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

1st Kevin Chan
Department of Electronic and Electrical Engineering
University of Bath
Bath, United Kingdom
cyc241@bath.ac.uk

2nd Pedram Asef
Department of Mechanical Engineering
University College London
London, United Kingdom
pedram.asef@ucl.ac.uk

3rd Alexandre Benoit
Department of Electronic and 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 (MA-DRL) 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

As the world grapples with the looming crisis of global warming — where human activities have warmed the planet by approximately 1°C above pre-industrial levels as of 2017 with an increase of 0.2C per decade [9] — the need for sustainable solutions has never been more pressing. Particularly worrisome is the transportation sector, responsible for 14% of global greenhouse gas (GHG) emissions, with road transport alone accounting for 75% of these emissions [10]. Given the Intergovernmental Panel on Climate Change's assertion that a 1.5°C increase poses heightened risks to natural and human systems [13], a significant transformation in this sector is imperative [11].

Electric Vehicles (EVs) offer a promising pathway to mitigating environmental hazards. A wealth of research, including a review covering over 4,000 studies, shows that a transition to EVs could substantially curb GHG emissions [12]. Legislative initiatives like the European Parliament's 2035 ban on petrol and diesel cars further underscore the critical role of EVs [15].

While discussions around EVs often focus on light EVs, there is an unexplored potential in commercial electric vehicles (CEVs). The commercial vehicles (CV) market, burgeoning with more than 36 million registered vehicles in the EU as of 2022 [16], serves as a largely untapped avenue for impactful change. Yet, despite this potential, EV adoption in this sector remains disappointingly low [11]. Nowadays, CEVs include light commercial vehicles (LCV), heavy-duty trucks, and buses. Recent surveys have shared that the total share of electric LCVs exceeded the passenger EVs with an increase of 90% in 2022 representing 310 000 sold. Numbers are becoming very promising for bus sales accounting for 66k sold globally in 2022 [17].

Comprehending the factors behind the low adoption rate is essential for grasping the challenges confronting the CV industry. These challenges can direct the focus of research efforts in the field. Several such challenges are outlined:

- **Scarcity of Charging Stations in Remote Areas**: EVs in particular face a shortage of charging infrastructure, especially in remote areas. This limits their utility for businesses that require quick turnaround times [18] [19].
- **Impact on Vehicles with Tight Schedules**: The scarcity of charging stations and unpredictable charging times severely affect vehicles that operate on tight schedules, such as delivery vans and shuttle buses [11].
- **Increasing Demand for Electricity**: With more CVs transitioning to CEVs, the electricity demand will rise, adding strain to the existing grid infrastructure [20].

- **Integration of Renewable Energy Sources**: Renewable energies like wind and solar photovoltaic (PV) are variable and intermittent, making it challenging to match electricity supply with demand. This intermittency can cause grid instability and increased costs [20].
- **High Power Requirements of Charging Stations**: The current grid infrastructure may not be well-equipped to meet the high power requirements of charging stations. High-voltage connections are necessary and could place significant strain on local distribution networks [21].
- **Cost and Time-Consuming Upgrades**: Upgrading the existing electrical grid to accommodate these challenges can be both costly and time-consuming.
- **Need for Intelligent Energy Management Systems**: Traditional grid systems might not be able to optimize charging times or balance supply/demand efficiently [22].

The report WHAT REPORT? Citation Missing focuses on Vehicle-to-Grid (V2G) technology, which serves as a two-way energy link between EVs and the grid. This technology can stabilize energy supply and offers new ways to store renewable energy. Additionally, these smart systems facilitate demand response and load management through dynamic pricing and incentives for off-peak charging. V2G technology can also help stabilize the grid during periods of high demand or supply fluctuations [20] [23] [24].

Various areas for improvement within the realm of smart grids (SGs) have been identified, promising not only to enhance user experience but also to deepen our understanding of the technologies that address existing issues. The EMS stand as one viable solution applicable to SGs. For the integration with the CV sector, our focus is on optimizing charging schedules and enhancing grid efficiency, thereby making a significant contribution to the reduction of GHG emissions [22]. The EMS focuses on both technical and economic aspects, aiming to improve system performance and power quality [25].

Energy management strategies in EMS are broadly categorized into centralized optimization and autonomous operation [26]. Balancing the supply and demand in the context of CEV charging stations has been both extensive and intricate [27]. However, as will be discussed in the literature review, the emergence of advanced optimization techniques like deep dynamic learning (DRL) stands out as a superior approach for optimizing EMS. DRL exhibits capabilities for complex decision-making and adaptability to environmental changes [28]. It is particularly effective in managing uncertainties related to renewable energy and EV charging. Compared to traditional machine learning, DRL excels in scalability, real-time operation, and long-term optimization when applied to hybrid SGs. Applying such techniques is indirectly improving some other challenges, such as charging time [29][18], reducing the charging cost [30] and peak load [27].

Limitations of current EMS solutions mainly revolve around the unpredictability of EV charging patterns, leading to inaccuracies in system designs [21] [31] [32] [33].

This study focuses on creating an advanced Intelligent Energy Management System (EMS) that effectively manages the daily power equilibrium within a hybrid photovoltaic (PV)-grid-connected microgrid (HPVGT), particularly tailored for CEV depots. The methodology hinges on a multi-agent system strategy within dynamic programming (DP) learning, with a spotlight on PPO as a sophisticated technique framed within an actor-critic network, treating the HPVGT as the agent's environment. The novelty is translated in the ability to build a robust model without priori information on the State of Charge or Departure Time. We suggest adopting a dynamic scheduling technique, which will serve as a benchmark to gauge the effectiveness of our Deep Reinforcement Learning (DRL) method. Our multi-agent system is expected to provide a state-of-the-art answer to the CEV charging scheduling challenge, showcasing theoretical and practical accuracy in the application of RL.

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

The progression of EMS methodologies from EV to CEV is reviewed, with a specific emphasis on the optimization of charging scheduling techniques. Also, we will explore several reviews about PPO and DRL. Recent studies have extensively explored diverse machine learning strategies to enhance Energy Management Systems (EMS) for various applications microgrid [35], including Vehicle-to-Home (V2H), Grid-to-Vehicle (G2V), and Vehicle-to-Grid (V2G). Traditional approaches have also been employed, utilizing more direct and heuristic techniques to address challenges within EMS. This study [36] offers a solution to improve EV charging efficiency using particle swarm optimization (PSO). A mathematical model is presented to reduce the total charging time. Another research [37] focuses on developing a dynamic hunting leadership (DHL) algorithm for optimizing EV smart charging (EVSC) strategies in power grids with high EV presence. The core aim is to enhance the grid's voltage stability and avoid power supply issues. The DHL method, set against a backdrop of increasing EV integration, aims to harmonize charging schedules without overloading the grid and to ensure a consistent energy supply. The study in [34] presents a compelling case for adopting PPO, a DRL method, over DP for complex industrial optimization problems. The research shows that PPO, when applied to the multi-item stochastic capacitated lot-sizing issue, not only approaches the optimal solutions for small-scale problems but also significantly outperforms traditional benchmarks in larger, more complex scenarios where DP falls short. This superiority, coupled with the PPO's linear growth in computation time, underlines its practicality for large-scale industrial applications. In essence, the research [3] underscores that PPO is an advanced DP learning technique which excels in complex stochastic environments with large state and action spaces. PPO stands out due to its simplicity, effectiveness in sample use, and its ability to iteratively optimize through environmental interactions. PPO achieves a balance between reliability, sample efficiency, and computational simplicity, making it an optimal choice for challenging RL tasks.

In [38] the authors provide a comprehensive analysis of how DRL can be applied to optimize EV dispatch to improve the integration of renewable energy sources and power system operations. They explore both single-agent and multi-agent DRL algorithms, evaluating their effectiveness in various EV-related tasks like G2V, V2H and V2G. The article [39] discusses the need for intelligent coordination within the EV charging network to manage the impacts of vehicle electrification on the electricity grid. V2G deployment is examined through a detailed approach, which involves utilizing data from EV batteries and AI to conduct cost-benefit analyses. The paper [40] reviews the application of RL as a tool for optimal control in energy systems, including EV charging stations, with a focus on model-free RL techniques. It analyzes over 80 highly cited studies from 2015 to 2023, showcasing the versatility of RL in managing the complex dynamics of these systems. The paper emphasizes the potential of RL to address these challenges, offering adaptability and optimization capabilities for complex environments. By considering factors such as energy availability, user demand, and pricing. The researchers in [41] propose a method using DRL to cut EV charging costs by adapting to changing electricity prices and user behaviour. It applies an advanced LSTM network to process price data and the deep deterministic policy gradient algorithm to optimize charging times, demonstrating up to a 70.2% cost reduction compared to other methods. Similarly, [42] introduces a DRL method to minimize long-term PEV charging costs, addressing unpredictable factors such as user behaviour and energy prices. It employs a novel combination of model-based and model-free RL, updating from both real and simulated experiences. Deep neural networks replace traditional lookup methods for efficiency in vast state spaces. The technique, which includes price prediction via LSTM, outperforms existing charging strategies in simulations, optimizing policy and preventing SOC depletion more effectively. This study [43] addresses the challenges of EV charging loads causing potential grid overload by proposing a decentralized, incentive-based demand response coordination for EV charging stations. It models the charging process as a Markov decision process, using DRL algorithms like deep deterministic policy gradient to balance the grid's demand response revenue with user satisfaction and peak load reduction. In [44] the researchers address the complex issue of scheduling charging for multiple EVs in SGs. By leveraging multi-agent DRL, the study provides a decentralized approach that adapts to each EV's status, offering rapid decision-making suitable for real-world applications. The DRL model used outperforms existing heuristic methods. A comparative study [46] investigates the forecasting of EV charging loads using six machine learning models: RNN, LSTM, Bi-LSTM, GRU, CNN, and transformers. This study [47] develops and validates a learning-based EV charging strategy that accounts for diverse travel patterns and does not require future electricity price information. Utilizing a deep Q-network (DQN)-based RL approach, the findings show that EVs can save over 98% on electricity costs compared to immediate charging methods.

While existing studies emphasize charging schedules centered on cost, they tend to rely on prior data for constructing DRL models. The literature reveals a noticeable gap in research specific to the CEV industry, which has distinct objectives compared to non-commercial EVs. Commercial entities prioritize operational efficiency in their charging schedules to maximize service provision rather than merely reducing peak electricity costs in SGs. The scheduling strategies for CEVs, therefore, diverge considerably from those discussed in earlier sections. The unique demands of the commercial sector, such as night-time charging needs that align with daytime photovoltaic energy production, have not been adequately explored in current research. This oversight presents an opportunity, as the CEV sector is a significant contributor to emissions and faces particular challenges, including the necessity for fast charging and substantial energy requirements, which may not be fully met by existing SG approaches.

III. Problem Formulation/Methodology

Microgrid management is a complex endeavor due to its diverse responsibilities, such as managing voltage and frequency, distributing loads, coordinating distributed energy resources (DERs), and overseeing power exchanges with the main grid. These functions require different levels of attention and operate on varying time scales, which calls for a layered control framework that is significantly different for CEVs as opposed to passanger EVs.

1) Primary control:
   - Objectives: maintains voltage and frequency stability, especially after the microgrid transitions to an islanded operating mode.
   - Components: zero level control hardware for internal voltage and current control loops of the DERs.
   - Challenges: must provide independent active and reactive power-sharing controls for DERs in the presence of both linear and non-linear loads.
2) Secondary control:
   - Objective: compensates for the voltage and frequency deviations caused by the operation of primary controls.
   - Components: advanced control algorithms and communication system for real-time adjustments.
   - Challenges: requires fast response times and robust algorithms to adapt to changing conditions.
3) Tertiary control:
   - Objective: manages the power flow between the microgrid and the main grid and facilitates economically optimal operation.

- Components: advanced optimization algorithms and market-based control strategies.
- Challenges: must consider long-term objectives, e.g., cost optimization and grid resiliency.

While the foundational primary control systems are established, the primary focus of this paper is on the secondary control. Since aspects of the primary control exert influence over the secondary control, we will briefly discuss pertinent elements of the primary layer.

A. Dynamic Programming Approach

Dynamic Programming (DP) is a time-honored and commonly utilized method for orchestrating the charging of EVs [8]. We employ a DP-based EMS as a benchmark to evaluate the performance of our multi-agent PPO framework. This comparison aims to determine the effectiveness of sophisticated algorithms in orchestrating energy distribution within a commercial microgrid tailored for EV charging stations.

The procedural logic of the DP-based EMS takes into account several pivotal elements, providing a thorough strategy for the allocation of energy resources:

- Initial Input Evaluation: The process begins by collecting vital input data for each EV, including the arrival status, any scheduled departure times, and both the present and needed state of charge (SoC).
- Charging Duration Calculation: The system estimates the charging time necessary for each EV to reach its target SoC from its current level.
- Prioritization for Timed Departures: For CEVs with set departure times, a priority score is computed. This score inversely relates to the time remaining until the vehicle's departure, minus the estimated charging time needed.
- Constrained Optimization for Timing: The system tackles a restricted optimization issue akin to the classic knapsack problem. Its goal is to maximize the total priority score within the limits of power availability, ensuring that the power output from the utility service provider matches or exceeds the demand from the charging stations.
- Priority Setting for Unscheduled Departures: CEVs without specific departure times receive a priority level based on the inverse of their present SoC.
- Optimization Considering SoC Limitations: Utilizing another variant of the knapsack problem, the algorithm seeks to maximize the combined priority scores. This optimization is constrained by the power supply from the utility service provider, which must meet or surpass the demand at the charging stations after accommodating all EVs with scheduled departures.
- **Distribution of Energy Resources**: In the optimization phase, energy resources are allocated to each EV based on the results, managing the charging station switches.
- **Iterative Refinement**: 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.

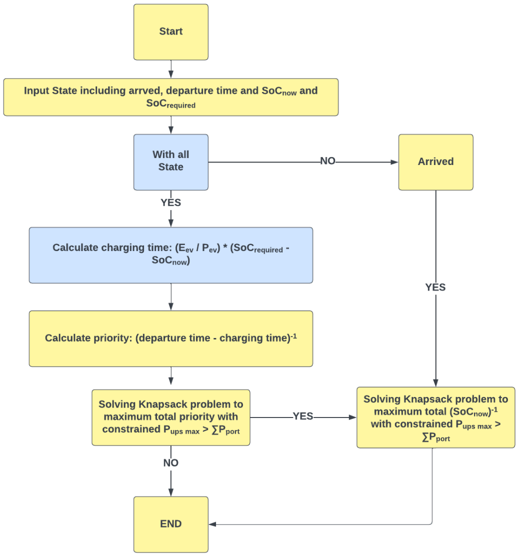

Start
Input State including arrved, departure time and SoC_{now} and $SoC_{required}$
With all State
NO
Arrived
YES
Calculate charging time: $(E_{ev} / P_{ev}) * (SoC_{required} - SoC_{now})$
YES
Calculate priority: (departure time - charging time)$^{-1}$
Solving Knapsack problem to maximum total priority with constrained $P_{ups\ max} > \sum P_{port}$
YES
Solving Knapsack problem to maximum total $(SoC_{now})^{-1}$ with constrained $P_{ups\ max} > \sum P_{port}$
NO
END

Fig. 1: Custom charging scheduling DP flowchart

B. Deep Reinforcement Learning

Our model harnesses the power of DRL, which synergizes the foundational methods of RL with the capabilities of deep neural networks. This integration equips agents with the proficiency to navigate and strategize within complex scenarios.

In the realm of DRL, an agent is methodically trained to interact with an environment to realize a specific aim. This process involves assimilating environmental cues and rewards, upon which the agent bases its strategic actions, directed by a well-defined policy. The intricacies of the policy and the neural network architecture are fine-tuned to handle the stochastic nature of the microgrid environment. In this algorithm, the agent operates within the confines of a microgrid system, undertaking the pivotal task of streamlining energy distribution. The agent's actions are calibrated to enhance the grid's operational efficiency, ensuring a balanced and sustainable energy flow.

The model is designed to navigate the complexities of the microgrid, learning from a diverse range of scenarios to achieve the highest echelons of distribution efficacy. Through iterative learning and adaptive decision-making, the agent aims to not only meet the immediate demands of energy distribution but also to anticipate and plan for future conditions, thereby optimizing the microgrid's performance over time. To illustrate, consider the agent's evolution through continuous learning cycles—each action is evaluated against the backdrop of environmental feedback, which shapes the subsequent decision-making processes. This dynamic learning environment encourages the development of robust strategies that can effectively respond to fluctuating energy demands and supply conditions.

C. 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 method). 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: the Actor-Critic method.

Now that we have seen the Actor Critic's big picture, let's dive deeper to understand how the Actor and Critic improve together during the training.

As we saw, with Actor-Critic methods, there are two function approximations (two neural networks):

- **Actor**: a policy function parameterized by θ: $\pi_\theta(s)$
- **Critic**: a value function parameterized by μ: V_μ

• Actor: This neural network outputs the policy, which is a distribution over actions. The actor takes the current state of the environment as input and outputs the probabilities of taking each possible action. • Critic: This neural network estimates the value of taking a particular action in a given state. It helps the actor to understand how good the action is in terms of future rewards.

Actor-Critic Interaction: The actor proposes actions, and the critic evaluates them. The critic's evaluations are used to update the actor's policy. This dual mechanism allows for more stable and faster convergence compared to traditional methods.

D. Multi-Agent Reinforcement Learning (MARL)

Our model is based on a multi-agent system in a decentralised environement, it means that no information is shared between the agents. It simplifyes 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.

E. Proximal Policy Optimization

The core strength of our approach lies in the implementation of Multi-Agent Reinforcement Learning (MARL) combined with the Proximal Policy Optimization (PPO) technique. This method excels in scenarios lacking prior knowledge, a common occurrence in commercial Electric Vehicle (EV) charging systems where user patterns are unpredictable.

PPO stands out in the realm of advanced Deep Reinforcement Learning (DRL) algorithms for its ability to enhance the learning process's efficiency and stability. It operates on dual neural networks—the actor, which determines the policy, and the critic, which assesses the value function [4].

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 the equation described underneath:

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

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 question 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 behavior 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.

The components of the equation function as follows:

- **Unrestricted Component**: This element functions as an alternative to the logarithmic probability typically employed in the policy's objective function, multiplying the probability ratio by the advantage estimate. Nevertheless, in the absence of limits, a scenario where the selected action is significantly more likely under the current policy than the previous one can cause an oversized step in the policy gradient, leading to an over-adjustment of the policy.
- **Constrained Component**: By applying a cap, the equation restricts the extent of policy modification, maintaining the new policy within a bounded range of deviation from the old one to prevent overly dramatic changes.

F. Reward Function Design

The reward function is engineered to align agents' behaviours with the specified objectives. The reward function comprises:

- Charging percentage on EVs over it's SoC:

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

- Voltage Direct Current (VDC):

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

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

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

- Penalisation metric for failing to achieve requisite SoC levels within specified timeframes

$$W_3 \times N_{NC} \times \text{Penalty} \tag{6}$$

G. Trust Region Constraints

Proximal Policy Optimization (PPO) incorporates a trust region limitation to maintain policy updates within a certain range. This precaution ensures that newly adopted policies don't deviate excessively from previous ones, thereby avoiding radical changes that might disrupt the stability of the learning progression [3].

H. Value Decomposition Networks into a MultiAgent DRL

In the expansion of the microgrid's capabilities, particularly with the addition of more charging ports, the system's complexity and the size of the action space have increased significantly, leading to a demand for a sophisticated Multiagent Deep Reinforcement Learning (MDRL) approach. This approach scales traditional deep reinforcement learning to complex, multi-agent environments, preparing the system for future growth and complexity [4]. Utilizing Proximal Policy Optimization (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 MDRL 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 [5]. 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 [6].

A representation of our model architecture can be seen in Fig 2

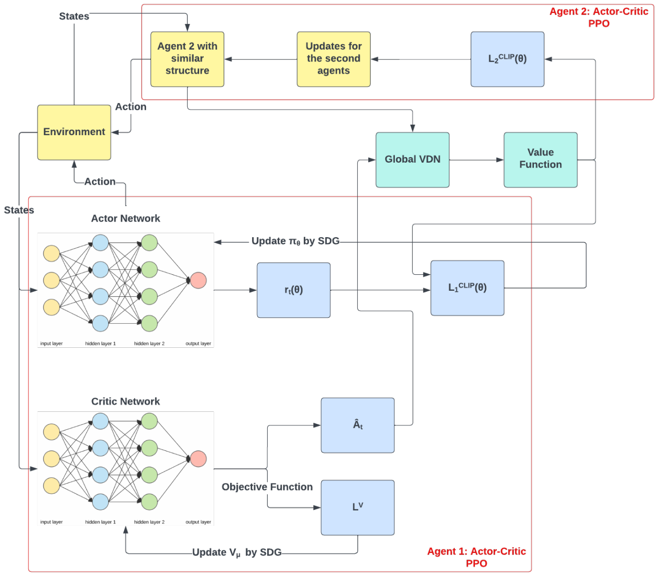


Fig. 2: Overall Architecture of our MARL system (implemented with 2 agents) using an Actor-Critic PPO process with the use of VDN [7]

IV. Experimental Set-up and Implementation

The model described has been trained using the Matlab Reinforcement Learning ToolBox where the model environments (described after) interacts seamlessly with the RL agents using MATLAB and SIMULINK. 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. Overcoming Information Gaps in Commercial EV Charging

In CEV charging scenarios, the predictability of user behavior is often uncertain, presenting a unique challenge for energy management systems. Traditional approaches rely heavily on prior information to function optimally, but our methodology breaks away from this constraint.

Our MARL framework, powered by PPO, demonstrates robust performance even in the absence of prior information. This is a significant advantage for commercial EV infrastructures where user behavior and charging patterns can be unpredictable and varied.

B. Optimizing the Training Environment for MARL PPO Agents

Our 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.

1) Environment Setup and Agent Objectives: The simulated microgrid environment 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 Energy Management System (EMS).
- The collective goal of these agents is twofold: ensuring that each EV achieves the required State of Charge (SoC) by the predetermined departure time, and reducing reliance on the central power grid.

2) Observation and Action Spaces: To facilitate these objectives, our agents operate within a well-defined observation and action space:

Observation Space:

- SoC levels of the UPS and each EV, along with the EVs' scheduled departure times.
- A set of thirteen floating-point values representing the percentage of power drawn from the central grid, formalized using MATLAB's `rlNumericSpec` object to allow for 13-dimensional float number inputs.

Action Space:

- Boolean variables representing the on/off status of each charging port.
- A Boolean switch to regulate the connection to the central grid, instantiated using MATLAB's `rlFiniteSetSpec` object, which encapsulates all possible actions.

Through this strategic environment setup, we ensure that our agents are trained in conditions that reflect the complexities of real-world commercial EV charging infrastructures while remaining computationally manageable. accommodates all potential actions.

C. Agent Architecture and Training Parameters

The input to the Actor network is the set of observations, and it yields a distribution of probabilities across various actions. Initially designed with a three-layered hidden structure comprising 120, 60, and 30 neurons [2], this setup did not successfully achieve convergence in initial tests.

In tandem with the Actor, the Critic network assimilates the state of observations and the actions performed, generating a predictive score for potential returns. Originally, it replicated the structural design of the Actor.

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.

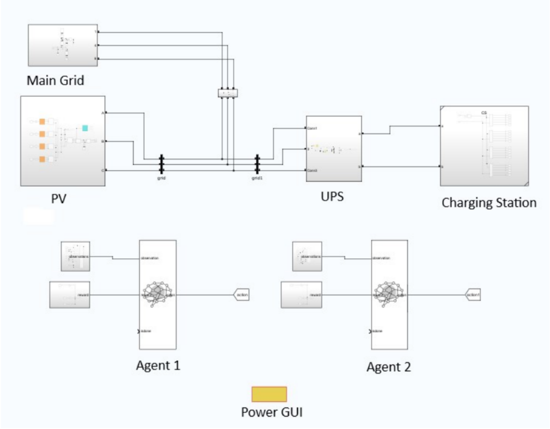


Fig. 3: Overall System Representation within Simulink

V. Experimental Results and Validation

In this analysis, we investigate the learning progress of PPO agents tasked with optimizing energy distribution in a microgrid setting. 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 4 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.

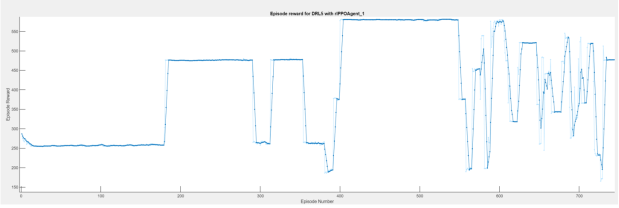


Fig. 4: Single Agent with Wide Trust Region

B. Single Agent with Narrower Trust Region

Fig 5 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.

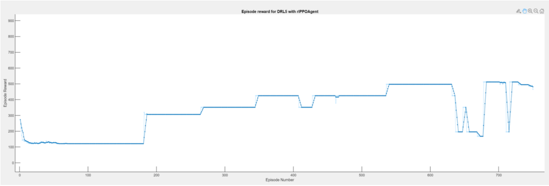
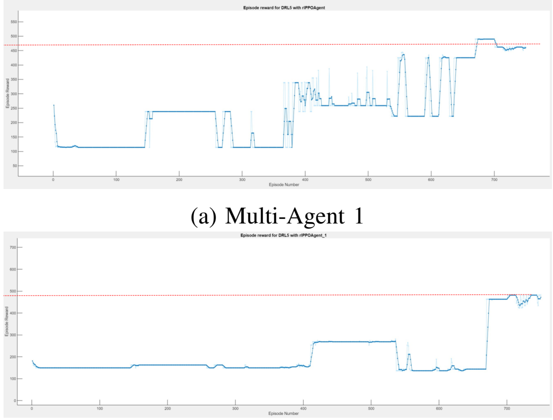


Fig. 5: Single Agent with Narrow Trust Region

C. Multi-Agent Training

Fig 6 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 converges 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.



(a) Multi-Agent 1

(b) Multi-Agent 2

Fig. 6: Multi-Agent Training Result

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.

D. Comparative Performance Analysis with DP Method

The concluding segment of the results analysis offers a side-by-side evaluation of the MARL using PPO against the DP strategies. This assessment uses several key indicators to ascertain the efficiency of each method in regulating the distribution of energy within the microgrid system. Crucial to this analysis is the inclusion of prior information such as anticipated energy demands and scheduled departure times. The findings are encapsulated in Table I and detailed as follows:

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	85	91	85	88

- **Performance**: 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.
- **Computation Time**: All assessed methodologies demonstrated rapid computational speeds, suggesting their potential suitability for applications necessitating immediate decision-making.
- **Penalty for Insufficient Charging**: 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.
- **Training Time**: 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.
- **PV to Total Energy Consumption Ratio**: Harnessing prior information, the MARL PPO method achieved an impressive 91% ratio of photovoltaic (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.

This is where you put your graphs, analyse and explain whatever behaviour

VI. Conclusion

Our study heralds a pioneering advancement in microgrid management through the strategic application of Deep Reinforcement Learning (DRL) within a multi-agent system, addressing the critical day-night energy imbalance with remarkable efficacy. The implementation of our Multi-Agent Reinforcement Learning (MARL) with Proximal Policy Optimization (PPO) agents demonstrated a striking 95% efficiency, far exceeding the capabilities of the traditional Dynamic Programming approach, which only managed a 45% effectiveness rate. This significant performance gap emphasizes the enhanced robustness and adaptability of our MARL PPO approach, particularly in managing the energy output of photovoltaic (PV) systems during peak daylight and optimizing its use for commercial EV charging after dusk.

The DRL's sophisticated algorithms for secondary control in our system have been integral in ensuring that the renewable energy harnessed during daylight does not go to waste, instead contributing to a more balanced, cost-effective, and grid-independent energy consumption during the night. This showcases the potential of DRL to significantly reduce grid dependency and operational costs.

Additionally, our work excels in the application of PPO agents for advanced multi-agent cooperation thanks to the VDN implementation, establishing new frontiers of adaptability and efficiency for smart microgrid management. The thorough comparative analysis reinforces the unique value of our approach in real-time energy systems, especially for scenarios that lack historical data, which is often the case in commercial EV charging station management.

In summary, our research consolidates its place at the forefront of energy management solutions, streamlining operational efficiency while optimizing sustainability. It proposes a method that is not just cost-effective but is equipped to adapt to and address the complex demands in the fast-growing commercial EV sector. The DRL-based methodologies we've developed offer actionable insights, positioning our study as a blueprint for future innovation in sustainable microgrid operations.

Acknowledgment

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] LaFarge, Nicholas & Miller, Daniel & Howell, Kathleen & Linares, Richard. (2020). Guidance for Closed-Loop Transfers using Reinforcement Learning with Application to Libration Point Orbits. 10.2514/6.2020-0458.

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

[4] 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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Haustein, K. et al. (2017) 'A real-time Global Warming index', Scientific Reports, 7(1). doi:10.1038/s41598-017-14828-5.

[10] 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.

[11] Frank, S. et al. (2023) Built for purpose: EV adoption in light commercial vehicles, McKinsey &amp; 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).

[12] 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.

[13] 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.

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

[15] 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

[16] European Automobile Manufacturers' Association. "Vehicles in use europe 2023." (2023), [Online]. Available: https://www.acea.auto/files/ACEA-report- vehicles-in-use-europe-2023.pdf

[17] Virta Ltd. (2023) The Global Electric Vehicle Market in 2023 – virta, Virta Global. Available at: https://www.virta.global/en/global-electric-vehicle-market (Accessed: 05 November 2023).

[18] 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

[19] V. S. Patyal, R. Kumar, and S. Kushwah, "Modeling barriers to the adoption of electric vehicles: An indian perspective," Energy, vol. 237, p. 121 554, 2021. DOI: 10.1016/j.energy.2021.121554

[20] 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)

[21] 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

[22] H. T. Mouftah, M. Erol-Kantarci, and M. Husain Rehmani, Transportation and Power Grid in Smart Cities: Communication Networks and Services. Wiley, 2019, pp. 293–312. DOI: 10.1002/9781119515154.ch11

[23] M. Meliani, A. E. Barkany, I. E. Abbassi, A. M. Darcherif, and M. Mahmoudi, "En- ergy management in the smart grid: State-of-the-art and future trends," Inter- national Journal of Engineering Business Management, vol. 13, 2021. DOI: 10 . 1177/18479790211032920

[24] L. Jian, Y. Zheng, X. Xiao, and C.-C. Chan, "Optimal scheduling for vehicle-to-grid operation with stochastic connection of plug-in electric vehicles to smart grid," Applied Energy, vol. 146, 2015. DOI: 10.1016/j.apenergy.2015.02.030

[25] A. Ovalle, A. Hably, S. Bacha, and M. Ahmed, "Voltage support by optimal inte- gration of plug-in hybrid electric vehicles to a residential grid," Oct. 2014. DOI: 10.1109/IECON.2014.7049169

[26] 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)

[27] 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

[28] 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

[29] T. S. Bomfim, “Evolution of machine learning in smart grids,” 2020. DOI: 10.1109/ SEGE49949.2020.9182023

[30] Q. Zhang, H. Li, L. Zhu, et al., “Factors influencing the economics of public charg- ing infrastructures for ev–a review,” Renewable and Sustainable Energy Reviews, vol. 94, 2018. DOI: 10.1016/j.rser.2018.06.022

[31] L. Yao, W. H. Lim, and T.-S. Tsai, “A real-time charging scheme for demand re- sponse in electric vehicle parking station,” IEEE Transactions on Smart Grid, vol. 8, no. 1, 2017. DOI: 10.1109/TSG.2016.2582749

[32] L. Jian, Y. Zheng, and Z. Shao, “High efficient valley-filling strategy for centralized coordinated charging of large-scale electric vehicles,” Applied Energy, vol. 186, 2017. DOI: https://doi.org/10.1016/j.apenergy.2016.10.117

[33] Y. Wu, A. Ravey, D. Chrenko, and A. Miraoui, “Demand side energy management of ev charging stations by approximate dynamic programming,” Energy Conver- sion and Management, vol. 196, 2019. DOI: 10.1016/j.enconman.2019.06. 058

[34] 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

[35] 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.

[36] 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.

[37] 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

[38] 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

[39] Nagaraju D, Kumar GN, Kumar SS, Suresh V, Prasad KMVV, Hossam K, AboRas KM. Impact of Plug-in Electric Vehicles on Grid Integration with Distributed Energy Resources: A Review. Frontiers in Energy Research. 2023; 10. DOI: 10.3389/fenrg.2022.1099890

[40] 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

[41] Li, Sichen, et al. ”Electric vehicle charging management based on deep reinforcement learning.” Journal of Modern Power Systems and Clean Energy 10.3 (2021): 719-730.

[42] F. Wang, J. Gao, M. Li and L. Zhao, ”Autonomous PEV Charging Scheduling Using Dyna-Q Reinforcement Learning,” in IEEE Transactions on Vehicular Technology, vol. 69, no. 11, pp. 12609-12620, Nov. 2020, doi: 10.1109/TVT.2020.3026004.

[43] 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.

[44] 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.

[45] A. Hafeez, R. Alammari and A. Iqbal, ”Utilization of EV Charging Station in Demand Side Management Using Deep Learning Method,” in IEEE Access, vol. 11, pp. 8747-8760, 2023, doi: 10.1109/ACCESS.2023.3238667.

[46] Koohfar S, Woldemariam W, Kumar A. Performance Comparison of Deep Learning Approaches in Predicting EV Charging Demand. Sustainability. 2023; 15(5):4258. https://doi.org/10.3390/su15054258

[47] Hao X, Chen Y, Wang H, Wang H, Meng Y, Gu Q. A V2G-oriented reinforcement learning framework and empirical study for heterogeneous electric vehicle charging management. Sustainable Cities and Society. 2023; 89:104345. https://doi.org/10.1016/j.scs.2022.104345.