Reinforcement Learning for Optimized EV Charging Through Power Setpoint Tracking

Yilmaz, Yunus Emre; Orfanoudakis, Stavros; Vergara, Pedro P.

DOI
10.1109/ISGTEUROPE62998.2024.10863457

Publication date
2024

Document Version
Final published version

Published in
IEEE PES Innovative Smart Grid Technologies Europe, ISGT EUROPE 2024

Reinforcement Learning for Optimized EV Charging Through Power Setpoint Tracking

Yunus Emre Yılmaz[1], Stavros Orfanoudakis[1], Pedro P. Vergara[1]
[1]Intelligent Electrical Power Grids, Delft University of Technology, The Netherlands
emails: yemreyilmaz8@gmail.com,{S.Orfanoudakis, P.P.VergaraBarrios}@tudelft.nl

2024 IEEE PES Innovative Smart Grid Technologies Europe (ISGT EUROPE) | 979-8-3503-9042-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISGTEUROPE62998.2024.10863457

***Abstract*—Decarbonizing the transportation sector involves adopting electric vehicles (EVs); a shift that introduces significant challenges in energy distribution management and raises concerns about grid stability. Charge Point Operators (CPOs) are important in this transition as they control the EV charging process by balancing the needs of EV users and the grid. This study presents a smart-charging model from the perspective of CPOs for handling EVs located in a commercial parking lot to minimize the Power Setpoint Tracking (PST) error. To solve this sequential decision-making problem, a Markov Decision Process (MDP) model is designed and solved using Deep Deterministic Policy Gradient (DDPG), a Deep Reinforcement Learning (DRL) algorithm. The proposed model can effectively manage the uncertainties associated with EV arrivals and fluctuating charging demands by structuring the action and state space to incorporate power constraints. The experimental evaluation using realistic EV behavior data shows that the proposed approach significantly outperforms uncontrolled charging, reducing PST error while effectively managing multiple EV chargers and EVs with varying battery capacities and power limitations.**

***Index Terms*—EV optimization, power setpoint tracking, deep reinforcement learning (DRL), charge point operators, DDPG.**

I. Introduction

As environmental concerns increase, the electric vehicle (EV) market is experiencing rapid growth, with EVs rising from under 5% of global sales in 2020 to 14% in 2022 [1]. This increase in EV adoption presents both challenges and opportunities for power grids, especially those reliant on renewable energy sources (RES). Technologies such as smart charging and Vehicle-to-Grid (V2G) [2] are vital, improving grid stability and providing economic advantages to EV users and Charge Point Operators (CPOs) [3]. However, the EV charging problem contains various uncertainties, including the timing of EV arrivals and fluctuating electricity prices, necessitating the implementation of fast and accurate optimization techniques for scheduling the charging of EVs in real time [4].

Various studies have shown that optimizing the charging schedule of EVs can yield promising results in terms of profitability and grid stability. For instance, a Mixed Integer Linear Programming (MILP) based approach at a public charging station incorporating a battery energy storage system (BESS) and forecasts from PV generation and EV arrival times resulted in an 82.8% increase in daily profits [5]. Similarly, another MILP formulation was applied to a community microgrid featuring PV and a BESS; the optimization resulted in a 33.4% reduction in operational costs by optimizing EV charging and discharging schedules [6]. However, mathematical optimization often struggles to handle complex larger-scale problems with uncertainties, especially in real-time settings.

Reinforcement Learning (RL) can handle uncertainties in sequential decision-making problems by repeatedly interacting with the problem's environment and adjusting to real-time state conditions [7]. In [8], a Deep Reinforcement Learning (DRL) algorithm was utilized to maximize EV owner profits by exploiting V2G capabilities at a residential EV charger. However, the use of discrete charging levels limited the representation of the EV charging problem. Similarly, [9] enhanced profit optimization for EV aggregators using a combination of Deep Deterministic Policy Gradient (DDPG) and prioritized experience replay, which excelled across many residential EV chargers. [10] proposed a recurrent DDPG approach that scaled efficiently without retraining the RL model, showing great scalability potential. In contrast, [11] focused on achieving targeted load schedules using a fitted Q-learning approach, optimizing charging stations in a centralized manner.

This paper introduces a Power Setpoint Tracking (PST) approach, solved using RL, to address the need for smart charging in large-scale infrastructures while ensuring real-time computation. In detail, PST is a dynamic smart charging method that empowers EV aggregators and CPOs to manage energy distribution in a controlled manner, unlike uncontrolled charging. For example, PST allows CPOs to effectively coordinate their EV fleets to meet power demands specified by market contracts or agreements with Distribution System Operators (DSOs). To evaluate the effectiveness of this approach, we conducted experiments using real-world data on EV behavior and specifications, providing insights into the practicality and performance of the system.

II. Power Setpoint Tracking

PST is a charging approach usually adopted by CPOs to match the contracted energy with consumption by scheduling EV charging (Fig. 1). Initially, CPOs purchase energy in the day-ahead market to supply their chargers and allocate it to EVs the following day. Based on this contracted energy, CPOs set power limits for specific times of the next day and schedule EV charging accordingly. It is crucial that this scheduling accurately matches the power setpoints to avoid the need for costly additional energy purchases in the intraday and balancing markets or risking unsatisfied customers with uncharged EVs. This paper focuses on how energy is distributed once a power setpoint is determined rather than on energy trading.

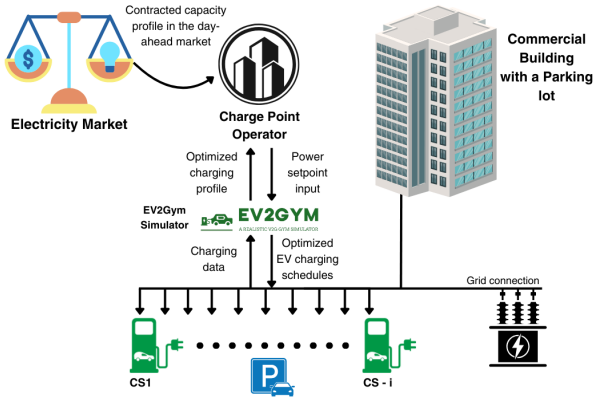

Fig. 1. Usual operations of CPOs controlling public and private parking lots.

A. EV2Gym Simulation Environment

Developing and evaluating different charging strategies for the formulated PST problem requires a realistic digital simulation environment. For the purposes of this paper, a realistic V2G simulator named EV2Gym [12] was utilized to run the simulations. EV2Gym is a flexible simulator that tests algorithms in EV smart charging and V2G scenarios. EV2Gym is a Gym environment [13]; therefore, it accelerates the development of RL algorithms while it is offering customizable settings for chargers, transformers, and EV specifications to fit various scenarios. The simulator utilizes open-source data, integrates custom data such as electricity prices or EV behavior, and supports saving replays to compare different algorithms or against optimal solutions.

B. Mathematical Model

A mathematical model of the PST problem is formulated using a Mixed Integer Nonlinear Programming (MINLP) formulation. The simulation consists of T discrete time steps t. Additionally, i represents each charging station for EVs to connect and is part of the set of charging stations C. The set of charging stations C are connected to a transformer to introduce transformer power limits ($\underline{P}^{\text{tr}}$, $\overline{P}^{\text{tr}}$) to the simulation. Furthermore, there is a set of EVs indicated by H, where each EV $j \in H$. Lastly, the binary variable u is introduced to show if an EV is connected to a charging station i at time step t.

The objective function is the total squared difference between P_t^{set} and P_t^{tot} in (1), where P_t^{set} is the contracted power and P_t^{tot} the actually consumed power (3) used at timestep t. The P_t^{set} are determined for each step with a 5% surplus from the total charging demand of a day and distributed to the simulation's steps by utilizing electricity prices as negative weights. Furthermore, the algorithm controls the charging current $I_{i,t}^{\text{ch}}$ $\forall t \in T$ and $i \in C$. At each time step t and charging station i, power is determined by the product of the controlling current $I_{i,t}^{\text{ch}}$, voltage V, phases ϕ, charging efficiency η, and a binary variable $u_{i,t}$ which indicates if an EV is connected, as described in (2). The total charging power P_t^{tot} at each step is the sum of the power from all stations as shown in (3), complying with the transformer's power limits in (4) and each station's current constraints (5). Moreover, the State of Charge (SoC) within each EV changes based on charging power and time interval Δt according to (6). Each EV's energy level at arrival in (9) and connection status at any time are also detailed and known in (10), ensuring compliance with respective EV model constraints on battery capacity and charging power, while $j \in H$, as outlined in (7) and in (8).

$$\min_{I_{i,t}^{\text{ch}}} \sum_{t \in T} \left(P_t^{\text{set}} - P_t^{\text{tot}}\right)^2 \tag{1}$$

Subject to:

$$P_{i,t}^{\text{ch}} = I_{i,t}^{\text{ch}} \cdot V \cdot \sqrt{\phi} \cdot \eta \cdot u_{i,t} \quad \forall i, \forall t \tag{2}$$

$$P_t^{\text{tot}} = \sum_{i \in C} P_{i,t}^{\text{ch}} \quad \forall i, \forall t \tag{3}$$

$$\underline{P}_t^{\text{tr}} \leq P_t^{\text{tot}} \leq \overline{P}_t^{\text{tr}} \quad \forall t \tag{4}$$

$$\underline{I}_i^{\text{ch}} \leq I_{i,t}^{\text{ch}} \leq \overline{I}_i^{\text{ch}} \quad \forall i, \forall t \tag{5}$$

$$E_{i,j,t} = E_{i,j,t-1} + P_{i,t}^{\text{ch}} \cdot \Delta t \quad \forall i, \forall j, \forall t \tag{6}$$

$$\underline{E}_{i,j} \leq E_{i,j,t} \leq \overline{E}_{i,j} \quad \forall i, \forall j, \forall t \tag{7}$$

$$\underline{P}_j^{\text{ch}} \leq P_{i,j,t}^{\text{ch}} \leq \overline{P}_j^{\text{ch}} \quad \forall i, \forall j, \forall t \tag{8}$$

$$E_{i,t} = E_{i,t}^{\text{arr}} \quad \forall i, \forall t | t = t_i^{\text{arr}} \tag{9}$$

$$u_{i,t} \in \{0, 1\} \quad \forall i, \forall t \tag{10}$$

III. Solving the PST Problem with DDPG

The uncertainties in EVs' arrival, departure times, and SoC at arrival make DRL a suitable alternative for solving this problem in real-time settings. DRL excel at fast, adaptive decision-making in uncertain environments, with the DDPG algorithm [14] being particularly effective for handling the continuous state and action spaces which are common in EV charging problems. To apply DDPG, the problem should first be formulated as a finite Markov Decision Process (MDP), characterized by (S, A, P, R, γ) [15]. The agent, as the decision maker, interacts with the environment by taking actions (a_t), shifting its state from s_t to s_{t+1}, and earning rewards based on a reward function (r_t). Hence, MDP has a state space $\mathcal{S}$, an action space $\mathcal{A}$, a state transition function $P(s_{t+1}|s_t, a_t)$, a reward R, and discount factor (γ) determining the significance of immediate and future rewards.

A. State Space and Action Space

Designing the state and action space is crucial in achieving the desirable performance in RL. The state vector $\boldsymbol{s}$ comprises three variables and an additional three for each controlled EV charger i. The EV charger variables default to zero when no EV is connected, maintaining a consistent state vector size of $|3 + 3C|$ at each time step t, where C represents the total number of chargers. The fixed variables include the normalized time step t/T, the contracted power setpoint P_t^{set}, and the previous total power usage P_{t-1}^{tot}. The EV charger state variables represent the normalized EV arrival and departure times $t^{\text{arr}}/T, t^{\text{dep}}/T$, and the SoC_t for each connected EV. Note that the departure time is being communicated only after

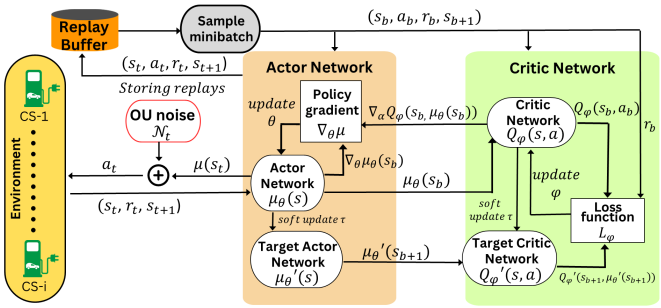

Fig. 2. Diagram of proposed DDPG-based approach for EV charging.

an EV connects to the charger. Normalizing time in the state vector simplifies the state space and improves time perception consistency, thereby enhancing the DDPG agent's learning efficiency. The state vector is formed as:

$$\mathbf{s}_t = [\tfrac{t}{T}, P_t^{\text{set}}, P_{t-1}^{\text{tot}}, \tfrac{t_i^{arr}}{T}, \tfrac{t_i^{dep}}{T}, SoC_{i,t}] \in S, i \in C$$

Actions within the environment are taken according to the constraints listed from Equations (2) to (10). The charging of EVs is regulated through the charging current $I_{i,t}^{\text{ch}}$. Actions assume continuous values between 0 and 1, with 0 indicating no charging and 1 representing full power charging, resulting in an action vector $\mathbf{a}_t \in [0,1]$ for each charger. Thus, the action vector for the environment corresponds to the total number of chargers, C, forming the action vector:

$$\mathbf{a}_t = [0,1]^{\text{C}} \in A.$$

B. Reward Function

The reward function calculates the squared difference between the power setpoints and total charging power at each step. Additionally, it incorporates a parameter named *charge power potential* P_t^{pot} for assisting the learning process considering the charging needs of EVs as shown in (11). This is represented by $I_{i,t}^{\text{pot}}$, which captures the potential charging capacity based on the number of connected EVs, their SoCs, and the power capabilities of the EVs and charging station i.

$$P_t^{\text{pot}} = \sum_{i \in C} I_{i,t}^{\text{pot}} \cdot V \cdot \sqrt{\phi} \cdot \eta \quad \forall i, \forall t \tag{11}$$

Therefore, the reward function is designed to minimize the gap between actual power usage and the lower of the power setpoints or the charge power potential:

$$R_t = -(min(P_{t-1}^{\text{pot}}, P_{t-1}^{\text{set}}) - P_{t-1}^{\text{tot}})^2. \tag{12}$$

C. DDPG Algorithm

The DDPG algorithm was selected to solve the aforementioned MDP as it can efficiently handle continuous action and state spaces. In the rest of this section, our DDPG-based approach will be thoroughly explained as illustrated in Fig. 2. Furthermore, Alg. 1 outlines the training process of the RL agent. The DDPG algorithm starts by initializing the actor network $\mu_\theta(s)$ and the critic network $Q_\varphi(s,a)$, parameterized by θ and φ. The actor maps states S to actions A, while the critic assesses the policy by estimating the Q-value of state-action pairs. Additionally, a replay buffer $\mathcal{R}$ is set up to store past transitions (s_t, a_t, r_t, s_{t+1}). In the exploration phase, at each time step t, the agent selects an action $(\mu_\theta(s_t))$ using the actor-network, by incorporating an Ornstein–Uhlenbeck (OU) action noise $(\mathcal{N}_t)$ to increase exploration while learning. The state (s_t), the executed action (a_t), its resulting state (s_{t+1}) and the achieved reward (r_t) are then stored in $\mathcal{R}$. After the replay buffer stores a predetermined amount of transitions, the algorithm samples a mini-batch of transitions (M) to update both the actor and critic networks. The target Q-value for these updates is calculated with (13).

Algorithm 1 DDPG Algorithm Training Process

```
Init. actor μ_θ, critic network Q_φ, and replay buffer R
Init. target networks Q' and μ' with θ' ← θ, φ' ← φ
for episode = 1, ..., U do
    Receive initial observation state s_1
    Initialize a random OU noise N
    for t = 1, ..., T do
        Select and execute action a_t = μ_θ(s_t) + N_t, observe
        reward r_t and new state s_{t+1}
        Store transition (s_t, a_t, r_t, s_{t+1}) in R
        Sample a minibatch M from R and find y_b (13)
        Update critic by minimizing L_φ (14)
        Update actor policy using policy gradient ∇_θ μ (15)
        Update the actor and critic networks μ_θ and Q_φ
        Update the target networks μ'_θ and Q'_φ
    end for
end for
```

$$y_b = r_b + \gamma Q'_\varphi(s_{b+1}, \mu'_\theta(s_{b+1})) \tag{13}$$

The algorithm then updates the critic network by minimizing the loss between its predicted Q-values, $Q_\varphi(s_b, a_b)$, and the target Q-values $Q'_\varphi(s_{b+1}, \mu'_\theta(s_{b+1}))$ by (14).

$$L_\varphi = \frac{1}{M} \sum_b (y_b - Q_\varphi(s_b, a_b))^2 \tag{14}$$

Subsequently, the actor policy is updated using the sampled policy gradient, aimed at actions that maximize the critic's predicted Q-values in (15).

$$\nabla_\theta \mu = \frac{1}{M} \sum_b \nabla_a Q_\varphi(s_b, \mu_\theta(s_b)) \nabla_\theta \mu_\theta(s_b) \tag{15}$$

After updating the actor policy, the weights for both the main and target networks of the actor and critic are updated. The learning rates for the gradient ascent process of the actor-network are applied by $\theta \leftarrow \theta + \alpha^\mu \nabla_\theta \mu$. Following that, the gradient descent process of the critic network is applied as $\varphi \leftarrow \varphi - \alpha^Q \nabla_\varphi L(\varphi)$. Lastly, the weights of the target networks μ'_θ and Q'_φ are updated towards the main networks using a soft update τ by $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$ and $\varphi' \leftarrow \tau\varphi + (1-\tau)\varphi'$. Main and target actor and critic networks undergo soft updates periodically to ensure that the policy improves steadily until the training concludes.

TABLE I
PST EVALUATION METRIC DEFINITIONS

Metric	Symbol	Equation
Squared Tracking Error (kW^2)	ϵ^{tr}	$\sum_{t\in T}(P_t^{set} - P_t^{tot})^2$
Energy Tracking Error (kWh)	$\lvert\epsilon^{tr}\rvert$	$\sum_{t\in T}\lvert P_t^{set} - P_t^{tot}\rvert \cdot \Delta t$
User Satisfaction $(\%)$	ϵ^{usr}	$\frac{1}{\lvert\mathcal{E}\rvert} \cdot \sum_{k\in\mathcal{E}} \frac{SoC_k}{SoC_k^*}$
Power Tracker Surplus (kW)	ϵ^{sur}	$\sum_{t\in T} \max((P_t^{tot} - P_t^{set}), 0)$

IV. EXPERIMENTAL EVALUATION

This section presents the experimental evaluation of the DDPG-based PST approach. In our case studies, a CPO manages the charging of EVs in a workplace parking lot with 10 EV chargers by applying the PST strategy, to minimize PST errors. The CPO purchases energy from the day-ahead market in advance, planning its use for charging sessions from 6 am to 6 pm, reflecting typical office hours. Each 15-minute interval is assigned a specific power setpoint, considering the wholesale market contracting intervals. The case study was modelled with the EV2Gym [12] simulator and was used to train and evaluate the RL agent.

EV arrival patterns, duration of stay, and charging needs specific to a workplace are obtained from ElaadNL [16]. Additionally, historic day-ahead electricity prices are retrieved from entso-e [17]. Combined with total EV registration in the Netherlands by 2023 data from RVO [18], these offer a practical foundation for solving the PST problem in workplace settings. Additionally, power setpoints are determined with a 5% flexibility margin, which means that the CPO purchases 5% more energy in the day-ahead market than the total charging demand for the next day.

Furthermore, the following evaluation metrics highlighted in Table I are considered. The squared tracking error ϵ^{tr}, aligns with the DDPG's reward function and the objective function of the mathematical formulation. Secondly, the energy tracking error $|\epsilon^{tr}|$ indicates PST error during a day's charging sessions in kWh. User satisfaction ϵ^{usr} measures the SoC increase against the desired SoC target of 80%, reflecting user experience. Lastly, power tracker surplus ϵ^{sur} indicates the extent of charging power that exceeds determined power setpoints in kW unit.

A. Training and Testing Settings

To train the agent, 1,200,000 time steps of 15 minutes are completed, equating to 25,000 episodes, i.e., 25,000 days of 12 hours each. The RL agent's learning is evaluated by the convergence and maximization of the mean reward and its performance considering the evaluation metrics in Table I in contrast to Optimal and CAFAP algorithms.

In our study, we employed a replay buffer $\mathcal{R}$ with a size of 1,000,000 transitions and a minibatch size M of 64. The discount factor γ was set at 0.99, with a soft update τ value of 0.0005 and a learning rate α of 0.001. For exploration, we used action noise $\mathcal{N}$ with a standard deviation 0.2. The actor networks $(\mu_\theta, \mu'_\theta)$ were configured with sizes of two fully connected layers with 128 nodes; similarly, both the main and target critic networks (Q_φ, Q'_φ) were configured with 64 nodes, respectively. In Fig. 3, the convergence of mean rewards of an episode can be observed, obtained from 10 training sessions using the selected hyperparameter set. The agent converges after the 5000^{th} episode. However, the training continued until the 25000^{th} episode to ensure convergence.

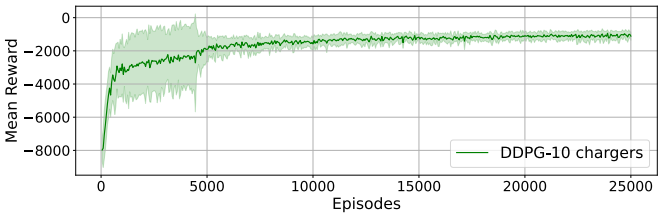

Fig. 3. Mean episode rewards from 10 training sessions.

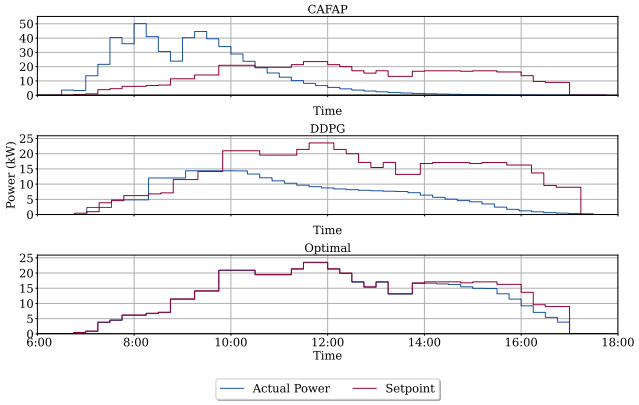

Fig. 4. Power setpoints and actual power usage in a single evaluation scenario.

B. Performance Evaluation of DDPG

To evaluate the performance of the DDPG algorithm, the trained model was tested with randomly generated 100 scenarios and was compared with two baseline approaches. The first benchmark is the optimal solution, derived from the mathematical formulation of the PST minimization problem, assuming complete knowledge at the start of the day—an impractical scenario due to uncertainties like EV arrival times and SoC. The second benchmark is a common charging strategy, *charge as fast as possible* (CAFAP), where EVs charge at maximum capacity upon connection without any charging control.

In Fig. 4, the charging operation for one replay can be observed. The CAFAP algorithm charged EVs as fast as possible upon their arrival, resulting in PST errors, which can be translated to potential risks for imbalances. Conversely, the DDPG algorithm scheduled charging times to better align with power setpoints, optimizing charging and minimizing associated costs and risks. Consecutively, the Optimal algorithm charged EVs while minimizing the PST error to an experimental minimum, however, such precision is impractical in real-world settings due to uncertainties.

Overall, the DDPG algorithm had better performance than the CAFAP algorithm as it exceeded the power setpoints by 78% less as shown in Table II. This indicates that the DDPG

TABLE II
AVERAGE PERFORMANCE IN 100 EPISODES

Algorithm	ϵ^{tr} (kW^2)	$\lvert\epsilon^{tr}\rvert$ (kWh)	ϵ^{usr} $(\%)$	ϵ^{sur} (kW)
CAFAP	11862.3 ± 4278.1	147.68 ± 25.38	99.80 ± 0.20	283.77 ± 49.30
DDPG	4972.0 ± 1753.9	97.62 ± 18.65	88.61 ± 3.00	61.73 ± 27.42
Optimal	189.5 ± 129.0	13.16 ± 5.06	98.52 ± 0.70	0.00025 ± 0.0005

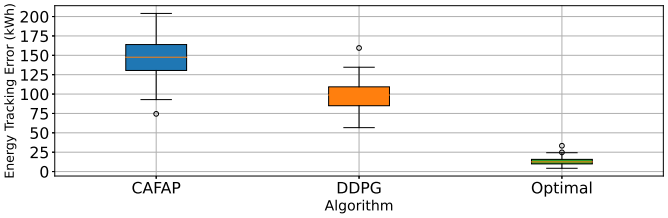

Fig. 5. Distribution of energy tracking error throughout 100 evaluated replays

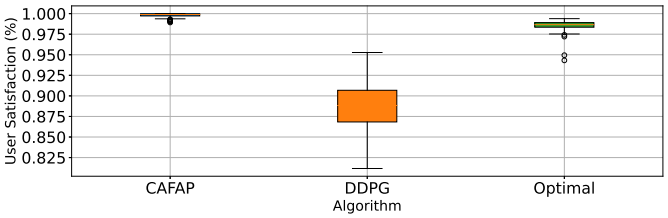

Fig. 6. Distributions of user satisfaction throughout 100 evaluated replays

algorithm can ensure better adherence to predetermined power levels, which is particularly important when the demand for charging exceeds the power capacity in alternative scenarios. Fig. 5 represents the averages and standard deviations of the compared algorithms. Energy tracking error was analyzed for three algorithms. The DDPG algorithm notably outperformed CAFAP, with energy tracking errors averaging about 50 kWh less per replay. However, the Optimal algorithm excelled further, achieving the best performance in minimizing the energy tracking error as expected due to its status.

In detail, the user satisfaction evaluated in Fig. 6, shows that both the CAFAP and Optimal algorithms nearly fully charged EV batteries to the desired SoC level. This is an expected result due to CAFAP's fast charging strategy and the Optimal algorithm's benchmark status. In contrast, the DDPG algorithm did not perform as well as the other benchmarks, achieving an average user satisfaction rate of 88.6% across 100 episodes. This indicates that while DDPG effectively reduces the energy tracking error, it does so at the expense of user satisfaction. This result shows that although the DDPG algorithm compromises user satisfaction, it maintains a relatively high baseline, reflecting a strategic balance between minimizing PST error and optimizing EV user satisfaction. Remarkably, achieving 88.6% average user satisfaction while reducing energy tracking error by 34% compared to CAFAP represents a significant achievement for the DDPG algorithm.

V. CONCLUSION

This study introduced a DDPG-based approach for EV smart charging at a workplace parking lot to minimize PST error by meeting predetermined power setpoints. The proposed approach was compared against two benchmark algorithms, CAFAP and Optimal derived from a MINLP formulation of the PST minimization problem. As a result, the DDPG algorithm managed to decrease the PST error by 34%, while keeping EV user satisfaction at 88.6%. Furthermore, it achieved a 78% reduction in exceeding power setpoints compared to CAFAP, highlighting the DDPG's promises for further research to operate in EV charging scenarios with restricted power capacity.

REFERENCES

[1] IEA. (2023) Global ev outlook 2023. Paris. License: CC BY 4.0. [Online]. Available: https://www.iea.org/reports/global-ev-outlook-2023

[2] K. M. Tan, V. K. Ramachandaramurthy, and J. Y. Yong, "Integration of electric vehicles in smart grid: A review on vehicle to grid technologies and optimization techniques," *Renewable and Sustainable Energy Reviews*, vol. 53, p. 720–732, Jan. 2016.

[3] K. Mahmud, M. J. Hossain, and J. Ravishankar, "Peak-load management in commercial systems with electric vehicles," *IEEE Systems Journal*, vol. 13, no. 2, p. 1872–1882, Jun. 2019.

[4] O. Sadeghian, A. Oshnoei, B. Mohammadi-ivatloo, V. Vahidinasab, and A. Anvari-Moghaddam, "A comprehensive review on electric vehicles smart charging: Solutions, strategies, technologies, and challenges," *Journal of Energy Storage*, vol. 54, p. 105241, Oct. 2022.

[5] A. Dukpa and B. Butrylo, "Milp-based profit maximization of electric vehicle charging station based on solar and ev arrival forecasts," *Energies*, vol. 15, no. 1515, p. 5760, Jan. 2022.

[6] E. Srilakshmi and S. P. Singh, "Energy regulation of ev using milp for optimal operation of incentive based prosumer microgrid with uncertainty modelling," *International Journal of Electrical Power & Energy Systems*, vol. 134, p. 107353, Jan. 2022.

[7] C. E. P. R. Institute, D. Zhang, X. Han, T. U. of Technology, C. Deng, and C. E. P. R. Institute, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE Journal of Power and Energy Systems*, vol. 4, no. 3, p. 362–370, Sep. 2018.

[8] Z. Wan, H. Li, H. He, and D. Prokhorov, "A data-driven approach for real-time residential ev charging management," in *2018 IEEE Power & Energy Society General Meeting (PESGM)*, Aug. 2018, p. 1–5.

[9] D. Qiu, Y. Ye, D. Papadaskalopoulos, and G. Strbac, "A deep reinforcement learning method for pricing electric vehicles with discrete charging levels," *IEEE Transactions on Industry Applications*, vol. 56, no. 5, p. 5901–5912, Sep. 2020.

[10] H. Li, G. Li, T. T. Lie, X. Li, K. Wang, B. Han, and J. Xu, "Constrained large-scale real-time ev scheduling based on recurrent deep reinforcement learning," *International Journal of Electrical Power & Energy Systems*, vol. 144, p. 108603, Jan. 2023.

[11] N. Sadeghianpourhamami, J. Deleu, and C. Develder, "Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, p. 203–214, Jan. 2020.

[12] S. Orfanoudakis, C. Diaz-Londono, Y. E. Yılmaz, P. Palensky, and P. P. Vergara, "Ev2gym: A flexible v2g simulator for ev smart charging research and benchmarking," 2024.

[13] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," Jun. 2016, arXiv:1606.01540 [cs]. [Online]. Available: http://arxiv.org/abs/1606.01540

[14] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31st International Conference on Machine Learning*, vol. 32, no. 1. PMLR, Jun. 2014, pp. 387–395.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning, Second Edition: An Introduction*, ser. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: Bradford Books, 2018, vol. Second edition.

[16] "Open datasets for electric mobility research — update april 2020," 2024. [Online]. Available: https://platform.elaad.io/analyses/ElaadNL_opendata.php

[17] ENTSO-E. (2024) Central collection and publication of electricity generation, transportation and consumption data and information for the pan-european market. [Online]. Available: https://transparency.entsoe.eu/

[18] Rijksdienst voor Ondernemend Nederland (RVO), "Statistics electric vehicles and charging in the netherlands up to and including september 2023," 2023.