Intelligent Energy Management using Multi-Agent Dynamic Learning for Scheduling Commercial Electric Vehicle Charging Stations

Kevin Chan
Dept of Electrical Engineering
University of Bath
Bath, United Kingdom
cyc241@bath.ac.uk

Pedram Asef
Dept of Mechanical Engineering
University College London
London, United Kingdom
pedram.asef@ucl.ac.uk

Alexandre Benoit
Dept of Electrical Engineering
University of Bath
Bath, United Kingdom
amgb20@bath.ac.uk

***Abstract*—For commercial electric vehicles (CEVs), an underexplored challenge is the complexity of demand and supply management, which is vital for the efficient operation and broader adoption of CEVs. By leveraging advanced smart grid technologies and intelligent energy management systems, the research endeavors to create a cost-effective software solution for optimizing the charging process. This study deploys proximal policy optimization (PPO) multi-agent deep reinforcement learning (MARL) within an actor-critic network architecture. Agents are responsible for managing the supply and demand of energy from two grids welcoming ten charging stations each pumping energy from the integrated uninterruptible power supply (UPS). Performance metrics are compared against a dynamic programming (DP) approach, serving as a benchmark. The DP model excels when prior information is readily available. In contrast, PPO agents exhibit remarkable robustness and adaptability in environments lacking such information obtaining 95% accuracy. These insights not only enrich the existing academic discourse but also establish new performance benchmarks for practical implementations.**

***Index Terms*—Electric vehicle, energy management, charging station, dynamic programming, multi-agent dynamic learning, proximal policy optimization, neural network, solar photovoltaic-integrated grid.**

I. Introduction

The transportation sector, responsible for 14% of global greenhouse gas emissions, must evolve to mitigate its environmental impact [7]. The push towards Electric Vehicles (EVs) is a response to the climate crisis, with studies indicating that a shift to EVs could markedly reduce emissions [9]. Legislative measures, such as the 2035 ban on combustion engine vehicles, emphasize this shift [12].

Challenges hindering the adoption of EVs, particularly Commercial Electric Vehicles (CEVs), include limited charging infrastructure and the increased electricity demand straining the grid [13]. This underscores the need for advanced energy management systems (EMS) that efficiently balance supply and demand, integrating renewable energy without destabilizing the grid [15].

Our research centers on Vehicle-to-Grid (V2G) technology, which optimizes this balance by providing a two-way energy exchange between EVs and the power grid, enhancing grid stability and renewable energy storage [14]. We explore the potential of Deep Reinforcement Learning (DRL) in EMS for its superior decision-making and adaptability, addressing challenges such as charging time and grid efficiency [17] [18]. Despite existing limitations in predicting EV charging patterns [15], the application of DRL in smart grids suggests a scalable and effective approach to optimizing energy distribution in the face of evolving energy landscapes.

This research aims to develop a sophisticated Intelligent EMS for managing power in hybrid photovoltaic-grid-connected microgrids, especially for CEV depots. Our approach leverages a multi-agent system focusing on Proximal Policy Optimization (PPO) within an actor-critic framework. The innovative aspect of our study lies in creating a robust model that operates effectively without needing prior knowledge of State of Charge or Departure Time, using dynamic programming scheduling to benchmark the efficacy of our DRL approach. This methodology positions our system at the forefront of addressing CEV charging schedule challenges with high theoretical and practical precision.

The rest of the paper is organised as follows. Section II delves into existing research focused on energy management systems, specifically addressing demand/supply balancing and EV charging schedules. Section III provides an examination of various methods employed to resolve constrained optimization problems. Section V executes numerical tests to validate the efficacy of the proposed method. The final section, Section VI, offers concluding remarks.

II. Related Works

Machine learning has significantly influenced EMS for microgrids [20], bringing advancements in vehicle to home (V2H), grid to vehicle (G2V), and V2G applications. Traditional direct and heuristic methods, such as particle swarm optimization for improved EV charging efficiency [21], have been complemented by dynamic algorithms like the Dynamic

Hunting Leadership (DHL) method to maintain grid voltage stability with a high presence of EVs [22].

A pivotal shift towards DRL is evident, with PPO highlighted as an effective method for industrial optimization, particularly in stochastic environments [19], [1]. This is due to its simplicity and ability to learn iteratively from interactions within complex systems. DRL's application extends to optimizing EV dispatch, enhancing renewable energy integration [23], and intelligent coordination within EV charging networks for grid impact management [23]. Model-free RL techniques are reviewed for their optimal control in energy systems [24] [26], and a decentralized, incentive-based demand response approach is proposed to manage EV charging loads [25].

The research also delves into multi-agent DRL for scheduling EV charging in solar grid (SGs), presenting a decentralized, adaptable solution for real-time application [26], and compares various machine learning models for forecasting EV charging loads.

Despite these developments, a gap persists in addressing the specific needs of the CEV industry. There is a significant divergence in priorities between commercial and non-commercial EVs, with the former focusing on operational efficiency over cost minimization. The alignment of charging schedules with renewable energy production and the high demands for rapid charging in the CEV sector are not fully met by current SG strategies, highlighting an area for further research.

III. Problem Formulation and Methodology

A. Microgrid Architecture

The microgrid architecture used in this paper is a hybrid system, primarily designed to cater to the charging needs of CEVs. The key components of this microgrid are:

1) **Photovoltaic Array:** This serves as the primary renewable energy source, enhancing solar power to generate electricity.
2) **Uninterruptible Power Supply (UPS):** The UPS system plays a crucial role in energy storage and power quality management. It consists of an AC-DC converter, a Voltage Source Converter (VSC), and two battery packs. The VSC regulates the incoming power and directs it to the battery packs based on their state of charge. The battery packs store excess energy generated by the PV array during the day and discharge it during periods of high demand or when the PV system is not generating power (e.g., at night).
3) **Charging Stations:** The microgrid includes two zones, each equipped with 10 charging ports, totaling 20 charging ports. These stations are where the CEVs connect to receive power for charging.
4) **Energy Management System (EMS):** The EMS is the brain of the microgrid, responsible for intelligently managing the energy flow between the PV array, UPS, and charging stations. This is where we utilize the multi-agent PPO algorithm to optimize energy distribution.

The control of the microgrid is hierarchical, with primary and secondary control levels (the secondary being the focus of this paper:

1) **Primary Control:** This level focuses on maintaining voltage and frequency stability. It utilizes Maximum Power Point Tracking (MPPT) to optimize the power output from the PV array and Voltage Source Converters (VSCs) to regulate the voltage and power flow within the microgrid.
2) **Secondary Control:** This level is responsible for higher-level energy management decisions. It employs the EMS to make real-time adjustments to the energy distribution, considering factors like EV charging demand, PV generation, and battery state of charge.

B. Dynamic Programming Approach

DP is a time-honored and commonly utilized method for orchestrating the charging scheduling of EVs [6]. We employ a DP-based EMS as a benchmark to evaluate the performance of our multi-agent PPO. This comparison aims to determine the effectiveness of sophisticated algorithms in orchestrating energy distribution within a commercial microgrid tailored for EV charging stations (CS).

The procedural logic of the DP-based EMS considers a thorough strategy for the allocation of energy resources as seen in Fig 1.

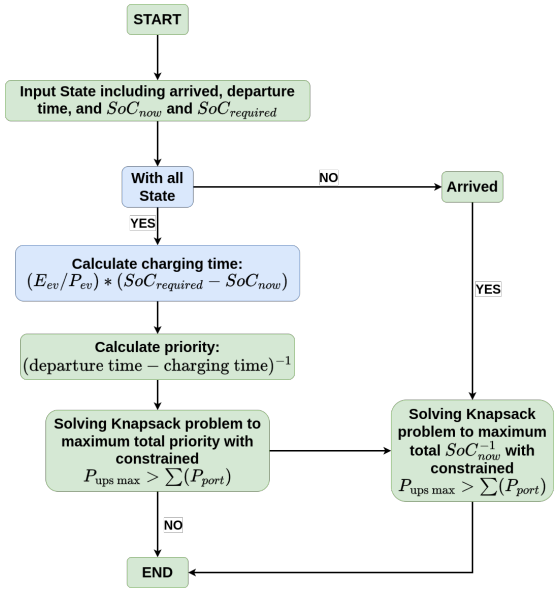


Fig. 1: Custom charging scheduling DP flowchart

In the optimization phase, the distribution of energy resources are allocated to each EV based on the results, managing the charging station switches. When the energy distribution cycle is complete, the algorithm updates inputs for the next cycle. This iterative process persists until CEVs are optimally charged or the energy reserves are exhausted.

C. Deep Reinforcement Learning

Our model utilizes DRL, combining RL with deep neural networks (DNN) to train agents in complex environments. In our microgrid application, the agent learns to optimize energy distribution, balancing the grid's efficiency and sustainability. It continuously adapts to environmental feedback, refining its decisions to improve the grid's performance over time. This enables the agent to effectively manage varying energy demands and supply conditions through advanced strategy development.

D. Actor-Critic Methods

Actor-Critic methods, a hybrid architecture combining value-based and policy-based methods that helps to stabilize the training by reducing the variance using an Actor that controls how our agent behaves (policy-based). A Critic that measures how good the taken action is (value-based method). The solution to reducing the variance of the Reinforce algorithm and training our agent faster and better is to use a combination of policy-based and value-based methods.

With actor-critic methods, there are two function approximations (two NNs). The Actor is a policy function parameterized by θ: $\pi_\theta(s)$ where the goal is to propose a probabilistic actions space. The Critic is a value function parameterized by μ: V_μ where it evaluates the actions by estimating the value of taking a particular action in a given state and updates the actor's policy. It helps the actor to understand how good the action is in terms of future rewards. This dual mechanism allows for more stable and faster convergence compared to traditional methods.

E. Multi-Agent Reinforcement Learning (MARL)

Our model is based on a multi-agent system in a decentralised environment, which means that no information is shared between the agents. It simplifies the system design but it does not know the state of other agents.

The DRL agent interacts with the microgrid by sending control signals to adjust energy distribution, charge or discharge energy storage systems, or connect/disconnect from the main grid. The microgrid, in turn, provides the agent with observations such as current load and battery status, which the agent uses to make future decisions.

F. Proximal Policy Optimization

The core strength of our approach lies in the implementation of MARL combined with the PPO technique. This innovative method excels in scenarios lacking prior knowledge, a common occurrence in commercial EV charging systems where user patterns are unpredictable.

PPO's critical innovation is its cautious approach to policy updates during training, aimed at ensuring stable convergence towards optimal solutions. The rationale is twofold: empirically, smaller policy adjustments tend to yield more consistent convergence, and excessive changes risk detrimental policy performance, from which recovery can be prolonged or even unachievable. PPO achieves this careful balance by calculating a ratio that reflects the extent of policy change from one iteration to the next. This ratio is then clipped within a specified range, denoted as $[1-\epsilon, 1+\epsilon]$, constraining the policy to remain proximate to the previous one—hence the term 'proximal policy.' This mechanism, embodied in the clipped surrogate objective function, strategically restricts the policy update, ensuring that changes stay within a conservative range to foster stable and reliable learning outcomes.

Avoiding large updates is the primary function of:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\Big[\min(\underbrace{\underbrace{r_t(\theta)}_{\text{Ratio Function}}\hat{A}_t}_{UnclippedPart}, \underbrace{\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)}_{ClippedPart}\Big] \quad (1)$$

The ratio functions is the following:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \quad (2)$$

The quantity in (2) represents the likelihood of choosing a specific action a_t given the current state s_t under the new policy relative to the old policy. This value symbolized as $r_t(\theta)$, acts as a gauge for the change in policy over time:

A value of $r_t(\theta)$ greater than 1 suggests that the new policy has a higher propensity to select action a_t in state s_t compared to the former policy. Conversely, a value of $r_t(\theta)$ less than 1 implies a reduced tendency for the action under the new policy in contrast to the old. This ratio serves as a straightforward metric for assessing the extent of deviation between the new and prior policy settings.

G. Reward Function Design

The reward function is engineered to align agents' behaviours with the specified objectives of minimizing this last, which is:

- Charging percentage on EVs over it's SoC:

$$W_1 \times \sum_{ZEV_s}\left(\frac{SoC_{N+1} - SoC_N}{SoC_N}\right) \quad (3)$$

- Voltage Direct Current (VDC):

$$W_{VDC} \times \sum_{AEV_s}\left(\frac{SoC_{N+1} - SoC_N}{SoC_N}\right) \quad (4)$$

- The ratio of energy sourced from PV panels to the total energy consumption from both PV and the central grid.

$$W_2 \times \left(\frac{P_{PV}}{(P_{PV} + P_M)}\right) \quad (5)$$

- Penalisation metric for failing to achieve requisite SoC levels within specified time frames.

$$W_3 \times N_{NC} \times \text{Penalty} \quad (6)$$

H. Trust Region Constraints

PPO incorporates a trust region limitation to maintain policy updates within a certain range. This precaution ensures that newly adopted policies do not deviate excessively from previous ones, thereby avoiding radical changes that might disrupt the stability of the learning progression [1].

I. Value Decomposition Networks into a MARL

Due to the complexity of the EV charging station, a sophisticated MARL approach is needed. This approach scales traditional DRL to complex, multi-agent environments, preparing the system for future growth and complexity [2]. Utilizing PPO, each agent is responsible for a section of the action space and works in concert with others to achieve collective goals like energy efficiency and cost reduction.

To address the high complexity of the system and improve optimization, value decomposition networks (VDN) have been integrated into the MARL framework. VDN breaks down the overall value function into individual components for each agent, facilitating the optimization of each agent's policy towards a common, overarching reward [3]. This method efficiently overcomes the issue of correlated policies, enabling agents to work with a degree of independence while maintaining overall alignment and coordination. Moreover, it endows the system with adaptability and robustness. Making it capable of handling various challenges, such as fluctuating demand, the unpredictability of renewable energy sources, and potential system faults [4]. A representation of our proposed model architecture can be seen in Fig 2 with the use of VDN [5].

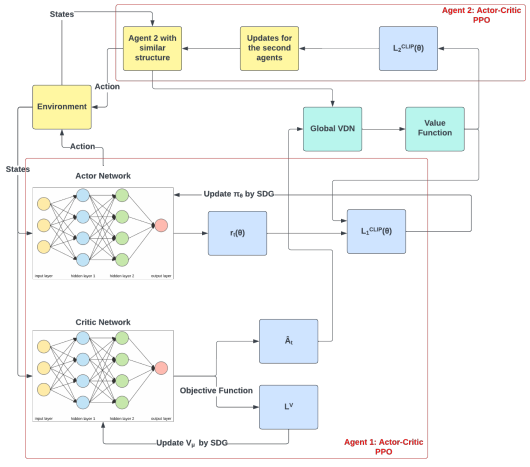


Fig. 2: Overall architecture of proposed MARL system (implemented with 2 agents) using an Actor-Critic PPO process.

IV. Experimental Set-up and Implementation

The model has been trained RTX A4000 GPU of a Dell Precision 5820 Tower Workstation with a Intel Xeon W-2245 (8 Core), 3.9 GHz (4.5 GHz Max Turbo) and 32 GB of memory.

A. Optimal Training Environment for MARL PPO Agents

The proposed approach meticulously calibrates the number of charging ports to maintain a balance between a realistic representation of a commercial EV charging setup and the computational tractability required for efficient agent training. By structuring the environment to allow the zones to function both independently and collectively within the microgrid, we facilitate an accurate assessment of the multi-agent PPO system's ability to dynamically allocate energy resources.

The simulated microgrid environment (Fig. 2) is designed to serve as the training ground for our agents. Within this environment, each PPO agent manages a specific array of EV charging ports, functioning as an individual EMS. The collective goal of these agents is ensuring that each EV achieves the required SoC by the predetermined departure time, and reducing reliance on the main power grid.

The defined agents operate within a defined observation space with SoC levels of the UPS and each EV, along with the EVs' scheduled departure times and A set of thirteen floating-point values representing the percentage of power drawn from the central grid. Agents require an action space of boolean variables representing the on/off status of each charging port and a boolean switch to regulate the connection to the central grid.

B. Agent Architecture and Training Parameters

Initially designed with a three-layered hidden structure comprising 120, 60, and 30 neurons, this setup did not successfully achieve convergence in initial tests.

Owing to the initial design's failure to converge, modifications were made to both the Actor and Critic networks, which involved increasing the neurons in each hidden layer to 256, 128, and 64, respectively. This enhancement allowed the networks to detect more complex patterns, thereby aiding in achieving convergence.

Additional hyperparameters, such as the rate of learning, the size of the batches, and the rate of discount were carefully optimized. A subsequent implementation employing PPO resulted in markedly better convergence and consistency across a multitude of episodes.

V. Results and Discussion

The learning progress of PPO agents tasked with optimizing energy distribution in a microgrid setting is studied. The exploration covers three distinct configurations to determine how effectively the PPO strategy performs under varying conditions.

A. Single Agent with Wide Trust Region

Fig 3 exhibits the episodic reward trajectory for a single agent operating under a policy with a wide trust region. This approach allows for larger updates to the policy during training. It appears that the agent experiences considerable volatility in performance, with significant fluctuations in episodic reward. Such variance suggests that while the wide trust region may accelerate learning in some episodes, it

may also introduce instability, leading to periods of reduced performance. To mitigate this, a more conservative approach or a dynamic adaptation of the trust region could be explored to balance learning speed and stability.

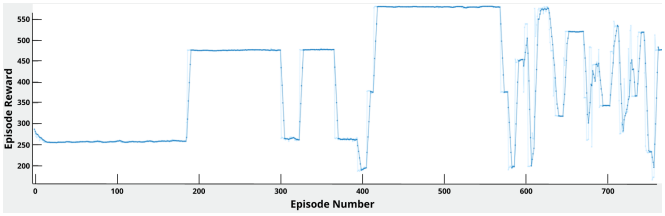


Fig. 3: Single Agent with Wide Trust Region

B. Single Agent with Narrower Trust Region

Fig 4 features a single agent adhering to a policy with a narrower trust region, constraining the magnitude of policy updates. The rewards here display less fluctuation compared to the wide trust region scenario, indicating a smoother learning process. However, there are still sharp drops in performance, which could imply that while the narrow trust region promotes stability, it might also slow down the agent's ability to adapt to more optimal policies. Refining the balance between exploration and exploitation might enhance the agent's performance consistency.

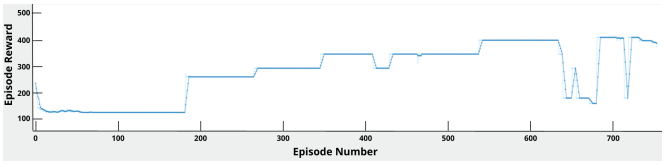
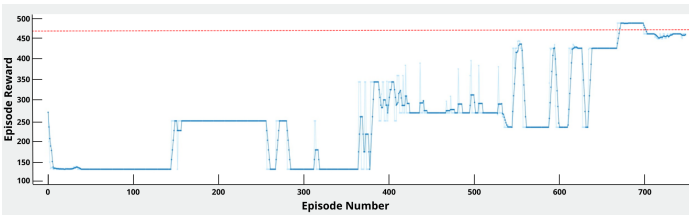


Fig. 4: Single agent with narrow trust region

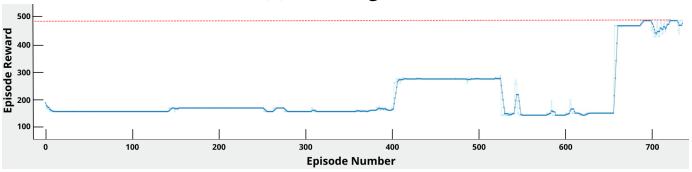
C. Multi-Agent Training

Fig 5 portrays the learning curves of multiple agents working collaboratively or competitively within the same environment. The presence of multiple agents introduces complexity due to the interactions between the agents' policies. Interestingly, the collective dynamics seem to produce more consistent reward patterns in some phases, potentially indicating that multi-agent collaboration can lead to more robust policy development and try to achieve globally optimal policy. We can see that both agents converge collaboratively to a reward of around 470. However, the increased complexity also leads to unpredictability, as seen in certain episodes with sharp reward declines. Implementing communication protocols or shared learning strategies could potentially improve coordination and result in more stable performance.

These findings offer meaningful observations regarding the flexibility and effectiveness of PPO agents when orchestrating intricate energy networks. Moreover, these results lay a foundational backdrop for the ensuing discussion segment, wherein these empirical revelations will be contemplated within the expansive scope of this investigative study.



(a) Multi-agent 1



(b) Multi-agent 2

Fig. 5: Multi-agent training result

D. Comparative Performance Analysis with DP Method

The analysis evaluates MARL using PPO against DP strategies, focusing on key indicators of efficiency in energy distribution within the microgrid. This includes prior information on energy demands and departure times. Table I summarizes the findings.

TABLE I: Comparative performance analysis with or without priori information (PI) compared to the DP method benchmark

Control Method	**DP w/ PI**	**MARL PPO w/ PI**	**DP w/o PI**	**MARL PPO w/o PI**
Performance	Baseline	102%	45%	95%
Computation Time	Fast	Fast	Fast	Fast
Penalty of Insufficient Charging	0	0	6 times	1 time
Training Time	n/a	40h	n/a	40hr
PV/Total Energy Consumption (kWh)	85	91	85	88

In reinforcement learning, efficiency is indicated by the cost minimized or maximized according to the reward function, described by different equations (Eq.3 to 6). Higher cumulative rewards imply better performance and higher efficiency. For example, a 2% improvement of MARL PPO with prior information over DP indicates a 2% higher cumulative reward, demonstrating superior EMS performance. The paper defines two primary objectives for the EMS: ensuring all EVs reach their required SoC by their designated departure times and minimizing the energy drawn from the central grid.

The MARL PPO method equipped with prior information marginally surpassed the baseline DP approach by 2%. In contrast, the MARL PPO method lacking prior information achieved a commendable 95% efficiency relative to the baseline. However, the DP method deprived of prior information was considerably less effective, realizing only 45% efficiency. All assessed methodologies demonstrated rapid computational

speeds, suggesting their potential suitability for applications necessitating immediate decision-making. The MARL PPO strategies demonstrated superior management in avoiding penalties for insufficient charging, maintaining zero penalties with the advantage of prior information and incurring just a single penalty without it. Conversely, the DP method without such information suffered 6 penalties. Approximately 40 hours were necessary to train the MARL PPO methods. This time investment is substantial but is justified as a one-off commitment to secure enduring performance enhancements. Harnessing prior information, the MARL PPO method achieved an impressive 91% ratio of PV energy to total energy consumption, indicative of more effective utilization of sustainable energy resources. This metric suggests room for improvement in optimizing energy sourcing, highlighting a potential area for further technological development or algorithmic refinement.

VI. Conclusion

Our study marks a significant advancement in microgrid management by effectively utilizing DRL within a multi-agent system to address the day-night energy imbalance. We successfully implemented MARL with PPO agents, achieving a remarkable 95% efficiency, significantly outperforming traditional Dynamic Programming approaches which only managed a 45% effectiveness rate. This highlights the superior adaptability and robustness of our MARL PPO method, especially in managing photovoltaic systems and optimizing CEV charging.

The DRL algorithms for secondary control in our system effectively utilize renewable energy during the day, enhancing grid independence and reducing operational costs at night. The implementation of PPO agents through VDN further advances multi-agent cooperation, setting new benchmarks in smart microgrid management. Overall, a pioneering approach in energy management is presented, particularly in the commercial EV sector, by combining operational efficiency with sustainability. Our DRL-based methodologies provide a foundation for future innovation in sustainable microgrid operations.

References

[1] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

[2] Hernandez-Leal, P., Kartal, B. & Taylor, M.E. A survey and critique of multiagent deep reinforcement learning. Auton Agent Multi-Agent Syst 33, 750–797 (2019). https://doi.org/10.1007/s10458-019-09421-1

[3] Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W.M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J.Z., Tuyls, K. and Graepel, T., 2017. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv:1706.05296.

[4] Papoudakis, G., Christianos, F., Rahman, A. and Albrecht, S.V., 2019. Dealing with non-stationarity in multi-agent deep reinforcement learning. arXiv preprint arXiv:1906.04737.

[5] Lim, Hyun-Kyo & Kim, Ju-Bong & Heo, Joo-Seong & Han, Youn-Hee. (2020). Federated Reinforcement Learning for Training Control Policies on Multiple IoT Devices. Sensors. 20. 1359. 10.3390/s20051359.

[6] Hajidavalloo, Mohammad & Shirazi, Farzad & Mahjoob, Mohammad. (2020). Performance of different optimal charging schemes in a solar charging station using DP. Optimal Control Applications and Methods. 41. 10.1002/oca.2619.

[7] Lamb, W.F. et al. (2021) 'A review of trends and drivers of greenhouse gas emissions by sector from 1990 to 2018', Environmental Research Letters, 16(7), p. 073005. doi:10.1088/1748-9326/abee4e.

[8] Frank, S. et al. (2023) Built for purpose: EV adoption in light commercial vehicles, McKinsey & Company. Available at: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/built-for-purpose-ev-adoption-in-light-commercial-vehicles (Accessed: 09 October 2023).

[9] Requia, W.J. et al. (2018) 'How clean are electric vehicles? evidence-based review of the effects of electric mobility on air pollutants, greenhouse gas emissions and human health', Atmospheric Environment, 185, pp. 64–77. doi:10.1016/j.atmosenv.2018.04.040.

[10] Intergovernmental Panel on Climate Change (IPCC) (2022) Global Warming of 1.5°C: IPCC Special Report on Impacts of Global Warming of 1.5°C above Pre-industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty. Cambridge: Cambridge University Press. doi: 10.1017/9781009157940.

[11] Schrijver, Alexander (2003). Combinatorial Optimization: Polyhedra and Efficiency. Algorithms and Combinatorics. Vol. 24. Springer. ISBN 9783540443896.

[12] European Environment Agency. "Electric vehicles and the energy sector - impacts on europe's future emissions." (2021), [Online]. Available: https://www.eea. europa.eu/publications/electric-vehicles-and-the-energy

[13] M. Ehsani, K. V. Singh, H. O. Bansal, and R. T. Mehrjardi, "State of the art and trends in electric and hybrid electric vehicles," Proceedings of the IEEE, vol. 109, no. 6, 2021. DOI: 10.1109/JPROC.2021.3072788

[14] H. Farhangi and G. Joos, Microgrid Planning and Design: A Concise Guide. John Wiley & Sons, 2019, pp. 1–24. [Online]. Available: https://ieeexplore.ieee. org/book/8671408 (visited on 03/18/2023)

[15] G. Chandra Mouli, M. Kefayati, R. Baldick, and P. Bauer, "Integrated pv charging of ev fleet based on energy prices, v2g, and offer of reserves," IEEE Transactions on Smart Grid, vol. 10, no. 2, Mar. 2019. DOI: 10.1109/TSG.2017.2763683

[16] H. Farhangi and G. Joos, Microgrid Planning and Design: A Concise Guide. John Wiley & Sons, 2019, pp. 57–63. [Online]. Available: https : / / ieeexplore . ieee.org/book/8671408 (visited on 03/18/2023)

[17] A. Luo, Q. Xu, F. Ma, and Y. Chen, "Overview of power quality analysis and control technology for the smart grid," Journal of Modern Power Systems and Clean En- ergy, vol. 4, no. 1, 2016. DOI: 10.1007/s40565-016-0185-8

[18] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep rein- forcement learning: A brief survey," IEEE Signal Processing Magazine, vol. 34, no. 6, 2017. DOI: 10.1109/MSP.2017.2743240

[19] L. van Hezewijk, N. Dellaert, T. Van Woensel, and N. Gademann, "Using the proximal policy optimisation algorithm for solving the stochastic capacitated lot sizing problem," International Journal of Production Research, vol. 61, no. 6, pp. 1955-1978, 2023. DOI: 10.1080/00207543.2022.2056540

[20] P. Asef, R. Taheri, M. Shojafar, I. MPoras, and R. Tafazolli, 2023. SIEMS: A Secure Intelligent Energy Management System for Industrial IoT Applications. IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 1039-1050, DOI: 10.1109/TII.2022.3165890.

[21] An, Y., Gao, Y., Wu, N., Zhu, J., Li, H., and Yang, J., 2023. Optimal scheduling of electric vehicle charging operations considering real-time traffic condition and travel distance. Expert Systems with Applications, Volume 213, Part B, 118941. ISSN 0957-4174. DOI: 10.1016/j.eswa.2022.118941.

[22] Ahmadi B, Shirazi E. A Heuristic-Driven Charging Strategy of Electric Vehicle for Grids with High EV Penetration. Energies. 2023; 16(19):6959. https://doi.org/10.3390/en16196959

[23] Qiu D, Wang Y, Hua W, Strbac G. Reinforcement Learning for Electric Vehicle Applications in Power Systems: A Critical Review. Renewable and Sustainable Energy Reviews. 2023; 173:113052. https://doi.org/10.1016/j.rser.2022.113052

[24] Vamvakas D, Michailidis P, Korkas C, Kosmatopoulos E. Review and Evaluation of Reinforcement Learning Frameworks on Smart Grid Applications. Energies. 2023; 16(14):5326. https://doi.org/10.3390/en16145326

[25] R. Jin, Y. Zhou, C. Lu, J. Song, "Deep reinforcement learning-based strategy for charging station participating in demand response," Applied Energy, vol. 328, 2022, 120140, ISSN 0306-2619, https://doi.org/10.1016/j.apenergy.2022.120140.

[26] K. Park, I. Moon, "Multi-agent deep reinforcement learning approach for EV charging scheduling in a smart grid," Applied Energy, vol. 328, 2022, 120111, ISSN 0306-2619.