Advances in Applied Energy 12 (2023) 100151

Contents lists available at ScienceDirect

Advances in Applied Energy

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

Identifying generalizable equilibrium pricing strategies for charging service providers in coupled power and transportation networks

Yujian Ye [a,*], Hongru Wang [b], Tianxiang Cui [c], Xiaoying Yang [c], Shaofu Yang [d,e], Min-Ling Zhang [d,e]

[a] School of Electrical Engineering, Southeast University, Sipailou 2, Nanjing, 210096, China
[b] School of Cyber Science and Engineering, Southeast University, Sipailou 2, Nanjing, 210096, China
[c] School of Computer Science, University of Nottingham Ningbo China, 199 Taikang E Rd, Ningbo, 315104, China
[d] School of Computer Science and Engineering, Southeast University, Sipailou 2, Nanjing, 210096, China
[e] Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China

ARTICLE INFO

Dataset link: https://dx.doi.org/10.17632/25vjrnz3gp.1

Keywords:
Coupled power-transportation network
Electric vehicles
Multi-agent reinforcement learning
Nash equilibrium
Pricing game analysis

ABSTRACT

Transportation electrification, involving large-scale integration of electric vehicles (EV) and fast charging stations (FCS), plays a critical role for global energy transition and decarbonization. In this context, coordination of EV routing and charging activities through suitably designed price signals constitutes an imperative step in secure and economic operation of the coupled power-transportation networks (CPTN). This work examines the non-cooperative pricing competition between self-interested EV charging service providers (CSP), taken into account the complex interactions between CSPs' pricing strategies, EV users' decisions and the operation of CPTN. The modeling of CPTN environment captures the prominent type of uncertainties stemming from the gasoline vehicle and EV origin-destination travel demands and their cost elasticity, EV initial state-of-charge and renewable energy sources (RES). An enhanced multi-agent proximal policy optimization method is developed to solve the pricing game, which incorporates an attention mechanism to selectively incorporate agents' representative information to mitigate the environmental non-stationarity without raising dimensionality challenge, while safeguarding the commercial confidentiality of CSP agents. To foster more efficient learning coordination in the highly uncertain CPTN environment, a sequential update scheme is also developed to achieve monotonic policy improvement for CSP agents. Case studies on an illustrative and a large-scale test system reveal that the proposed method facilitates sufficient competition among CSP agents and corroborates the core benefits in terms of reduced charging costs for EV users, enhancement of RES absorption and cost efficiency of the power distribution network. Results also validate the excellent generalization capability of the proposed method in coping with CPTN uncertainties. Finally, the rationale of the proposed attention mechanism is validated and the superior computational performance is highlighted against the state-of-the-art methods.

1. Introduction

1.1. Background and motivation

As a promising approach to achieve efficient energy transition and decarbonization, governments worldwide have taken significant initiatives towards decarbonization of both generation and demand sides of the energy systems [1]. Alongside these initiatives, however, a plethora of techno-economic challenges on the operation and development of electricity systems emerge. At the generation side, large-scale integration of renewable energy sources (RES) is witnessed. Approximately 3.1 TW of RES capacity has been installed around the world by of the end of 2021, and this capacity is projected to further increase by 2.4 TW between 2022 and 2027 [2]. RES is intrinsically characterized by high intermittency and limited controllability, challenging the cost-efficient balancing of the electricity system. Furthermore, the unmatched electrical demand and RES generation gives rise to high RES curtailment, significantly limiting their carbon reduction potential [3]. At the demand side, the decarbonization agenda and advancement of vehicular

* Corresponding author.
E-mail addresses: yeyujian@seu.edu.cn (Y. Ye), whr_ee@seu.edu.cn (H. Wang), tianxiang.cui@nottingham.edu.cn (T. Cui), scyxy3@nottingham.edu.cn (X. Yang), sfyang@seu.edu.cn (S. Yang), zhangml@seu.edu.cn (M.-L. Zhang).

https://doi.org/10.1016/j.adapen.2023.100151
Received 16 May 2023; Received in revised form 7 September 2023; Accepted 10 September 2023
Available online 18 September 2023

Nomenclature

A. Abbreviations

TN	Transportation Network
GV	Gasoline Vehicle
EV	Electric Vehicle
SoC	State-of-Charge
O-D	Origin-Destination
FCS	Fast Charging Station
CSP	Charging Service Provider
TAP	Traffic Assignment Problem
UE	User Equilibrium
PDN	Power Distribution Network
DNO	Distribution Network Operator
DG	Distributed Generator
RES	Renewable Energy Source
ACOPF	Alternating Current Optimal Power Flow
SOCP	Second Order Conic Program
DLMP	Distributional Locational Marginal Price
CPTN	Coupled Power-Transportation Network
POMG	Partially Observable Markov Game
MARL	Multi-Agent Reinforcement Learning
CTDE	Centralized Training and Decentralized Execution

B. Indices and sets

$i, j \in J$	Index and set of TN nodes.
$w \in W$	Index and set of O-D pairs for EVs.
$w^g \in W^g$	Index and set of O-D pairs for GVs.
$p \in \Xi_w$	Index and set of paths of EV O-D pair w.
$p^g \in \Xi_w^g$	Index and set of paths of GV O-D pair w.
$a \in A$	Index and set of links in TN.
$A_p \subset A$	Subset of links in path p.
$k \in K$	Index and set of FCSs.
$K_z \subset K$	Subset of FCSs operated by CSP z.
z, Z	Index and number of CSPs/agents.
$n, m \in M$	Index and set of PDN nodes.
$M_m \subset M$	Subset of PDN nodes connected to node m.
$M^{im} \subset M$	Subset of PDN nodes connected to the transmission grid.
$M^{wg} \subset M$	Subset of PDN nodes with wind generators connected.
$M^{dg} \subset M$	Subset of PDN nodes with DGs connected.
$M^{ev} \subset M$	Subset of PDN nodes with FCSs connected.

C. Parameters

ω	Monetary value of unit time of EV and GV users (\$/h).
c^{tm}	Cost associated with charging time at FCSs (\$/MW).
t_a^0	Free travel time of EV and GV flows on link a (h).
t_k^0	Average charging time of EV flows at FCS k (h).
Γ	Parameter of waiting time function.
$\overline{x}_a$, $\overline{x}_k$	Capacities of link a and FCS k (veh).
$\overline{d}_w$	Upper limit of EV O-D travel demand d_w (veh).
d_w^g	GV O-D travel demand (veh).
ε_w^s, ε_w^i	Slope and interception of the EV O-D demand cost elasticity function (\$/veh and \$).
Δ^{TN}	Node-link incidence matrix of TN.
I_w	Vector with all zero entries except for the entries corresponding to the origin (equals to 1) and destination (equals to -1).
L_a	Distance of link a (km).
ΔE	Travel distance dependent, average reduction rate of SoC of EV flows (MWh).
$\overline{E}$	Maximum SoC of EV flows (MWh).
$\underline{E}_w$	Minimum SoC of EV flows in O-D pair w driven by EV users' differentiated range anxiety (MWh).
E_w^0	Initial SoC of EV flows of O-D pair w (MWh).
$\overline{F}_i$	Maximum charging electricity of EV flows at TN node i (MW).
c_n^l, c_n^q	linear and quadratic cost coefficients of DG at PDN node n (\$/MW^2 and \$/MW).
λ_n^{im}	Price of purchasing electricity from the transmission grid at PDN node n (\$/MWh).
$\overline{P}_n^g, \underline{P}_n^g$	Maximum and minimum active power limits of DG at PDN node n (MW).
$\underline{Q}_n^g, \overline{Q}_n^g$	Maximum and minimum reactive power limits of DG at PDN node n (MVaR).
P_n^{wg}	Maximum available wind power generation at PDN node n (MW).
P_n^d, Q_n^d	Active and reactive power demand at PDN node n (MW, MVaR).
$r_{n,m}$, $x_{n,m}$	Resistance and reactance of line (n,m) (Ω).
$z_{n,m}$	Impedance of line (n,m) (Ω).
$\overline{S}_{n,m}$	Thermal limit of line (n,m) (MVA).
$\overline{U}_n, \underline{U}_n$	Maximum and minimum limits of voltage magnitude square at PDN node n (kV^2).

D. Variables

π_k	Charging price at FCS $k \in K_z$ set by the CSP z (\$/MWh).
$\delta_{p,w,a}$	Binary variable indicating whether link a is in the path p of EV O-D pair w ($\delta_{p,w,a}=1$) or not ($\delta_{q,w,a}=0$).
$\delta_{p,w,a}^g$	Binary variable indicating whether link a is in the path p of GV O-D pair w ($\delta_{p,w,a}^g=1$) or not ($\delta_{q,w,a}^g=0$).
$\sigma_{p,w,k}$	Binary variable indicating whether FCS k is in path p of EV O-D pair w ($\sigma_{p,w,k}=1$) or not ($\sigma_{p,w,k}=0$).
$f_{p,w}$	EV flow on path p of O-D pair w (veh).
x_a, x_a^g	EV and GV flows on link a (veh).
x_k	EV flows at FCS k (veh).
d_w	Elastic EV travel demand of O-D pair w (veh).
C_w^{tot}	Total travel costs of O-D pair w (veh).
t_a, t_a^g	EV and GV traveling time on link a (h).
t_k	EV charging time at FCS k (h).
$F_{p,w,k}^e$	Charging electricity of EV flows on path p of O-D pair w at FCS k (MW).
P_k^{evd}	Aggregate EV charging demand at FCS k (MW).
$y_{p,w,a}$	Binary variable indicating whether path p of O-D pair w traverses link a ($y_{p,w,a}=1$) or not ($y_{p,w,a}=0$).
$E_{w,i}$	Remaining SoC of EV flows in O-D pair w at TN node i (MWh).
$C_{p,w}^{ch}$	Charging electricity cost of EV flows in path p of O-D pair w (\$).
$C_{p,w}^{tm}$	Charging time cost of EV flows in path p of O-D pair w (\$).
P_n^g	Active power generation of DG at PDN node n (MW).
Q_n^g	Reactive power generation of DG at PDN node n (MVaR).
P_n^{wg}	Active wind power generation at PDN node n (MW).
P_n^{im}	Electricity import from the transmission grid at PDN node n (MW).
$P_{n,m}$	Active power flow on line (n,m) (MW).
$Q_{n,m}$	Reactive power flow on line (n,m) (MVaR).
$I_{n,m}$	Squared magnitude of current on line (n,m) (kA^2).
U_n	Squared magnitude of voltage at node n (kV^2).

Table 1
Review and Comparison of Relevant Literature.

Ref	TN model	Vehicle flows	PDN model	FCS management	O-D travel demand	SoC of EV flows?	Sources of uncertainty in TN	Sources of uncertainty in PDN
[11]	-	EV	SOCP	Monopoly	Inelastic	No	-	-
[12]	-	EV	-	Monopoly	Inelastic	Yes	Initial SoC	-
[13]	TAP-UE	EV	-	Competition	Elastic	Yes	-	-
[14]	TAP-UE	EV&GV	SOCP	Monopoly	Inelastic	No	-	-
[15]	TAP-UE	EV	SOCP	Monopoly	Inelastic	No	Charging prices	-
[16]	TAP-UE	EV	SOCP	Monopoly	Inelastic	Yes	-	-
[17]	TAP-UE	EV	SOCP	Monopoly	Inelastic	Yes	EV O-D demand	RES
[18]	TAP-UE	EV	-	Monopoly	Inelastic	Yes	-	-
[19]	TAP-UE	EV&GV	SOCP	Monopoly	Inelastic	Yes	-	-
[20]	TAP-UE	EV	DCOPF	Monopoly	Inelastic	No	-	-
[21]	TAP-UE	EV	lin-dist-flow	Monopoly	Inelastic	No	-	-
[22]	TAP-UE	EV&GV	SOCP	Monopoly	Inelastic	No	EV O-D demand	RES
[23]	TAP-UE	EV	ACOPF	Competition	Elastic	No	-	-
[24]	TAP-UE	EV	-	Competition	Inelastic	No	Arrival time	-
[25]	TAP-UE	EV	-	Competition	Inelastic	No	-	-
[26]	TAP-UE	EV&GV	-	Competition	Inelastic	No	-	-
[27]	TAP-UE	EV	lin-dist-flow	Competition	Inelastic	No	-	RES
[28]	TAP-UE	EV	-	Competition	Inelastic	Yes	Initial SoC	-
[29]	TAP-UE	EV&GV	SOCP	Monopoly	inelastic	No	-	-
[30]	-	EV	-	Monopoly	Inelastic	No	Arrival time	-
[31]	-	EV	-	Monopoly	Elastic	Yes	Arrival time	RES
[32]	-	EV	SOCP	Monopoly	Inelastic	Yes	Arrival time	-
[33]	-	EV	ACOPF	Monopoly	Elastic	No	Arrival time	RES
This paper	TAP-UE	EV&GV	SOCP	Competition	Elastic	Yes	EV O-D demand cost elasticity, EV initial SoC, GV O-D demand	RES

technologies have paved the way for the electrification of the transport sector through large-scale integration of electric vehicles (EV) and fast charging stations (FCS), alongside the conventional gasoline vehicles (GVs), coupling more closely the operation of the transportation network (TN) and power distribution network (PDN) [4,5]. Latest reports reveal that the global sales of EVs were doubled in 2021 from 2020, reaching a new record of 6.6 million. In China, an adaptation of approximately 2.6 million EVs and 3.3 million FCSs is witnessed in 2020 [6], and a projection of 50% vehicle sales will be electric by 2030 [7].

However, the limited ranging capability and charging requirement distinguish the behavior of EV users from the users of GV based on internal combustion engines. The former can substantially alter the traffic flow pattern and may aggravate the level of congestion on the roads and at the FCSs of the TN [8]. Furthermore, the EV charging power (which can be up to 400 kW for FCSs with DC chargers [9]) is significantly higher than that of conventional residential loads. Thus, uncoordinated charging behavior of EVs (and its significant stochasticity) could disproportionately increase the peaks of overall demand, creating voltage deviation and network congestion in the PDN [10]. Consequently, suitably coordination of the EV charging routes and demand constitutes an imperative step in secure and economic operation of the *coupled power-transportation networks* (CPTN).

1.2. Review of previous work

A large number of papers on coordination of EV routing and charging in the CPTN and pricing have been published in the last several years, a detailed comparison of them in terms of several key aspects is conducted in Table 1.

In terms of EV routing and charging coordination, authors in [11] investigate the EV users' driving paths and charging locations selection problem, targeted to minimizing the driving and charging time costs and the overall power supply cost. A combined operation scheme for battery swapping station and autonomous mobility-on-demand (AMoD) system is developed in [12], where the swapping and scheduling decisions of the EV fleet are optimized from the social perspective of the AMoD system. The above works assume the EVs are controllable by a central coordinator, and neglect the EV users' self-interested routing and charging behaviors and the resultant traffic condition. Another stream of works captures the individual rationality of individual GV users, and describes the steady-state distribution of traffic flows, known as the *traffic assignment problem* (TAP), the solution of which is referred to as an *User Equilibrium* (UE) [34]. The TAP-UE problem is later calibrated to meet modeling requirement of EV users, by including relevant operating constraints related to the charging behavior of the EV users and/or their State-of-Charge (SoC) levels, some of them also take into account the uncertainties, such as the charging prices, EV Origin-Destination (O-D) travel demand (Table 1). However, the solution of the TAP-UE problem faces multitudes of challenges, mainly reflected in i) the necessity of exploring all feasible paths for all EV O-D demands, ii) severe non-convexities and non-linearities associated with the operating models of traffic links and nodes and iii) multi-source uncertainties associated with the GV/EV users' diversified traveling requirements and preferences.

Assuming rational behavior of EV users, an efficient charging pricing mechanism could alter the EV charging routing decisions and the distribution of their charging loads, and subsequently affects the operation of the PDN. In this context, previous works focusing on pricing either focus on a single FCS [30,31] or make the assumption that the FCSs are part of municipal infrastructure and their charging prices are set by a monopolistic authority (as labelled as "Monopoly" in Table 1), including the distribution network operator (DNO) and a charging service provider (CSP).

In the former, the charging prices are established by the DNO, which is set as a fixed price or set as the distribution locational marginal prices (DLMP) by solving the dual problem of the ACOPF problem (or the convexified version as an SOCP), targeted to minimizing the overall operating costs of the PDN (Table 1). Evidently, fixed prices do not reflect the locational operating status of the PDN, which is affected by the spatial distribution of the EV's charging demand. On the other hand, although DLMPs convey reflective information regarding the operating status of the PDN [35], they do not capture the operating status of the TN [29]. In the latter, the FCSs are privately owned and operated by multiple CSPs acting as profit-driven entities (as labelled as "Competition" in Table 1). Apart from the government subsidy, the charging price is a dominant factor influencing the profit of the CSPs. Consequently, self-interested CSPs will compete to attract EV users through

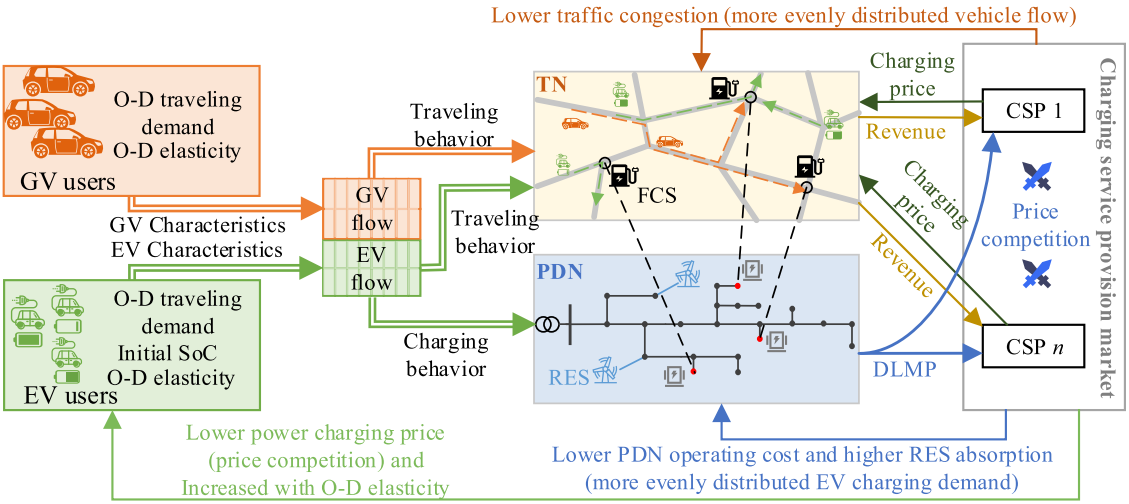

Fig. 1. Illustration of the overall formation and settings of the examined problem and its various value streams in supporting energy transition and transportation electrification.

their strategic price signals, to maximize their own economic profits. Suitable CSP competition is envisaged to drive down the charging price for EV users. These benefits for FCS operators and EV users will further escalate the momentum for transportation electrification [36]. The realization of these benefits, however, calls for an adequate pricing mechanism which holistically capture the complex interactions of competing CSPs' pricing strategies, EV users charging behavior as well as the resultant operating status of the CPTN.

In this context, the interaction between any single strategic CSP and the operation of CPTN can be modeled as a bi-level optimization problem (with CSPs as decision-makers in the upper level and CPTN operation modeled in the lower level problems); while the non-cooperative pricing game involving multiple self-interested CSPs can be formulated as a multi-leader-common-follower (with CSPs as leaders and CPTN as follower) game with the objective of identifying the non-cooperative Nash Equilibrium (NE) resultant from the price competition. The solution to the above optimization problems, however, face unprecedented challenges:

First, the derivation of the analytical solution to the bi-level problem by converting it to an Mathematical Program with Complementarity Constraints (MPCC) proves to be intractable, since the lower-level optimization problems (i.e. TAP-UE and ACOPF) are highly non-convex and non-linear, which prevents the conversion to the Karush–Kuhn–Tucker (KKT) optimality conditions. This, in turn, prohibits the analytical computation of the NE of the pricing game through the solution of the Equilibrium Program with Complementarity Constraints (EPCC) [37]. The assumptions of complete information (i.e. each player in the pricing game has full knowledge of the market dynamics and its competitors' strategies) made by both MPCC and EPCC models are also impractical [38,39]. Secondly, the competition among multiple CSPs leads to a *non-stationary* market environment in which each CSP's pricing strategy has a profound influence on others. However, each CSP's strategy can only be formed with *incomplete information* on the market dynamics and its rivals' strategies, driven by business confidentiality restrictions. This renders the estimation of the NE (which necessitates global/complete information) challenging. Lastly, both the bi-level problem and the NE estimation problem face multi-source uncertainties (Table 1) and therefore strive for decision robustness.

Driven by the aforementioned three-fold challenges, limited efforts on CSP pricing make a large number of impractical assumptions (Table 1): i) the TN is not modeled, i.e. the routing and charging behavior of EV and/or GV users in the TN is not realistically represented, ii) the PDN is not modeled, i.e. the impact of EV charging demand on the operation of PDN is neglected and iii) the sources of uncertainty in the TN and/or PDN are partially or completely neglected, i.e. the capability of the pricing strategies in coping with CPTN uncertainties may be limited.

Alternatively, the CSP pricing game can be formalized as a coordinated decision-making process under uncertainties involving multiple strategic agents. In this case, each agent needs to take into account and interact with not only the CPTN environment but also other learning agents, which can be modeled as a Markov Game (MG) and approached with Multi-Agent Reinforcement Learning (MARL) methods. MARL method constitutes a promising alternative to study the competition among multiple agents in a non-stationary environment. MARL also constitutes a computationally efficient method to estimate and analyze the NE of game with incomplete information or when the closed form solution is unattainable [40]. However, successful application of MARL in the examined pricing game necessitates reconciling the conflicts between learning under incomplete information, resolving environmental non-stationarity (which requires each agent be informed of the global information) and protecting the commercial confidentiality of the CSPs.

To this end, the *Centralized Training and Decentralized Execution* (CTDE) MARL paradigm provides an effective remedy for tackling the non-stationarity. However, despite several past CTDE methods such as Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [41] and Multi-Agent Proximal Policy Optimization (MAPPO) [42] achieve good performance in multi-agent environments, they generally cannot guarantee consistent policy improvement even with the correct gradients, making the convergence unstable [43]. In addition, both MADDPG and MAPPO necessitate taking the concatenation of the observations and actions of all agents as input for centralized training of the critic networks, which may lead to dimensionality challenge. Furthermore, direct information exchange involving local/private observations and actions breaks the business confidentiality of the CSPs, who in practice, will not disclose/exchange such information with their rivals [13,28] in a price competition.

1.3. Contributions

This paper attempts to fill the knowledge gap in modeling the non-cooperative pricing game among CSPs and develops a tailored MARL method for its effective solution. Fig. 1 illustrates the overall formation and settings of the pricing of charging services and EV/GV coordination in CPTN problems, the various value streams achieved by the developed methodological framework for EV users, PDN and TN is also highlighted, demonstrating its significance in supporting energy transition and transportation electrification.

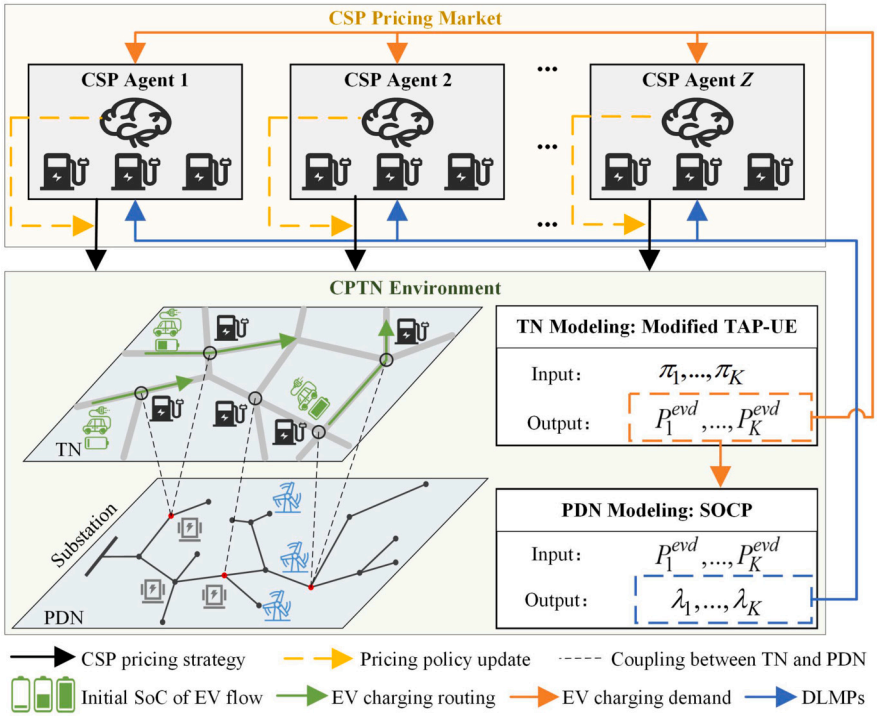

Fig. 2. Architecture of the proposed CSP pricing game.

The following novel contributions can be identified:

- A modified TAP-UE model is developed to identify the UE of the TN, comprising the vehicle flows of GV and EV users and the charging load of the EV users. The proposed model takes into account the most prominent types of uncertainties associated with EV O-D demand cost elasticity and initial SoC levels as well as the GV O-D demand (Table 1). An efficient solution procedure based on iterative solution of TAP-UE and an active path generation (APG) algorithm is developed for UE identification.

- The CPTN environment is constructed which involves sequential solution of the TAP-UE and SOCP problems for a given set of CSP pricing strategies. The non-cooperative pricing game among CSPs is then formulated as a partially observable MG (POMG) with incomplete information and encompasses the uncertainties stemming from both the operation of TN and PDN.

- An improved MAPPO method is proposed to determine an optimal pricing strategy for each FCS and adaptively learn the NE to the pricing game. The performance of MAPPO is leveraged with the attention mechanism (namely, Att-MAPPO) to selectively incorporate agents' representative information for estimating the Q-values, which eliminates the environmental non-stationarity without exploding the input dimension of the critic networks, while safeguarding the confidentiality of the FCSs. To foster more efficient learning coordination in the highly uncertain CPTN environment, Att-MAPPO also uses sequential update scheme to achieve monotonic policy improvements for the agents.

- Case studies on an small illustrative example and a large-scale CPTN validate the effectiveness of the proposed Att-MAPPO method in estimating the NE of the CSP pricing game while adaptively cope with the multi-source uncertainties. Results also validate the capability of Att-MAPPO in facilitating adequate and sufficient competition among CSPs, promising a range of benefits for CPTN including significant cost savings for EV users, enhanced RES absorption and overall cost efficiency for PDN.

2. Problem formulation

2.1. Overall problem setup and assumptions

The investigated CSP pricing game falls within the category of multi-leader-common-follower games and exhibits a bi-level architecture, as depicted in Fig. 2. The CSPs are regarded as active market participants competition in a *CSP pricing market*, and the FCSs they operate (which may locate differently in the TN) are not regarded as municipal infrastructures. Under the assumption of elastic EV travel demand, CSPs compete among themselves through suitable FCS-specific charging prices (the pricing strategy will be obtained by training the corresponding CSP agent) to attract EV users and influence their routing and charging decisions (constrained by the TN), targeted to acquire higher electricity sales revenue. This distinguishes CSPs from conventional electrical demand aggregators, whose revenue is not affected by the dynamics of the TN. On the other hand, the competition of CSPs also lies in reducing their electricity purchase costs from the PDN, e.g. the CSPs may adjust their prices to guide EV users to charge at FCSs connected to/located closely to the RES generators.

At the upper level, the CSP agents compete with each other to establish its charging prices. At the lower level, the EVs users adjust their routing and charging options in the TN in response to the price signals, targeted to minimize their total travel costs, which eventually forms the charging load as the solution to the modified TAP-UE problem. Based on the load distribution, the DNO schedules the generators (including RES) in the PDN by solving the SOCP problem, and subsequently calculates the DLMPs at different FCSs. It can be observed that the CSP pricing, EV routing/charging and DNO power dispatching decisions are coupled by the charging prices, charging loads and DLMPs, which then makes up the overall CSP profit. The following subsections detail the settings and composition of the CPTN environment, including the modified TAP-UE model of the TN, its solution procedure, and the SOCP model of the PDN.

2.2. TN modeling: modified TAP-UE model

The TN can be represented as a connected directed graph $G = [N, A]$, where N and A denote the sets of nodes and links/roads. Each EV/GV user travels in G, departing from origin O and arrives at destination D. A path is composed of multiple end-to-end connected links, while a link can be shared by multiple different paths. Each O-D pair $w \in W$ is connected by a set of paths Ξ_w, where the EV flows will be distributed to, and same for the GV flows. Additionally, a path p is called *feasible* if the EV flow is capable of completing the journey with or without recharging. Moreover, a feasible path is called *active*/inactive if the EV flow on it is strictly greater than/equal to zero. All paths are assumed to be feasible for GV users, i.e. they are able to complete their journeys without refueling.

In practice, users' travel demands are inherently elastic, suggesting that EV users will reduce their demand to some extent in the event of higher travel costs, by using for example, public transportation or adjust their departure times. Furthermore, the EV O-D demand and the initial SoC levels of EV are considered to be stochastic rather than deterministic (Table 1), capturing the inherent uncertainties associated with different users' charging requirement and preferences. Finally, individual rationality of EV users is assumed, i.e., the users will respond to charging prices and make optimal decisions individually. The UE is reached, if for any users, the total travel costs on all active paths for a given O-D demand are equal, and are less or equal to that the user would incur on any inactive paths. The modified TAP-UE problem is provided in Appendix A.1.

2.3. Iterative solution procedure for UE identification

It can be observed that the solution of TAP-UE requires exploring of all paths $p \in P_w$ for all O-D pairs to determine the optimal routing and charging decisions of all EV users. However, it can be anticipated that most of the paths may be infeasible or feasible but not active, which are both redundant to explore, from the solution point of view.

To this end, the UE can be identified through a two-level iterative process outlined in Algorithm 1. At iteration l and at the EV cluster level (correspond to each EV O-D pair), provided the traffic flows obtained in the previous iteration, an active path generation (APG) algorithm is developed and thus executed to identify possible active paths that can contribute to the reduction of total travel costs between any O-D pairs, through the solution of the optimization formulated in Appendix A.2, for each O-D pair w.

Algorithm 1 Iterative Solution Procedure for UE

Input: Charging price π_k at FCS k
Initialize: Initialize traffic flows as x_a, x_a^g and $x_k = 0$
1: **for** l = 1, 2,...,L **do**
2: **for** $w = 1,...,W$ **do**
3: Given $x_k^{[l-1]}$ and $x_k^{[l-1]}$, execute APG by solving problem (A.16)-(A.24) for $\delta_{p,w,a}^{[l]}$, $\sigma_{p,w,k}^{[l]}$, and $F_{p,w,k}^{e,[l]}$, and then evaluate $C_{p,w}^{ch,[l]}$ and $C_{p,w}^{tm,[l]}$
4: If $\exists p \in \Xi_w : C_{w,p}^{tot}$ is reduced, add new paths p to the active paths $\Xi_w^{[l]}$ for O-D pair w; otherwise terminate the procedure.
5: **end for**
6: Given $\Xi^{[l]}$, $C_{p,w}^{ch,[l]}$ and $C_{p,w}^{tm,[l]}$, solve the TAP-UE problem (A.1)-(A.15) for $x_a^{[l]}$, $x_k^{[l]}$ and $f_{p,w}^{[l]}$.
7: **end for**

2.4. PDN modeling: SOCP model

The aggregate charging demand at each FCS is served by PDN, which is usually a radial network, and can be represented by a tree topology. The power flow distribution over the PDN is generally obtained by solving the ACOPF problem, which is non-convex and NP-hard. Pursuing tractability, convex relaxation is performed, the resulting formulation of the following SOCP problem [44] is provided in Appendix A.3.

3. Multi-agent coordination for profit-driven CSPs interacting in CPTN

3.1. Non-cooperative pricing game among CSPs

Having detailed the coupled optimization models pertaining to the CPTN operation, the objective of a self-interested CPS z operating multiple FCSs (and as a market player in the upper level problem of the pricing game described in Section 2.1) lies in identifying suitable charging prices $\pi_k, \forall k \in K_z$ so as to maximize its profit Ω_z:

$$\max_{\boldsymbol{\pi}_z=[\pi_k,\forall k\in K_z]} \Omega_z(\boldsymbol{\pi}_z, \boldsymbol{\pi}_{z-}) = \sum_{k\in K_z} (\pi_k - \lambda_k^*) P_k^{evd,*} \tag{1}$$

where:

$$[P_k^{evd,*}, \forall k] = \arg\min U(\pi_k, \pi_{k-}) \tag{2}$$

where $\boldsymbol{\pi}_z$ and $\boldsymbol{\pi}_{z-}$ represent the prices at FCSs operated by CSP z and other CSPs, respectively; λ_k^* correspond to the DLMPs, i.e. optimal value of the dual variable associated with constraint (A.26) (Section 2.4) and $P_k^{ev,*}$ is the optimal aggregate EV charging demand at FCS k encoded in the UE solution. Two observations can be drawn: i) a CSP's payoff function depends on its own and other CPSs' pricing strategies and ii) the evaluation of the payoff function necessitates the complete knowledge of the TAP-UE and SOCP models (including the associated uncertainties). Regarding the former, barriers pertaining to commercial confidentiality often prevent a CSP from assessing its competitors' pricing strategies. Concerning the latter, $P_k^{ev,*}$ cannot be expressed in closed-form, since it involves successive solution of an iterative process and the SOCP problem.

Based on the above analysis, it can be concluded that conventional model-based analytical methods, such as MPCC and EPCC (Section 1.2), relying on the prerequisite of full and accurate knowledge of rivals' strategies and the complex market dynamics and the assumption of convex and linear market models, may not applicable to address the examined pricing game. Furthermore, even if conventional methods can return the optimal solution for a single scenario, the optimization process needs to be re-engineered for a new scenario, and therefore no incremental knowledge has been gained by solving multiple optimization problems for different scenarios. Namely, conventional methods do not exhibit the desirable generalization capability towards uncertainty adaptability.

To this end, the CPTN optimization model actually describes a black-box non-stationary CSP pricing market condition for each CSP. The nature of the game is non-cooperative and is based on incomplete information, which can be denoted as $\mathbb{G} = \{Z, \{\boldsymbol{\pi}_z\}_{z\in Z}, \{\boldsymbol{\Omega}_z\}_{z\in Z}\}$. Let us denote the set of strategy $\boldsymbol{\Psi} = \{\boldsymbol{\pi}_z^*\}_{z\in Z}$ as an NE strategy, under which no CSPs can further increase its payoff by unilaterally changing its pricing strategy, as described by in the following set of constraints:

$$\Omega_z(\boldsymbol{\pi}_z^*, \boldsymbol{\pi}_{z-}^*) \geq \Omega_z(\boldsymbol{\pi}_z, \boldsymbol{\pi}_{z-}^*), \forall \boldsymbol{\pi}_z, \forall z \tag{3}$$

As discussed in the literature, existence and uniqueness of NE are not generally guaranteed [45], the No-regret Index (NI) adopted in [46] is thus employed as a suitable metric to assess the proximity of the obtained set of pricing strategies to NE. In the examined problem, the NI is expressed as the ratio of the sum of the payoffs given by the current set of pricing strategies to the sum of the payoffs given by the set of best response strategies (provided that other players retain their current strategies).

$$NI = \frac{\sum_{z\in Z} \Omega_z(\boldsymbol{\pi}_z, \boldsymbol{\pi}_{z-})}{\sum_{z\in Z} \max_{\boldsymbol{\pi}_z} \Omega_z(\boldsymbol{\pi}_z, \boldsymbol{\pi}_{z-})} \tag{4}$$

According to (4), if NI is equal to 1, the pricing strategies satisfy the NE condition, signaling NE identification.

3.2. Markov game formulation

The examined pricing game (Fig. 2) resembles a decision-making process that involves multiple agents. In this multi-agent scenario, agents need to take into account and interact with not only the CPTN environment but also other learning agents. This type of multi-agent coordination problem is usually modeled through a Partially Observable Markov Game [47]. Formally, POMG is defined by Z agents (corresponding to the Z CSPs of the pricing game) with a set of states S describing the global state, a collection of private observations $\{\mathcal{O}_{1:Z}\}$, a collection of action sets $\{\mathcal{A}_{1:Z}\}$, a collection of reward functions $\{\mathcal{R}_{1:Z}\}$ and a state transition function $\mathcal{T}$. At time period t, each CSP agent z chooses an action a_z according to its policy μ_z based on its current observation o_z (the partial observability mainly stems from the fact that the private observation only contains partial information from the global state). The environment then moves into the next state according to the state transition function conditioned on the actions of all agents. Each agent z obtains a reward $r_{z,t}$ and a local observation for the next period $o_{z,t+1}$, and aims to maximize its cumulative discounted reward $R_z = \sum \gamma^t r_{z,t}$, where $\gamma \in [0,1)$ is the discount factor.

The local observation $o_{z,t}$ of an CSP agent is defined as $o_{z,t} = [x_k, t_k, \sum_{w,p} F^e_{p,w,k}, \sum_w E_{w,k}, \lambda_k, \forall k \in K_z]$, which encodes the EV flows, charging time, charging electricity of EV flows, remaining SoC of EV flows and electricity purchasing price from PDN at FCS k, respectively. Taking local observations of all agents at step t into account, we have $o_t = (o_{1,t}, o_{2,t}, ..., o_{Z,t})$, for simplicity, the global state s_t is set to be o_t. The agent's action $a_{z,t}$ corresponds to the charging price $[\pi_k, \forall k \in K_z]$ in range $[\underline{\pi}, \overline{\pi}]$. The reward of each agent is defined as the profit Ω_z in (1).

4. Proposed multi-agent deep reinforcement learning method

4.1. Overview of state-of-the-art MARL frameworks

Multi-Agent Reinforcement Learning (MARL) addresses sequential decision-making problems with more than one agent operating in a common environment. Each agent aims to optimize its own long-term reward by interacting with the environment and other agents [48]. One natural approach to tackle MARL problems is concurrent learning. The idea is to let each agent learn an individual policy independently [49], while considering other agents as a part of the environment. In this context, concurrent learning framework offers decentralized execution for agents which is aligned with the examined pricing game. However, as all agents are learning and adapting their policies, the frequent change in these policies renders the environment dynamics non-stationary, contributing to instability or even divergence. Furthermore, significant computational and memory burdens arise when each agent trains its own policy, and represent it with a complex model like a deep neural network. Finally, because the agents do not share experiences, this approach is often criticized by its low sampling and learning efficiency.

In contrast, centralized learning takes as input the observations and actions of all agents and learns actions jointly for all agents, both training and execution stages are centralized. In this context, centralized learning is not suitable for the examined non-cooperative CSP pricing game which involves training of multiple self-interested agents where their actions are executed in a decentralized fashion. In additional, this framework exhibits significant communication requirement during execution, as the computation of the central policy and the distribution of individual actions necessitates input of all observations from agents. Furthermore, the implementation of this approach violates the CSP agents' commercial confidentiality, since in practice, they will not expose their pricing strategies and exchange private/local information (such as the EV flows/charging demand at the FCSs they own) directly with their rivals. Finally, this approach offers limited scalability potential, since the dimension of the joint action space increases exponentially with the number of agents and each agent's action dimension.

To tackle environmental non-stationarity, an established MARL paradigm, referred to as Centralised Training with Decentralised Execution (CTDE), is proposed [41]. In CTDE, each agent has access to the publicly accessible, global information (usually represented as a concatenation of all agents' observations and actions) to construct their critic networks at the training stage, whereas this information is not incorporated at execution (i.e. the actions are executed in a decentralized manner). As a result, the CTDE paradigm enables several successful developments of single-agent algorithm extensions in energy and power systems related applications [13,39,50,51] (several representative ones are employed as baseline MARL method in the Case studies). However, these methods generally do not offer monotonic improvement guarantee, which may impede agents from improving policies in a stable manner [43]. Furthermore, since the training of critic networks is centralized, these methods generally share the same dimensionality and confidentiality preservation limitations with the centralized learning framework.

In light of the above, a tailored MARL method combining the strength of the multi-agent extension of PPO and the attention mechanism, namely Att-MAPPO, is developed in this paper to address the examined CSP pricing coordination problem. Att-MAPPO also belongs to the CTDE paradigm, but as opposite to MADDPG/MAPPO which indifferently incorporates the private information (observations and actions) of all agents to train their critics, Att-MAPPO enables learning of all agents' critics jointly by sharing a set of learnable parameters among agents. Furthermore, the employment of the attention mechanism [52] enables selectively paying attention to the relevant information (an abstract embedding of agents' private information) of other agents during training. Driven by the above, Att-MAPPO offers greater scalability, significantly lower communication and computational complexity, and is able to safeguard the commercial confidentiality of CSPs in the pricing market.

4.2. Proposed Att-MAPPO method

Since the Att-MAPPO method is founded on a multi-agent extension of the Proximal Policy Optimization (PPO) method, the latter is briefly introduced in this subsection. PPO constitutes a representative single-agent RL algorithm that is simple to implement and has been applied in various applications and shows stable performance [53]. PPO is a method to simplify the complex calculation of Trust Region Policy Optimization (TRPO) [54], which updates policies by taking the largest step possible and uses KL-Divergence to control the distance between the probability distributions of the new policy $\mu^{\phi^{[l+1]}}$ and old policy $\mu^{\phi^{[l]}}$ at the same time:

$$\phi^{[l+1]} = \arg\max_{\phi} \mathcal{L}\left(\phi^{[l]}, \phi\right) \tag{5}$$

subject to:

$$\bar{D}_{KL}\left(\phi \| \phi^{[l]}\right) \le \delta \tag{6}$$

where l denotes the iteration number and $\mathcal{L}\left(\phi^{[l]}, \phi\right)$ is the *surrogate advantage* that is used to assess the quality of the new policy μ^{ϕ} with respect to the old policy $\mu^{\phi^{[l]}}$:

$$\mathcal{L}\left(\phi^{[l]}, \phi\right) = \mathop{\mathbb{E}}_{a\sim\mu^{\phi^{[l]}}, s\sim\rho\left(\mu^{\phi^{[l]}}\right)} \left[\kappa \hat{A}_{\mu}(s,a)\right] \tag{7}$$

$\hat{A}_{\mu}(s,a)$ is an estimation of the state-action advantage function; $\kappa = \mu^{\phi}(a|s)/\mu^{\phi^{[l]}}(a|s)$ is the policy ratio; and $\bar{D}_{KL}\left(\phi \| \phi^{[l]}\right)$ is an average KL-divergence between the current and old policies among the states that the old policy visited:

$$\bar{D}_{KL}\left(\phi \| \phi^{[l]}\right) = \mathop{\mathbb{E}}_{s\sim\rho\left(\mu^{\phi^{[l]}}\right)} \left[D_{KL}\left(\mu^{\phi}(\cdot|s) \| \mu^{\phi^{[l]}}(\cdot|s)\right)\right] \tag{8}$$

TRPO searches for the new policy parameter $\phi^{[l+1]}$ within the trust region at each iteration l using heuristic approach. To lower the computational cost on $\bar{D}_{KL}\left(\phi\|\phi^{[l]}\right)$ when updating μ^{ϕ}, PPO utilizes a clipped objective function to penalize parameter updates to ensure that the new policy is in close proximity of the old one:

$$L^{\mathrm{PPO}}\left(\mu^{\phi}\right)=\mathop{\mathbb{E}}_{a\sim\mu^{\phi^{[l]}},s\sim\rho\left(\mu^{\phi^{[l]}}\right)}\min\left[\kappa\hat{A}_{\mu}(s,a),\mathrm{clip}\left(\kappa,1\pm\epsilon\right)\hat{A}_{\mu}(s,a)\right] \tag{9}$$

where the threshold interval ϵ is used to control how far the new policy can diverge from the previous one, the policy ratio κ beyond ϵ is clipped, which allows PPO to manage the size of policy updates effectively.

The original MAPPO [42] extends the trust region approach in the multi-agent setting which utilizes agents' aggregated trajectories to carry out policy optimization at each iteration. The joint policy $\boldsymbol{\mu^{\phi}}=[\mu^{\phi_1},...,\mu^{\phi_Z}]$ is optimized by maximizing the objective function as:

$$L^{\mathrm{MAPPO}}\left(\boldsymbol{\mu^{\phi}}\right)$$

$$=\sum_{z=1}^{Z}\mathop{\mathbb{E}}_{a_z\sim\mu^{\phi_z^{[l]}},o_z\sim\rho\left(\mu^{\phi_z^{[l]}}\right)}\min\left[\kappa_z\hat{A}_{\boldsymbol{\mu}}(\boldsymbol{o},a_z),\mathrm{clip}\left(\kappa_z,1\pm\epsilon\right)\hat{A}_{\boldsymbol{\mu}}(\boldsymbol{o},a_z)\right] \tag{10}$$

where $\boldsymbol{o}=[o_1,...,o_Z]$ and $\boldsymbol{a}=[a_1,...,a_Z]$ denote the collection of all agents' local observations and actions, respectively; $\kappa_z=\mu^{\phi_z}\left(a_z|o_z\right)/\mu^{\phi_z^{[l]}}\left(a_z|o_z\right)$ denotes the policy ratio of agent z, $\hat{A}_{\boldsymbol{\mu}}(\boldsymbol{o},a_z)=Q_{\boldsymbol{\mu}}(\boldsymbol{o},\boldsymbol{a})-Q_{\boldsymbol{\mu}}(o,a_{z-})$ signifies the multi-agent *advantage function* which compares the value of a specific action a_z to the mean value averaged over all actions of agent z, which therefore indicating whether the current action a_i will cause an increase in expected return; $Q_{\boldsymbol{\mu}}(\boldsymbol{o},\boldsymbol{a})$ represents the state-value function which is estimated by the critic network of each agent.[1]

However, the exploration of other agents may have an significant impact on the state value function estimations of MAPPO, resulting in learning instability. It has been shown that even when the gradients are correct, MAPPO cannot ensure consistent policy improvement [43]. To resolve this, we adopt a modified MAPPO approach that uses stochastic update schemes of agents' gradient directions to ensure monotonic improvement. To this end, each agent's individual policy is updated in a sequential manner and each agent has a distinct optimization objective during the sequential update that considers the updates from other agents. Each agent needs to wait for the other agents to complete updating due to the update schemes and such an update can guarantee a monotonic improvement of the policy. Practically, the joint policy $\boldsymbol{\mu^{\phi}}$ is updated as:

$$\boldsymbol{\phi^{[l+1]}}=\arg\max_{\boldsymbol{\phi}}\left(\kappa_z-1\right)\left(\mu_{j\in z-}\kappa_j\right)\hat{A}_{\boldsymbol{\mu}}(\boldsymbol{o},\boldsymbol{a}) \tag{11}$$

subject to:

$$\bar{D}_{KL}\left(\phi_z\|\phi_z^{[l]}\right)\le\delta,\forall z \tag{12}$$

where $z-$ represents the set of all agents other than agent z and it is indexed with j. Analogous to the single-agent PPO approach, in order to reduce the computational cost of $\bar{D}_{KL}\left(\phi_z\|\phi_z^{[l]}\right)$ when updating $\mu^{\phi^{[l]}}$, equations (11) and (12) can be simplified by using first-order derivatives as:

$$L^{\mathrm{Att-MAPPO}}\left(\boldsymbol{\mu^{\phi}}\right)$$

$$=\sum_{z=1}^{Z}\mathop{\mathbb{E}}_{a_z\sim\mu^{\phi_z^{[l]}},o_z\sim\rho\left(\mu^{\phi_z^{[l]}}\right)}\min\left[\kappa_z\left(\mu_{j\in z-}\kappa_j\right)\hat{A}_{\boldsymbol{\mu}}(\boldsymbol{o},\boldsymbol{a}),\right.$$

$$\left.\mathrm{clip}\left(\kappa_z,1\pm\epsilon\right)\left(\mu_{j\in z-}\kappa_j\right)\hat{A}_{\boldsymbol{\mu}}(\boldsymbol{o},\boldsymbol{a})\right] \tag{13}$$

In the original MAPPO method, the estimation of the state-value function $Q_{\boldsymbol{\mu}}$ (and subsequently the advantage function $A_{\boldsymbol{\mu}}$) takes the concatenation of all agents' local observations $\boldsymbol{o}$ and actions $\boldsymbol{a}$ as input in the centralized training. During execution, agent z's decisions can be performed in a fully decentralized manner through the deployed actor network μ^{ϕ_z} only using its local observation o_z. It can be observed that the critic's input dimension increases exponentially with the action/observation dimension of each agent, as well as the number of agents, contributing to limited scalability. In addition, $\boldsymbol{a}$ may expose CSPs' pricing strategies to their competitors, while $\boldsymbol{o}$ may reveal the private/local information of the FCSs operated by each CSP agent (such as the EV flows and charging demands) which closely tied up with the business strategy of the CSP in the pricing market. To address this, the CSP pricing market platform is introduced and acts as a trusted third party which provides each CSPs with information that reflect the collective behavior of other CSPs during centralized training, without knowing their specific local information.

To this effect, the estimation of $Q_{\boldsymbol{\mu}}^{\theta_z}(\boldsymbol{o},\boldsymbol{a})$ under Att-MAPPO only takes as input agent z's local observation o_z and action a_z, and other agents' *contributions* to agent z, $\boldsymbol{x_z}$, which is an abstract, learned representation of other agents' local information:

$$Q_{\boldsymbol{\mu}}^{\theta_z}(\boldsymbol{o},\boldsymbol{a})=f_z(g_z(o_z,a_z),\boldsymbol{x_z}) \tag{14}$$

where f_z is a two-layer multi-layer perceptron (MLP) and $g_z(o_z,a_z)=e_z$ is a one-layer MLP embedding function. The contributions from other agents x_z are evaluated as a weighted sum of each agent's *value* v_z as:

$$x_z=\sum\nolimits_{j\in z-}\eta_j v_j=\sum\nolimits_{j\in z-}\eta_j h(H_j^{v}e_j) \tag{15}$$

where H_j^{value} represents a learnable parameter matrix which transforms its embedding e_j into a *value*, h is a non-linear activation function, and η_j denotes the *attention weight* assigned to each agent j's value (i.e. the extent of attention that agent z paid to the value of agent j). It is obtained by comparing the similarity between agent z's embedding e_z with other agents' embeddings e_j, and then passes the similarity value to a softmax operator:

$$\eta_j=\exp^{(H_k e_j)^T H_q e_i}/\sum\nolimits_{j=1}^{N}\exp^{(H_k e_j)^T H_q e_i} \tag{16}$$

where H_j^{key} and H_j^{query} represent learnable parameter matrices that transform e_j/e_z into a *key/query*. Subsequently, the advantage function can be estimated based on the Q-value.

The training and execution architecture of the Att-MAPPO method is depicted in Fig. 3. During training, each agent z communicates its embedded information e_z to the pricing market platform, the latter then calculates and informs agent z the contributions of all other agents, x_z, in their estimations of Q-values (equation (14)). It can be observed that the agents communicate with each other implicitly through the market platform and without directly exchanging private information among them. During execution, the critics are no longer required (Fig. 3), and the weights of the actors are fixed, the trained actors are deployed for pricing of each CSP participant in the market.

5. Case studies

This section conducts numerical experiments on two test systems. We first validate the effectiveness of the proposed MARL method in estimating the NE of the CSP pricing game on an illustrative system with 5-node TN and 6-node PDN. Then, we consider a commonly employed

[1] It is clear that the estimation of the critic necessitates the input of all agent's actions and local observations.

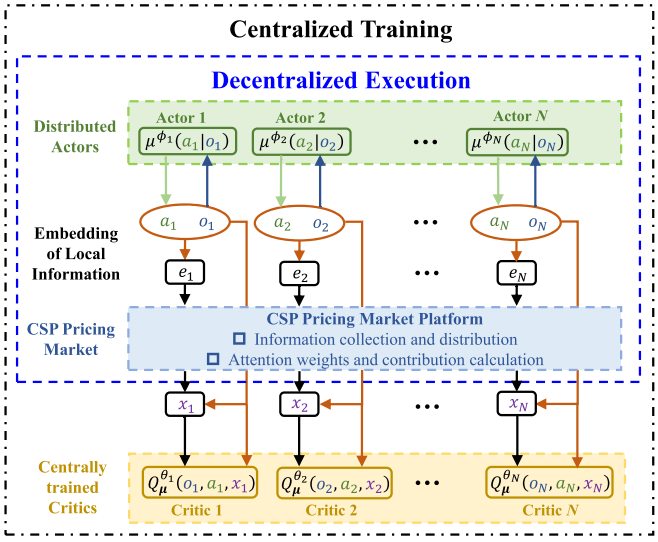

Fig. 3. Training and execution architecture of Att-MAPPO.

test system involving a 20-node TN and a 17-node PDN, and corroborate the benefits associated with price competition among CSPs in reducing the charging costs for elastic EV users in the TN, while improving the absorption of RES and the cost efficiency of the PDN. Although the formulation of CPTN environment and the design of CSP agents indeed allow the consideration of multiple FCSs (which may locate differently in the TN) being operated by the same CSP agent, and the charging price is offered at each FCS (i.e. the proposed MADRL method facilitates learning of FCS-specific prices that exhibit spatial diversity), it is assumed in the case studies that each CSP operates one FCS, for the ease of visualization and analysis of results.

The developed iterative solution procedure for UE identification and the SOCP model have been coded and solved using the optimization software Gurobi^{TM} [55] in Python, while the proposed Att-MAPPO and baseline MADRL methods have been implemented in Python with Tensorflow [56] on a computer with a 10-core 3.70 GHz Intel(R) Core(TM) i9-10900K processor and 128 GB of RAM. Parallel processing is implemented for the examined MARL methods, with one CPU core responsible for the learning of one agent, to leverage learning efficiency [57].

To evaluate the performance of the proposed approach, we compare our results with four baseline MARL methods including MAPPO, MADDPG, MADQN and Conc-PPO. The first three baseline methods all fall within the CTDE paradigm, each FCS agent estimate its Q-value by directly taking as input the local/private observations and actions of all the CSP agents, whereas the proposed method employs the attention mechanism to selective incorporates relevant and embedded information of other agents in the each agent's Q-value estimate. In Conc-PPO, CSP agents compete in CPTN environment by employing the concurrent learning approach without any input information from other CSPs. The learning is performed in parallel in Conc-PPO.

The first three baseline methods all fall within the CTDE paradigm, each CSP agent estimate its Q-value by directly taking as input the local/private observations and actions of all the CSP agents, whereas the proposed Att-MAPPO methods employ the attention mechanism to selective incorporates relevant and embedded information of other agents in the each agent's Q-value estimate.

5.1. Small network example

The data and settings of the small network example involving a 5-node TN and 5-node PDN is provided in Appendix B.1 for space limitation reasons. Fig. 4 illustrates the episodic moving average NI of the pricing game as well as the episodic moving average reward (or profit) for each CSP agent. It is observed that the NI under proposed Att-MAPPO method converges to approximately 1, which suggests successful identification of the NE of the non-cooperative pricing game, achieving 3.21%, 5.17% and 9.45% higher NI compared to MAPPO, MADDPG and MADQN, respectively. This superior learning performance is also attributed to the developed sequential update scheme (Section 4.2) which is able to guarantee consistent policy improvement for the agents, which is deemed imperative in multi-agent coordination in the highly uncertain CPTN environment. By contrast, Conc-PPO method exhibits the highest variance and an unstable learning behavior, eventually failing to reach convergence at termination (500 episodes). As discussed in Section 4.1, this is because all the CSP agents are adapting their pricing strategies independently, rendering the dynamics of the CPTN environment non-stationary for any agents.

As discussed in Section 1.2, the FCSs are generally regarded as PDN assets, and the charging prices are established by the DNO as fixed prices or as the DLMPs. Conversely, the proposed CSP pricing strategies coordination comprehensively captures the interdependence among the charging prices, EV routing and charging, and PDN operation with RES. To this end, we compare the learning behavior of Att-MAPPO in the following four cases:

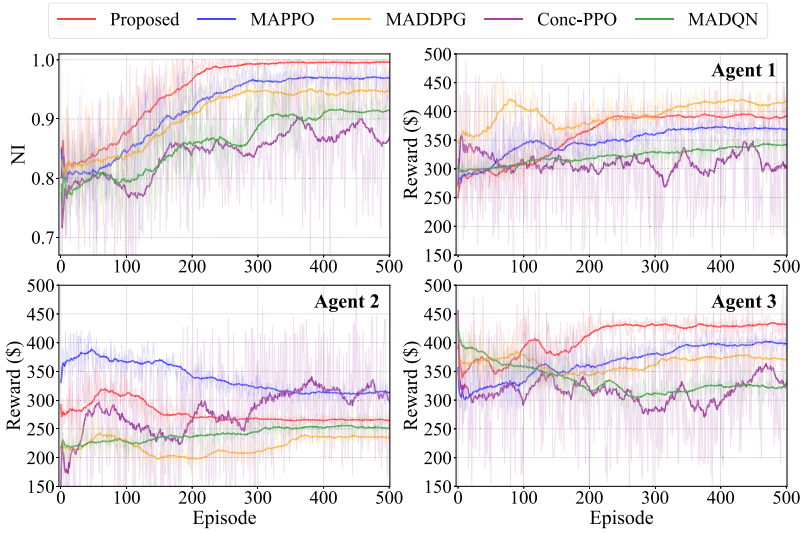

Fig. 4. Episodic moving average NI of the pricing game and episodic moving average reward of each CSP agent for the small network example.

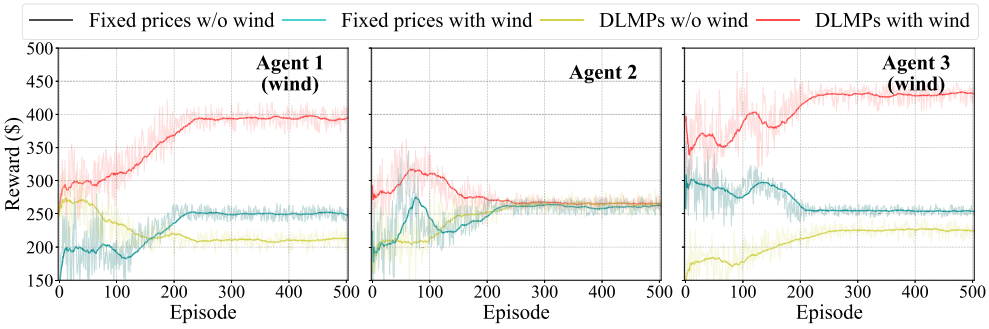

Fig. 5. Episodic moving average of economic profit of each CSP for the four cases of the small network example.

Table 2
Charging demand F_k^e (in MW), charging prices π_k (in $/MWh) and electricity purchasing prices (i.e. fixed prices or DLMPs) (in $/MWh) for each CSP under Att-MAPPO at convergence for the four cases of the small network example.

	CSP 1 (wind)			CSP 2			**CSP 3 (wind)**		
Case	F_1^e	π_1	λ_1	F_2^e	π_2	λ_2	F_3^e	π_3	λ_3
i) & ii)	5.12	150	100	5.12	150	100	5.12	150	100
iii)	5.65	147	102	5.70	150	101	5.61	147	102
iv)	6.04	144	78	5.34	150	97	6.42	142	74

Table 3
Charging demand F_k^e (in MW), charging prices π_k (in $/MWh) and electricity purchasing prices (in $/MWh) under Att-MAPPO at convergence under different settings of the CSP agents.

	CSP 1 (wind)			CSP 2			**CSP 3 (wind)**		
	F_1^e	π_1	λ_1	F_2^e	π_2	λ_2	F_3^e	π_3	λ_3
Setting 1	6.04	144	78	5.34	150	97	6.42	142	74
Setting 2	6.10	146	78	5.52	149	100	6.14	145	76

i) Fixed prices w/o wind: no wind generators are connected to the PDN and the electricity purchase prices λ_k in the FCS profit calculation (1) is fixed at 100$/MWh;

ii) Fixed prices with wind: the same setting of λ_k as in i), F1 and F3 are co-located with wind (Fig. B.1);

iii) DLMPs w/o wind: the same setting of wind generation as in i), but λ_k is set as the DLMPs calculated as at the optimal solution of SOCP (A.25)-(A.37); and

iv) DLMPs with wind: the same setting of λ_k as in iii) and the same setting of wind generation as in ii).

Fig. 5 illustrates the episodic moving average of charging prices while Table 2 reveals the economic conditions of each CSP corresponding to the above four cases. As depicted in Fig. 5, the learning curves corresponding to cases i) and ii) overlap with each other, and all three CSPs set their charging prices at 150$/MWh (i.e. the price cap). This is because that the purchase costs of the CSPs are calculated with fixed prices, and therefore the operating status of the PDN has no impact on the purchase costs and therefore on the pricing strategies of the CSPs. Furthermore, since the capacities of the links and FCSs corresponding to each of the three feasible paths are identical, and given the fact that one pair of O-D demand is considered, the EV flow and charging demand are evenly distributed to each path and each FCS (i.e. 5.12 MW) in the TAP-UE (Table 2). As a result, each CSP will cap its charging price in order to maximize its revenue, regardless of whether its charging demand is supplied by the wind generation or not. Consequently, all three CSPs attain identical profit of $253 in cases i) and ii) (Fig. 5).

On the other hand, in case iv) when the CSP purchase cost is calculated with DLMPs, CSPs/Agents 1 and 3 enjoy much cheaper purchase prices (78$/MWh and 74$/MWh) since their charging demand is supplied with the co-located, free wind generation, compared to case iii) where the demand is supplied by the DG at a quadratic cost. As a result, CSPs 1 and 3 will reduce their charging prices to 144$/MWh and 142$/MWh in order to stay competitive and secure higher amount of EV charging demand (6.04 MW and 6.42 MW). Conversely, CSP 2 still prices at 150$/MWh to maintain its revenue at a high level, in compensation for its dropped charging demand (since EV users prefer to route and charge their vehicles to cheaper FCSs) (Table 2). In other words, CSPs 1 and 3 become more competitive compared to CSP 2 in the pricing game, driven by the reduced purchase cost resultant from the co-located wind generator and the increase charging demand resultant from the reduced charging prices. Consequently, CSPs 1 and 3 acquire significantly higher economic profits than CSP 2 in case iv) than the value they acquire in case iii) (Table 2).

Finally, we consider two different settings of the CSP agents. Setting 1: each CSP agent operates one FCS and Setting 2: CSP agent 1 operates FCS 1 and FCS 3, while CSP agent 2 operates FCS 2, although the charging prices are still FCS-specific prices that can exhibit spatial diversity. Table 3 reveals the economic conditions of each CSP corresponding to the two different settings. As revealed in Table 3, in Setting 2, since the two wind-collocated FCSs are both managed by CSP agent 1, a visible reduced extent of price competition and is witnessed, reflected in the increased charging prices (π_1 and π_3) but still attracting significant charging demand of EV users, compared to the values observed in Setting 1. Conversely in Setting 1, since each FCS is operated by different CSP agents, agents 1 and 3 are in higher extent of price competition, which tends to reduce the charging prices at both sites resultant from such competition. To this end, it can be observed that the proposed Att-MAPPO method is capable of modeling different settings of the CSP agents, which often resulting in their different gaming behaviors in the charging pricing market.

5.2. *Large network example*

The data and settings of the large network example involving a 20-node TN and 33-node PDN is provided in Appendix B.2 for space limitation reasons. Fig. 6 and Fig. 7 illustrate, respectively, the episodic moving average NI of the pricing game and the episodic moving average reward of each CSP agent.

As shown in Fig. 6, under Conc-PPO, the CSP agents are learning and adapting their policies individually, frequent change in their policies renders the environment dynamics non-stationary, contributing to unstable and non-convergent learning performance after 15,000 episodes of training. Similar oscillative and divergent behavior is observed in each agent's individual learning process (Fig. 7).

On the other hand, convergence is prevalent under MADRL methods within the CTDE paradigm, among them, the proposed Att-MAPPO method improves the NI in a stable fashion with a decreasing variance (as also observed in individual learning process (Fig. 7)). At convergence, Att-MAPPO significantly outperforms baseline MADRL methods, achieving the highest moving average NI of 0.94, which is 8.12%, 10.57% and 18.73% higher than the NI achieved under MAPPO, MAD-

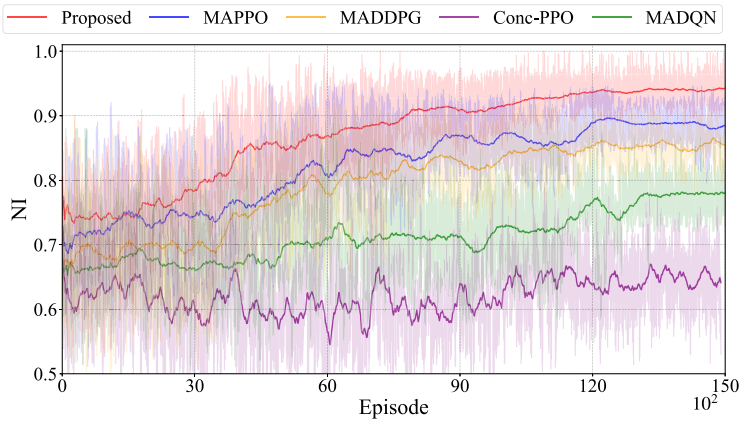

Fig. 6. Episodic moving average NI of the pricing game for the large network example.

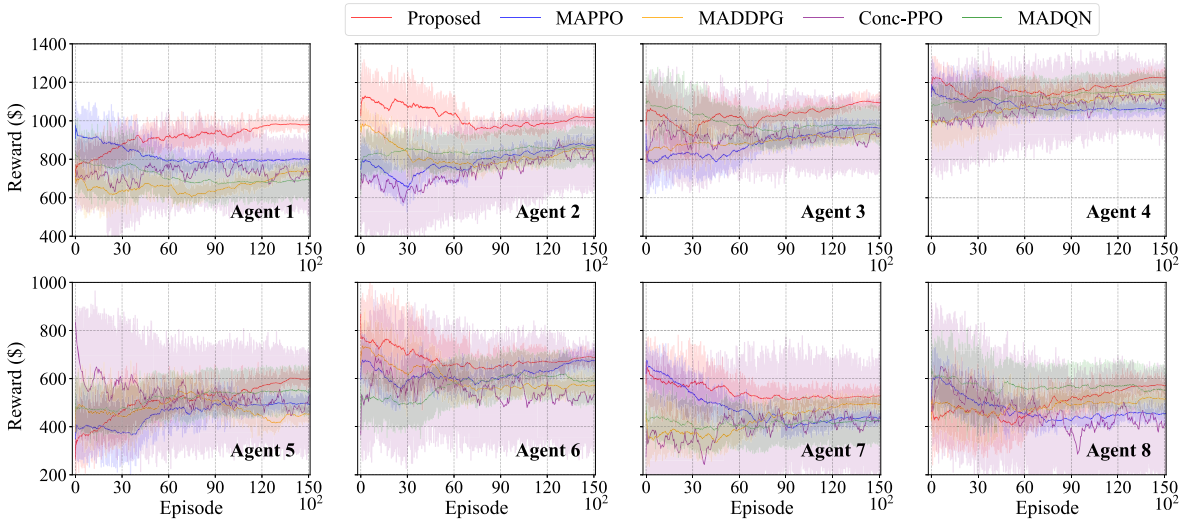

Fig. 7. Episodic moving average reward of each CSP agent for the large network example.

DPG and MADQN, respectively. Att-MAPPO's superiority in terms of learning stability and convergence is also attributed to the developed sequential update scheme, which guarantees consistent policy improvement for all agents (as reflected with a decreased variance, as depicted clearly in the learning process of agents 1 and 2 in Fig. 7)). Its usefulness in foster more efficient learning coordination in the highly uncertain CPTN environment is again demonstrated in this larger network example. Furthermore, the attention mechanism enables selective incorporation of relevant information of agents (i.e. the contribution in equation (15)) to aid centralized training of Att-MAPPO. Comparing with directly importing the local observations and actions of all agents in MAPPO, MADDPG and MADQN, the proposed method not only substantially reduces the input dimension of the critic network, but also allows each agent to make more informed decisions based on the influence of the collective behavior of other agents participated in the pricing market, but without knowing their specific local information, safeguarding the business confidentiality of the CSPs in the market. Conversely when the attention mechanism is not used, each agent treats other agents' information as equal. This means that, for a particular CPTN operating scenario, each agent may not be able to identify a subset of other agents which constitutes its key competitors, whose pricing policies are of highest relevance and is critical for the said agent to adapt its pricing policy to.

In order to further highlight the superior convergence behavior of Att-MAPPO, the standard deviation of the episodic reward for all eight agents measured every 3000 episodes is compared under Att-MAPPO and Conc-PPO, as summarized in Table 4. The standard deviations under Conc-PPO fluctuate and maintain at a large level during the entire learning process of 15000 episodes for all eight agents, no decreasing trend can be observed. Conversely, a clear trend of gradually reducing standard deviation of the episodic reward is observed under Att-MAPPO for all the 8 agents. Furthermore, it can be observed in the columns of Table 4 corresponding to episode 15,000 or at termination, the standard deviations only account for less than 2% of the average reward under Att-MAPPO for all agents. On the contrary, these ratios range from approximately 18% to 30% under Conc-PPO.

5.2.1. *Impact of EV coordination on RES absorption, CSP profits and PDN cost*

Similarly as in the small network example, we examine the same four cases to analyze the interactions of CSP pricing, EV users routing and charging as well as the operating cost efficiency and RES absorption ratio of the PDN. In cases i) and ii), a fixed price of 120$/MWh is assumed. Fig. 8 illustrates episodic moving average of charging prices while Table 5 reveals the economic profit of each CSP corresponding to the examined four cases.

It can be observed in Table 5 that certain trends are prevailing in both small and large network examples. In cases i) and ii), the prevalent strategies are to cap the charging price around the maximum limit (180$/MWh) regardless of whether the FCS charging demand can be meet with free wind generation or not, since the purchase costs of the CSPs are settled with fixed prices. In other words, co-locating with RES

Table 4
Standard deviation of episodic reward for each agent under Att-MAPPO and Conc-PPO.

	Att-MAPPO					Conc-PPO				
Episode	3000	6000	9000	12000	15000	3000	6000	9000	12000	15000
Agent 1	49.1	40.8	38.1	29.2	10.4 (1.3%)	149.0	142.2	135.9	130.6	132.3 (18.1%)
Agent 2	65.5	48.6	27.4	24.2	11.9 (1.1%)	143.8	133.2	135.7	138.9	134.0 (18.4%)
Agent 3	61.8	42.7	24.6	20.9	12.3 (1.2%)	147.2	130.1	127.7	123.6	128.1 (18.3%)
Agent 4	49.9	40.3	28.8	25.7	10.7 (0.8%)	149.1	139.2	128.7	135.4	138.3 (25.2%)
Agent 5	61.2	46.6	41.0	31.2	12.6 (1.6%)	157.4	149.3	143.9	124.0	133.6 (27.3%)
Agent 6	71.3	36.9	25.4	24.1	11.7 (1.8%)	157.7	150.8	145.0	141.9	145.1 (29.5%)
Agent 7	80.1	51.2	34.9	27.2	9.3 (1.2%)	154.2	140.3	129.8	136.7	139.1 (26.3%)
Agent 8	66.3	50.7	34.9	27.4	8.2 (1.6%)	152.2	149.9	140.2	138.6	135.5 (31.2%)

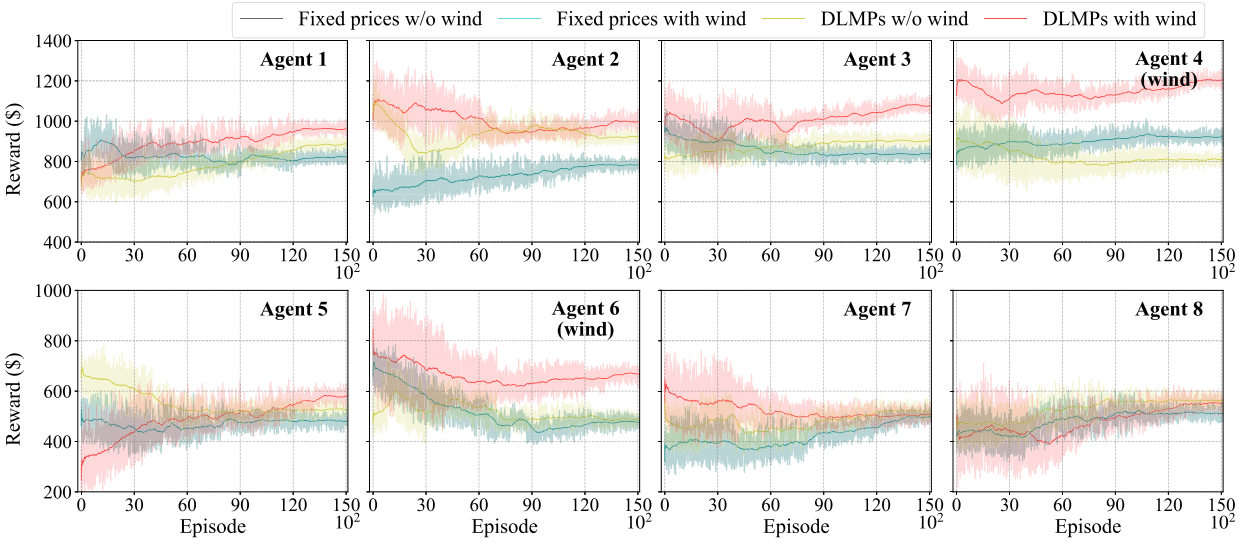

Fig. 8. Episodic moving average of economic profit of each CSP for the four cases of the large network example.

Table 5
Charging demand F_k^e (in MW), charging prices π_k (in $/MWh) and electricity purchasing prices (i.e. fixed prices or DLMPs) (in $/MWh) for each CSP under Att-MAPPO at convergence for the four cases of the large network example.

	CSP 1			CSP 2			CSP 3			**CSP 4 (wind)**		
Case	F_1^e	π_1	λ_1	F_2^e	π_2	λ_2	F_3^e	π_3	λ_3	F_4^e	π_4	λ_4
i) & ii)	13.6	180	120	13.4	180	120	14.1	178	120	13.8	180	120
iii)	14.3	178	122	15.1	176	118	14.9	176	125	13.9	177	121
iv)	17.3	173	114	16.4	175	111	16.7	175	112	19.6	169	107
	CSP 5			**CSP 6 (wind)**			CSP 7			CSP 8		
Case	F_5^e	π_5	λ_5	F_6^e	π_6	λ_6	F_7^e	π_7	λ_7	F_8^e	π_8	λ_8
i) & ii)	8.3	179	120	8.1	180	120	7.6	180	120	9.1	177	120
iii)	8.3	174	119	7.5	175	120	8.2	174	124	7.9	176	113
iv)	9.9	169	114	10.7	166	102	9.2	172	113	9.7	170	111

does not create a business advantage for CSPs in the pricing market. Nevertheless, in the large network example, when the size of the TN is larger and O-D travel demands are higher and more diversified, although all CSPs set identical charging prices in cases i) and ii), the charging demands at CSPs 1-4 (inner ring of the TN) are significantly larger than the demands at CSPs 5-8 (outer ring of the TN) (Table 5), since TN nodes T4 and T5 are set as the main origins for the EV O-D demand, the charging activities are thus more prominent in a close proximity around T4 and T5, creating profit deviations between CSPs 1-4 and CSPs 5-8 (Fig. 8).

Furthermore, comparing cases iii) and iv), CSPs 4 and 6 are motivated to reduce their charging prices in order to attract more EV users and thus higher charging demand (19.6 MW and 10.7 MW, respectively), since the later can be supplied from the co-located and free wind generation. As a result, CSP 4/6 is more competitive with respect to other CSPs in the inner/outer ring of the TN. This result, once again, validates the capability of the proposed method in comprehensive encapsulation of the inter-related relationship across CSP pricing strategies, EV routing and charging load distribution in PDN, and DLMPs are reflective on the purchase costs influenced by the absorption of wind generation. On the other hand, such cost reduction is unattainable for other CSPs, and therefore their price reductions (in response to the price reduction of CSPs 4 and 6) are very limited, driven by the need to maintain their revenue at a high level, albeit at the expense of some demand curtailment (Table 5). Overall, CSPs 4 and 6 acquire significantly higher profits than CSPs in their close proximity in case iv) (Fig. 8).

Fig. 9 depicts the boxplot of wind absorption ratio for cases ii) and iv) and boxplot of overall PDN operating costs for cases i)-iv) for 100 test episodes. It is observed that the average wind absorption ratio is

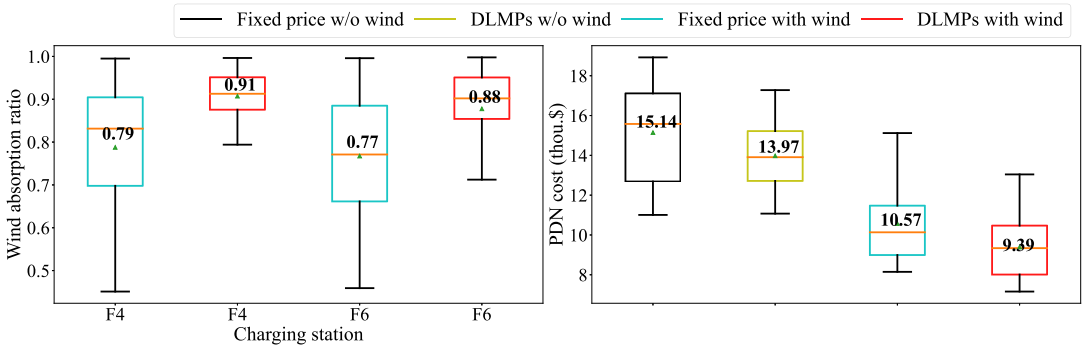

Fig. 9. Boxplot of wind absorption ratio and overall PDN operating costs with 100 test episodes.

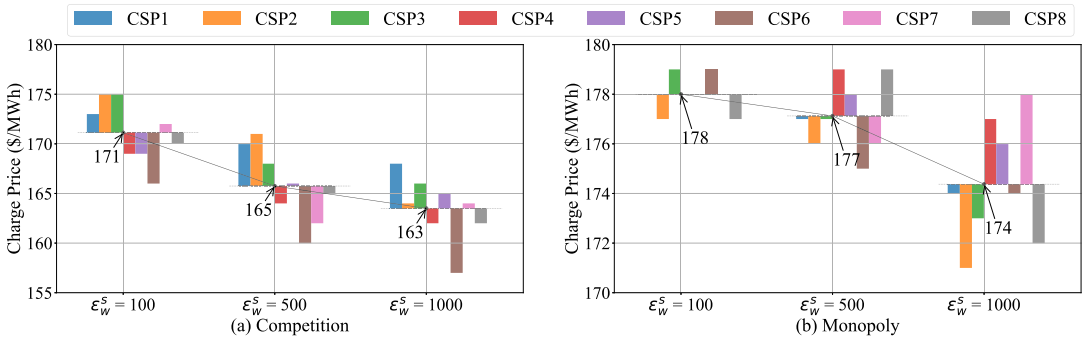

Fig. 10. Optimal charging prices of each CSP obtained at the convergence of the proposed Att-MAPPO method, for an increasing degrees of cost elasticity and under competition and monopoly scenarios.

significantly higher in case iv) than in case ii), while its variance is significantly lower. The reduced mean is mainly associated with the increased EV charging demand at CSPs 4 and 6 (driven by the reduced charging prices). The size of EV charging demand depends on the wind generation scenario, i.e. in scenarios with high/low wind generation, more/less EV charging demand will be attracted to CSPs 4 and 6 to absorption more/less wind, this is reflected in the reduced variance in wind absorption. This suggests that the learnt coordinated pricing policies and the resultant ultimate charging demand of all EV users can generalize well to different wind (or RES) generation scenarios. As a result, the overall cost efficiency of the PND is substantially improved with higher RES penetration, and the DLMPs are incorporated as guiding signals for EV routing and charging activities.

5.2.2. *Impact of cost elasticity of EV O-D demand*

The aim of this section lies in exploring the impact of cost elasticity of EV O-D demand on the pricing strategies of the CSPs. To this effect, we examine the following two scenarios regarding FCS management (Table 1). The first scenario "competition" involves self-interested CSPs competing against each other in the pricing market, which is also the scenario described in Section 2.1. In the second scenario "monopoly", it is assumed that a single CSP determines the prices for all CSPs and its pricing strategy is trained by the single-agent PPO method [53].

Fig. 10 illustrates the individual charging prices of each CSP and average prices obtained at the convergence of the proposed Att-MAPPO method, for an increasing degrees of cost elasticity (i.e. higher values for parameter ε_w^s) and under competition and monopoly scenarios. The charging prices are much higher in the monopoly scenario, since the single CSP occupies the entire market and it therefore strategically leverages the prices to attain the maximum profit; whereas the CSPs tend to lower their prices to attract more EV charging demand for higher market shares in the competition scenario. Furthermore, as the elasticity increases, it can be observed that the average charging prices are reduced under both scenarios, but the extent of the reduction in more prominent in competition than in monopoly. This reveals the benefits of price competition among CSPs in lower the charging costs for the EV users, and the benefit is more significant considering the intrinsic and increasing elasticity of the users (given the rapid development of public transportation infrastructures and diversified means for user commuting). On the other hand, the EV users do not enjoy cost reduction under monopolistic managing of FCSs, even with higher cost elasticity, hindering the progress for higher adoption of EVs.

5.2.3. *Impact of initial SoC and GV O-D demand*

This section analyzes in more depth the impact of initial SoC of EV O-D demand and GV O-D demand on the routing and charging activities of EV users as well as the overall traffic flows in the TN. To this effect, we examine one of the long distance travel demand from node T2 to node T11 for both EV and GV users, considering 4 different scenarios regarding initial SoC of EV O-D demand: VL SoC, L SoC, M SoC and H SoC (in ascending order) and 3 different scenarios regarding GV O-D demand: no GV, L GV and H GV (in ascending order).

Fig. 11 depicts the EV flows and charging demand as well as GV flows in O-D pair T2-T11, for different initial SoC of EV O-D demand and GV O-D demand scenarios. The paths selected by the majority of EV/GV users are highlighted using black arrows in Fig. 11. First of all, let us examine the scenario without GV O-D demand, and thus the analysis focuses on the impact of initial SoC of EV O-D demand, this refers to the four sub-figures in the 1st row of Fig. 11. Since origin T2 is closest to the C6 (co-locate with wind generator) and benefits from its lower charging prices, an active path involving C6 constitutes the most economic option for most EV users when their initial SoC/range anxiety is low/high (scenarios: VL SoC and L SoC). On the other hand, for higher initial SoC, an active path connecting T2 and T11 involving C4 (also co-locate with wind generator) is dominant choice for most EV users, where the charging activities can be delayed given less range anxiety.

Next, let us analyze the combined impact of initial SoC and GV O-D demand, the relevant results are depicted in the sub-figures in the 2nd-5th rows of Fig. 11. In scenario L GV (i.e. 2nd-3rd rows of Fig. 11), in order to avoid traffic congestion, the GV users predominately select

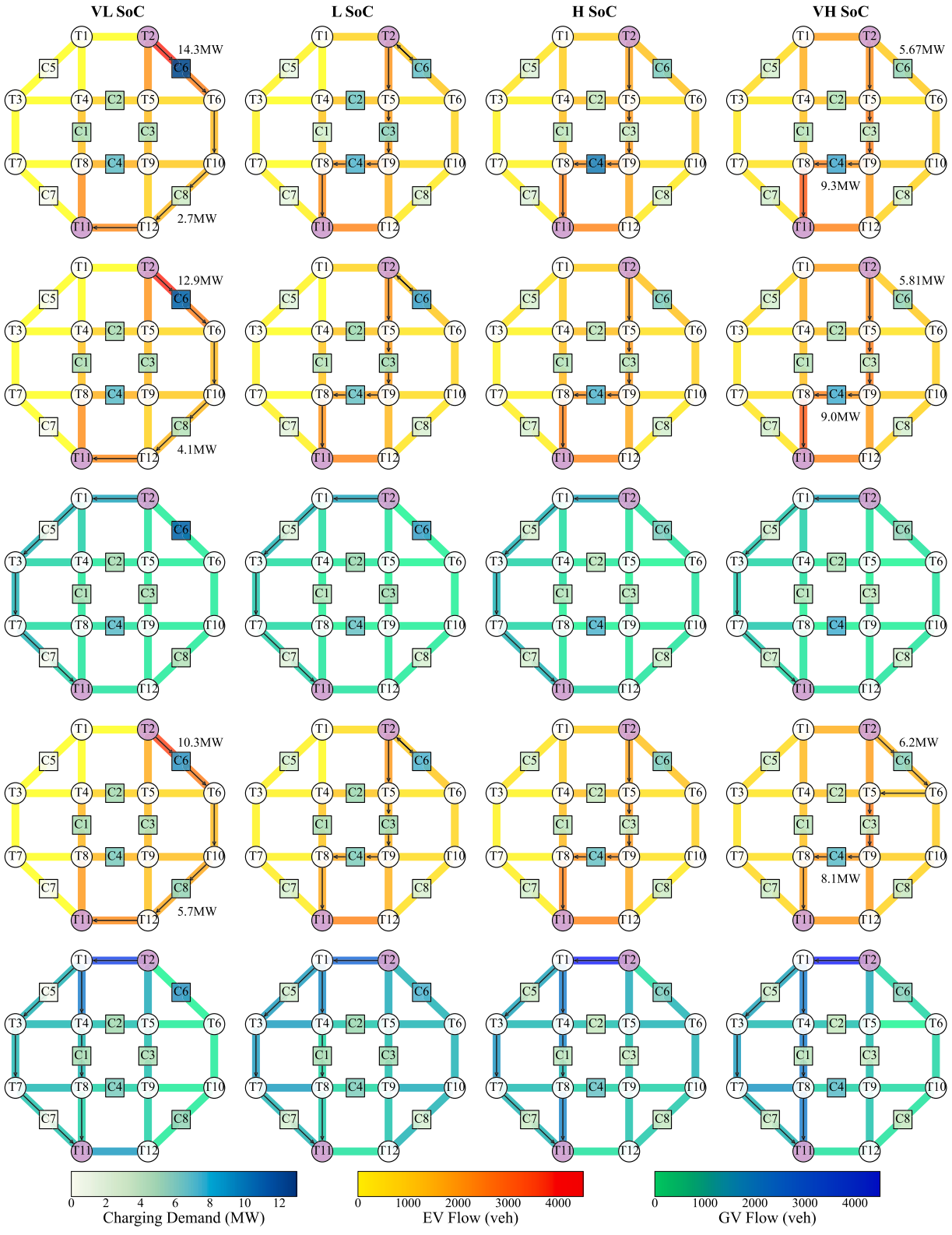

Fig. 11. EV flows and charging demand for no GV scenario (1st row), EV flows and charging demand/GV flows for L GV (2nd-3rd rows) and H GV scenarios (4th-5th rows) and for different initial SoC scenarios. The black arrows highlight the paths selected by the majority of EV/GV users.

a path that corresponds to the lowest travel time and does not overlap with the most prominent EV charging path, in this case path T2 > T1 > T4 > T8 > T11. As a result, the EV routing and charging behavior does not change with respect to the no GV scenario. Conversely, in scenario H GV (i.e. 4th-5th rows of Fig. 11), when the initial SoC is low, it can be observed that the EV users still choose the same charging path as the one in no GV scenario, but the spatial distribution of EV charging demand is affected. Specifically, the EV users will choose to acquire less energy at C6 and then acquire more energy at C8, in order to reduce the congestions at C6 and therefore the charging time costs, despite the

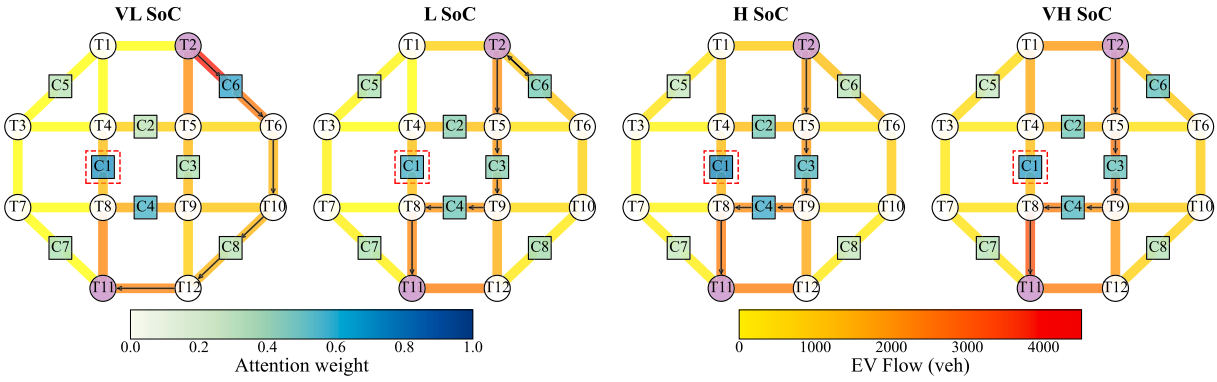

Fig. 12. Illustration of attention weights for agent C1 and EV flows (assuming no GV) for different initial SoC scenarios.

Table 6
Computational performance of the examined MADRL methods.

Method	Att-MAPPO	MAPPO	MADDPG	Conc-PPO	MADQN
Average training time per episode (s)	1.33	1.94	1.97	1.71	1.78
Number of episode	12000	12500	12800	15000[a]	13000
Total training time (h)	4.43	6.74	7.00	7.12	6.43
Average execution time per CSP (ms)	19.25	18.88	18.85	17.29	23.17

[a] Failure to attain convergence.

charging price is lower at C6 than at C8. Furthermore, when the initial SoC is high, the high penetration of GV users affects the EV charging path. Specifically, the EV users will choose to acquire energy at the two CSPs with the lowest prices (first at F6 and then at F4), despite the fact that the travel energy requirement for EV users is the lowest. As a result, it shifts the charging burden at C4 to C6, and thus reduces the charging time costs of EV users at C4 and alleviates the traffic congestion on the links in the inner ring of the TN. Analogously to scenario L GV, to alleviate traffic congestion, paths T2 > T1 > T4 > T8 > T11 and T2 > T1 > T3 > T7 > T11 constitute the most prominent routing choices for GV users, despite a portion of GV users route commonly with EV users. This analysis validates the capability of the proposed Att-MAPPO method in generalizing to different scenarios of initial SoC and GV O-D demand scenarios, by identifying the optimal charging routing and/or charging activities of GV/EV users.

5.2.4. Rationale of attention mechanism

Fig. 12 illustrates the learned attention weights of agent 1 and the associated EV flows (assuming no GV) for different initial Soc scenarios (the same scenario as depicted in the 1st row of Fig. 11). The darker the color gets, more attention is paid to the respective agents. It can be observed that when the range anxiety is severe as observed in the VL SoC scenario, C1 pays more attention to the embedded information of C6 and C4 (which attract the highest and second highest EV flows and thus charging demand) to set its own pricing strategy, even though C6 locates geographical far away for C1. On the other hand, when the range anxiety is low as observed in the VH SoC scenario, a great proportion of EV flows are shifted towards C4 for charging, C1 thus pays more attention to the neighboring agents (C2, C3 and C4). In other words, in this scenario, the four CSPs in the inner ring of the TN exhibit higher competitiveness in the pricing market in terms of attracting EV users than the four CSPs in the outer ring, and consequently each of these agents will pay more attention to the rest three competitors. Consequently, the proposed attention mechanism facilitates each agent to selectively paying attention to the information of important and relevant agents during training, which not only improves the learning efficiency but also lightens the training burden (Section 5.2.5).

5.2.5. Computational performance

Table 6 summarizes the computational evaluation of the investigated five MADRL methods, with regard to a) the average training time per episode, b) the number of episodes and c) the total training time required to attain convergence, and d) the average execution time per CSP. As previously discussed, Conc-PPO fails to reach convergence and thus the reported number of episodes and total training time corresponds to the 15,000 episodes executed in our experiments.

As shown in Table 6, the average training time per episode is the highest in MADDPG and MAPPO, which is driven by their requirements to indifferently incorporation of the observations and actions of all agents to train their critics, which increases the training burden at each time step. The number of episodes and the total training time required to reach convergence are highest in Conc-PPO (since convergence is not achieved at termination), lower in MAPPO, MADDPG and MADQN, and the lowest in Att-MAPPO owning to the employment of attention mechanism and the sequential update scheme. The former allows the learning of each agent's policy benefiting from other agents' experiences, while keeping a low computational burden by paying attention only to relevant agents' information. The latter facilitates more efficient learning coordination in the highly uncertain CPTN environment, and guarantee monotonic policy improvements for the agents, and thus accelerate the overall coordination of all agent's pricing policies. Furthermore, it can be observed that all investigated MADRL methods exhibit a similar average execution time for each CSP, which is in the order of milliseconds, implying that they can be effectively deployed in practice for all participated CSP agents, and for various uncertain CTPN environment operating status.

6. Conclusion and future work

This work examines the non-cooperative pricing coordination between profit-driven CSPs, targeted to derive suitable locational price signals to guide the desirable routing and charging behaviors of EV users, so as to unleash their significant flexibility potential and deliver benefits towards enhancement of CPTN operation efficiency. The pricing game simultaneously takes into account the complex interactions between CSPs' pricing strategies, users' decisions, as well as the operation of CPTN. Prominent types of uncertainties stemming from the EV

O-D demand cost elasticity, initial SoC and GV O-D demand are encapsulated in modeling of environment.

A novel MARL method, founded on the CTDE paradigm and combines the strength of MAPPO and the attention mechanism, is developed to solve the pricing game. The incorporated attention scheme enables selecting relevant encoded information to estimate the state-value and advantage functions, as opposite to indifferently incorporating the local information of all agents for such estimations. As a result, Att-MAPPO overcomes the non-stationarity, computational complexity, and confidentiality drawbacks of state-of-the-art MARL approaches. Furthermore, to foster more efficient learning coordination in the highly uncertain environment, a sequential update scheme is developed to achieve monotonic policy improvement and promise better convergence property in the learning stage.

Case studies on an illustrative and a large-scale test systems reveal that the proposed method facilitates sufficient competition among CSPs, and reveal the core benefits in terms of reduced charging costs for EV users by examining a range of O-D demand cost elasticity scenarios, enhancement of RES absorption and cost efficiency improvement of the power distribution network. Results also corroborate the generalization capability of Att-MAPPO in coping with the uncertain EV O-D demand cost elasticity, EV initial SoC level and GV O-D demand, by identifying equilibrium pricing strategies and the associated optimal charging routing and charging decisions of users.

Despite the advantages featured by the proposed MADRL methodological framework, it assumes a perfect environment without communication failures are not examined. In practice, the embedded information that one agent (or several agents) communicates to the pricing market platform may be lost due to communication failures or packet loss, which will in turn affects the training of the pricing strategy of all the agents. In light of this challenge, future work will explore Bayesian inference techniques that allow the normal agents to estimate/update beliefs on the embedded information of agents experiencing communication failure, based on historical learning experiences.

CRediT authorship contribution statement

Yujian Ye: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. **Hongru Wang:** Data curation, Formal analysis, Software, Validation, Visualization. **Tianxiang Cui:** Formal analysis, Investigation, Methodology, Resources, Validation, Writing – review & editing. **Xiaoying Yang:** Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization. **Shaofu Yang:** Formal analysis, Investigation, Validation. **Min-Ling Zhang:** Investigation, Project administration, Writing – review & editing.

Declaration of competing interest

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

Data availability

Datasets related to this article can be found at https://dx.doi.org/10.17632/25vjrnz3gp.1, an open-source online data repository hosted at Mendeley Data [58].

Acknowledgements

This work has been supported in part by the National Natural Science Foundation of China under Grant 52207082, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20220842, in part by State Key Laboratory of Power Systems Operation and Control under Grant SKLD23KM08, and in part by Ningbo Natural Science Foundation under Grant 2023J194.

Appendix A

A.1. Modified TAP-UE model in Section 2.2

$$U(\boldsymbol{\pi}) = \min \sum_{a\in A} \omega \int_0^{x_a+x_a^g} t_a(z)dz + \sum_{k\in K} \omega \int_0^{x_k} t_k(z)dz + \sum_{w\in W, p\in \Xi_w} (C_{p,w}^{ch} + C_{p,w}^{tm}) f_{p,w} - \sum_{w\in W} \int_0^{d_w} g_w^{-1}(z)dz \tag{A.1}$$

subject to:

$$C_{p,w}^{ch} = \sum_{k\in K} \pi_k F_{p,w,k}^e, \forall w \in W, \forall p \in \Xi_w \tag{A.2}$$

$$C_{p,w}^{tm} = \sum_{k\in K} c^{tm} F_{p,w,k}^e, \forall w \in W, \forall p \in \Xi_w \tag{A.3}$$

$$d_w = g_w(C_w^{tot}, \epsilon_w), \forall w \in W \tag{A.4}$$

$$0 \le d_w \le \overline{d}_w, \forall w \in W \tag{A.5}$$

$$x_a = \sum_{w\in W, p\in \Xi_w} f_{p,w}\delta_{p,w,a}, \forall a \in A \tag{A.6}$$

$$x_a^g = \sum_{w\in W, p^g\in \Xi_w} f_{p,w}^g \delta_{p,w,a}^g, \forall a \in A \tag{A.7}$$

$$x_k = \sum_{w\in W, p\in \Xi_w} f_{p,w}\sigma_{p,w,k}, \forall k \in K \tag{A.8}$$

$$t_a\left(x_a, x_a^g\right) = t_a^0\left(1 + 0.15\left((x_a + x_a^g)/\overline{x}_a\right)^4\right), \forall a \in A \tag{A.9}$$

$$t_k\left(x_k\right) = t_k^0 \Gamma x_k/(\overline{x}_k - x_k), \forall k \in K \tag{A.10}$$

$$\sum_{p\in \Xi_w} f_{p,w} = d_w, \forall w \in W \tag{A.11}$$

$$\sum_{p^g\in \Xi_w^g} f_{p,w}^g = d_w^g, \forall w^g \in W^g \tag{A.12}$$

$$f_{p,w} \ge 0, \forall w \in W, \forall p \in \Xi_w \tag{A.13}$$

$$f_{p,w}^g \ge 0, \forall w^g \in W^g, \forall p^g \in \Xi_w^g \tag{A.14}$$

$$P_k^{evd} = \sum_{w\in W, p\in P_w} f_{p,w} F_{p,w,k}^e, \forall k \in K \tag{A.15}$$

The objective function (A.1) is composed of four cost components: i) total travel time cost on all links of EV and GV users, ii) total waiting time cost at all FCSs, iii) charging electricity cost (A.2) and charging time cost (A.3) of all EV users and iv) the cost reduction attributed to the reduction of EV O-D travel demand. Constraint (A.4) expresses the EV O-D demand function, parameterized by cost elasticity ϵ_w, capturing the negative correlation between O-D demand w and the total costs C_w^{tot}. Constraint (A.5) imposes an upper bound for the elastic EV travel demand d_w. Constraints (A.6)-(A.8) describe the relationships between the EV/GV path flow $f_{p,w}/f_{p,w}^g$ and EV/GV flows on link a and EV flows at FCS k. Constraint (A.9) penalizes the EV/GV flows on link a through a sharp increase in the traveling time. Constraint (A.10) represents the queuing time of the EV flows at FCS k. Constraints (A.11)-(A.12) specify the equality between the total EV/GV path flows and the O-D demand for each O-D pair. Constraints (A.13)-(A.14) express the non-negativity of EV/GV path flows. Finally, constraint (A.15) expresses the EV charging demand at each FCS.

A.2. APG optimization model in Section 2.3

$$\min \sum_{a\in A_p} \omega t_a(x_a^*, x_a^{g*}) y_{w,a} + \sum_{k\in K} \left(\omega t_k(x_k^*)\epsilon_{k,w} + c^{tm} F_{p,w,k}^e + \pi_k F_{p,w,k}^e\right) \tag{A.16}$$

subject to:

$$\Delta^{TN} \boldsymbol{y_w} = \boldsymbol{I_w} \tag{A.17}$$

$$E_{w,j} - E_{w,i} + L_a \Delta E - F^e_{p,w,i} = \rho_{w,a}, \forall p \in \Xi_w, \forall (i,j) = a \in A_p \tag{A.18}$$

$$-M(1 - y_{w,a}) \le \rho_{w,a} \le M(1 - y_{w,a}), \forall a \in A_p \tag{A.19}$$

$$0 \le E_{w,i} \le \overline{E}, \forall i \in J \tag{A.20}$$

$$E_{w,i} - L_a \Delta E \ge -M(1 - y_{w,a}) + \underline{E}_w, \forall (i,j) = a \in A_p \tag{A.21}$$

$$E_{w,i} = E^0_w, \forall i = O_w \tag{A.22}$$

$$0 \le F^e_{p,w,i} \le \overline{F}_i, \forall p \in \Xi_w, \forall i \in J \tag{A.23}$$

$$F^e_{p,w,i}/M \le \varepsilon_{i,w} \le F^e_{p,w,i} M, \forall p \in \Xi_w, \forall i \in J \tag{A.24}$$

The objective function (A.16) minimizes the total travel cost of O-D pair w, x^*_a, x^{g*}_a and x^*_k are the optimal EV and GV flows obtained by the TAP-UE problem (A.1)-(A.15) solved in the previous iteration. $y_{w,a}$ and $\varepsilon_{i,w}$ are binary variables indicating whether path p traverses link a and whether the EV flows of O-D pair w charge at TN node i. Constraint (A.17) specifies the path origin and the destination. Constraints (A.18) and (A.19) reveal the relationship between the EV user's charging routing and demand on its remaining SoC level on each link, where M denotes a large positive number and $\rho_{w,a}$ indicates the auxiliary variable that distinguishes a link that is included in the path of a certain O-D demand or otherwise. Constraints (A.20) and (A.21) describe the maximum and minimum SoC levels of EV flows, the latter is associated with the EV users' range anxiety. Constraint (A.22) specifies the SoC level at the Origin O_w of O-D pair w, realistically capturing the differentiated battery status for EV users, as opposed to [17–19] which assume identical initial SoC for all O-D pairs or for all EV users. Constraints (A.23) and (A.24) regulate the charging demand of EV flows at each FCS-located TN node, based on the charging power capacity of relevant FCSs.

A.3. SOCP model in Section 2.4

$$\min \sum_{n \in M} \left[c^l_n P^g_n + c^q_n (P^g_n)^2 \right] + \sum_{n \in M^{im}} \lambda^{im}_n P^{im}_n \tag{A.25}$$

subject to:

$$P_{n,m} + P^g_n + P^{wg}_n - r_{n,m} I_{n,m} = \sum_{m \in M_m} P_{n,m} + P^d_n + P^{evd}_n \ : \lambda_n, \forall n, m \in M \tag{A.26}$$

$$Q_{n,m} + Q^g_n - x_{n,m} I_{n,m} = \sum_{m \in M_m} Q_{n,m} + Q^d_n, \forall n, m \in M \tag{A.27}$$

$$U_m - U_n = -2(r_{n,m} P_{n,m} + x_{n,m} Q_{n,m}) + (z_{n,m})^2 I_{n,m}, \forall n, m \in M \tag{A.28}$$

$$I_{n,m} U_n \ge (P_{n,m})^2 + (Q_{n,m})^2, \forall n, m \in M \tag{A.29}$$

$$\underline{P}^g_n \le P^g_n \le \overline{P}^g_n, \forall n \in M^{dg} \tag{A.30}$$

$$\underline{Q}^g_n \le Q^g_n \le \overline{Q}^g_n, \forall n \in M^{dg} \tag{A.31}$$

$$0 \le P^{wg}_n \le \overline{P}^{wg}_n, \forall n \in M^{wg} \tag{A.32}$$

$$P^{evd}_n = \sum_{w \in W, p \in P_w} f_{p,w} F^e_{p,w,n} \forall n \in M^{ev} \tag{A.33}$$

$$\sqrt{(P_{n,m})^2 + (Q_{n,m})^2} \le S_{n,m}, \forall n, m \in M \tag{A.34}$$

$$\underline{U}_n \le U_n \le \overline{U}_n, \forall n \in M \tag{A.35}$$

$$P_{n,m} - r_{n,m} I_{n,m} \ge 0, \forall n, m \in M \tag{A.36}$$

$$Q_{n,m} - x_{n,m} I_{n,m} \ge 0, \forall n, m \in M \tag{A.37}$$

The objective function (A.25) consists of the total generation costs of DGs and the electricity purchasing costs from the transmission grid. Constraints (A.26) and (A.27) express the nodal active and reactive power balance constraint for each PDN node, respectively. The Lagrangian multiplier associated with constraint (A.26) is denoted as λ_n, which constitutes the DLMPs. Constraint (A.28) describes the drop of voltage magnitude for each distribution line. Constraints (A.29) indicate the second-order cone relaxation. Constraints (A.30) and (A.31) detail the minimum and maximum limits for active and reactive power output of DG. Constraint (A.32) depicts the minimum and maximum wind generation output, where $\overline{P}^{wg}$ is modeled as random variables, capturing the inherent intermittency of wind generation. Constraint (A.33) expresses the equivalent EV charging demand, as described in the UE (Section 2.3). Constraints (A.34) and (A.35) represent the thermal and square voltage magnitude limits, respectively. Finally, constraints (A.36) and (A.37) prevent the reverse power flows, owing to the PDN relay protection requirements [59].

Appendix B

B.1. Data for small network example

The topology of the coupled TN and PDN is depicted in Fig. B.1. One pair of elastic O-D demand traveling from TN node 1 to 5 is considered. The O-D pair is connected through 3 separated paths, each consists of 2 end-to-end connected links. The capacity of each link is 4000 veh, the adjacent two links in a path are joined with a FCS with 7200 veh capacity for EV flows and 90 kW capacity for EV charging demand. The monetary value of unit time is 1.5$/h. The DG is located at node 2 of the PDN with cost coefficients $c^q = 0.2\$/\text{MW}^2$ and $c^l = 40\$/\text{MW}$, while the wind generators are co-located with FCSs 1 and 3. It is assumed that each FCS is operated by a separate CSP, and the pricing range is set to [110,150]$/MWh. The coefficients of the elasticity function, the initial SoC and the maximum available wind power generation are modeled as random variables. Other relevant techno-economic parameters of TAP-UE and SOCP models are organized in a supplementary datasheet [58].

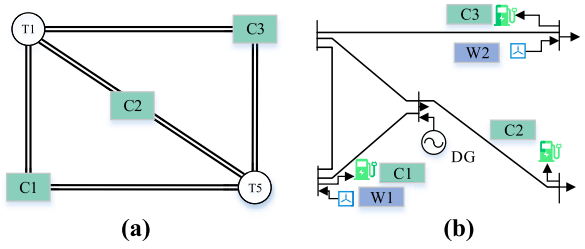

Fig. B.1. Topology of the coupled (a) 5-node TN and (b) 5-node PDN.

B.2. Data for large network example

The topology of the 20-node TN and 33-node PDN is shown in Fig. B.2. Fourteen pairs of elastic O-D travel demands are examined, their origin and destination are summarized in Table B.1. In order to capture the diversified travel needs of different users, center TN nodes T4 and T5 are assumed to be the main origins (which represents a business area located in the city center), while edge nodes T2, T7, T10, T11 and T12 are assumed to be the main destinations (which represent residential districts). Additionally, we also model the O-D demand of users traveling from the residential areas back to the city center (e.g. T1 to T9) and long-distance travel demand in the residential districts (e.g. T2 to T11). The O-D demand exhibits a linear cost elasticity function $d_w = g_w(C^{tot}_w, \epsilon_w) = -\varepsilon^s_w C^{tot}_w + \varepsilon^i_w$ and an initial SoC level E^0_w. To capture the intrinsic variability of EV users, coefficients ε^s_w, ε^i_w (which govern the extent of cost elasticity) are considered as random variables sampled from Gaussian distributions $\varepsilon^s_w \sim \mathcal{N}(3{,}000{,}900)\$/\text{veh}$ and $\varepsilon^l_w \sim \mathcal{N}(1{,}000{,}300)$ veh. Random variables E_{w,O_w} are sampled from a truncated Gaussian distribution $E_{w,O_w} \sim \mathcal{N}(9,2.5)$ MWh, the maximum SoC $\overline{E}$ is set as 20 MWh whereas the minimum SoC $\underline{E}_w$ is O-D

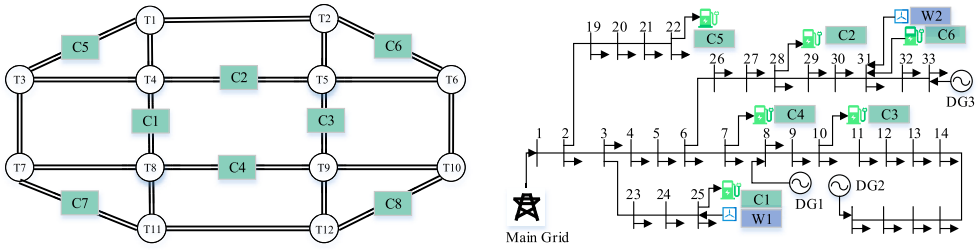

Fig. B.2. Topology of the coupled (a) 20-node TN and (b) 33-node PDN.

Table B.1
Origin, designation and minimum SoC (in MWh) of EV flows in each O-D pair for the large network example.

Origin	T4	T4	T4	T4	T4	T1	T3
Destination	T2	T7	T10	T11	T12	T12	T10
$\underline{E}_w$	3	3	3	3	3	4	4
Origin	T2	T2	T7	T6	T1	T5	T5
Destination	T11	T10	T12	T11	T9	T11	T7
$\underline{E}_w$	4	4	4	5	4	3	3

Table B.2
Distance (in km) and free travel time (in h) of EV and GV flows on each link for the large network example.

Link	L_a	t_a^0	Link	L_a	t_a^0
T1-T2	200	5	T1-T4	98	3
T2-T5	79	2	T3-T4	85	3
T5-T6	82	2	T3-T7	190	5
T6-T10	200	4	T7-T8	89	3
T9-T10	91	3	T8-T11	98	3
T9-T12	90	3	T1-F5	180	2
T3-F5	180	1	T2-F6	170	1
T6-F6	170	2	T4-F2	135	3
T5-F2	135	3	T4-F1	140	3
T8-F1	140	3	T8-F4	132	3
T9-F4	132	3	T5-F3	138	3
T9-F3	138	3	T7-F7	175	2
T11-F7	175	1	T12-F8	182	1
T10-F8	182	2	T11-T12	200	5

pair dependent and driven by EV users' range anxiety, their values are depicted in Table B.1.

The capacity of each link is 7,500 veh. The free travel time of EV flows on each link of the TN t_a^0 are summarized in Table B.2. The average charging time of EV flows at each FCS is set as $t_k^0 = 2$ h, each FCS has a 12,000 veh capacity for EV flows and a 90 kW capacity for EV charging demand. It is assumed that each FCS is operated by a separate CSP, and the pricing range is set to [140,180]$/MWh. The DGs are located at PDN nodes 8, 18 and 33 with cost coefficients $c^q = 0.3\$/MW^2$ and $c^l = 260, 145, 155\$/MW$ respectively, the price of importing electricity from transmission grid at PDN node 1 is set as $\lambda^{im} = 400\$/MWh$. The wind generators W1 and W2 are co-located with FCSs F4 and F6 at PDN nodes 25 and 31, the maximum available wind power generation is also modeled as a random variable sampled from a Weibull distribution $P^{wg} \sim \mathcal{W}(5.4, 1.64)$, which are re-scaled to 8 and 3 MW, respectively. The value of travel time is $\omega = 6\$/h$. The parameters related to the SOCP implementation of the PDN, associated with DGs, fixed demand, network topology and technical limits are organized in a supplementary datasheet [58].

References

[1] Zhang R, Hanaoka T. Deployment of electric vehicles in China to meet the carbon neutral target by 2060: provincial disparities in energy systems, co2 emissions, and cost effectiveness. Resour Conserv Recycl 2021;170:105622. https://doi.org/10.1016/j.resconrec.2021.105622.

[2] IEA. Renewables 2022. Available from: https://www.iea.org/reports/renewables-2022, 2022. [Accessed 13 September 2023].

[3] Chen S, Li Z, Li W. Integrating high share of renewable energy into power system using customer-sited energy storage. Renew Sustain Energy Rev 2021;143:110893. https://doi.org/10.1016/j.rser.2021.110893.

[4] Das H, Rahman M, Li S, Tan C. Electric vehicles standards, charging infrastructure, and impact on grid integration: a technological review. Renew Sustain Energy Rev 2020;120:109618. https://doi.org/10.1016/j.rser.2019.109618.

[5] Wang H, Ye Y, Wang Q, Tang Y, Strbac G. An efficient lp-based approach for spatial-temporal coordination of electric vehicles in electricity-transportation nexus. IEEE Trans Power Syst 2023;38:2914–25. https://doi.org/10.1109/TPWRS.2022.3189482.

[6] Hove A, Sandalow D. Electric vehicle charging in China and the United States. Available from: https://www.energypolicy.columbia.edu/publications/electric-vehicle-charging-china-and-united-states, 2019. [Accessed 13 September 2022].

[7] IEA. Global ev outlook 2022. Available from: https://www.iea.org/reports/global-ev-outlook-2022, 2023. [Accessed 13 September 2023].

[8] Geng L, Lu Z, He L, Zhang J, Li X, Guo X. Smart charging management system for electric vehicles in coupled transportation and power distribution systems. Energy 2019;189:116275. https://doi.org/10.1016/j.energy.2019.116275.

[9] Tahir Y, Khan I, Rahman S, Nadeem MF, Iqbal A, Xu Y, et al. A state-of-the-art review on topologies and control techniques of solid-state transformers for electric vehicle extreme fast charging. IET Power Electron 2021;14:1560–76. https://doi.org/10.1049/pel2.12141.

[10] Fu Q, Du W, Wang H, Xiao X. Stability analysis of dc distribution system considering stochastic state of electric vehicle charging stations. IEEE Trans Power Syst 2021;37:1893–903. https://doi.org/10.1109/TPWRS.2021.3121316.

[11] Zhang H, Hu Z, Song Y. Power and transport nexus: routing electric vehicles to promote renewable power integration. IEEE Trans Smart Grid 2020;11:3291–301. https://doi.org/10.1109/TSG.2020.2967082.

[12] Ding Z, Tan W, Lee W-J, Pan X, Gao S. Integrated operation model for autonomous mobility-on-demand fleet and battery swapping station. IEEE Trans Ind Appl 2021;57:5593–602. https://doi.org/10.1109/TIA.2021.3110938.

[13] Qian T, Shao C, Li X, Wang X, Chen Z, Shahidehpour M. Multi-agent deep reinforcement learning method for ev charging station game. IEEE Trans Power Syst 2021;37:1682–94. https://doi.org/10.1109/TPWRS.2021.3111014.

[14] Lv S, Chen S, Wei Z, Zhang H. Power–transportation coordination: toward a hybrid economic-emission dispatch model. IEEE Trans Power Syst 2022;37:3969–81. https://doi.org/10.1109/TPWRS.2021.3131306.

[15] Yuan Q, Ye Y, Tang Y, Liu Y, Strbac G. A novel deep-learning based surrogate modeling of stochastic electric vehicle traffic user equilibrium in low-carbon electricity–transportation nexus. Appl Energy 2022;315:118961. https://doi.org/10.1016/j.apenergy.2022.118961.

[16] Yuan Q, Ye Y, Tang Y, Liu X, Tian Q. Low carbon electric vehicle charging coordination in coupled transportation and power networks. IEEE Trans Ind Appl 2023;59:2162–72. https://doi.org/10.1109/TIA.2022.3230014.

[17] Qian T, Shao C, Li X, Wang X, Shahidehpour M. Enhanced coordinated operations of electric power and transportation networks via ev charging services. IEEE Trans Smart Grid 2020;11:3019–30. https://doi.org/10.1109/TSG.2020.2969650.

[18] He F, Yin Y, Lawphongpanich S. Network equilibrium models with battery electric vehicles. Transp Res, Part B, Methodol 2015;67:306–19. https://doi.org/10.1016/j.trb.2014.05.010.

[19] Wei W, Wu L, Wang J, Mei S. Network equilibrium of coupled transportation and power distribution systems. IEEE Trans Smart Grid 2018;9:6764–79. https://doi.org/10.1109/TSG.2017.2723016.

[20] Alizadeh M, Wai H-T, Chowdhury M, Goldsmith A, Scaglione A, Javidi T. Optimal pricing to manage electric vehicles in coupled power and transportation networks. IEEE Trans Control Netw Syst 2017;4:863–75. https://doi.org/10.1109/TCNS.2016.2590259.

[21] Shao C, Li K, Qian T, Wang X, Shahidehpour M. Generalized user equilibrium for coordinated operation of power-traffic networks. In: 2022 IEEE power & energy society general meeting (PESGM). IEEE; 2022. p. 1–5.
[22] Sun G, Li G, Li P, Xia S, Zhu Z, Shahidehpour M. Coordinated operation of hydrogen-integrated urban transportation and power distribution networks considering fuel cell electric vehicles. IEEE Trans Ind Appl 2021;58:2652–65. https://doi.org/10.1109/TIA.2021.3109866.
[23] Lai S, Qiu J, Tao Y, Zhao J. Pricing for electric vehicle charging stations based on the responsiveness of demand. IEEE Trans Smart Grid 2023;14:530–44. https://doi.org/10.1109/TSG.2022.3188832.
[24] Yuan W, Huang J, Zhang YJA. Competitive charging station pricing for plug-in electric vehicles. IEEE Trans Smart Grid 2017;8:627–39. https://doi.org/10.1109/TSG.2015.2504502.
[25] Zavvos E, Gerding EH, Brede M. A comprehensive game-theoretic model for electric vehicle charging station competition. IEEE Trans Intell Transp Syst 2022;23:12239–50. https://doi.org/10.1109/TITS.2021.3111765.
[26] Sohet B, Hayel Y, Beaude O, Jeandin A. Hierarchical coupled driving-and-charging model of electric vehicles, stations and grid operators. IEEE Trans Smart Grid 2021;12:5146–57. https://doi.org/10.1109/TSG.2021.3107896.
[27] Li K, Shao C, Zhang H, Wang X. Strategic pricing of electric vehicle charging service providers in coupled power-transportation networks. IEEE Trans Smart Grid 2023;14:2189–201. https://doi.org/10.1109/TSG.2022.3219109.
[28] Tan J, Wang L. Real-time charging navigation of electric vehicles to fast charging stations: a hierarchical game approach. IEEE Trans Smart Grid 2017;8:846–56. https://doi.org/10.1109/TSG.2015.2458863.
[29] Cui Y, Hu Z, Duan X. Optimal pricing of public electric vehicle charging stations considering operations of coupled transportation and power systems. IEEE Trans Smart Grid 2021;12:3278–88. https://doi.org/10.1109/TSG.2021.3053026.
[30] Zhao Z, Lee CK. Dynamic pricing for ev charging stations: a deep reinforcement learning approach. IEEE Trans Transp Electrif 2022;8:2456–68. https://doi.org/10.1109/TTE.2021.3139674.
[31] Wang S, Bi S, Zhang YA. Reinforcement learning for real-time pricing and scheduling control in ev charging stations. IEEE Trans Ind Inform 2021;17:849–59. https://doi.org/10.1109/TII.2019.2950809.
[32] Affolabi L, Shahidehpour M, Gan W, Yan M, Chen B, Pandey S, et al. Optimal transactive energy trading of electric vehicle charging stations with on-site pv generation in constrained power distribution networks. IEEE Trans Smart Grid 2022;13:1427–40. https://doi.org/10.1109/TSG.2021.3131959.
[33] Luo C, Huang Y-F, Gupta V. Stochastic dynamic pricing for ev charging stations with renewable integration and energy storage. IEEE Trans Smart Grid 2018;9:1494–505. https://doi.org/10.1109/TSG.2017.2696493.
[34] Sun Y, Zhao P, Wang L, Malik SM. Spatial and temporal modelling of coupled power and transportation systems: a comprehensive review. Energy Convers Econ 2021;2:55–66. https://doi.org/10.1049/enc2.12034.
[35] Papavasiliou A. Analysis of distribution locational marginal prices. IEEE Trans Smart Grid 2018;9:4872–82. https://doi.org/10.1109/TSG.2017.2673860.
[36] Lu Z, Shi L, Geng L, Zhang J, Li X, Guo X. Non-cooperative game pricing strategy for maximizing social welfare in electrified transportation networks. Int J Electr Power Energy Syst 2021;130:106980. https://doi.org/10.1016/j.ijepes.2021.106980.
[37] Lu Y, Liang Y, Ding Z, Wu Q, Ding T, Lee W-J. Deep reinforcement learning-based charging pricing for autonomous mobility-on-demand system. IEEE Trans Smart Grid 2021;13:1412–26. https://doi.org/10.1109/TSG.2021.3131804.
[38] Ye Y, Qiu D, Sun M, Papadaskalopoulos D, Strbac G. Deep reinforcement learning for strategic bidding in electricity markets. IEEE Trans Smart Grid 2020;11:1343–55. https://doi.org/10.1109/TSG.2019.2936142.
[39] Ye Y, Qiu D, Li J, Strbac G. Multi-period and multi-spatial equilibrium analysis in imperfect electricity markets: a novel multi-agent deep reinforcement learning approach. IEEE Access 2019;7:130515–29. https://doi.org/10.1109/ACCESS.2019.2940005.
[40] Nguyen TT, Nguyen ND, Nahavandi S. Deep reinforcement learning for multiagent systems: a review of challenges, solutions, and applications. IEEE Trans Cybern 2020;50:3826–39. https://doi.org/10.1109/TCYB.2020.2977374.
[41] Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative-competitive environments. In: Proceedings of the 31st international conference on neural information processing systems; Dec. 2017. p. 6382–93. Available from: https://dl.acm.org/doi/10.5555/3295222.3295385.
[42] Yu C, Velu A, Vinitsky E, Gao J, Wang Y, Bayen A, et al. The surprising effectiveness of PPO in cooperative multi-agent games. In: Thirty-sixth conference on neural information processing systems datasets and benchmarks track; 2022.
[43] Kuba JG, Chen R, Wen M, Wen Y, Sun F, Wang J, et al. Trust region policy optimisation in multi-agent reinforcement learning. In: International conference on learning representations; 2022.
[44] Low SH. Convex relaxation of optimal power flow—part II: exactness. IEEE Trans Control Netw Syst 2014;1:177–89. https://doi.org/10.1109/TCNS.2014.2309732.
[45] Rosen JB. Existence and uniqueness of equilibrium points for concave n-person games. Econometrica 1965:520–34. https://doi.org/10.2307/1911749.
[46] Berg K, Sandholm T. Exclusion method for finding Nash equilibrium in multiplayer games. In: Proceedings of the AAAI conference on artificial intelligence, vol. 31; 2017.
[47] Littman ML. Markov games as a framework for multi-agent reinforcement learning. In: Machine learning proceedings 1994. Elsevier; 1994. p. 157–63. Available from: https://dl.acm.org/doi/10.5555/284860.284934.
[48] Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multiagent reinforcement learning. IEEE Trans Syst Man Cybern 2008;38:156–72. https://doi.org/10.1109/TSMCC.2007.913919.
[49] Tan M. Multi-agent reinforcement learning: independent versus cooperative agents. In: International conference on machine learning; 1997. Available from: https://api.semanticscholar.org/CorpusID:260435822.
[50] Ye Y, Tang Y, Wang H, Zhang X-P, Strbac G. A scalable privacy-preserving multi-agent deep reinforcement learning approach for large-scale peer-to-peer transactive energy trading. IEEE Trans Smart Grid 2021;12:5185–200. https://doi.org/10.1109/TSG.2021.3103917.
[51] Ye Y, Papadaskalopoulos D, Yuan Q, Tang Y, Strbac G. Multi-agent deep reinforcement learning for coordinated energy trading and flexibility services provision in local electricity markets. IEEE Trans Smart Grid 2023;14:1541–54. https://doi.org/10.1109/TSG.2022.3149266.
[52] Jiang J, Lu Z. Learning attentional communication for multi-agent cooperation. In: Proc. advances in neural information processing systems (NIPS 2018); 2018. p. 7254–64.
[53] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint. Available from: https://doi.org/10.48550/arXiv.1707.06347, 2017.
[54] Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. In: Bach F, Blei D, editors. Proceedings of the 32nd international conference on machine learning. Proceedings of machine learning research, vol. 37. 2015. p. 1889–97.
[55] Gurobi Optimization. Gurobi optimizer reference manual. Available from: https://www.gurobi.com, 2023.
[56] Abadi M, Agarwal A, Barham P, et al. Tensorflow, large-scale machine learning on heterogeneous systems. Available from: https://dl.acm.org/doi/10.5555/3026877.3026899, 2015.
[57] Masini S, Bientinesi P. High-performance parallel computations using python as high-level language. In: Euro-Par 2010 parallel processing workshops. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. p. 541–8. Available from: https://dl.acm.org/doi/10.5555/2031978.2032051.
[58] Ye Y, Wang H. Dataset of resr csp pricing coordination and analysis submission. Available from: https://dx.doi.org/10.17632/25vjrnz3gp.1, 2023.
[59] Sun Y, Chen Z, Li Z, Tian W, Shahidehpour M. Ev charging schedule in coupled constrained networks of transportation and power system. IEEE Trans Smart Grid 2019;10:4706–16. https://doi.org/10.1109/TSG.2018.2864258.