus b and $p^d_{b,t}$, $q^d_{b,t}$ are active/reactive power demand at bus b. The current magnitude of the line between bus b and b' is formulated as (3).

$$\begin{cases} p^l_{b,t} = \sum_{b' \in \mathscr{B}} G_{bb'} v_{b,t} v_{b',t} \cos\left(\phi_{bb',t}\right) + B_{bb'} v_{b,t} v_{b',t} \sin\left(\phi_{bb',t}\right) \\ q^l_{b,t} = \sum_{b' \in \mathscr{B}} - B_{bb'} v_{b,t} v_{b',t} \cos\left(\phi_{bb',t}\right) + G_{bb'} v_{b,t} v_{b',t} \sin\left(\phi_{bb',t}\right) \end{cases} \tag{1}$$

$$\begin{cases} p^g_{b,t} + p^l_{b,t} = p^d_{b,t} \\ q^g_{b,t} + q^l_{b,t} = q^d_{b,t} \end{cases}, \forall b \in \mathscr{B}, t \in \mathscr{T} \tag{2}$$

$$i_{l(b \to b'),t} = \left| Y_{bb'} \left(v_{b,t} - v_{b',t}\right) \right|, \forall l \in \mathscr{L}, t \in \mathscr{T} \tag{3}$$

2.3. EV charging model

The EV charging model is formulated as (4), where $p_{n,t}$ is the charging power at charging port n, $\Omega_{n,t}$ is the connection of the port n (binary equals to 1 if an EV is connected). T^{arr}_n and T^{dep}_n are respectively the arrival and departure time of the EV connected to the port n, $\Delta E_{n,t}$ denotes the delivered energy since connection and $\Delta E_{n,t}$ shall not exceed the EV demand $\overline{E}_n$.

$$\begin{cases} 0 \le p_{n,t} \le \overline{p}_n \times \Omega_{n,t} \\ \Omega_{n,t} = 1, \forall t \in \left[T^{arr}_n, T^{dep}_n\right], else\ \Omega_{n,t} = 0 \\ \Delta E_{n,t} = \Delta E_{n,t-1} + p_{n,t} \times \Delta t \\ 0 \le \Delta E_{n,t} \le \overline{E}_n \end{cases} \tag{4}$$

2.4. Transformer thermal model

The normalized electric loading of the transformer $p^{tr}_{c,t}$ at charging station c is calculated as (5), where $\overline{p}^{tr}_c$ is the nominal power of transformer. According to the IEC standard (IEC, 2018), the transformer thermal loading is linked to the oil temperature $\theta^o_{c,t}$ and hotspot temperature $\theta^h_{c,t}$, which are respectively calculated by (6) and (7), where $\left[x, y, r, \kappa_{11}, \kappa_{21}, \kappa_{22}, \tau_0, \tau_w, \Delta\theta^{or}, \Delta\theta^{hr}\right]$ are the transformer parameters, δt is the calculation time resolution (1 min resolution), while $\Delta\theta^1_{c,t}$ and $\Delta\theta^2_{c,t}$ are two auxiliary variables.

$$p^{tr}_{c,t} = \frac{1}{\overline{p}^{tr}_c} \sum_{n \in \mathscr{N}} p_{n,t} \tag{5}$$

$$\begin{cases} \Delta\theta^o_{c,t} = \dfrac{1}{\kappa_{11}\tau_0}\left[\theta^a_t + \Delta\theta^{or}\left(\dfrac{1 + r\left(p^{tr}_{c,t}\right)^2}{1 + r}\right)^x - \theta^o_{c,t-1}\right] \\ \theta^o_{c,t} = \theta^o_{c,t-1} + \Delta\theta^o_{c,t} \times \delta t \end{cases} \tag{6}$$

$$\begin{cases} \Delta\theta^1_{c,t} = \dfrac{1}{\kappa_{22}\tau_w}\left[\kappa_{21}\Delta\theta^{hr}\left(p^{tr}_{c,t}\right)^y - \Delta\theta^1_{c,t-1}\right] \\ \Delta\theta^2_{c,t} = \dfrac{\tau_0}{\kappa_{22}}\left[(\kappa_{21} - 1)\Delta\theta^{hr}\left(p^{tr}_{c,t}\right)^y - \Delta\theta^2_{c,t-1}\right] \\ \theta^h_{c,t} = \theta^o_{c,t} + \left(\Delta\theta^1_{c,t} - \Delta\theta^2_{c,t}\right) \times \delta t \end{cases} \tag{7}$$

2.5. Control problem

The control objective is to maximize the energy delivered to the EVs (when connected), while avoiding any electric or thermal overloading. With the models above, the control problem can be summarize as (8), where $\underline{v}/\overline{v}$ are lower and upper limits of voltage, $\overline{i}$ is the maximum line loading, $\overline{\theta}^o/\overline{\theta}^h$ are maximum oil and hotspot temperature.

$$\max_{p_{n,t}} \sum_{t \in \mathscr{T}} \sum_{c \in \mathscr{C}} \sum_{n \in \mathscr{N}} \Delta E_{n,t}, \text{s.t.} \begin{cases} (1) \sim (7) \\ \underline{v} \le v_{b,t} \le \overline{v}, i_{l,t} \le \overline{i} \\ p^{tr}_{c,t} \le 1.5, \theta^o_{c,t} \le \overline{\theta}^o, \theta^h_{c,t} \le \overline{\theta}^h \end{cases} \tag{8}$$

However, the control problem (8) is too complex to be solved analytically. To begin with, the power flow model is nonconvex. Next, in actual deployment, the EV charging model suffers from multiple uncertainties, especially for the arrival time T^{arr}_n, departure time T^{dep}_n and energy demand $\overline{E}_n$ for the future connected EVs are hard to predict. The second source of uncertainties relate to the connected EVs, for which the accurate T^{dep}_n and $\overline{E}_n$ are not available for the controller. Lastly, the transformer model is nonlinear and partially observable – i.e. the auxiliary variables $\Delta\theta^1_{c,t}$ and $\Delta\theta^2_{c,t}$ do not physically exist. To solve the control problem (8), we then proposed a model-free SDRL approach to solve the control problem (8).

3. Methodology

3.1. Partially-Observable Markov Decision Process

A Partially-Observable Markov Decision Process can handle unobservability (in our case: $T^{dep}_n, \overline{E}_n, \Delta\theta^1_{c,t}, \Delta\theta^2_{c,t}$) in an uncertain environment

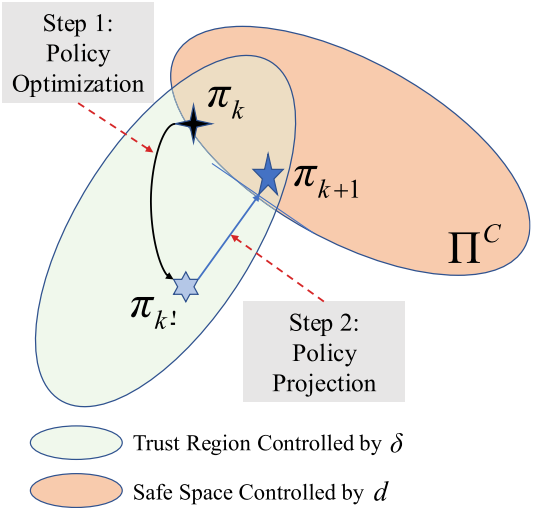

Fig. 2. PCPO Illustration.

with the following formulation:

State: The state vector contains all variables of the environment. It contains the information from the EV and the grid, and it is formulated as (9):

$$s_t = \Big[\mathbf{p_t^d}, \mathbf{q_t^d}, \mathbf{i_{l,t}}, \mathbf{v_{b,t}}, \left(p_{c,t}^{tr}, \theta_{c,t}^{o}, \theta_{c,t}^{h}, \Delta\theta_{c,t}^{1}, \Delta\theta_{c,t}^{2}\right)_{c\in\mathscr{C}}, \left(T_n^{arr}, T_n^{dep}, \overline{E}_n, \Omega_{n,t}, p_{n,t}, \Delta E_{n,t}\right)_{n\in\mathscr{N}}\Big] \tag{9}$$

Observations: As discussed above, $T_n^{dep}, \overline{E}_n$ of the connecting EVs are unknown, together with $\Delta\theta_{c,t}^1, \Delta\theta_{c,t}^2$ of the transformer that are not observable. Therefore, the observations vector is formulated as (10):

$$o_t = \Big[\mathbf{p_t^d}, \mathbf{q_t^d}, \mathbf{i_{l,t}}, \mathbf{v_{b,t}}, \left(p_{c,t}^{tr}, \theta_{c,t}^{o}, \theta_{c,t}^{h},\right)_{c\in\mathscr{C}}, \left(T_n^{arr}, \Omega_{n,t}, p_{n,t}, \Delta E_{n,t}\right)_{n\in\mathscr{N}}\Big] \tag{10}$$

Actions: The control actions are charging power setpoint at each charging port, namely $a_t = [p_{n,t}]$

Reward: The reward function r_t represents the objective of the G2V problem (8). It calculates the total delivered energy to the EV when the certain EV at charging port n departs, and is formulated as (11).

$$r_t = \sum_{c\in\mathscr{C}}\sum_{n\in\mathscr{N}} \Delta E_{n,t}, \forall t = T_n^{dep} \tag{11}$$

Costs: The costs denote the violation of inequality constraints of the G2V problem (8). In the implementation, the cost functions noted as c_j quantify the violation of j- th safety constraint. For constraint j, we note it as a universal form as $x_{t,j} \le \overline{x}_j$, as (12) shows and the total cost is calculated as (13)

$$\begin{cases} \mathbf{x_t} = \left[-v_{b,t}, v_{b,t}, i_{l,t}, p_{c,t}^{tr}, \Delta E_{n,t}, \theta_{c,t}^{o}, \theta_{c,t}^{h}\right] \\ \overline{\mathbf{x}} = \left[-\underline{v}, \overline{v}, \overline{i}, 1.5, \overline{E}_n, \overline{\theta}^{o}, \overline{\theta}^{h}\right] \end{cases} \tag{12}$$

$$c_t = \sum_{x_{t,j}\in\mathbf{x_t}} \frac{|x_{t,j} - \overline{x}_j|}{\overline{x}_j} \times \max\left(0, x_{t,j} - \overline{x}_j\right) \tag{13}$$

3.2. Safe deep reinforcement learning

The expected reward and cost are computed as (14), where $\tau \in \pi$ denotes the control sequences of (s_t, o_t, a_t) computed with a given policy π and γ is the discount factor. The objective of the Safe Deep Reinforcement Learning (SDRL) is to obtain the best control policy π^* with maximal $J^R(\pi)$ and feasible $J^C(\pi)$. This can be mathematically formulated as (15), where d is the tolerance of violation. Therefore, we adopt SDRL to handle the constrained optimal control problem (8).

$$\begin{cases} J^R(\pi) = \mathbb{E}_{\tau\in\pi}\left[\sum_{t\in\mathscr{T}} \gamma^t r_t\right] \\ J^C(\pi) = \mathbb{E}_{\tau\in\pi}\left[\sum_{t\in\mathscr{T}} \gamma^t c_t\right] \end{cases} \tag{14}$$

$$\pi^* = \text{argmax}_\pi J^R(\pi), \text{s.t. } J^C(\pi) \le d \tag{15}$$

Furthermore, the advantage function of reward and cost are defined as (16), where Q_π^R, Q_π^C are the action-value function and V_π^R, V_π^C are the state-value functions. The advantage functions measure how good a new policy is over an existing policy.

$$\begin{cases} A_\pi^R(o_t, a_t) = Q_\pi^R(o_t, a_t) - V_\pi^R(o_t) \\ A_\pi^C(o_t, a_t) = Q_\pi^C(o_t, a_t) - V_\pi^C(o_t) \end{cases} \tag{16}$$

In the proposed control scheme a Projection-based Constrained Policy Optimization (PCPO) is envisioned (Yang et al., 2020). In the k-th episodes of the training process, the parameterized policy π_k is projected into the feasible space which is defined as $\Pi^C : \{\pi | J^C \le d\}$, and ϑ_k is the policy parameters vector. The PCPO then consists of two steps, with the policy optimization followed by its projection (Fig. 2).

The PCPO then consists of two steps, with the policy optimization followed by its projection. The first step is to find an intermediate policy $\pi_{k'}$ over current policy π_k, by trust region policy optimization (Schulman et al., 2015) using advantage function. The policy is updated within a given range δ from the current policy, as (17) shows, D_{KL} is the Kullback-Leibler (KL) divergence (The KL divergence can be achieved by Pytorch distribution toolbox). The second projection step can be formulated as (18). The target is to find a policy π_{k+1} within the feasible space Π^C and with the smallest distance to $\pi_{k'}$.

$$\pi_{k'} = \underset{\vartheta}{\text{argmax}} \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^R\right], s.t. \underset{o\sim\pi_k}{\mathbb{E}} \left[D_{KL}(\pi \| \pi_k)\right] \le \delta \tag{17}$$

$$\pi_{k+1} = \underset{\vartheta}{\text{argmin}} D_{KL}(\pi \| \pi_{k'}), s.t.\ J^C(\pi_k) + \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^C\right] \le d \tag{18}$$

Obviously, (17) and (18) cannot be solved analytically. Therefore, the Taylor expansion approximation are adopted, similar with (Yang et al., 2020), (Achiam et al., 2017). First, we linearize the expectation of A_π^R, A_π^C in around ϑ_k and its first order Taylor expansion can be noted as (19). Because the advantage function measures the difference between two different policies, then we have $\underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_{\pi_k}^R\right] = 0$, $\underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_{\pi_k}^C\right] = 0$ as these are measurement of the same policy. We note $g = \nabla_\vartheta \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^R\right]$ and $a = \nabla_\vartheta \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^C\right]$. These gradient operations can be achieved by autograd function of Pytorch.

$$\begin{cases} \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^R\right] \approx \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_{\pi_k}^R\right] + \left(\nabla_\vartheta \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^R\right]\right)^T (\vartheta - \vartheta_k) \\ \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^C\right] \approx \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_{\pi_k}^C\right] + \left(\nabla_\vartheta \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} \left[A_\pi^C\right]\right)^T (\vartheta - \vartheta_k) \end{cases} \tag{19}$$

To approximate the KL-Divergence, we adopted the second order Taylor series expansion used in (Schulman et al., 2015) (because the first and second term are cancelled out, details in appendix A), which can be

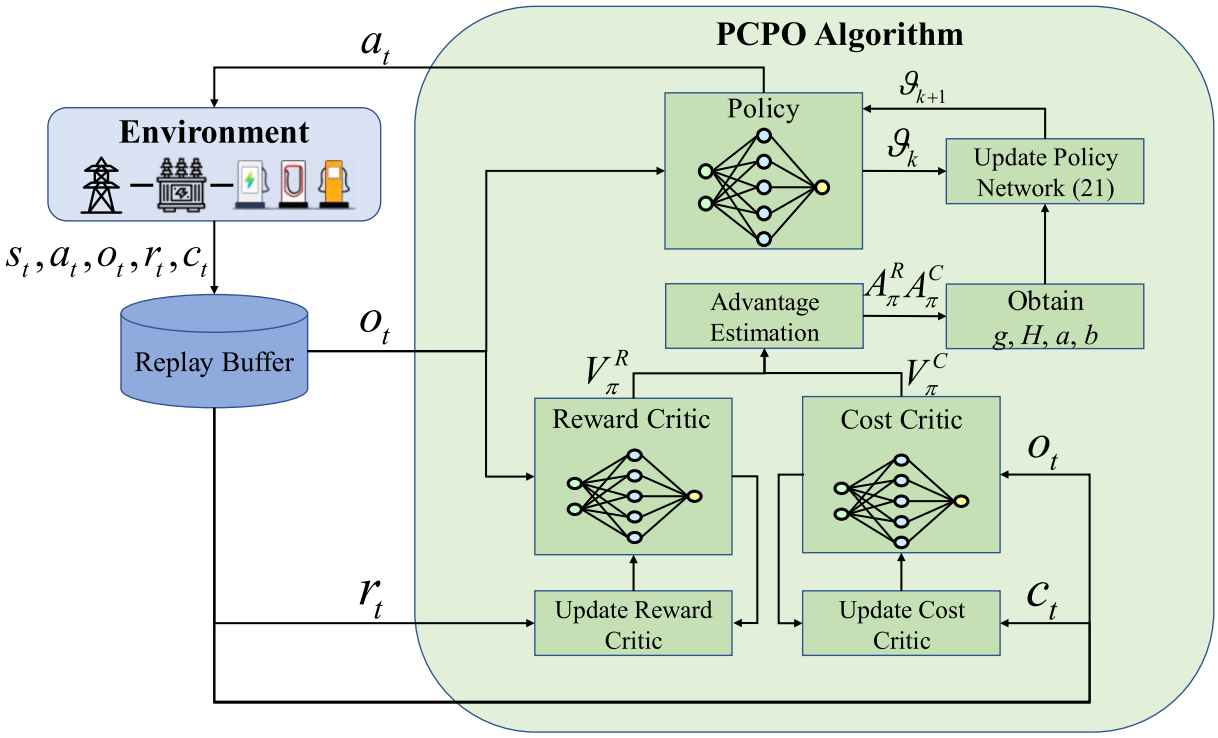

Fig. 3. PCPO training diagram.

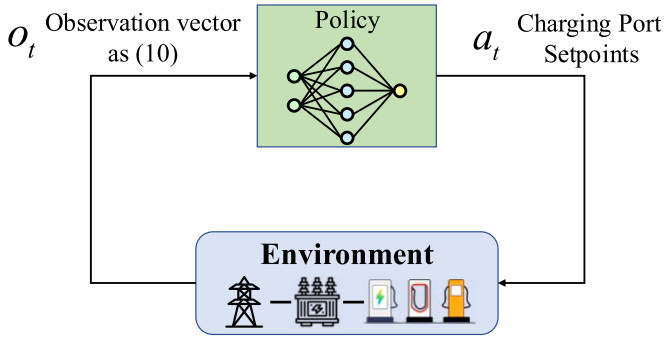

Fig. 4. Online deployment.

expressed as (20), where $H = \left[\nabla^2_{\vartheta} D_{KL}(\pi \| \pi_k)\right]_{\vartheta=\vartheta_k}$. The H is achieved by autograd function of Pytorch.

$$D_{KL}(\pi \| \pi_k) \approx \frac{1}{2}(\vartheta - \vartheta_k)^T H (\vartheta - \vartheta_k) \tag{20}$$

Therefore, the policy parameter update in PCPO can be approximated as (21), where $b = J^C(\pi_k) - d$. This can be executed by subtracting the tolerance of violation d from the episode cost.

$$\begin{cases} \vartheta_{k'} = \underset{\vartheta}{\text{argmax}} g^T(\vartheta - \vartheta_k), s.t. \frac{1}{2}(\vartheta - \vartheta_k)^T H(\vartheta - \vartheta_k) \le \delta \\ \vartheta_{k+1} = \underset{\vartheta}{\text{argmax}} \frac{1}{2}(\vartheta - \vartheta_{k'})^T H(\vartheta - \vartheta_{k'}), s.t. a^T(\vartheta - \vartheta_{k'}) + b \le 0 \end{cases} \tag{21}$$

The approximated form (21) can be analytically solved by Karush–Kuhn–Tucker conditions and the policy parameter update can be formulated as (22). The process of algorithm and training is shown in Alg. 1 and Fig. 3. The parameter update of the critic network and estimation of advantage are omitted here as it is same with the Proximal Policy Optimization (PPO) (Schulman et al., 2017). What's more, in practice, the term $H^{-1}g$ (and $H^{-1}a$) is calculated through conjugate gradient method, which is same with (Schulman et al., 2015) and the details are in appendix B. Together with the g, a and b, the parameter of the policy network is updated.

$$\vartheta_{k+1} = \vartheta_k + \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g - \max\left(0, \frac{\sqrt{\frac{2\delta}{g^T H^{-1} g}} a^T H^{-1} g + b}{a^T H^{-1} a}\right) H^{-1} a \tag{22}$$

Alg. 1. Algorithm of PCPO

Inputs: Hyperparameters δ, d, Total Episodes $\mathscr{K}$, Replay Buffer $\mathscr{D}$
Outputs: Trained Policy π^*
For $k \in \mathscr{K}$:
  Initialize environment;
  For $t \in \mathscr{T}$:
    Run $a_t \sim \pi_k$ and store transitions $(s_t, a_t, o_t, r_t, c_t)$ in $\mathscr{D}$;
  Sample transitions from $\mathscr{D}$, then compute g, a, H, b;
  Update ϑ_{k+1} follows (22), and update reward & cost critic network.

After the policy network in Fig. 3 is trained, it is deployed for online testing, as shown in Fig. 4. During each timestep, the trained policy network takes the observations as input and then generates actions. These actions involve setting the charging port power setpoint, and the charging port delivers this power to the connected EV.

4. Experiments & results

4.1. Benchmark algorithms

1) Uncontrolled Charging

The uncontrolled charging (Unctrl) follows a simple rule-based algorithm that delivers maximum charging capacity $\bar{p}_n$ regardless of transformer temperature and network constraints.

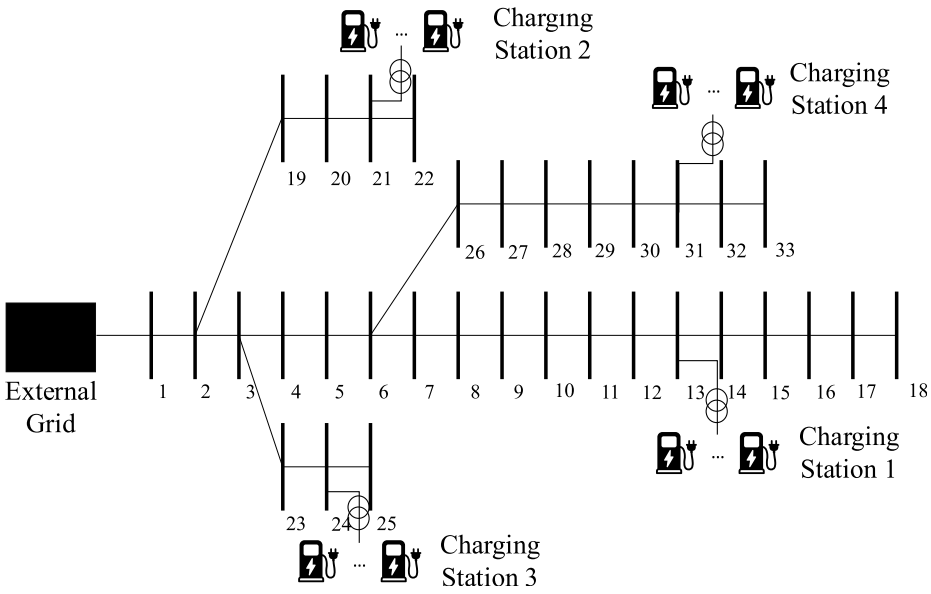

Fig. 5. IEEE 33 buses with charging stations.

Table 2
Values of transformer parameters.

Parameter	x	y	κ_{11}	κ_{21}	κ_{22}
Value	0.8	1.6	1.0	1.0	2.0
Parameter	τ_o	τ_w	$\Delta\theta^{or}$	$\Delta\theta^{hr}$	r
Value	180 min	4 min	55 K	23 K	5

2) Proximal Policy Optimization

The Proximal Policy Optimization (PPO) is a classic on-policy algorithm (Schulman et al., 2017), and it combines the reward and cost by (23) before the policy is updated following (24), where L is the loss function. The PPO is a well-studied algorithm, and applied in the EV charging management problem, such as (Qi et al., 2023) and (Zhang et al., 2024b). In this work, the PPO benchmark is realized with open-sourced software Stable-baseline 3 (Raffin et al., 2021).

$$A_{\pi_k} = A^R_{\pi_k} - A^C_{\pi_k} \tag{23}$$

$$\begin{aligned} \vartheta_{k+1} &= \underset{\vartheta}{\operatorname{argmax}} \underset{\substack{o\sim\pi_k \\ a\sim\pi}}{\mathbb{E}} [L(\vartheta)] \\ s.t.\ L(\vartheta) &= \min\left(\frac{\pi(a|o)}{\pi_k(a|o)}A_{\pi_k}, \operatorname{clip}\left(\frac{\pi(a|o)}{\pi_k(a|o)}, 1-\varepsilon, 1+\varepsilon\right)A_{\pi_k}\right) \end{aligned} \tag{24}$$

3) Lagrangian-based PPO

The Lagrangian-based PPO (PPOLag) (Paternain et al., 2023) hybridizes primal-dual decomposition and PPO algorithm. The algorithm calculates the surrogated advantage with a non-negative Lagrangian multiplier λ, as in (25), then updates the policy with combined objective with PPO follows (24). Lagrangian multiplier λ is updated if the new policy is outside the feasible space, as in (26) with η an hyperparameter that controls the update speed of λ. Recently, the PPOLag is also studied for the case of EV fleet charging management, such as (Ding et al., 2022) and (Zhang et al., 2023b). In this work, the PPOLag is performed with open-sourced software Omnisafe (Ji et al., 2024).

$$A_{\pi_k} = A^R_{\pi_k} - \lambda_k A^C_{\pi_k} \tag{25}$$

$$\lambda_{k+1} = \max\left(\lambda_k + \eta\left(J^C(\pi_{k+1}) - d\right), 0\right) \tag{26}$$

4) Constrained Policy Optimization

The Constrained Policy Optimization (CPO) (Achiam et al., 2017) is similar to the PCPO proposed in the paper. The major difference is that the CPO do not calculate an intermediate policy $\pi_{k'}$ but directly updates the policy by taking safety into consideration, as (27) shows. During the training, if the (27) is infeasible, the parameter is updated with (28). The CPO is studied in (Li et al., 2020) when the safety are considered in EV fleet charging management. In this work, the CPO is performed with open-sourced software Omnisafe (Ji et al., 2024).

$$\begin{aligned} &\vartheta_{k+1} = \underset{\vartheta}{\operatorname{argmax}} g^T(\vartheta - \vartheta_k) \\ &s.t. \begin{cases} a^T(\vartheta - \vartheta_k) + b \le 0 \\ \frac{1}{2}(\vartheta - \vartheta_k)^T H(\vartheta - \vartheta_k) \le \delta \end{cases} \end{aligned} \tag{27}$$

$$\vartheta_{k+1} = \vartheta_k - \sqrt{\frac{2\delta}{a^T H^{-1} a}} H^{-1} a \tag{28}$$

4.2. Case configuration

Fig. 5 shows the configuration of the power distribution grid with charging stations considered as a case study here. There is a total of 4 charging stations, each with 13 charging ports. The grid load time series is from Simbench (Meinecke et al., 2020), the EV charging records are from Palo Alto EV Charging Record (Wong, 2021). Parameters of transformer thermal loading model are shown in Table II, which are from the IEC standard (IEC, 2018). The nominal power for each transformer is 40 kW. The maximum power limit for each charging port is set to 7.7 kW. The thermal limitation $\overline{\theta}^o$ and $\overline{\theta}^h$ for the transformer are 102 °C and 120 °C. The bus voltage limitation is [0.9, 1.1] p. u. and the maximum line loading is 0.25 kA.

4.3. Dataset description

The Simbench (Meinecke et al., 2020) provides the normalized and realistic load profiles, with 15-min time resolution. The load profile

Table 3
Comparison of experiment setups.

Section	Objective	Experiments Setups	Compared Methods	Metrics
IV.C(1)	**Online Performance**	Moderate Ambient Temperature (Sept.2019)	PCPO vs. SOTA Benchmarks	Energy Delivery & Constraint Violations
IV.C(2)	**Unknown Information Impact**	Training Configuration in Appendix.C	PCPO (full info) vs. PCPO (partial info)	Training Rewards
IV.C(3)	**Ambient Temperature Impact**	Various Ambient Temperature	PCPO vs. SOTA Benchmarks	Energy Delivery & Constraint Violations
IV.C(4)	**Transformer Aging**	Various Ambient Temperature	PCPO vs. Uncontrolled Charging	Loss-of-Life for Transformers
IV.C(5)	**Hyperparameter Analysis**	Training Configuration in Appendix.C	Various Hyperparameter of PCPO	Training Rewards

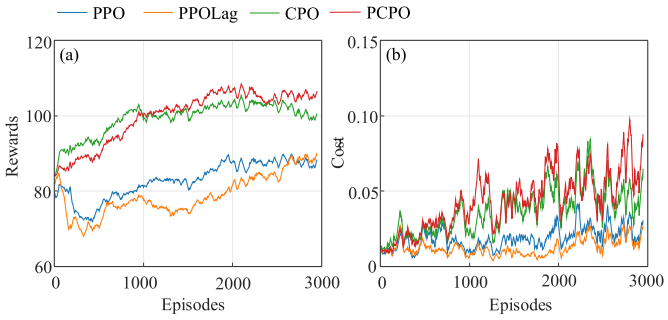

Fig. 6. Sdrl training curves (a) rewards (b) costs.

Table 4
Peak rewards and costs of SDRL algorithms.

	PPO	PPOLag	CPO	PCPO
Rewards	94.56	93.98	110.32	114.05
Costs	0.03	0.03	0.19	0.21
Net Rewards	94.53	93.35	110.13	**113.84**

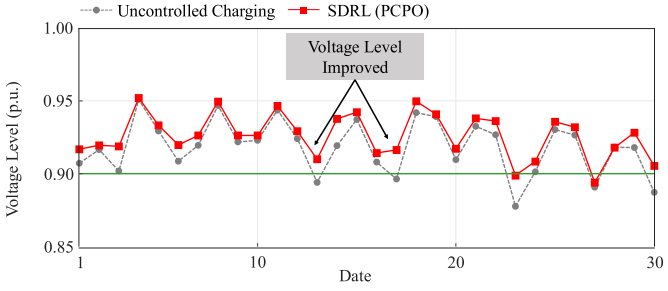

Fig. 7. Lowest voltage of the grid throughout testing.

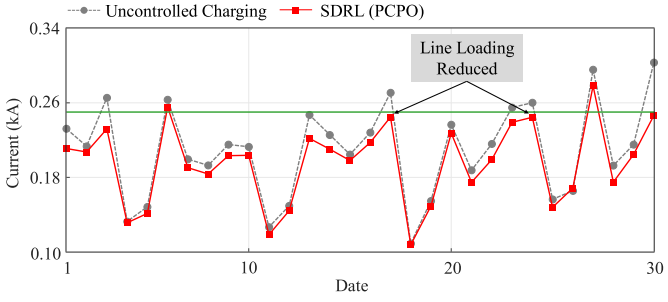

Fig. 8. Peak current of the grid throughout testing.

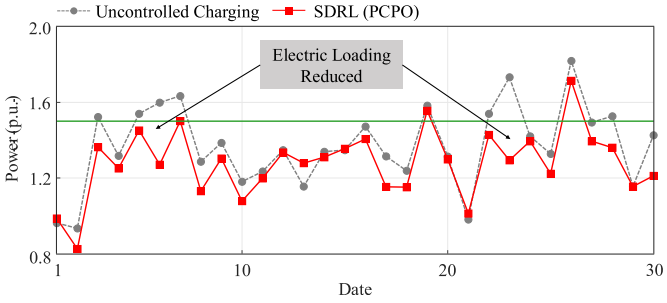

Fig. 9. Peak power of transformers throughout testing.

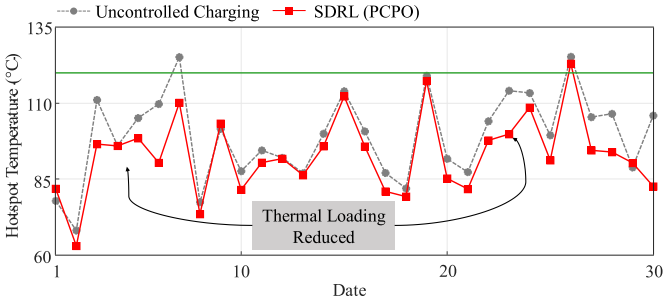

Fig. 10. Peak hotspot temperatures of transformers.

Table 5
Numerical results of entire sept. 2019.

	E (kWh)	c_I	c_V	c_P	c_{oil}	c_{hst}
Unctrl	31710.13	1.72	0.48	0.82	0	0.13
PPO	24654.42	0.03	0.14	0.19	0	0
PPOLag	23532.14	0.09	0.01	0.13	0	0
CPO	28493.08	0.35	0.10	0.28	0	0
PCPO	29186.82	0.36	0.03	0.35	0	0.04

contains variables of active power demand $p^d_{b,t}$ and reactive power demand $q^d_{b,t}$. The Simbench contains 4 main types of load profiles, namely rural, semi-urban, urban and commercial, and each contains different profiles. These load profiles are from anonymized recorded power measurement of 1-year measurement. The grid voltage level varies from 0.4 kV to 380 kV. In total, the Simbench provides around 30 different normalized load profiles. We adopted the commercial load profile because the EV charging records are from public charging stations.

The Palo Alto EV Charging Record (Wong, 2021) contains the variables of the arrival time T^{arr}_n , departure time T^{dep}_n and energy demand E_n of historical charging processes. The Palo Alto EV Charging Record contains 259415 charging records from 2011. From these charging records, there are 169408 valid records (e.g., no NaN values). On average, in each day, there are around 110 EV hosted by the charging station, and

Table 6
Results of Friedman and Nemenyi post Hoc test.

Test Data	χ^2	p-value	Mean Ranks				Significantly Different Pairs at CD = 1.048
Daily Rewards of 20 Days	43.98	1.52×10^{-9}	PPO	PPOLag	CPO	PCPO	PPO-PCPO
			2.95	3.70	2.25	1.10	PPOLag-PCPO
							CPO-PCPO

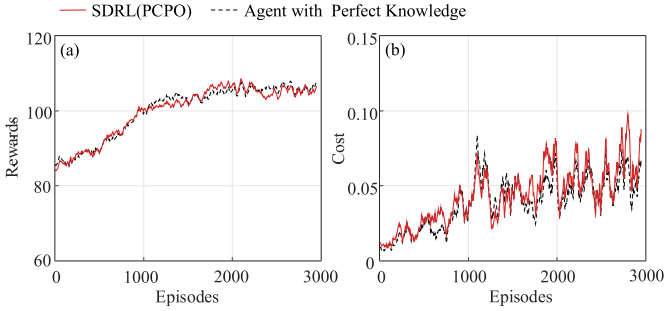

Fig. 11. Agents w./w.r.t. EV knowledge (a) Rewards (b) Cost.

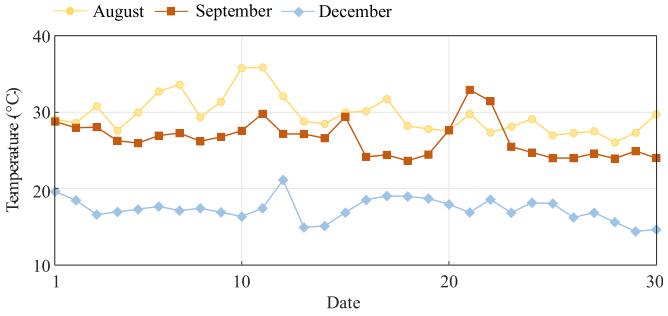

Fig. 12. Daily peak temperature of August and September.

Table 7
Influence of ambient temperature.

Test	Strategy	E (kWh)	c_I	c_V	c_P	c_{oil}	c_{hst}
Aug.	Unctrl	32197.01	4.06	0.96	1.75	0	0.88
	PPO	25135.06	0.21	0.05	0.20	0	0
	PPOLag	24542.40	0.83	0.08	0.03	0	0
	CPO	28646.80	1.85	0.31	0.53	0	0.07
	PCPO	28574.51	1.18	0.12	0.25	0	0.07
Sept.	Unctrl	31710.13	1.72	0.48	0.82	0	0.13
	PPO	24654.42	0.03	0.14	0.19	0	0
	PPOLag	23532.14	0.09	0.01	0.13	0	0
	CPO	28493.08	0.35	0.10	0.28	0	0
	PCPO	29186.82	0.36	0.03	0.35	0	0.04
Dec.	Unctrl	29715.15	5.75	0.56	1.55	0	0
	PPO	23681.46	1.46	0.24	0.87	0	0
	PPOLag	22959.35	0.56	0.20	0.03	0	0
	CPO	29253.47	4.24	0.38	1.33	0	0
	PCPO	29253.47	4.80	0.49	0.51	0	0

the charging station delivers 1030.73 kWh energy to the EVs. We resampled the Palo Alto EV Charging Record into 15-min time resolution, in order to adapt the Simbench dataset.

4.4. Experiment setups

The experiments environment is done in Python, and the simulation of the case study in section IV.B is achieved with Gymnasium (Towers et al., 2024). Notably, a clamping function in PyTorch ensures the charging port power limit $\bar{p}_n$, as the Gymnasium requires a fixed action space. We normalized the charging port power into $[-1, 1]$ to be compatible with action space and ensure better performance. The training and testing process are run on our own PC. The configuration of our CPU is 13th Gen Intel(R) Core (TM) i7-13700H 2.40 GHz with 32 GB RAM.

For the initial hyperparameters, the cost limit d for all the algorithms (except PPO) are set to 0.5. Other hyperparameters can be found in Table IX in appendix C. All the agents are trained with historical data from Jan. 1st, 2019 to Jun. 30th, 2019. The agents are tested in different months in order to test the robustness against different ambient temperature scenario, namely Aug.2019, Sept.2019 and Dec.2019. The results are obtained with entire testing months.

We conducted five experiments for verification. First, we conducted a real-time online test of the PCPO against other benchmarks to verify its effectiveness. Next, we compared the PCPO with full knowledge of EV charging information with the proposed framework (unknown departure and energy demand). Third, we considered higher and lower ambient temperature environments to test the robustness of the system. To further emphasize the benefits of the proposed charging management strategy, we compared the impact of EV charging on transformer ageing. Ultimately, we present a hyperparameter analysis. The evaluation metrics are total energy delivery and total constraint violations. The experiments are summarized in Table 3.

Table 8
Transformer aging effect (loss-of-life).

Test	Alg.	Station 1	Station 2	Station 3	Station 4
Aug.	Unctrl	44.8	1312.8	401.6	264.8
	PCPO	18.1 **(-59 %)**	136.8 **(-89 %)**	48.4 **(-88 %)**	62.4 **(-76 %)**
Sept.	Unctrl	47.4	138.4	186.7	89.4
	PCPO	19.8 **(-57 %)**	107.8 **(-22 %)**	87.4 **(-53 %)**	49.6 **(-44 %)**
Dec.	Unctrl	13.58	26.63	91.21	23.98
	PCPO	11.57 **(-8.6 %)**	17.87 **(-32.9 %)**	65.85 **(-27.8 %)**	21.99 **(-8.3 %)**

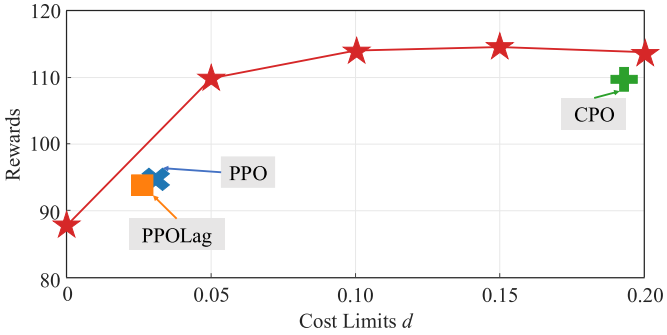

Fig. 13. Cost limits influence on the proposed agent.

4.5. Results

5) Training & Test Results

The reward and cost curves during training are shown in Fig. 6. The

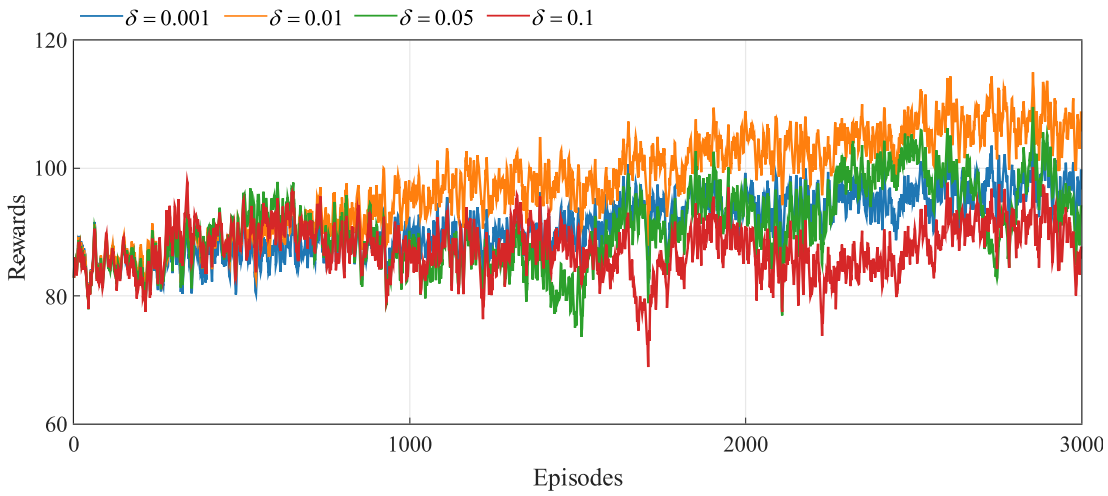

Fig. 14. KL-divergence tolerance influence on the proposed agent.

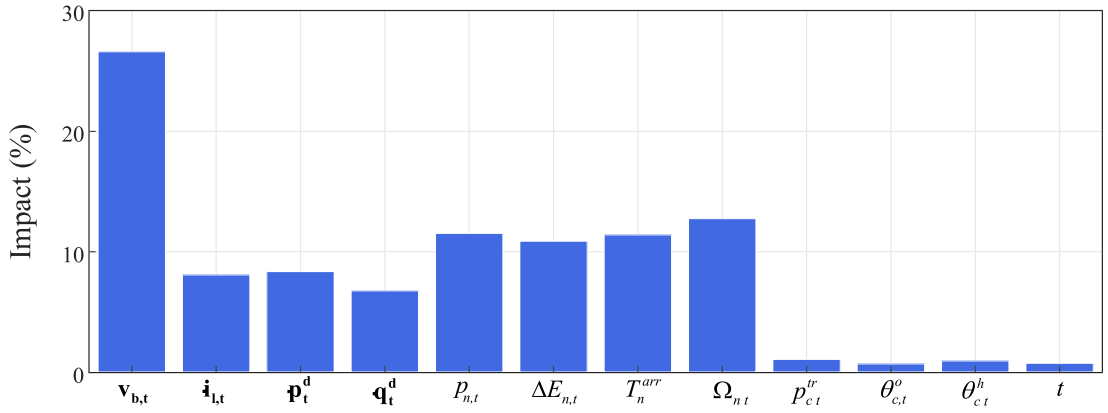

Fig. 15. Impact Index of each feature.

curves are smoothed using a moving average with a window size of 50 episodes. The reason for the cost increase with the reward is that with higher energy delivery to the EV fleet, the transformers and grid lines' power will be greater, with a higher possibility of overloading. Nevertheless, all algorithms ensure that the cost is lower than the cost limit d, i.e., constraints fulfilled. Meanwhile, the PCPO agent achieves the highest rewards, as shown in Table IV. The results show that the PCPO has the highest net rewards, 3.4 % higher than the CPO and 21.9 % higher than the PPOLag. Compared with the CPO, the PCPO agent converges more slowly yet finally reaches a higher reward. These results show that the proposed algorithm has the best understanding of constraints, as it can achieve the highest rewards while ensuring the cost limit is met.

Figs. 7 and 8 display the grid's lowest voltage and peak current for the proposed PCPO and uncontrolled charging, along with daily values during the test period. From Fig. 7, the proposed methods can reduce the voltage violation derivations, as the lowest voltage of the proposed charging is higher than that of uncontrolled charging. The results show that the proposed method can avoid current overloading. What's more, the average runtime for the proposed method is 0.0002 s per timestep, much smaller than the control time resolution.

The peak power and peak hotspot temperature of the four transformers throughout testing are shown in Figs. 9 and 10. According to the results, the proposed methods can effectively avoid power overloading and thermal overloading. Meanwhile, the proposed method can reduce thermal loading.

The numerical results of the benchmark and proposed methods are summarized in Table V. The metrics are the total energy delivery and total constraint violations, where $c_I, c_V, c_P, c_{oil}, c_{hst}$ are cost of current, voltage, transformer power overloading, oil temperature and hotspot temperature overloading, which are defined in (13). From the results obtained, uncontrolled charging leads to the highest constraint violation. All benchmarks and proposed methods can reduce constraint violations, albeit at the cost of less energy delivery. The PPO and PPOLag can deliver 77.7 % and 74.2 % energy, considerably lower than the proposed PCPO. Meanwhile, the proposed method can reduce the current violation by 79.0 % and voltage violation by 93.7 %, improving the grid's safe operation. For the transformer aspect, the proposed methods reduced the power overloading by 57.3 % and the hotspot temperature overloading by 69.2 %.

To analyze whether there are statistically significant differences between the PCPO and the benchmarks (PPO, PPOLag and CPO), we performed nonparametric statistical test method, namely Friedman Test and Nemenyi Post Hoc Test (Terpilowski, 2019). The test data are daily rewards of 20 days. From the results in Table 6, the p-value from Friedman test is much smaller than 0.05. What's more, the Nemenyi Post Hoc Test results shows that the PCPO is significantly different from the other benchmarks (mean ranks difference more than 1.048).

6) Influence of Uncertain EV Information

In order to analyze the effectiveness of the proposed control, which has no access to the actual departure and energy demand of EVs, we ran a comparison experiment between the proposed implementation and an

agent that could access $\bar{E}_n, T_n^{dep}$ as (29) shows. The rewards and costs curves are shown as Fig. 11. From the results, the proposed method can reach similar performance with the agent that have full and perfect knowledge of EV information.

$$o_t = \left[\mathbf{p_t^d}, \mathbf{q_t^d}, \mathbf{i}_{\mathbf{l},\mathbf{t}}, \mathbf{v}_{\mathbf{b},\mathbf{t}}, \left(p_{c,t}^{tr}, \theta_{c,t}^{o}, \theta_{c,t}^{h},\right)_{c\in\mathscr{C}}, \left(T_n^{arr}, T_n^{dep}, \bar{E}_n, \Omega_{n,t}, p_{n,t}, \Delta E_{n,t}\right)_{n\in\mathscr{N}}\right] \tag{29}$$

7) Influence of the Ambient Temperature

The transformer thermal loadings (6) and (7) are highly related to the ambient temperature. The previous test period covered the entire month of September. Additional tests are performed with greater temperature (August) and cooler values (December) as displayed in Fig. 12.

The numerical results are shown in Table VII. With higher ambient temperatures, the proposed method reduces its energy delivery to 88.7 %, around 3 % less than the energy delivery ratio in September. In this scenario, avoiding constraint violation becomes a priority for the proposed method. The proposed methods significantly reduced current overloading by 70.9 %, voltage violation by 87.5 %, power overloading by 85.7 %, and hotspot temperature overloading by 92.0 %. Compared with the benchmarks, the results are similar with the previous test in section IV.C(1), that the PPOLag deliver much less energy to the EV fleet.

When the ambient temperature is low (December), there will be no thermal overloading. In this case, the proposed method tends to satisfy the energy demand of the EV fleet (delivers 98.4 % of energy), and still reduces 67.1 % of transformer power overloading compared to the uncontrolled charging. Due to the low temperature, the thermal loading is alleviated. Therefore, the proposed method delivers 5 % more energy to the EV fleet. The PPOLag cannot understand the change of the temperature, still delivers much less energy. Therefore, the proposed method is effective under both higher and lower ambient temperatures.

8) Transformer Aging Reduction

According to (IEC, 2018), the thermal loading of the transformer can significantly affect its aging on the transformer. The transformer aging model is defined as (30), where V_t is the normalized aging and Λ is the total loss of life with the unit of hour. In Table VIII, the total loss of life for each transformer of the case study is displayed. The results show that uncontrolled charging will lead to a greater loss of life, especially in high ambient temperatures (August). This is due to the high thermal loading. With the proposed method, the loss of life associated with each transformer is reduced, particularly in August. If the ambient temperature is low, for example, in December, the proposed method can still reduce the aging, but the amount of reduction is smaller.

$$\begin{cases} V_t = 2^{\left(\theta_t^{hst}-98\right)/6} \\ \Lambda(h) \approx \sum_{t\in\mathscr{T}} (V_t \times \Delta t) \end{cases} \tag{30}$$

9) Hyperparameter Analysis

In Fig. 13, we show the influence of cost limitation d on the proposed agent. During this experiment, d is set from 0 to 0.2 with 0.05 step size. From the results, if the value of d is too low, the agent will reach lower reward because of the smaller feasible action space. With increasing d, the agent returns better performances. It is noteworthy to mention that, when the d is set to 0.05, the proposed agent reached a similar cost level compared with the PPO and PPOLag agents, while still achieving a higher reward. This further proves that the proposed agent has a better conception of constraints and can better balance reward and cost.

From the (17), another key hyperparameter for the SDRL is the KL-divergence Tolerance δ. This hyperparameter controls the policy search range. In Fig. 14 we provide the training performance under different δ. From the results, if the δ is too small, the agent learning curve converges much slower, and results in sub-optimal with given training episodes. If the δ is too high, the agent also results in sub-optimal, as the search step is too big that the agent may by-pass the optimal policy.

5. Conclusion

This paper proposes a model-free, safe DRL for EV fleet charging while considering grid and transformer constraints. To be realistic, this work considers the unknown departure time and energy demand, as well as the partial observability of the transformer thermal models. Several benchmark SDRL algorithms are considered to assess the effectiveness of the proposed algorithms. Among these benchmarks, the Lagrangian-based PPO results in the lowest constraint violation, yet it delivers less than 80 % of the energy to the EV Fleet. On the other hand, the proposed methods can deliver around 92 % of energy to the EV fleet while reducing constraint violations, especially in power grid line loading and transformer thermal loading, which demonstrates that the proposed methods can effectively charge the EV fleet while ensuring safe constraints. To analyses the influence of unknown EV information, we compared the proposed strategy with the SDRL formulation that has full information. The results show that the proposed strategy can successfully overcome the drawback of unknown EV information. To analyze the influence of weather, the proposed method was tested in both high and low ambient temperature environments, and the results show that it remains effective, as the proposed method deliver more than 90 % energy to the EV. To further illustrate the effectiveness of the proposed strategy, we analyzed the transformer aging, and the proposed method can reduce aging by at most 89 %, compared with uncontrolled charging. The major limitation of the proposed method is that it does not eliminate constraint violations. Therefore, one interesting research direction is to develop an approach that ensures no constraint violation.

The proposed strategy offers a solution for integrating EVs into a distribution grid, particularly in scenarios where the EV's energy demand and departure time are uncertain. Among potentials for further investigation: 1) considering a more complex transformer thermal model, such as integration of a cooling system; 2) considering the integration of transportation networks with power distribution grid, such as (Han et al., 2025) 3) considering integration of storage and renewable energy generation, and consequently their forecasting.

CRediT authorship contribution statement

Zhewei Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Rémy Rigo-Mariani:** Writing – review & editing, Visualization, Supervision, Investigation, Funding acquisition. **Nouredine Hadjsaid:** Supervision, Project administration, Funding acquisition. **Yan Xu:** Supervision, Funding acquisition.

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.

Acknowledgement

This work was supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme, of the grant DESCARTES.

Appendix

M. Explanation of The KL Divergence Approximation

The second order Taylor Expansion can be derived as (31). Obviously, the first item $D_{KL}(\pi_k\|\pi_k)$ is zero as this measure the distance of the same policy. Meanwhile, $D_{KL}(\pi\|\pi_k)$ reach the minimum when $\vartheta \to \vartheta_k$, then we have the first order gradient $[\nabla_\vartheta D_{KL}(\pi\|\pi_k)]_{\vartheta=\vartheta_k} = 0$. Therefore, the KL divergence is approximated as (20).

$$D_{KL}(\pi\|\pi_k) \approx D_{KL}(\pi_k\|\pi_k) + [\nabla_\vartheta D_{KL}(\pi\|\pi_k)]_{\vartheta=\vartheta_k}(\vartheta-\vartheta_k) + \frac{1}{2}(\vartheta-\vartheta_k)^T\left[\nabla^2_\vartheta D_{KL}(\pi\|\pi_k)\right]_{\vartheta=\vartheta_k}(\vartheta-\vartheta_k) \tag{31}$$

N. Explanation of Conjugate Gradient Method

The Conjugate Gradient Method is introduced in (Hestenes and Stiefel, 1952). It is a numerical method to solve linear equations such as $H\chi = g$ when H is symmetric. This algorithm is shown in Alg. 2.

Alg. 2. Iterative Conjugate Gradient Method

Input: Initial Solution χ_0 Tolerance, Maximum Iteration $\bar{k}$
Output: Final Solution χ^*
$\nu_0 = g - H\chi_0$
$\rho_0 = r_0$
While $\|\nu_k\|_2 > \xi$ and $\leq \bar{k}$:
$\alpha_k = \frac{\nu_k^T \nu_k}{\rho_k^T H \rho_k}$
$\chi_{k+1} = \chi_k + \alpha_k \rho_k$, $\nu_{k+1} = \nu_k - \alpha_k^T H \alpha_k$
$\beta_k = \frac{\nu_{k+1}^T \nu_{k+1}}{\nu_k^T \nu_k}$
$\rho_{k+1} = \nu_{k+1} + \beta_k \rho_k$
$k = k + 1$
$\chi^* = \chi_k$

O. Hyperparameters for Algorithms

Table IX
Hyperparameters for Algorithms

Parameters	PPO	PPOLag	CPO	PCPO
Hidden Layer	[64, 64]	[64, 64]	[64, 64]	[64, 64]
Activation Function	Tanh	Tanh	Tanh	Tanh
Buffer Size	1e6	1e6	1e6	1e6
Batch Size	64	64	64	64
Learning Rate	0.003	0.003	0.003	0.003
KL Divergence Tolerance δ	0.02	0.02	0.02	0.02
Discount Factor γ	0.99	0.99	0.99	0.99
Clip Range ε	0.2	0.2	–	–
Lagrangian Initial	–	10	–	–
Lag Learning Rate η	–	0.035	–	–
Tolerance ξ	–	–	1e-10	1e-10
Maximum Iteration $\bar{k}$	–	–	10	10

P. Feature Analysis of Proposed Methods

In order to compare the different feature impact, we studied the sensitivity of policy network with given features. The sensitivity is defined as (32), which is the expectation of L2-norm of gradient between the policy network output (action vector a) against input element (j-th element of observation vector o) The sensitivity is analyzed as Alg. 3. In each iteration, the observation vector o are randomly generated with each element within [0,1]. We first calculate get a Jacobian matrix ($\mathbb{R}^{|a|\times|o|}$). Next, we compute the impact of each input by L2-norm. Then, we group the impact vector regarding to each feature vector. After enough iterations, we calculate the average of these impact vectors and normalize the final results.

$$\mathscr{I}_j = \mathbb{E}\left[\left\|\frac{\partial a}{\partial o_j}\right\|_2\right] \tag{32}$$

Alg. 3. Sensitivity Calculation of Policy Network

Input: Trained Policy Network, Total Iterations K
Output: Normalized Impact Vector
For $k \in K$:
  Randomly generate observation vector o and get the action vector a by trained policy network;
  Calculate the Jacobian Matrix J, where $J_{i,j} = \frac{\partial a_i}{\partial o_j}$;
  Calculate the Impact $\mathscr{I}$ by (32) and get Impact vector $\mathscr{I}$;
  Group and sum the $\mathscr{I}$ by the features in (10);
Calculate the Average of $\mathscr{I}$ in each iteration, and normalize the result.

We run 1000 iterations, and the results are shown in Fig. 15. From the results, the bus voltage has the most impact on the policy network, while the charging power, port occupancy, delivered energy and arrival time has also very high impact. The transformer temperature has less impact on the policy network. As the transformer temperature is related to the charging power, the policy network focus more on the charging power, and implicitly ensure no thermal overloading.

Data availability

The data that has been used is confidential.

References

World Energy Outlook 2024," IEA. Accessed: February. 20, 2025. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2024.

Achiam, J., Held, D., Tamar, A., Abbeel, P., 2017. Constrained policy optimization. In: Proceedings of the 34th International Conference on Machine Learning. PMLR, pp. 22–31. https://doi.org/10.48550/arXiv.1705.10528.

Baran, M.E., Wu, F.F., 1989. Network reconfiguration in distribution systems for loss reduction and load balancing. IEEE Trans. Power Deliv. 4 (2), 1401–1407. https://doi.org/10.1109/61.25627.

Chen, G., Yang, L., Cao, X., 2025. A deep reinforcement learning-based charging scheduling approach with augmented lagrangian for electric vehicles. Appl. Energy 378, 124706. https://doi.org/10.1016/j.apenergy.2024.124706.

Ding, H., Xu, Y., Chew Si Hao, B., Li, Q., Lentzakis, A., 2022. A safe reinforcement learning approach for multi-energy management of smart home. Electr. Power Syst. Res. 210, 108120. https://doi.org/10.1016/j.epsr.2022.108120.

Gnanavendan, S., et al., 2024. Challenges, solutions and future trends in EV-Technology: a review. IEEE Access 12, 17242–17260. https://doi.org/10.1109/ACCESS.2024.3353378.

Görges, D., 2017. Relations between model predictive control and reinforcement learning. IFAC-Pap. 50 (1), 4920–4928. https://doi.org/10.1016/j.ifacol.2017.08.747.

Han, Q., Li, X., He, L., 2025. A hierarchical deep reinforcement learning method for coupled transportation and power distribution system dispatching. Eng. Appl. Artif. Intell. 145, 110264. https://doi.org/10.1016/j.engappai.2025.110264.

Hestenes, M.R., Stiefel, E., 1952. Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand. 49 (6), 409. https://doi.org/10.6028/jres.049.044.

Hussain, A., Bui, V.-H., Kim, H.-M., 2022. Deep reinforcement learning-based operation of fast charging stations coupled with energy storage system. Electr. Power Syst. Res. 210, 108087. https://doi.org/10.1016/j.epsr.2022.108087.

IEC, 2018. Power trans. – Part 7: Load. guide for min.-oil-immers. power trans. IEC 60076-7:2018 https://webstore.iec.ch/en/publication/34351. (Accessed 9 December 2024). 1 89.

Ji, J., et al., 2024. OmniSafe: an infrastructure for accelerating safe reinforcement learning research. J. Mach. Learn. Res. 25 (285), 1–6. https://doi.org/10.5555/3722577.3722862.

Li, H., Wan, Z., He, H., 2020. Constrained EV charging scheduling based on safe deep reinforcement learning. IEEE Trans. Smart Grid 11 (3), 2427–2439. https://doi.org/10.1109/TSG.2019.2955437.

Li, Y., Zhang, Z., Xing, Q., 2025. Real-time online charging control of electric vehicle charging station based on a multi-agent deep reinforcement learning. Energy 319, 135095. https://doi.org/10.1016/j.energy.2025.135095.

Liu, D., Zeng, P., Cui, S., Song, C., 2023. Deep reinforcement learning for charging scheduling of electric vehicles considering distribution network voltage stability. Sensors 23 (3). https://doi.org/10.3390/s23031618, 3.

Martins, H., Henriques, C.O., Figueira, J.R., Silva, C.S., Costa, A.S., 2023. Assessing policy interventions to stimulate the transition of electric vehicle technology in the european union. Socioecon. Plann. Sci. 87, 101505. https://doi.org/10.1016/j.seps.2022.101505.

Meinecke, S., et al., 2020. SimBench—A benchmark dataset of electric power systems to compare innovative solutions based on power flow analysis. Energies 13 (12). https://doi.org/10.3390/en13123290. Art. no. 12.

Paternain, S., Calvo-Fullana, M., Chamon, L.F.O., Ribeiro, A., 2023. Safe policies for reinforcement learning via primal-dual methods. IEEE Trans. Autom. Control 68 (3), 1321–1336. https://doi.org/10.1109/TAC.2022.3152724.

Qi, T., Ye, C., Zhao, Y., Li, L., Ding, Y., 2023. Deep reinforcement learning based charging scheduling for household electric vehicles in active distribution network. J. Mod. Power Syst. Clean Energy 11 (6), 1890–1901. https://doi.org/10.35833/MPCE.2022.000456.

Qiu, D., Wang, Y., Hua, W., Strbac, G., 2023. Reinforcement learning for electric vehicle applications in power systems:a critical review. Renew. Sustain. Energy Rev. 173, 113052. https://doi.org/10.1016/j.rser.2022.113052.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N., 2021. Stable-Baselines3: reliable reinforcement learning implementations. J. Mach. Learn. Res. 22 (268), 1–8.

Rezaee, S., Farjah, E., Khorramdel, B., 2013. Probabilistic analysis of Plug-In electric vehicles impact on electrical grid through homes and parking lots. IEEE Trans. Sustain. Energy 4 (4), 1024–1033. https://doi.org/10.1109/TSTE.2013.2264498.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal Policy Optimization Algorithms. https://doi.org/10.48550/arXiv.1707.06347. *arXiv*: arXiv:1707.06347.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P., 2015. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning. PMLR, pp. 1889–1897. https://doi.org/10.48550/arXiv.1502.05477.

Terpilowski, M.A., 2019. scikit-posthocs: pairwise multiple comparison tests in python. J. Open Source Softw. 4 (36), 1169. https://doi.org/10.21105/joss.01169.

Towers, M., et al., 2024. Gymnasium: a Standard Interface for Reinforcement Learning Environments. https://doi.org/10.48550/arXiv.2407.17032. *arXiv*: arXiv:2407.17032.

Tuchnitz, F., Ebell, N., Schlund, J., Pruckner, M., 2021. Development and evaluation of a smart charging strategy for an electric vehicle fleet based on reinforcement learning. Appl. Energy 285, 116382. https://doi.org/10.1016/j.apenergy.2020.116382.

Wong, E., 2021. "Electric Vehicle Charging Station Usage. July 2011 - Dec 2020).". https://data.cityofpaloalto.org/dataviews/257812/electric-vehicle-charging-station-usage-july-2011-dec-2020/.

Wu, T., Scaglione, A., Surani, A.P., Arnold, D., Peisert, S., 2023. Network-constrained reinforcement learning for optimal EV charging control. In: 2023 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (Smartgridcomm), pp. 1–6. https://doi.org/10.1109/SmartGridComm57358.2023.10333926.

Yang, T.-Y., Rosca, J., Narasimhan, K., Ramadge, P.J., 2020. Projection-based constrained policy optimization. Presented at the Eighth International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.03152.

Yang, H., Xu, Y., Guo, Q., 2024. Dynamic incentive pricing on charging stations for real-time congestion management in distribution network: an adaptive model-based safe deep reinforcement learning method. IEEE Trans. Sustain. Energy 15 (2), 1100–1113. https://doi.org/10.1109/TSTE.2023.3327986.

Zhang, Y., Rao, X., Liu, C., Zhang, X., Zhou, Y., 2023a. A cooperative EV charging scheduling strategy based on double deep Q-network and prioritized experience replay. Eng. Appl. Artif. Intell. 118, 105642. https://doi.org/10.1016/j.engappai.2022.105642.

Zhang, S., Jia, R., Pan, H., Cao, Y., 2023b. A safe reinforcement learning-based charging strategy for electric vehicles in residential microgrid. Appl. Energy 348, 121490. https://doi.org/10.1016/j.apenergy.2023.121490.

Zhang, A., Liu, Q., Liu, J., Cheng, L., 2024a. CASA: cost-effective EV charging scheduling based on deep reinforcement learning. Neural Comput. Appl. 36 (15), 8355–8370. https://doi.org/10.1007/s00521-024-09530-3.

Zhang, J., Guan, Y., Che, L., Shahidehpour, M., 2024b. EV charging command fast allocation approach based on deep reinforcement learning with safety modules. IEEE Trans. Smart Grid 15 (1), 757–769. https://doi.org/10.1109/TSG.2023.3281782.

Zhang, Z., Rigo-Mariani, R., Hadjsaid, N., 2025. Comparative of control strategies on electrical vehicle fleet charging management strategies under uncertainties. Energy AI 21, 100522. https://doi.org/10.1016/j.egyai.2025.100522.

Zheng, X., Ju, C., Yang, G., Chu, J., 2025. Multi-agent modeling for energy storage charging station scheduling strategies in the electricity market: a cooperative

learning approach. J. Energy Storage 106, 114226. https://doi.org/10.1016/j.est.2024.114226.
Electric Vehicle Integration into Power Grids, 2021. ENTSO-E. https://www.entsoe.eu/2021/04/02/electric-vehicle-integration-into-power-grids/. (Accessed 21 February 2025).
Trends in Electric Cars – Global EV Outlook 2024 – Analysis," IEA. Accessed: September. 23, 2024. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2024/trends-in-electric-cars.