Surrogate-Assisted Fictitious Play for EV Charging Station Pricing Game

Bo-Lin Zheng
South China University of Technology
Guangzhou, China
bolinz0707@qq.com

Feng-Feng Wei
South China University of Technology
Guangzhou, China
fengfeng_scut@163.com

Wei-Neng Chen
South China University of Technology
Guangzhou, China
cschenwn@scut.edu.cn

***Abstract*—Pricing competition among Electric Vehicle Charging Stations (EVCSs) under Dynamic Traffic Assignment (DTA) poses a challenging game-theoretic problem with continuous action spaces and black-box objectives. Existing methods rely on static traffic models or discretized pricing, failing to capture the coupling between pricing strategies and congestion dynamics. This paper proposes a Surrogate-Assisted Fictitious Play (SAFP) framework that integrates belief-based Fictitious Play with Multi-Agent Deep Reinforcement Learning. The learned critic network serves as a differentiable surrogate to approximate best responses, enabling gradient-based NashConv estimation for convergence monitoring. An asynchronous master-worker architecture further decouples training from simulation to mitigate DTA latency. We construct DTA-integrated testbeds by extending benchmark traffic networks with EV charging scenarios, and validate the framework's convergence and economic interpretability with three MADRL instantiations (MADDPG, MFDDPG, IDDPG).**

***Index Terms*—Multi-Agent Deep Reinforcement Learning, Fictitious Play, EV Charging Pricing, Game Theory**

I. Introduction

With the rapid adoption of electric vehicles (EVs), charging stations (EVCSs) have become profit-oriented market participants that compete by optimizing pricing schedules [1]–[5]. Pricing decisions reshape the spatio-temporal distribution of traffic flows, while the resulting congestion in turn alters charging demand and station revenues, creating a tightly coupled feedback loop that must be accurately modeled for effective operational policies.

However, most existing studies rely on BPR-based static traffic models [3]–[5], neglecting dynamics such as queue spill-back, while the majority discretize action spaces [3], [4], limiting pricing precision. Closing these gaps requires integrating continuous pricing with Dynamic Traffic Assignment (DTA), yet this raises two fundamental challenges:

(1) **Intractable Continuous Search:** Continuous pricing creates an infinite search space within a non-differentiable, black-box DTA simulator where analytical gradients are unavailable.

This work was supported by the National Key Research and Development Project No. 2023YFE0206200, the National Natural Science Foundation of China under Grant 62376097, and the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515012028. (Corresponding author:Wei-Neng Chen)

(2) **Prohibitive Evaluation Cost:** Each joint strategy evaluation requires running a full DTA simulation until user equilibrium converges, severely bottlenecking iterative learning.

Drawing on surrogate-assisted learning [6] and asynchronous parallel optimization [7], we propose Surrogate-Assisted Fictitious Play (FP) via Asynchronous MADRL (SAFP-MADRL), a framework that addresses both challenges through three mechanisms: (1) **Surrogate Modeling:** In our Actor-Critic architecture with γ=0, the Critic directly approximates the immediate revenue landscape. Inspired by [8], we reuse it as a differentiable surrogate to approximate best responses, enabling gradient-based NashConv [9] estimation as a principled convergence criterion. (2) **Asynchronous Parallelism:** A master-worker architecture decouples high-frequency training from time-consuming DTA evaluation to mitigate simulation latency. (3) **Belief-Based FP:** A Global Belief Matrix, updated via exponential moving average of historical joint prices, serves as the constructed Markov state, filtering out exploration noise and stabilizing the learning trajectory toward equilibrium.

Our contributions are threefold:

1) **Bi-Level Dynamic Game Formulation:** We integrate continuous pricing with Restricted Dynamic User Equilibrium (DUE), capturing the coupling between pricing strategies and physical traffic dynamics.
2) **SAFP-MADRL Framework:** We propose an asynchronous master-worker architecture with belief-based FP, where the surrogate critic approximates best responses and enables gradient-based NashConv estimation as a convergence criterion.
3) **Testbed Construction & Validation:** We extend benchmark traffic network datasets with EV charging scenarios, including charging stations and virtual charging links, and validate the framework's convergence and economic interpretability with three MADRL instantiations (MADDPG, MFDDPG, IDDPG).

II. Related Work

A. Equilibrium Computation and Fictitious Play

FP is a classical iterative scheme where agents update strategies based on the empirical distribution of opponents' historical actions, with proven convergence in specific game

classes [10]. However, classical FP and Best Response [11] require analytical payoff functions and become computationally prohibitive in high-dimensional continuous action spaces. Surrogate-assisted optimization [6] offers a promising alternative by replacing expensive evaluations with learned models. Our work extends FP into a deep learning framework [8], using the critic network as a differentiable surrogate to enable tractable best-response approximation.

B. Evolution from RL to MADRL

Single-agent Reinforcement Learning (RL) treats other participants as part of a stationary environment, failing to capture strategic interdependence in multi-player settings. Multi-Agent Deep Reinforcement Learning (MADRL) [12] addresses this limitation by jointly modeling interacting agents. Representative paradigms include independent DDPG (IDDPG), MADDPG [13] and Mean-Field DDPG (MFDDPG) [14].

III. Problem Formulation

We consider competition between EVCSs in a transportation network modeled as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over a discretized time horizon $\mathcal{T} = \{1, \dots, T\}$. The problem is formulated as a bi-level game-theoretic framework, as illustrated in Fig. 1. The upper level represents the pricing competition, while the lower level captures the Dynamic User Equilibrium (DUE) of travelers. The two levels are implicitly coupled through charging demand, which is evaluated via traffic simulation.

A. Upper-Level: EVCS Pricing Game

The competition among EVCSs is modeled as a static non-cooperative game with complete information, denoted by $\mathbb{G} = \langle \mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{R_i\}_{i\in\mathcal{N}} \rangle$, and defined as follows:

- **Players:** $\mathcal{N} = \{1, \dots, N\}$ denotes the set of EVCS operators acting as independent agents.
- **Strategies:** Each EVCS $i \in \mathcal{N}$ selects a pricing schedule $\mathbf{p}_i = [p_{i,1}, \dots, p_{i,T}]^\top$, where $p_{i,t}$ represents the charging price at time slot t. The feasible strategy space is given by $\mathcal{A}_i = [p_{\min}, p_{\max}]^T$.
- **Payoffs:** Given a joint pricing profile $\mathbf{p} = (\mathbf{p}_i, \mathbf{p}_{-i})$, the revenue of EVCS i is defined as:

$$R_i(\mathbf{p}_i, \mathbf{p}_{-i}) = \sum_{t=1}^{T} p_{i,t} \cdot q_{i,t}\big(\mathbf{f}(\mathbf{p})\big), \tag{1}$$

where $q_{i,t}$ denotes the charging demand at station i and time t, and $\mathbf{f}(\mathbf{p})$ is the equilibrium traffic flow induced by the pricing profile.

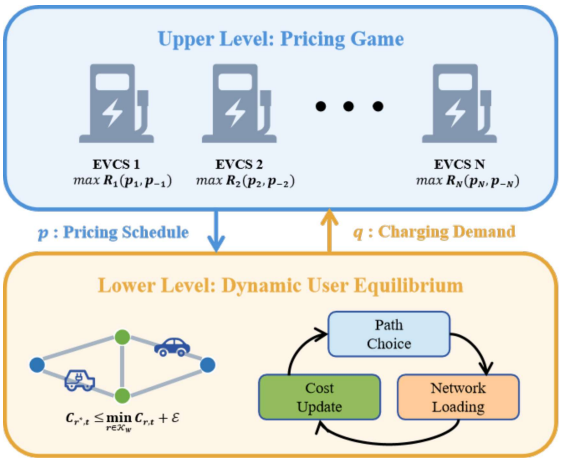

Fig. 1: The proposed bi-level game theoretic framework.

Since each EVCS selects its entire pricing schedule as a single decision variable without inter-temporal state transitions, the pricing competition is treated as a static game.

The solution concept is the Nash Equilibrium (NE) [15]. A joint strategy profile $\mathbf{p}^* = (\mathbf{p}_1^*, \dots, \mathbf{p}_N^*)$ is a NE if no EVCS can increase its revenue by unilateral deviation:

$$R_i(\mathbf{p}_i^*, \mathbf{p}_{-i}^*) \geq R_i(\mathbf{p}_i, \mathbf{p}_{-i}^*), \quad \forall \mathbf{p}_i \in \mathcal{A}_i, \ \forall i \in \mathcal{N}. \tag{2}$$

B. Lower-Level: Restricted Dynamic User Equilibrium

At the lower level, travelers respond to the charging prices by choosing routes to minimize their generalized travel costs. We consider a heterogeneous traffic flow consisting of charging vehicles and non-charging vehicles, defined over a set of origin-destination (OD) pairs $\mathcal{W}$. To ensure computational tractability, route choices are restricted to a precomputed set of candidate paths $\mathcal{K}_w$ for each OD pair $w \in \mathcal{W}$ (e.g., k-shortest paths). Specially, for charging vehicles, $\mathcal{K}_w$ is constructed to strictly include paths passing through at least one EVCS.

For a path $r \in \mathcal{K}_w$ with departure time t, the generalized travel cost is defined as:

$$C_{r,t}(\mathbf{f}, \mathbf{p}) = \alpha \cdot \tau_{r,t}(\mathbf{f}) + \delta_r \cdot \big(p_{s(r),\varphi(r,t)} \cdot D\big), \tag{3}$$

where $\tau_{r,t}(\mathbf{f})$ denotes the flow-dependent travel time, α is the value-of-time coefficient, $\delta_r \in \{0, 1\}$ indicates whether charging is required along path r, $s(r)$ denotes the charging station associated with the path, $\varphi(r,t)$ is the arrival time at the station, and D represents the charging demand of each car.

We adopt an ε-approximate Wardrop equilibrium [16] over the restricted path set. A traffic flow $\mathbf{f}$ is said to satisfy the restricted DUE condition if, for any OD pair w and departure time t, all utilized paths $r^* \in \mathcal{K}_w$ satisfy:

$$C_{r^*,t} \leq \min_{r\in\mathcal{K}_w} C_{r,t} + \varepsilon, \tag{4}$$

where ε is a small tolerance parameter.

The equilibrium flow $\mathbf{f}(\mathbf{p})$ thus defines an implicit mapping from pricing strategies to traffic distribution, which lacks a closed-form expression and is evaluated numerically via DTA simulation. Since the revenue function lacks analytical gradients, we adopt a MADRL-based framework for equilibrium approximation, detailed in the next section.

IV. METHODOLOGY

We propose a unified framework that reformulates the static EVCSs pricing game into a repeated Markov game based on *FP* [8], and integrates an asynchronous evaluation architecture to decouple training from the costly DUE simulation, enabling agents to iteratively approximate a NE.

A. Simulation-Based Environment: Restricted DUE

We define the simulation environment as a mapping $\Psi : \mathcal{A} \to \mathbb{R}^N$ that interfaces with the mesoscopic traffic simulator **UXsim** [17]. Given the pricing profile $\mathbf{p}$, the environment simulates the day-to-day evolution of the route choices of travelers until a ε-approximate restricted DUE is reached.

1) State-Augmented Graph for Charging Constraints: To rigorously enforce the constraint that every charging EV must visit exactly one charging station, we construct a state-augmented graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$. The original node set $\mathcal{V}$ is duplicated into two states: uncharged ($v^{uc} \in \mathcal{V}$) and charged ($v^c \in \mathcal{V}^c$).

- **State Transitions:** Each EVCS at node s is modeled as a unidirectional virtual link connecting s^{uc} to s^c. Adapted from [18], we take advantage of the traffic dynamics of the link to physically simulate charging behavior: the **free-flow travel time** on this link represents the required charging service time, while **flow-dependent congestion** naturally captures the queuing delay at the station.
- **Restricted Routing:** Since there are no edges from $\mathcal{V}^c$ to $\mathcal{V}^{uc}$, any path from an origin o^{uc} to a destination d^c strictly ensures exactly one charging event.
- **Candidate Generation:** We precompute a set of candidate paths $\mathcal{K}_w$ using the K-shortest paths algorithm on $\mathcal{G}'$ for charging vehicles.

2) Mesoscopic Day-to-Day Evolution: Unlike traditional macroscopic assignment methods, we adopt a evolutionary mechanism rooted in individual vehicle behavior, following the stochastic process frameworks discussed in [19], [20].

Let $r_v^{(n)}$ be the path chosen by vehicle v at iteration n, and $C_v^{(n)}$ be its generalized travel cost. At the end of each iteration, each vehicle evaluates potential alternative paths. Let C_v^* be the estimated cost of the best alternative path. The vehicle computes a relative cost gap $\delta_v^{(n)} = (C_v^{(n)} - C_v^*)/C_v^{(n)}$. To model the bounded rationality, we refine the switching probability as:

$$P_{\text{switch}}^{(n)} = \min(1, \lambda \cdot \delta_v^{(n)}) \cdot \frac{1}{1 + \mu \cdot n} \tag{5}$$

where λ controls the cost sensitivity, and the decay term $\frac{1}{1+\mu \cdot n}$ ensures convergence.

This process repeats until the average relative gap falls below ε, yielding an approximate equilibrium that serves as a consistent behavioral response for evaluating pricing strategies. The mapping Ψ then returns the realized revenues based on the equilibrium flows $\mathbf{f}$.

B. Belief-Based Markov Game Formulation

Since $\mathbb{G}$ is a static game with no inherent state transitions, we reformulate it as a **constructed repeated game** inspired by *NFSP* [21], where NE is approximated by iteratively optimizing against opponents' historical average behavior. We define the learning process as a Markov Game $\langle \mathcal{N}, \mathcal{O}, \mathcal{A}, \mathcal{R}, \gamma \rangle$, where step k denotes the learning iteration:

- **Agents ($\mathcal{N}$):** The set of N EVCS operators acting as independent learning agents.
- **Observation ($\mathcal{O}$):** Each agent observes a shared **Global Belief Matrix** $\mathcal{B}_k \in [p_{\min}, p_{\max}]^{N \times T}$, where row $\mathcal{B}_k[j]$ is the Exponential Moving Average (EMA) of agent j's pricing history. All agents share the same observation: $o_{i,k} = \mathcal{B}_k$.
- **Action ($\mathcal{A}$):** Agent i's action is its pricing schedule $\mathbf{a}_{i,k} = \mathbf{p}_i \in [p_{\min}, p_{\max}]^T$, with $\mathcal{A} = \prod_{i \in \mathcal{N}} \mathcal{A}_i$. We denote the **Joint Action Matrix** as $\mathbf{A}_k \in \mathbb{R}^{N \times T}$.
- **Belief Transition:** Upon executing $\mathbf{A}_k$, the belief matrix updates via EMA:

$$\mathcal{B}_{k+1} = (1 - \rho) \cdot \mathcal{B}_k + \rho \cdot \mathbf{A}_k \tag{6}$$

  where $\rho \in (0, 1)$ is the smoothing factor that filters exploration noise.
- **Reward ($\mathcal{R}$):** Consistent with Eq. (1): $r_{i,k} = R_i(\mathbf{A}_k) = [\Psi(\mathbf{A}_k)]_i$.
- **Discount (γ):** $\gamma = 0$, as the underlying game is one-shot and agents maximize immediate payoff.

C. Unified MADRL Framework

The simulation-based environment presents two fundamental challenges: the non-differentiability of the black-box reward function and the computational expense of the DUE simulator. To address these, we propose a unified Actor-Critic framework instantiated via Deep Deterministic Policy Gradient (DDPG) [22]. Crucially, this framework serves two purposes: it optimizes the pricing strategy and, as detailed in the subsequent section, constructs a differentiable surrogate model for convergence verification.

1) Surrogate-Assisted Optimization: Each EVCS agent i maintains an actor μ_{θ_i} and a critic Q_{ϕ_i}. With $\gamma = 0$, the critic acts as a surrogate that approximates the immediate revenue landscape.

Critic Update (Surrogate Learning): The critic minimizes the regression loss between the predicted value and the realized revenue r_i obtained from the DUE simulator:

$$\mathcal{L}(\phi_i) = \mathbb{E}_{\mathcal{D}} \left[\left(r_i - Q_{\phi_i}(\Omega_i, \mathbf{a}_i) \right)^2 \right] \tag{7}$$

where $\mathcal{D}$ denotes the replay buffer and Ω_i is the information structure. By minimizing this loss, Q_{ϕ_i} evolves into a differentiable approximation of the true black-box objective, i.e., $Q_{\phi_i} \approx R_i$.

Actor Update (Policy Gradient): The actor is updated by ascending the deterministic policy gradient flow provided by the learned critic:

$$\nabla_{\theta_i} J \approx \mathbb{E}_{\mathcal{D}} \left[\nabla_{\mathbf{a}_i} Q_{\phi_i}(\Omega_i, \mathbf{a}_i) \big|_{\mathbf{a_i} = \mu_{\theta_i}(o_i)} \right] \tag{8}$$

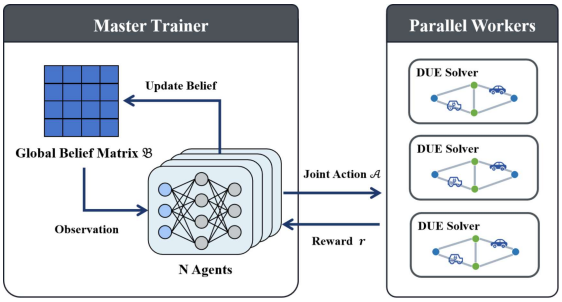

Fig. 2: The asynchronous evaluation architecture of SAFP-MADRL.

2) Scalable Instantiations via Information Structures: We instantiate the framework into three variants based on the input structure Ω_i of the critic, trading off between information completeness and scalability:

- **MADDPG (Full Information):** $\Omega_i = \{o_i, \mathbf{a}_1, \ldots, \mathbf{a}_N\}$. Captures exact pairwise couplings but scales linearly with N.
- **MFDDPG (Mean-Field):** $\Omega_i = \{o_i, \mathbf{a}_i, \bar{\mathbf{a}}_{-i}\}$. Uses the average action of opponents to decouple input dimension from population size [3].
- **IDDPG (Local Information):** $\Omega_i = \{o_i, \mathbf{a}_i\}$. Ignores explicit coordination, treating opponents as part of the environment dynamics.

D. Exploitability-Based Convergence Criterion

In general-sum games, standard RL metrics such as cumulative rewards are often misleading due to the cyclic dynamics of multi-agent interactions. To rigorously assess the quality of the learned strategies in our black-box environment, we operationalize **NashConv** [9] as an algorithm-dependent metric, employing the learned Critic networks Q_{ϕ_i} (defined in Sec. IV-C) as implicit payoff models.

Formally, we evaluate the stability of the current joint action profile $\mathbf{a} = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ produced by the learned policies. The estimated regret for agent i is the potential gain from unilateral deviation:

$$\text{Regret}_i \approx \max_{\mathbf{a}'_i} Q_{\phi_i}(\Omega_i, \mathbf{a}'_i, \mathbf{a}_{-i}) - Q_{\phi_i}(\Omega_i, \mathbf{a}_i, \mathbf{a}_{-i}) \quad (9)$$

and NashConv is defined as:

$$\text{NashConv} = \sum_{i=1}^{N} \text{Regret}_i \quad (10)$$

To approximate the best response (the max operator) and mitigate local optima, we employ a **multi-start optimization strategy**: K candidate actions (comprising the current action and $K-1$ random samples) are initialized and optimized via Adam for M steps with respect to the Critic's input gradients. The training terminates only when the normalized exploitability, defined as NashConv$/N$, falls below a strict threshold ε.

Algorithm 1 SAFP-MADRL

Input: N agents, M workers, $\gamma = 0$, ρ, intervals T_l, T_c, T_w, threshold ε.
Output: Optimized Pricing Schedule $\mathbf{a}^*$.

Init: $\mu_\theta, Q_\phi, \mu'_\theta, Q'_\phi$; $\mathcal{D} \leftarrow \emptyset$; $\mathcal{B} \leftarrow 0.5$; $t \leftarrow 0$.
Dispatch M initial tasks to workers.
while not converged **and** $t < t_{\max}$ **do**
  Wait for any worker result $\mathbf{r}$.
  Store $(\mathcal{B}, \mathbf{a}, \mathbf{r}, \mathcal{B}')$ in $\mathcal{D}$; $t \leftarrow t+1$.
  if $t \mod T_l = 0$ **then**
    Sample batch from $\mathcal{D}$.
    Update Q_ϕ (Eq. 7), μ_θ (Eq. 8).
    Soft-update: $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$.
  end if
  $\mathbf{a} \leftarrow \mu_\theta(\mathcal{B})$; $\mathcal{B}' \leftarrow (1-\rho)\mathcal{B} + \rho\mathbf{a}$.
  Dispatch $\langle \mathcal{B}, \mathbf{a}, \mathcal{B}' \rangle$ to freed worker.
  if $t \geq T_w$ and $t \mod T_c = 0$ **then**
    if NashConv$(/N < \varepsilon$ **then**
      Break
    end if
  end if
end while
return $\mathbf{a}^*$ derived from $\mathcal{B}$.

E. Asynchronous Implementation Protocol

To resolve the computational bottleneck caused by the time-consuming DUE simulation, we propose **Surrogate-Assisted FP via Asynchronous MADRL (SAFP-MADRL)**. This framework implements a **Decoupled Master-Worker Architecture**, illustrated in Fig. 2, which separates the high-frequency learning loop from the evaluation oracle.

Crucially, to preserve the semantic validity of the belief-based game against asynchronous latency, we enforce a **Belief Snapshot Mechanism**. When the Master dispatches a task, it records the current belief $\mathcal{B}$ and the post-interaction belief $\mathcal{B}'$ alongside the action $\mathbf{a}$. This ensures that the experience tuple $\langle \mathcal{B}, \mathbf{a}, \mathbf{r}, \mathcal{B}' \rangle$ stored in the buffer remains mathematically consistent with the FP formalism, regardless of simulation delays. The complete training procedure is summarized in Algorithm 1.

TABLE I: Statistics of Experimental Networks

Dataset	Nodes	Links	N	OD Pairs	Vehicles
Sioux Falls	24	76	4	528	33,005
Berlin-Friedrichshain	191	451	20	506	22,285
Anaheim	416	914	40	1,406	47,550

TABLE II: Hyperparameters and Network Configuration

Category	Parameter	Value
Network	Actor / Critic Hidden	(64,64)/(128,64)
	Optimizer / Activation	Adam / ReLU
Training	Learning Rate	10^{-3}
	γ / τ	0.0 / 0.01
	Buffer / Batch Size	10,000 / 64
	Noise σ_0 / Decay	0.2 / 0.995
NashConv	Check Interval	10 steps
	Starts (K) / Steps	5 / 50
	Threshold (ε) / Max Eval.	0.01 / 2000

TABLE III: Convergence Statistics across Datasets

Algorithm	**Sioux Falls** ($N = 4$)		**Berlin-Friedrichshain** ($N = 20$)		**Anaheim** ($N = 40$)	
	Min Exploit.	Eval. Step	Min Exploit.	Eval. Step	Min Exploit.	Eval. Step
MADDPG	**0.0072**	1279	**0.0080**	379	**0.0030**	729
IDDPG	0.0209	609	**0.0093**	549	**0.0092**	699
MFDDPG	0.0104	1789	0.0364	979	0.0336	1149

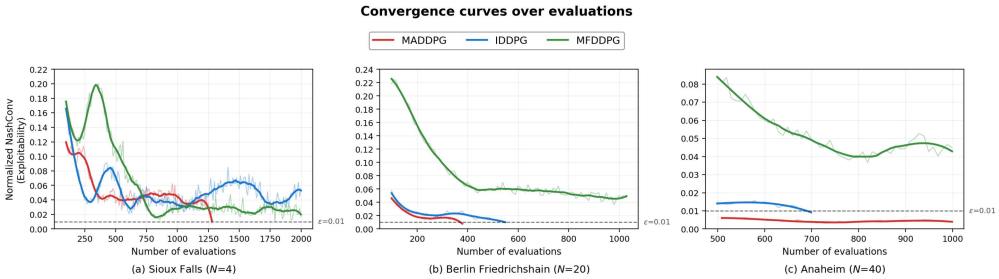

Fig. 3: Convergence curves of normalized NashConv (Exploitability) over evaluations.

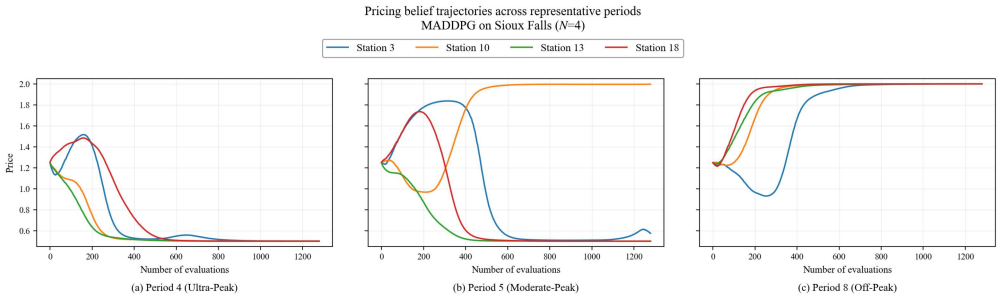

Fig. 4: Pricing belief trajectories of the four EVCS agents in Sioux Falls.

V. Experiment and Result Analysis

A. Experimental Setup

1) Simulation Environment and Datasets: The experiments are conducted on benchmark transportation networks adapted from [23]. To model competition among EVCSs, each network is augmented by designating a subset of nodes as charging nodes with virtual charging links.

We consider three networks of increasing scale and structural complexity: Sioux Falls (Small), Berlin-Friedrichshain (Medium), and Anaheim (Large). Key statistics of the three networks are summarized in Tab. I.

We set the penetration rate of EVs requiring charing to 10%. The pricing horizon is divided into $T = 8$. UXsim [17] is adopted as lower-level traffic simulator. The DUE convergence threshold (average relative gap) is set to $\varepsilon = 0.01$ for Sioux Falls, and adjusted to 0.03 for Berlin-Friedrichshain and Anaheim to balance computational efficiency.

2) Baselines & Hyperparameters: We evaluate the three framework instantiations (MADDPG, MFDDPG, IDDPG) defined in Section IV-C. All variants share the same SAFP architecture and neural network configuration (Table II), with Ornstein-Uhlenbeck exploration noise (decay 0.995) and NashConv evaluation every 10 iterations ($K = 5$ multi-start).

B. Convergence and Scalability Analysis

We evaluate the convergence performance of the three framework variants by monitoring the exploitability. The convergence curves are visualized in Fig. 3, and the quantitative statistics are detailed in Table III.

1) Impact of Information Structure: As observed in Fig. 3, **MADDPG (Full Information)** demonstrates superior stability and solution quality across all network scales. It consistently achieves the lowest exploitability scores (0.0072, 0.0080, and 0.0030 respectively) and is the only variant to strictly satisfy the convergence condition NashConv$/N < 0.01$ in the Sioux Falls network. This suggests that explicitly modeling the joint action space effectively mitigates the non-stationarity, enabling agents to converge to a strategy profile with significantly lower estimated exploitability.

2) Scalability and the "Independence" Phenomenon: A counter-intuitive finding is the performance of **IDDPG (Independent Learning)**. In the small-scale Sioux Falls network ($N = 4$), IDDPG exhibits oscillatory behavior and fails to reach the strict threshold (Min Exploitability ≈ 0.02). However, as the network scale increases to Berlin ($N = 20$) and Anaheim ($N = 40$), IDDPG successfully converges (≈ 0.009). This aligns with the theoretical insight that in large-population games, the influence of any single opponent

diminishes, allowing the aggregated behavior to be modeled as stationary environmental noise.

3) Limitations of Mean-Field Approximation: **MFDDPG** performs competitively in the small network but fails to converge in larger scenarios, plateauing at an exploitability of approx. 0.03. We attribute this to the *spatial heterogeneity* inherent in transportation networks: a charging station is primarily affected by competitors in its immediate topological vicinity rather than the global average. Consequently, as the network scale grows, the global mean action becomes a diluted signal that fails to capture local competitive dynamics.

C. Economic Interpretability of Learned Strategies

To validate the economic rationality, we visualize the **Belief Prices** evolution in Fig. 4. The results reveal strategies aligning with **Bertrand Competition** under varying demand elasticities across three representative periods:

- **High-Volume Price War (Period 4, Ultra-Peak):** As shown in Fig. 4(a), high traffic volume increases demand sensitivity. Agents adopt a **Volume Strategy**, undercutting prices to $p_{\min} = 0.5$ to capture the massive flow. The equilibrium collapses to the marginal floor, reflecting a fierce competition for utilization rather than per-unit margin.
- **Spatial Differentiation (Period 5, Moderate-Peak):** In Fig. 4(b), peripheral stations drop prices while Station 10 maintains $p_{\max}$, leveraging its strategic location on a critical traffic artery to sustain dominant market power.
- **High-Margin Monopoly (Period 8, Off-Peak):** In Fig. 4(c), free-flow conditions allow drivers to strictly prioritize proximity, creating rigid **Spatial Segmentation**. With inelastic demand from "captive" local users, agents switch to a **Margin Strategy**, converging to $p_{\max} = 2.0$ to harvest maximum revenue from their local monopolies.

VI. Conclusion

This paper addresses the EVCSs pricing game with continuous strategies under DTA environments. We introduced SAFP-MADRL, combining a differentiable surrogate critic for gradient-based updates with an asynchronous architecture to overcome DTA simulation latency.

Experiments validate stable NE convergence. MADDPG provides superior stability, while IDDPG becomes effective in larger networks due to the "independence phenomenon." MFDDPG struggles with spatial heterogeneity in traffic networks. The learned strategies exhibit economically interpretable behaviors, shifting between price wars and margin maximization based on traffic conditions.

Future work includes: (1) topology-aware mechanisms (e.g., GCNs) to capture local competitive dynamics; and (2) extending to coupled transportation-power systems.

References

[1] J. Tan and L. Wang, "Real-time charging navigation of electric vehicles to fast charging stations: A hierarchical game approach," *IEEE transactions on smart grid*, vol. 8, no. 2, pp. 846–856, 2015.

[2] W. Lee, R. Schober, and V. W. Wong, "An analysis of price competition in heterogeneous electric vehicle charging stations," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3990–4002, 2018.

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

[4] X. Yang, T. Cui, H. Wang, and Y. Ye, "Multiagent deep reinforcement learning for electric vehicle fast charging station pricing game in electricity-transportation nexus," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 4, pp. 6345–6355, 2024.

[5] Y. Wang, Q. Wu, Z. Li, S. Tao, S. Xie, X. Zhang, and W. K. V. Chan, "Federated multi-agent deep reinforcement learning-based competitive pricing strategy for charging station operators," *IEEE Transactions on Energy Markets, Policy and Regulation*, 2025.

[6] X.-Q. Guo, F.-F. Wei, J. Zhang, and W.-N. Chen, "A classifier-ensemble-based surrogate-assisted evolutionary algorithm for distributed data-driven optimization," *IEEE Transactions on Evolutionary Computation*, 2024.

[7] F.-F. Wei, W.-N. Chen, and J. Zhang, "Aiea: An asynchronous influence-based evolutionary algorithm for expensive many-objective optimization," *IEEE Transactions on Cybernetics*, 2024.

[8] N. Kamra, U. Gupta, K. Wang, F. Fang, Y. Liu, and M. Tambe, "Deep fictitious play for games with continuous action spaces," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2042–2044, 2019.

[9] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel, "A unified game-theoretic approach to multiagent reinforcement learning," *Advances in neural information processing systems*, vol. 30, 2017.

[10] G. W. Brown, "Iterative solution of games by fictitious play," *Act. Anal. Prod Allocation*, vol. 13, no. 1, p. 374, 1951.

[11] A. Matsui, "Best response dynamics and socially stable strategies," *Journal of Economic Theory*, vol. 57, no. 2, pp. 343–362, 1992.

[12] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE transactions on cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[13] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International conference on machine learning*, pp. 5571–5580, PMLR, 2018.

[15] J. Nash, "Non-cooperative games," *Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951.

[16] J. G. Wardrop, "Road paper. some theoretical aspects of road traffic research.," *Proceedings of the institution of civil engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[17] T. Seo, "Uxsim: lightweight mesoscopic traffic flow simulator in pure python," *Journal of Open Source Software*, vol. 10, no. 106, p. 7617, 2025.

[18] L. Graf, T. Harks, and P. Palkar, "Dynamic traffic assignment for electric vehicles," *Transportation Research Part B: Methodological*, vol. 195, p. 103207, 2025.

[19] M. Ishihara and T. Iryo, "Dynamic traffic assignment by markov chain," *Journal of Japan Society of Civil Engineers, Ser. D3 Infrastructure Planning and Management*, vol. 71, no. 5, pp. I503–I509, 2015. in Japanese.

[20] T. Iryo, D. Watling, and M. Hazelton, "Estimating markov chain mixing times: convergence rate towards equilibrium of a stochastic process traffic assignment model," *Transportation science*, vol. 58, no. 6, pp. 1168–1192, 2024.

[21] J. Heinrich and D. Silver, "Deep reinforcement learning from self-play in imperfect-information games," *arXiv preprint arXiv:1603.01121*, 2016.

[22] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[23] Transportation Networks for Research Core Team, "Transportation networks for research." https://github.com/bstabler/TransportationNetworks. Accessed: July 7, 2025.