

Safe Reinforcement Learning-based System for Connected and Autonomous Vehicle Charging Infrastructure

by

Md. Saharan Evan
20201020

Akil Rahman Efad
20201041

Nusrat Jahan Shukti
21101003

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2024

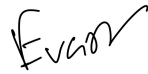
© 2024. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:



Md. Saharan Evan
20201020



Akil Rahman Efad
20201041



Nusrat Jahan Shukti
21101003

Approval

The thesis titled “Safe Reinforcement Learning-based System for Connected and Autonomous Vehicle Charging Infrastructure” submitted by

1. Md. Saharan Evan
2. Akil Rahman Efad
3. Nusrat Jahan Shukti

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 27, 2024.

Examining Committee:

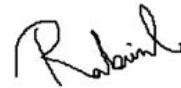
Supervisor:
(Member)



Dr. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Program Coordinator:
(Member)



Dr. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor; Chairperson
Department of Computer Science and Engineering
School of Data and Sciences
Brac University

Abstract

This paper is about creating a system that helps to manage the charging of electric vehicles that are connected and can drive autonomously taking into consideration the safe reinforcement learning outcomes in this process. The system is regarded as an intelligent decision support system (IDSS). In this system, a holding corporation that works the whole charging infrastructure, installs charging equipment for both regular electric vehicles driven by humans and autonomous vehicles. The problem arises when human-driven vehicles ask for more charging time and energy than they really need and to success charging request competition, which can particularly lead to cause issues. To address this problem, a proposed solution aims to make sure the charging equipment is used efficiently minimizing the risk of not having enough power available as well as considering all the safety of the charging equipment. Here a system will be introduced where it encourages human-driven vehicles to make rational charging requests based on data and noting down the parameters which are the number of DSOs (Distribution System Operator), the nearest finding of EVSE(Electrical Vehicle Supply Equipment), the association among the EVSEs, the starting and ending time of the plugin, energy absorption, time duration, request for charging for the CAV, CV and AVs. Furthermore, the introduction of a learning system where the charging equipment learns how to schedule charging sessions based on the procession from the main operator or the distribution system operator. The conducted experiments will show that this system improves the charging rate, active charging time, and energy usage compared to existing systems ensuring all the protection of the electrical and connected autonomous vehicles. Therefore, the study will contribute to making transportation systems smarter and addressing the challenges and safeties of connected and autonomous vehicles.

Keywords: Safe Reinforcement Learning, EV, CAV, AV, CAV, EVSE Charging System

Acknowledgement

First of all, we acknowledge Almighty Allah, without whose help our thesis may not have been completed.

Second, thanks go to kind support and guidance during our effort, Md. Golam Rabiul Alam, Ph.D., our advisor. He gave us support whenever we needed assistance. And at last, without our parent's unwavering support, it could not have been possible.

Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	1
1 Introduction	2
1.1 Problem statement	3
1.2 Research Contribution	4
1.3 Thesis Organization	5
2 Related Work	6
3 Proposed Method	17
3.0.1 Actor-Critic Method	18
3.0.2 Centralized Training Decentralized Execution	19
3.0.3 Centralized Attentive Critic	19
3.0.4 Integration of Future Charging Competition	20
3.0.5 Multi-Objective Enhancement	22
3.1 Laxity Estimation	24
3.2 Conditional Value at Risk	25
4 Preliminary analysis	27
4.1 Requested Charge vs Actual Delivered Charge	27
4.2 Laxity Sum of Each EVSE	28
4.3 Laxity(Sum) vs PDF(Sum) of each EVSE	29
5 Performance Evaluation for SRL-CAVCI	31
6 Conclusion	36
Bibliography	39

List of Figures

3.1	Top Level Overview of the proposed SRL-CAVCI	17
3.2	Decentralized execution of active agents	21
4.1	Requested Charge vs Actual Delivered Charge	27
4.2	Laxity of Each EV	28
4.3	Laxity Sum of Each EVSE	28
4.4	Summary of Laxity Sum of Each EVSE	29
4.5	Laxity(Sum) vs PDF(Sum) of each EVSE	29
4.6	Summary of Laxity(Sum) vs PDF(Sum) of each EVSE	30
5.1	Comparison of Different Algorithms	33
5.2	Regression Rate: Success Rate vs Reward	34
5.3	Distribution of n_rec and n_recsc	34
5.4	Comparison of Success Rates and Rewards	35
6.1	Pair plot of (Success rate, Derived success rate and Reward) Vs In- dication	36

Chapter 1

Introduction

The integration of connected and autonomous vehicle (CAV) technology has become a significant catalyst in the pursuit of smarter and more efficient urban living in an era dominated by intelligent transportation systems (ITS). The significance of integrating connected vehicles (CVs), as well as autonomous vehicles (AVs) into the connected and autonomous vehicle (CAV) ecosystem, is of utmost importance as we move away from conventional vehicles. It is projected that by the year 2040, a significant proportion of ICE automobiles, probably 55 percent, will be substituted by electric vehicles (EVs). This transition is expected to necessitate roughly 350 terawatt-hours (TWh) of electric power to adequately fulfill the energy requirements of these EVs. The management of electric vehicle supply equipment (EVSE) has become a crucial design challenge for adequately meeting the enlarging energy quantity for CAV charging systems. The effective management of energy in charging systems for connected and autonomous vehicles (CAVs) heavily relies on the adoption of electric car charging scheduling strategies that optimize the utilization of accessible power resources. This undertaking is mostly based on rational decision-making procedures, with the objective of selecting options that result in the highest possible advantage else usage, whether for a singular or for a whole system. The concept of a logical way of behaving posits that individuals engage in behaviors with the intention of maximizing their benefits, whereas any actions that deviate from this objective are considered irrational. Enhancing the effectiveness of electrical resource use among Electric Vehicle Supply Equipment (EVSEs) is the main goal of this study, with the goal of benefiting distribution system operators (DSOs). However, human-operated connected automobiles (CVs) provide challenges to the quest of reasonable electric vehicle (EV) charging scheduling. These CVs usually lead to irrationality concerning the amount of energy required and the duration of charging. Unlike autonomous cars (AVs), which rely on user input and can thus be subject to inaccurate demands, conventional vehicles (CVs) rely on automated analysis to accurately estimate and request specific quantities of energy and charging durations. Autonomous cars frequently have a propensity to require more energy and longer charging times than are really required, which results in less efficient use of energy. It is feasible to include a rational decision support system (RDSS) into the CAV charging infrastructure as a means of addressing the problem of rationalizing EV charging (CAV) systems. Individual electric cars (both conventional and autonomous) and their electric vehicle supply equipment (EVSE) responses to energy supply demands are considered by the RDSS. By examining and categorizing irrational behaviors

related to energy demand and supply, and measuring them as a tail risk within the context of the charging system for Connected and Autonomous Vehicles (CAVs), it becomes feasible to build a robust correlation between the behaviors of energy demand and supply inside CAV charging systems.

The main purpose of the study is to present the Safe Reinforcement learning-based system examining the risk associated with illogical demand of energy as well as supply in the context of Connected and Autonomous Vehicles in Charging Infrastructure (CAV-CI). The aim is to improve the efficiency of energy management. The safe RL-based system will capture the behavioral characteristics that arise from the parameters of energy demand and supply in connected and autonomous vehicles (CAV-CI). It utilizes data-driven methods to effectively handle rationality. This research makes substantial contributions in the following important areas: Developing a Intelligent Decision Support System (IDSS): The study will introduce a Intelligent decision support system (IDSS) designed for scheduling electric vehicle (EV) charging sessions in connected and automated vehicle (CAV) charging infrastructure (CI). This IDSS considers the unpredictable tail-risk of CV-AV-EVSE interactions. This method will assess remissness risk for each electric vehicle (EV) in the CAV-CI architecture. Optimizing rational rewards in Connected and Automated Vehicles with Cooperative Intelligence is the focus of this study. This research aims to lessen the hazards of negligence. The research advises using simulate-driven and intuitive methods to attain this goal. A consolidated risk adversarial agent (RAA), local self-learning agents for each Electric Vehicle Supply Equipment will make up the system. The autonomous EVSE-LAs will acquire knowledge of Electric Vehicle Supply Equipment (EVSE) systems. These agents will decide on energy supply, charging rate, and charge duration. Lack of diligence risk, determined from Conditional Value at Risk tail distributions, affects decision-making. In CAV-CI, the system will utilize a actor-critic architecture to communicate between EVSE-LAs and RAAs. As autonomous learners, EVSE-LAs set their own EV session scheduling policies while the RAA is the main teaching center. Moreover, each EVSE will be assigned a learning agent which is EVSE-LA that will update the policy of each EVSE and based on the behavior of the vehicle, scheduling indicator will perform. As our main goal is to increase the efficiency of energy utilization, the proper scheduling policy can help to increase the energy utilization.

1.1 Problem statement

A robust and scalable system that can effectively manage and optimize the charging process, while accommodating a wide range of user preferences and ensuring grid stability, is the main challenge in the context of safe reinforcement learning-based electric vehicles and connected autonomous vehicles (EVs and CAVs) charging infrastructure. Additionally, it aims to address important problems like managing irrational charging requests, success in compitative charging request based on scheduling indicator, laxity of each EVs, CVaR of each time session, and finding the best policy for the EVSEs and lastly proper indicator for each of the EVs. To build a secure and effective charging infrastructure for electric and linked autonomous vehicles, these issues must be resolved. In addition to encouraging the widespread use of AEVs and CAVs, addressing these issues would help create a more dependable and sustainable transportation ecology.

The main definitions used in the paper and the formal statement of the EV Charging Indicator problem are presented in this section.

Consider a set of N charging stations $\mathcal{C} = \{c^1, c^2, \dots, c^N\}$, By considering each day as an episode, we start by defining a charging request as follows.

Definition 1 *Charging Request of EVs:* A charging request $q_t = \langle l_t, T_t, T_t^c \rangle \in Q$ represents the t -th request (i.e., step t) in a single day. In this context, l_t indicates the location where the charging request q_t is made, T_t is the specific real-world time when the request is initiated, and T_t^c is the real-world time when the request is concluded. A charging request is considered complete if the vehicle either successfully charges or ultimately fails to do so (i.e., the vehicle abandons the attempt to charge). The notation $|Q|$ represents the total number of charging requests within the set Q . Furthermore, we may use q_t interchangeably to refer to the electric vehicle making the charging request q_t to simplify the discussion.

Definition 2 *Waiting period for charging (WPC):* The Waiting period for charging is defined as the subtraction between the requested location time l_t to done charging time of q_t to the target charging station $EVSE^i$.

Definition 3 *Cost of charging (CC):* The cost of charging is stated to be the cost per kWh of any specific EVSE. Usually, this would include a cost for electricity and a service fee.

Definition 4 *Failure rate of charging (FRC):* The total number of charging request came to the DSO and who accepted the indication and failed to charge, the ratio is considered as failure rate.

Problem 1 *Electric Vehicle (EV) Charging indicator :* In each day there are many request for charging Q a DSO gets, the porpose of the research is to provide indication each $q_t \in Q$ to the most proper charging station r_{evse} in $\|EVSE\|$, based on various aspects with the intention is to optimize using Safe RL approach to achieve the long-term goals of continuously minimizing the overall WPC, average CC, and the FRC for the electric vehicles $q_t \in Q$ who accept the indicator.

1.2 Research Contribution

In order to construct intelligent and adaptable electric cars and connected autonomous vehicles (EVs and CAVs) charging infrastructure, the main goal of this research is to use actor-critic network to find the best policy for each EVSE which is considered as an individual learning agent so that it can meet the requirements of each EVs and safe reinforcement learning techniques to optimize the actor-critic network. The following goals will be emphasized as this infrastructure strives to thoroughly address the highlighted challenges:

- First challenge is to handle the large state and action space which will create problem because of huge numbers of publicly available stations. Directly centralized learning will induce many inefficiency and scalability problem.

- Second challenge is to calculate the reward function, we used a safe reinforcement learning process so that we can skip the risk for the best policy among all the EVSE and to calculate the initial policy for each EVSE.
- Thirdly the most significant challenge is to collaborate between EVSEs so that only one EVSE can serve a single EV and rest of the EVSEs should wait for the better indicator.
- To execute the framework, a linear model has been used and this model has also been used to calculate the scheduling indicator which is dependent on the framework.

By attaining these goals, our research will help to establish a safe, efficient, and long-term charging infrastructure for connected and autonomous cars, supporting their wider acceptance and inclusion into future transportation systems.

1.3 Thesis Organization

In this section we have discussed about the organization of this research that how this explained based on the chapter.

- Chapter 2 discussed about related works and domain and the existing work.
- Chapter 3 introduced about the proposed Method and the Formulation of the architecture which described the algorithms along with the training and execution process.
- Chapter 4 described about the preliminary analysis of data which proposed to use in optimization considering hard constraint.
- Chapter 5 discussed performance Evaluations for SRL-CAVCI method compared with some baselines.
- Chapter 6 summarize and conclude the research along with proposing future work.

Chapter 2

Related Work

In the research paper [16] the authors embark on a journey to innovate and improve the design of charging stations for autonomous vehicles (AVCS) by employing a scientometrics-based approach. They began by analyzing past designs and their inherent challenges, which they gleaned from a meticulous review of English articles in the Lens database. The primary issues pinpointed from prior designs encompassed the diverse and non-standardized charging interfaces across different vehicle manufacturers and the inefficiencies of existing wireless charging solutions. Their novel design proposal seeks to address and rectify these challenges. The new AVCS concept is envisioned as a green energy solution, evidenced by its solar chargers and the adoption of a leaf-shaped design for the frame, emphasizing its commitment to sustainability. The station is planned to be versatile, accommodating various charging methods including underbody wireless, traditional plug-in methods, and even battery-swapping capabilities. To further refine the user experience, a communication system is proposed to discern the type and needs of an arriving vehicle. A standout feature is the robot arm, designed with a five-degree of freedom, ensuring it can flexibly connect to various vehicle charging ports. Through the combination of these features, the authors aim to greatly elevate the convenience, efficiency, and sustainability of AV charging stations.

Energy-Saving Local Route Scheduling for a Self-Directed Car Taking into Account the Suggested Load Position One of the main factors influencing the energy consumption of Self-Guided Vehicles (SGV) is the local path planning phase of navigation. This paper [14] suggests a way to use load position to increase the energy efficiency of the local path planning step. The results of the study show that compared to a general planner, the recommended one generates faster and more efficient routes across corridors and around obstacles. Thus, taking into account the load effect reduces the energy usage of SGV. They employed two models to do this objective. In order to take into account the change in the SGV's Centre of Mass (CoM) caused by the load properties, a kinetic model of the differential drive SGV is first created. Second, two learning models for online estimation of the position of CoM (PoCoM) and prediction of necessary energy of sample trajectories are created using machine learning techniques. As a result, the learning models are trained using the SGV's generated kinetic model. employing a dynamic model of SGV, creating a dataset for machine learning techniques. To comprehend torque requirements under diverse circumstances, the dataset comprises a variety of scenarios with variable in-

ertial characteristics and reference velocities. The dataset contains information on motor torques, angular and linear velocities, and PoCoM coordinates, together with additional Gaussian noise to account for errors in the industrial context. There are two learning models employed, and each has unique input and output features. In a controlled lab setting, the experimental validation of simulations is discussed using an industrial SGV. The SGV is put through its paces in several scenarios where loads are carried through waypoints and unforeseen barriers are encountered. The purpose of these tests is to evaluate how different load scenarios affect SGV’s energy usage. There are six distinct load locations and mass attempts made. 28 times are added to each attempt for a total of almost an hour of continuous mobility while taking location and mapping uncertainties into account. The findings demonstrate that the DWA approach, which is energy-efficient, produces smoother routes and optimizes trajectories around obstructions. These trials’ specifics and findings—including load information, trip distances, energy use, and energy efficiency comparisons between the suggested technique and the general DWA—are outlined. This paper presents a method for SGV path planning that is energy-efficient. It builds a dataset, trains two machine learning models for CoM and torque estimates, and constructs a kinetic model that takes into account the weight and position of the load. These models are included into the Dynamic Window Approach (DWA) for navigation, which improves energy efficiency over conventional DWA. This strategy supports real-time operations by optimizing SGV movements while accounting for load deployments in corridors and obstacle avoidance.

Alighanbari, S.et al. (2021) [10] describes Reinforcement learning (RL) plays a key part in allowing intelligent decision-making for autonomous cars, and the potential of autonomous driving to minimize traffic accidents brought on by human mistake is highlighted. With an emphasis on employing Model Predictive Control (MPC) as a filter to direct an RL agent’s exploration, the specific topic addressed is safe exploration in RL. The Deep Deterministic Policy Gradient (DDPG) agent’s exploration in autonomous driving situations is improved by the introduction of a Novel Model Predictive Control (NMPC) filter. In comparison to the baseline DDPG technique, the NMPC filter greatly enhances the performance of the DDPG agent, resulting in a large rise in the mean reward.confirming the heuristic rules’ success in directing the DDPG agent, even if they restrict the exploration space and produce better incentives. Utilizing SUMO as a traffic simulator and hybrid testing, which involves assessing a real car in an enclosed space using simulation scenarios and sensor inputs, is the co-simulation framework for the development and performance verification of autonomous vehicles in crucial situations. This study shows that noise and uncertainty have a limited effect on automobile systems, negating the need for robust approaches. illustrating the advantages of adaptive learning, which results in better rewards and less overfitting by just using one training sample per trajectory. The report makes indicator s for future research directions, including the incorporation of more realistic vehicle dynamics, steering and acceleration control, and the use of more accurate simulation settings and sensor data for training. Additionally, it recognizes a brand-new class of hazards known as automation risks that are connected to autonomous cars, placing a strong emphasis on the necessity of continuing research and development to solve technological, sensor-related, and safety issues in autonomous driving.

Author Kim et. al.[18] proposed the trajectory planning and control technique presented in this study for autonomous cars operating in multi-vehicle complicated urban situations. For safe and efficient trajectory following and obstacle avoidance in urban conditions, the integration of motion planning and control components with an emphasis on lateral and longitudinal MPC. It was tested on real world vehicles. The algorithm makes use of the ideas of a "free spaces" and "safe drivable envelope" to evaluate the operable zone in urban driving circumstances and to deal with on-road impediments efficiently. In order to keep a safe distance from earlier cars, for example, The velocity planner produces reference transverse and lateral conditions for the ego vehicle. In the event of in-lane obstacle avoidance, the lateral motion planner selects the desired lateral offset to align with the center of the drivable envelope. If there is a possibility of a side lane accident and the lane's free space is limited, the vehicle may undertake an evasive movement by entering the side lane or coming to a halt behind the obstruction. The vehicle may slow down to continue a safe space when the object completely fills the lane or try to change lanes depending on the side lane risk assessment. Model Predictive Control (MPC) issues that are both longitudinal and lateral in nature can be solved to provide the control inputs necessary for monitoring the reference states with safety assurances. Actual car experiments have proven the practicality of this motion control and planning system for urban autonomous driving. Future research aims to expand the suggested frame to different driving situations in cities, such as junctions that have both signals and no signals, where the algorithm would need to actively react to merging vehicles and anticipate the intentions of preceding vehicles, resembling human behavior in avoidance maneuvers.

The paper [17] introduces an advanced reinforcement learning framework aimed at enhancing safety and recovery during autonomous robot navigation. Utilizing a grid-world environment of various sizes for experiments, the authors propose two core methodologies: a Safety Shield and a Self-Recovery Mechanism. The Safety Shield acts as a filter to prevent the robot from taking risky actions, whereas the Self-Recovery Mechanism allows the robot to revert to a prior safe state should it encounter obstacles. These components were integrated into a Safe and Self-Recoverable Reinforcement Learning (SSRL) framework, which was then compared to traditional Q-learning algorithms. Results showed that SSRL not only converged faster but also registered fewer collisions with obstacles. Furthermore, the paper also addresses the framework's practical implications in real-world challenges like deep-sea and cave explorations, where resetting to an initial state may not be feasible, and conditions like lighting and terrain may change over time. Overall, SSRL demonstrated improved safety and efficiency, with the authors suggesting future work on enhancing the predictive capabilities of the Safety Shield and adapting the framework to dynamic environments.

Ge, Y. et al [7] addressed when utilizing reinforcement learning in industrial applications, safety issues are a major problem. Traditional approaches make an effort to alter the agent's goals and exploration techniques, but frequently fail to stop harmful states from occurring. To this end, a secure Q-learning technique based on constrained Markov decision processes is proposed. This approach, incorporating safety constraints as conditions, ensures that the agent always acts in a safe

environment while trying to find the optimal responses. Experimental results have shown the success of this strategy. There are now two main ways to deal with agent safety issues in reinforcement learning. The first approach includes changing the agent’s objective function to lessen the possibility that it would enter risky conditions, but it doesn’t offer a permanent solution to the safety issue. The second approach concentrates on enhancing exploration by obtaining data by randomly exploring the status and action spaces. Although this approach can improve algorithm performance, it doesn’t fundamentally address safety concerns, as agents may still get into hazardous situations due to inadequate information. There is a Q-learning technique presented based on limited Markov decision processes. This technique uses multidimensional constraints to limit each action to a subset of safe actions. We guarantee the agent’s safety during the initial exploration stage by limiting the agent’s possible states to a set of safe states. The algorithm has many uses, including boosting game player performance and promoting safety in robotics and driverless vehicles. It also emphasizes the possibility for various reinforcement learning methods to safely employ the Lagrange multiplier technique. This method is useful for future applications since it may be expanded to handle a variety of limited challenges.

Another research paper [11] introduces a novel three-layer charging system design that caters to both static and dynamic wireless charging while seamlessly integrating with existing wired charging infrastructure and standards within Intelligent Transportation Systems (ITS). The system leverages IoT technology and a handshake protocol, facilitated by vehicle-to-infrastructure (V2I) and vehicle-to-grid (V2G) communications, to efficiently fulfill charging requests for connected and autonomous electric vehicles (CAEVs) while optimizing trip routes. Key features include the dynamic distribution of charging requests across various charging equipment, secure billing using encrypted virtual currency, and the ability to detect and correct hardware-related issues like misalignment on wireless charging pads and speed errors in dynamic wireless charging systems. The system also excels in trip planning, reducing waiting times, travel costs, and energy consumption, achieving an impressive 90.25 percent charge delivery efficiency. The paper begins by highlighting the environmental benefits of electric vehicles (EVs) and autonomous electric vehicles (AEVs) while acknowledging challenges related to their limited driving ranges and longer charging times. Initiatives to build charging infrastructure have been undertaken in various countries to encourage EV adoption. Wireless charging solutions are explored, with magnetic resonance coupling identified as a promising wireless power transfer technique. The proposed architecture involves CAEVs communicating with infrastructure like roadside units (RSUs) and smart grids through V2I and V2G communication. The three-layer hierarchical charging system aims to enhance CAEV trip efficiency by reducing waiting times, travel costs, and energy consumption. It emphasizes secure billing through two proposed payment schemes using encrypted virtual currency. The system is also equipped to detect and correct misalignment and speed errors in wireless charging systems and prevent charge theft. Additionally, it can automatically configure a custom DWC infrastructure for testing.

The paper concludes by describing a simulator that allows users to predict and ana-

lyze the system’s performance, highlighting its remarkable charge delivery efficiency and efficiency in minimizing waiting times, travel costs, and energy consumption compared to manual EVSE searches by CAEVs. This research paper presents an innovative charging system design, integrating static and dynamic wireless charging with existing wired infrastructure for ITS. It offers efficient charging, route optimization, secure billing, and error detection capabilities, contributing to the advancement of smart city IoT applications and sustainable transportation.

In order to promote environmental sustainability, the use of electric vehicles (EVs) is rapidly increasing around the world, according to a study paper titled "Intelligent Charging Infrastructure Design for Connected and Autonomous Electric Vehicles in Smart Cities." [12] In order to meet the growing number of connected and autonomous EVs (CAEVs), it underlines the necessity for smart charging infrastructures and solves lengthy charging times by taking into account dynamic wireless charging. The research presents a three-layer hierarchical charging infrastructure concept that allows for communication between current wired charging systems and foreseeable wireless options. For effective scheduling of CAEV charging reservations over various networks, it suggests charging request and reservation message frames. For quick calculation and low latencies, the system uses error detection, vehicle-to-infrastructure (V2I), and vehicle-to-grid (V2G) connections. For both shared and nonshared CAEVs in smart cities, it also analyzes a dynamic wireless charging network (DWCN) suggestion tool to maximize charge delivery performance at the lowest possible cost. The environmental issues with present transportation methods are highlighted in the introduction along with the potential advantages of EVs, driverless cars, and mobility-on-demand made possible by the IoT. It emphasizes how crucial smart charging infrastructure is to overcoming these difficulties. In the paper, a flexible architecture for a smart charging infrastructure is proposed that supports both wired and wireless charging, efficiently manages CAEV charging schedules, and keeps backward compatibility with both existing standards and new wireless systems. Additionally, it presents a suggestion tool for dynamic wireless charging networks that is affordable, adding to the sustainability of smart cities.

Again, in [15] delves into innovative strategies to enhance the Electric Vehicle Supply Equipment (EVSE) infrastructure. As the popularity of electric vehicles surges, there’s a pressing need for advanced charging solutions. This research encompasses aspects like demand management, integration with the power grid, risk evaluation, and enhancements to the charging process, shedding light on the latest advancements in the domain. The global pivot towards eco-friendly transportation has accelerated the growth of electric vehicles. For these vehicles to reach their full potential, the EVSE infrastructure needs to be top-notch. In this academic work, cutting-edge techniques for boosting the EVSE infrastructure are explored. To improve EVSE performance, risk management is essential, particularly in situations when charging demand is erratic. To handle erratic charging demands from human-driven cars, the RAMALS system incorporates advanced risk assessment techniques. The system makes significant operational gains, including a decrease in policy mistakes and an increase in billing proficiency, by employing entropy regularization to guarantee constant training. Improving the caliber of charging experiences has been a main area of study. Existing systems frequently struggle to keep up with the excessive

energy demands made by connected cars, which can result in energy losses and ineffective charging.

Moreover, [6] shows the rapid urbanization and population growth of metropolitan areas increase transportation demand, which leads to frequent congestion. An ATSC system dynamically adjusts the signal timings according to real-time traffic conditions to relieve such congestion. In the past, widely implemented ATSC solutions, such as SCOOT and SCATS, used optimization techniques to coordinate traffic signals efficiently. However, other more complex systems, like OPAC and PRODYN, are much less applied because of their high computational complexity.

A number of interdisciplinary methods have been employed in ATSC for a long time. Early applications of fuzzy logic, genetic algorithms, and immune network algorithms showed innovation but posed challenges of scalability and adaptability. The development of Reinforcement Learning (RL), particularly within the framework of Markov Decision Processes (MDPs), provided an alternative, data-driven way of solving ATSC. Unlike the traditional optimization methods, RL approaches do not rely on heuristic assumptions and pre-defined models. Instead, these policies learn to optimize the control strategy through interactions with the traffic environment.

Earlier applications of RL in ATSC adopted simple models, such as a piece-wise constant table and linear regression, which were limited by scalability and sub-optimality. The incorporation of Deep Neural Networks into RL granted an enormous ability to deal with complex and high-dimensional tasks. Subsequently, a variety of RL methods have been applied, which, in general, can be categorized as Value-based, Policy-based, and Actor-critic methods.

Off-policy methods, which combine value estimation with off-policy exploration, are popular because they allow for efficient updates through experience replay. However, it requires reliance on one-step temporal difference updates, which exposes them to the sensitivity to the stationarity of the environment—a condition rarely met in dynamic traffic systems. Directly optimizing the policy based on sampled returns, policy-based methods like REINFORCE accommodate non-stationary transitions within each episode at the cost of high variance. Actor-critic methods combine the advantages of both value-based and policy-based approaches by using separate models for the policy and value functions to reduce bias and variance.

Among the actor-critic methods, the Advantage Actor-Critic (A2C) algorithm is popular for dealing with continuous action spaces, leveraging the power of DNNs to approximate the policy and value functions. Deploying centralized RL algorithms, such as A2C, is impractical in large-scale traffic networks, since it requires global state information and the joint action space grows exponentially.

MARL is a viable solution to address the scalability issue of distributing control at local RL agents at each individual intersection. The decentralization can result in each agent making a decision based on local observations, which eventually reduces the computational burden and improves scalability. Traditional MARL approaches primarily focus on Q-learning variants, like Independent Q-learning (IQL), in which each agent learns its policy independently by treating other agents as part of the environment. Though scalable, IQL often suffers from convergence due to increased partial observability and nonstationarity of the environment as agents update their policies independently.

Recent advances have attempted to stabilize MARL systems through mechanisms

of efficient communication and coordination among agents. This includes enabling experience replay in deep MARL to deal with nonstationarity. However, so far, the extension of the actor-critic methods, in particular A2C, within a decentralized multi-agent framework for ATSC remains largely unexplored.

We first apply independent A2C (IA2C) to ATSC, extending the principles of IQL to the A2C algorithm. We have further proposed two novel enhancement techniques to ensure better stability and robustness of the IA2C system: the incorporation of observations and fingerprints of other agents to improve state observability, and the introduction of a spatial discount factor with the main objective of maximizing improvements in local traffic. It thus balances the fitting power and fitting difficulty of the multi-agent A2C algorithm, leading to a more robust and scalable MA2C algorithm.

The efficacy of MA2C is tested through exhaustive evaluation on both synthetic and real-world traffic networks and found to excel over state-of-the-art decentralized MARL algorithms on robustness, optimality, and sample efficiency.

Another thing is seen in [9] feature selection is one of the most important processes in machine learning and data mining, where from a big feature space, the most relevant features will be determined with the purpose of improving model performance and reducing computational complexity. Traditional feature selection methods can be divided into three categories: filter methods, wrapper methods, and embedded methods. Filter methods select features based on their individual relevance, using statistics such as chi-square, ANOVA, or mutual information. Univariate feature selection is a filter method, which is effective due to its low computational resources, but it lacks the analysis of the interaction between features, thus not being efficient when it comes to choosing the optimal combination of features for a particular problem. Wrapper methods, like forward selection, backward elimination, and branch-and-bound algorithms, evaluate subsets of features based on their performance with a specific predictive model. While a better performance in terms of accuracy is achieved by considering the interaction among features using wrapper methods, they are computationally expensive and may not scale to the feature size of large datasets. Embedded methods integrate feature selection into the model training process, and the most common methods are regularization techniques. For example, LASSO (Least Absolute Shrinkage and Selection Operator) adds a term into the penalty of the model loss function, thereby forcing the coefficient of less important features to zero. These methods offer a balance in computational efficiency and model performance by selecting features during model training.

Recent breakthroughs in RL have brought a new dimension into feature selection, treating feature selection as a problem in sequential decision-making. RL-based methods automate the feature selection process by training an agent to move in the feature space guided by rewards from the environment. These methods seem promising in dealing with complex feature spaces but are mostly inefficient because of high exploration requirements. Interactive Reinforcement Learning augments traditional RL by using external trainers in order to guide the agent. This approach dramatically accelerates the process of learning by leveraging domain knowledge and human expertise. IRL has been successfully used in a lot of applications for making agents learn more effectively through interaction with skilled trainers. Therefore, the proposed Interactive Reinforced Feature Selection framework will be used to confront the computational dilemma of balancing effectiveness and efficiency in feature

selection. In formulating the feature selection problem as an interactive reinforcement learning paradigm, IRFS introduces a hybrid teaching strategy that integrates both self-exploration and guidance from external trainers. The framework utilizes multiple trainers with different skills, including a K-Best-based trainer and a Decision Tree-based trainer, for giving multiple perspectives on the relevance of the features. Besides, IRFS personalizes the process of teaching by categorizing agents into assertive and hesitant groups and giving advice adapted to the needs of the learners, further improving learning. Thus, IRFS combines the strengths of traditional methods for feature selection and the adaptive features of reinforcement learning to achieve a balanced and effective process of feature selection. Extensive experiments on real-world datasets demonstrate the superiority of IRFS over existing approaches in terms of efficiency and effectiveness. This new framework offers a promising solution for the long-standing challenge of optimizing feature selection in machine learning.

Forester et. al. [2] state that multi-agent communication and coordination have come a long way in current times, keeping in line with the advancement of deep learning and reinforcement learning techniques. Multi-agent systems were traditionally based on heuristic-based approaches and pre-defined communication protocols, which lacked adaptability and were limited in handling the complexity of real-world scenarios. Recent advances, especially in deep reinforcement learning, have opened up new opportunities for the development of autonomous communication strategies among the agents. Filter methods, wrapper methods, and embedded methods have been extensively used for feature selection. Filter methods consider feature relevance independently, while wrapper methods utilize predictive models to evaluate feature subsets. Embedded methods insert feature selection into model training, such as techniques like LASSO, adding penalty terms to feature coefficients and shrinking the coefficients of less important features. On the other hand, in multi-agent systems, RIAL and DIAL are two significant advancements. RIAL makes use of deep Q-learning, incorporating recurrent networks that handle partial observability by treating other agents as part of the environment and enabling agents to learn communication protocols. This also includes variations like independent Q-learning and shared network parameters among agents.

On the other hand, DIAL exploits centralised learning and decentralised execution. This allows for real-valued messages to be passed between agents at training time by the treatment of communication actions as differentiable connections. Therefore, the gradients can be back-propagated through the channel of communication, and the agents are end-to-end trainable. During execution, these messages are discretized and fit within the limited bandwidth communication constraints of the environment. The empirical studies of these methods have shown that deep learning can effectively discover and optimize protocols for communication in complex, partially observable environments. RIAL and DIAL have been deployed successfully on sequential decision-making tasks and raw input processing, which shows the robustness and adaptability of the agents. By integrating deep learning into reinforcement learning, the agents are enabled to develop sophisticated and efficient communication strategies with a significant outperformance of the state of the art. The presented developments underline the potential of deep reinforcement learning to revolutionize multi-agent communication and coordination. The possibility of learning and adapting communication protocols in an autonomous manner repre-

sents an important step toward the creation of more intelligent, flexible, and scalable multi-agent systems. This research contributes to the theoretical understanding of multi-agent communication and moreover paves the way for practical applications in robotics, autonomous vehicle guidance, or distributed sensor networks.

This paper [3] says EVs are of great interest to policymakers and researchers because of their possible economic and environmental benefits; therefore, they can be considered as having great potential for replacing traditional fuel-engine vehicles. The large-scale integration of EVs has greatly increased focus on developing efficient charging scheduling mechanisms to optimize the operations of systems and ensure that they address the main challenges caused by long refueling times and significant charging power demands of EVs. EVs charging scheduling mechanisms may be classified into two: temporal and spatial scheduling mechanisms. Temporal scheduling mechanisms aim to provide a indicator regarding suitable charging time to minimize the total cost of EVs charging and discharging for a day. For instance, researchers have proposed the globally optimal scheduling scheme and the locally optimal scheduling scheme to achieve this cost minimization. Additionally, some study optimal charging strategies based on drivers' self-interested behaviors, traffic congestion, operating expenses of the CSs, and pricing models. Temporal scheduling also aims at maximizing the operating profits of electric taxis through consideration of the uncertainties of the electricity prices and time-varying incomes. Spatial scheduling, on the other hand, aims to provide a indicator regarding geographically distributed charging stations in order to minimize the time for traveling and queuing. Spatial scheduling is very important, especially in an urban setup where the distribution of the charging stations and the demand for charging services will most likely greatly influence the overall efficiency and experience of the end users. This effectiveness of spatial scheduling can be further boosted by incorporating game-theoretical approaches, which ensure that the indicators are fair.

In this context, game-theoretical approaches have been put forward to develop fair and efficient spatial scheduling algorithms for EVs. They take into account the strategic interactions of EV users with the charging infrastructure in order to optimize system operation and user satisfaction. Numerical results of the studies by methods based on these approaches demonstrate their effectiveness in reducing the idle rate of the charging piles, minimizing EV queuing time, and eventually saving time for the users. To sum up, the integration of temporal and spatial scheduling mechanisms with game-theoretical approaches provides a holistic framework for the optimization of EV charging operations. This way, not only will EV charging be effective and economical, but fair and user-friendly, overcoming all the challenges linked with large-scale EV integration into the grid.

However in [4] the progress in the domain of multi-agent reinforcement learning (MARL) has evolved prominently in developing algorithms that will facilitate agents to learn cooperative behavior in complex, partially observable environments. Traditional methods on multi-agent cooperation mainly focus on centralized training, with decentralized execution to solve decentralized partially observable Markov decision processes (Dec-POMDPs). However, exact solutions to Dec-POMDPs are computationally intractable. Recent progress has been made on applying deep reinforcement learning (DRL) techniques to multi-agent systems. DRL merges deep learning with reinforcement learning, which uses the power of neural networks for handling large and complex observation spaces. The combination has thereby en-

abled single-agent reinforcement learning to solve difficult domains, such as playing Atari games and robotic locomotion, and showed significant success in learning policies for complex tasks. Researchers in MARL have developed several approximation methods to deal with the challenges of Dec-POMDPs. These include reinforcement learning techniques like Deep Q-Networks, policy gradient methods like Trust Region Policy Optimization, and actor-critic methods like Deep Deterministic Policy Gradient.

Empirically, it has been shown that policy gradient methods often outperform the temporal-difference and actor-critic methods, especially when using feed-forward neural architectures. On the other hand, recurrent neural networks have shown better performance in environments that need the memory of past observations, despite the difficulty of training them. This shows how important neural architecture choices are in developing efficient MARL algorithms. The introduction of the decentralized parameter sharing neural network policies has moved the field even further. Each agent can now develop emergent cooperation without explicit communication among agents. This approach enables solutions that scale to environments with continuous action spaces and a large number of agents. For example, promising results with PS-TRPO demonstrate the feasibility of scaling up multi-agent control tasks to dozens of agents cooperating toward common objectives. Generally, DRL within the framework of MARL is a significantly further step toward solving complex, cooperative tasks. The development and further refinement of these algorithms may allow solving a wide range of real-world applications, from robotic teams to distributed sensor networks. This research underlines the transformative impact that deep learning has on multi-agent systems and offers strong frameworks for developing cooperative policies in a variety of challenging environments.

Lastly in the paper [1] electric vehicles have been attaining high growth rates because of their environmental benefits and low operating costs compared to conventional gasoline vehicles. Their steep growth rate has made a large, strategically deployed network of public charging stations crucial in supporting them. The deployment of these charging stations has to be done effectively so that the time spent by drivers traveling to and waiting at the charging points is minimized and, in turn, improves the overall experience and further adoption of EVs. The present research in the field of the deployment of charging stations generally follows two approaches: temporal scheduling and spatial scheduling. Temporal scheduling suggests the optimum times for EV charging, managing the demand for reducing costs. However, the spatial distribution of these charging infrastructures is not considered—a critical factor to cut down the traveling and waiting time for EV users. Spatial scheduling, on the other hand, works on the strategic placement of these charging stations with their charging points to enhance their accessibility and reduce congestion.

The station siting problem has long been studied in the context of gas stations and hydrogen filling stations; it provides a foundation for understanding the complexity of deploying EV charging stations. However, unique characteristics, such as a longer duration for charging and higher variability in demand for EVs, require specialized models. The facility location models proposed earlier fail to capture these aspects and generally demand trip origin-destination data, which is difficult to acquire. To address these challenges, optimization frameworks for the deployment of electric vehicle charging infrastructures have been proposed recently. The frameworks accommodate historical trajectory data, road network information, and existing charging

station data to come up with optimal deployment strategies. Examples of this include integer programming models and polynomial-time approximation algorithms that have been developed to solve the problem of charging station placement in a manner that minimizes the travel and wait times of EV users. A major breakthrough in this respect is the Optimal Charging Station Deployment (OCSD) framework, which integrates Optimal Charging Station Placement (OCSP) and Optimal Charging Point Assignment (OCCA). The OCSD framework leverages real-world EV taxi trajectory data to extract seeking, charging, and traveling behavioral patterns, thus directing the strategic placement of new charging stations and the allocation of charging points. It ensures that the deployments not only meet current demand but are also scalable to accommodate future growth. The empirical performance of the OCSD framework is evaluated and considerable improvements over the baseline methods are shown. For instance, it reduces the average time to find a charging station by 26 percent to 94 percent, and it significantly reduces the waiting time before charging. Moreover, the results provide valuable guidelines on the optimal configuration of charging stations, indicating that when many charging points are provided, it is optimal to place a larger number of smaller stations, while when the number of charging points is limited, fewer, larger stations are more effective. From what the literature indicates, the importance of taking into consideration the temporal and spatial perspectives of the deployment of EV charging infrastructure is highlighted. The development of sophisticated optimization frameworks—such as the OCSD—represents a significant step forward toward an attempt at answering the peculiar challenge of EV charging station deployment and, therefore, toward the sustainable growth of electric mobility.

Chapter 3

Proposed Method

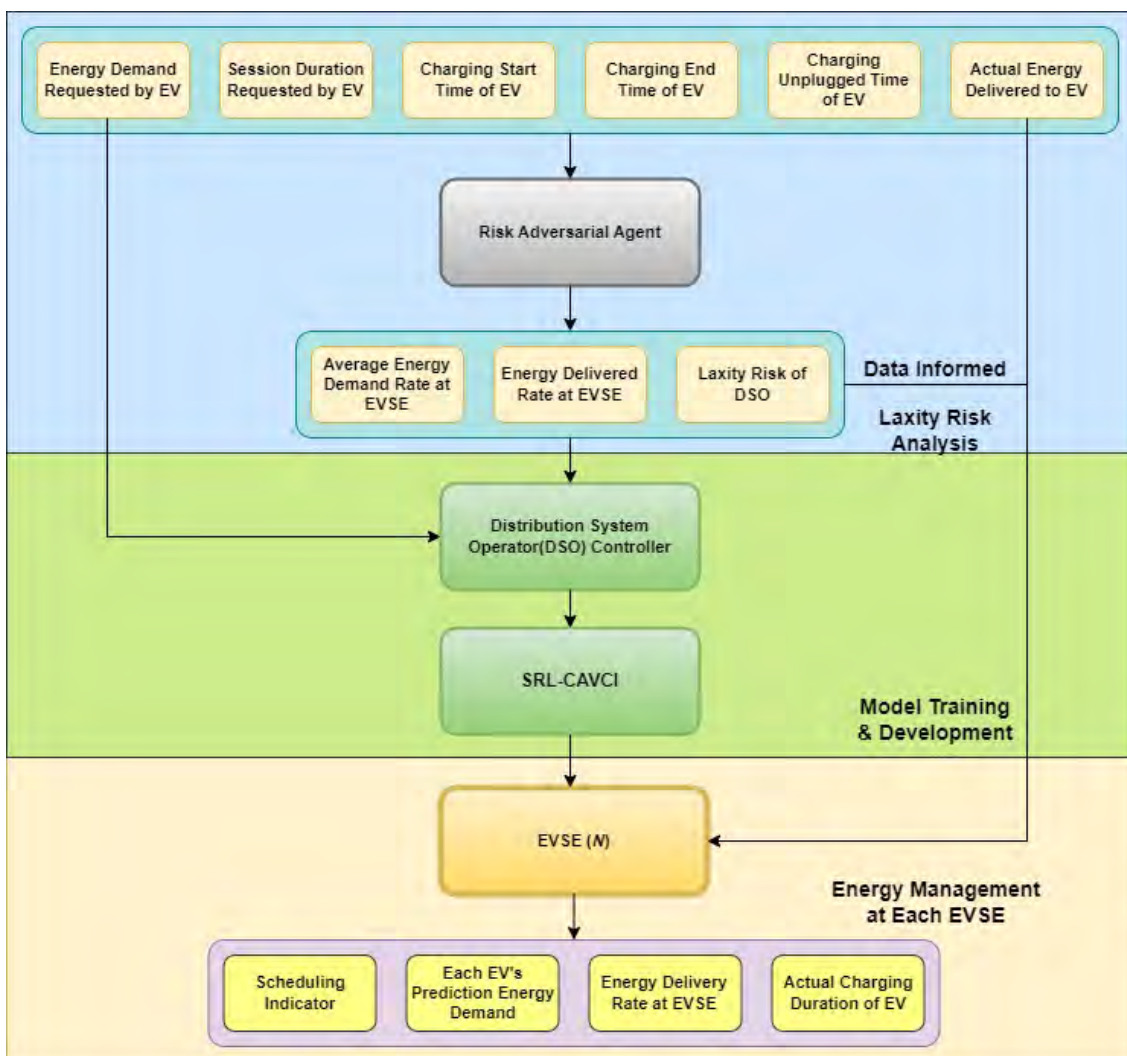


Figure 3.1: Top Level Overview of the proposed SRL-CAVCI

In this research we worked with simulation based actor-critic network which is used to find the best policies and indicator. Moreover we will use ACN Dataset [8] which holds the information about all Electric vehicles that took services from JPL and Caltech site for charging. Basically this data we are planning to use for optimize the actor-critic network using safe RL. Both of the data has all information available for

each electric vehicle like User ID, requested charge, available time, charging site, station ID, delivered charge, connection time, done charging time, disconnection time etc. From this dataset we can initially identify the inefficiency of charging. To solve this problem we proposed a Safe RL AVCI model that can handle the scheduling policy without any risk.

We provide the formulation of the SRL-CAVCI problem for the task of the EV charging indicator and detail our SRL-CAVCI framework with centralize training, decentralize execution. More than that, this research Generalize the multi-critic architecture to multiple modes and the Goals is to optimization.

3.0.1 Actor-Critic Method

Initially we discussed about some of the core aspects about the proposed Actor-Critic formulation for the EV charging indication task.

- **Agent** $EVSE$: The research considered a charging stations as individual agents $EVSE$. These individual agents provide real time indication decisions for a series of charging requests that came all over the day and each $EVSE$'s aim to achieve many long-term optimization objectives.
- **Observation** o_t^i : When a charging request q_t arrives, the observation o_t^i for agent $EVSE^i$ contains the following information: the index of $EVSE^i$, the actual time T_t , currently available charging spots at $EVSE$, the number of charging requests that are scheduled to occur soon near $EVSE$ (future demand), the available charging power of $EVSE$, the estimated time of arrival (ETA) from the location of request l_t to $EVSE$, and the capacity constraint (CC) of $EVSE$ at the next ETA. The set of observations determine the state of all agents at step t is further defined as $s_t = \{o_t^1, o_t^2, \dots, o_t^N\}$.
- **Action** a_t^i : A basic design for the action of agent $EVSE^i$ given an observation o_t^i is a binary decision on whether to advise the recommendation of q_t to itself for charge. However, the coordination of multiple agents will be challenging because one q_t can only indicate one specific station to charge. We designed each agent $EVSE^i$ to provide a scalar value as a "bid" for q_t , which is represented by its action a_t^i . This design is inspired by the bidding process. The agent provide the highest "bid" value, $r_c = EVSE^i$, where $i = \arg \max(u)$, is allocated q_t after the joint action is defined as $u_t = \{a_t^1, a_t^2, \dots, a_t^N\}$.
- **Transition.** The observation transition for each agent $EVSE^i$ is defined as the change from the current charging request q_t to the subsequent charging request q_{t+j} following the completion of q_t . Let us begin to highlight it with an example. Suppose at $T_t = 13:00$, a charge request q_t is made. Currently, each agent $EVSE^i$ acts a_t^i , following its observation o_t^i , and they decide together on the suggested station r_{c_t} . The next charge request, which is q_{t+j} , will take place at $T_{t+j} = 13 : 20$, after the request completion time of $T_t^c = 13 : 18$. The observation transition in this example for agent $EVSE^i$ is defined by $(o_t^i, a_t^i, o_{t+j}^i)$, where o_t^i represents the observation that is under progress, and o_{t+j}^i corresponds to the observation of q_{t+j} .

- **Reward.** Three objectives are combined into two natural reward functions in our SRL-CAVCI formulation, along with a delayed reward settlement mechanism. That is, it will receive the negative of WPC and the negative of CC from the environment as part of reward $r^{WPC}(s_t, u_t)$ and reward $r^{CC}(s_t, u_t)$, respectively, if the successful charge request q_t is obtained. If the WPC of q_t is above a certain level, the environment will give agents much smaller rewards to penalize them, which is considered to encourage them to reduce the FRC. In sum, we define two instant reward systems for three objectives as

$$r^{WPC}(s_t, u_t) = \begin{cases} -\text{WPC}, & \text{charging success} \\ \epsilon_{WPC}, & \text{charging failure} \end{cases} \quad (1)$$

$$r^{CC}(s_t, u_t) = \begin{cases} -\text{CC}, & \text{charging success,} \\ \epsilon_{CC}, & \text{charging failure} \end{cases} \quad (2)$$

where the penalty rewards are ϵ_{WPC} and ϵ_{CC} . In our the framework agents collaborate to determine the indicators as they share the same advantages. The proposed model determine the cumulative discounted reward by adding the rewards of all the indicated charging requests as the observation transition from o_t^i to o_{t+j}^i may span several lazy rewards (such as T_{t-h}^c and T_t^c). $T_{t'}^c$ for $q_{t'}$ (e.g., q_{t-h} and q_t) is between T_t and T_{t+j} , as indicated by

$$R_{t:t+j} = \sum_{T_t < T_{t'}^c \leq T_{t+j}} \gamma^{(T_{t'}^c - T_t - 1)} r(s_{t'}, u_{t'}), \quad (3)$$

where, depending on the learning objectives, $r(\cdot, \cdot)$ can represent either of the two reward functions or their average, and γ denotes the discount factor which we considered as 0.99.

3.0.2 Centralized Training Decentralized Execution

The SRL-CAVCI method for teaching agents to coordinate policies and solve non-stationarity is Centralized Training Decentralized Execution, CTDE. The three modules that make up the SRL-CAVCI are the centralized attentive critic, the delayed access information strategy to include forthcoming charging competition, and a decentralized process for execution. With respect to indication of EV charging, CTDE provides two-fold advantages. First, centralized training helps because of the use of the bigger, global view and retroactive incorporation of knowledge from the future enables the collaboration of different agents to learn specific regulation. However, since the process of execution is fully decentralized and does not need all the data involved in training, the online indicator application is bound to be effective and adaptive.

3.0.3 Centralized Attentive Critic

We build a multi-agent actor-critic architecture with a centralized attentive critic to learn a deterministic policy, enabling the agents to supply indicators jointly. In[5], a similar method based on the CTDE architecture is proposed, which feeds into the critic the full state s_t and the collective action u_t of all agents for it to learn

coordinated and cooperative policies. However, such a method in our assignment is subject to massive state and action space difficulties.

In reality, EVs are used to go to nearby stations to get charged. For this reason, once we receive a charging request, only a few agents that are nearest to the request, say top-nearest, are active following the indicator. Since other agents are far away, we set them to inactive and exclude them from the set of agents that participate in the indicator for. In this way, learning cooperation for better indicators is a problem that involves relatively few active agents. However, the active agents for different are most often different, and this is an intermediate problem. To address this, we propose combining the information of the active agents through a permutation-invariant attention mechanism. That is, the attention system automatically counts the impact of each active agent through

$$e_t^i = v^\top \tanh(W_a [o_t^i \oplus a_t^i \oplus p_t^i]), \quad (4)$$

where \oplus is the concatenation operation and v and W_a are learnable parameters. To find the each active agent’s impact weight d_t^i can help develop an attentive representation of all active agents $EVSE^i \in C_t^a$.

$$x_t = \text{ReLU} \left(W_c \sum_{i \in C_t^a} d_t^i [o_t^i \oplus a_t^i \oplus p_t^i] \right), \quad (5)$$

where W_c are learnable parameters.

The policy of actor network for every agent $EVSE^i \in C_t^a$ updated by the gradient of the anticipated return according to the chain rule that have provided to given the state s_t , joint action u_t , and the future demand p_t of active agents. This can be expressed as

$$\nabla_{\theta^i} J(b^i) = E_{s_t, u_t \sim D} \left[\nabla_{\theta^i} b_{\theta^i}^i(o_t^i) \nabla_{u_t^i} Q_\phi(x_t) |_{u_t^i = b_{\theta^i}^i(o_t^i)} \right], \quad (6)$$

where the transition tuples $(s_t^a, u_t^a, p_t^a, r_{t:t+j}, R_{t:t+j+1})$ are included in the learning replay buffer D , and θ^i are the learnable parameters of the actor policy b^i of agent $EVSE^i$. Based on the gradients that spread from the centralised attentive critic, each agent modifies its policy. The agents are encouraged to learn policies in a coordinated and cooperative manner as a result of the centralised attentive critic’s perception of more comprehensive knowledge about all active agents. By minimising the subsequent loss, the centralised attentive critic Q_ϕ is updated:

$$L(\theta) = E_{s_t, u_t, p_t, r_{t:t+j}, R_{t:t+j} \sim D} [(Q_\phi(x_t) - y_t)^2], \quad (7)$$

$$y_t = R_{t:t+j} + \gamma^{(T_{t+j}^c - T_t)} Q_\phi(x_{t+j}) |_{u_{t+j}^a = b_{\theta^i}^i(o_{t+j}^i)}, \quad (8)$$

where θ_ϕ represent the critic Q_ϕ ’s learnable parameters. With delayed parameters θ^i and Q_ϕ , respectively, b^i and Q_ϕ^i represent the target actor policy of $EVSE^i$ and target critic function.

3.0.4 Integration of Future Charging Competition

Since first-come, first-served policies apply to public charging stations, concurrently arriving EVs may eventually compete with one another. If charging requests are

recommended without taking into account this upcoming competition, WPC may rise or there may even be charging failures. Nevertheless, integrating future charging competition is difficult as it need precise forecasts of approaching electric vehicles and open charging locations in the future.

In this study, this research expand the centralised attentive critic by incorporating a delayed access approach that makes use of future charging competition data after the fact. More specifically, we provide a scoring function to assess the effect of competition in the future up until the charge is finished for a charging request q_t . To account for anticipated future competition for q_t , represented as \tilde{N} , we estimate the number of available charging slots at each $EVSE^i$ at incremental minutes following T_t . (\tilde{N}) may be negative, which denotes the quantity of EVs waiting in queue at the station. A fully-connected layer is used to acquire the future competition information for each $EVSE^i$:

$$p_t^i = \text{ReLU}(W_p \tilde{p}_t^i), \quad (10)$$

where the parameters that can be learned are W_p . To support the agents' cooperative policy learning, the p_t^i is included into the centralised attentive critic (Eqs. (4)–(6)).

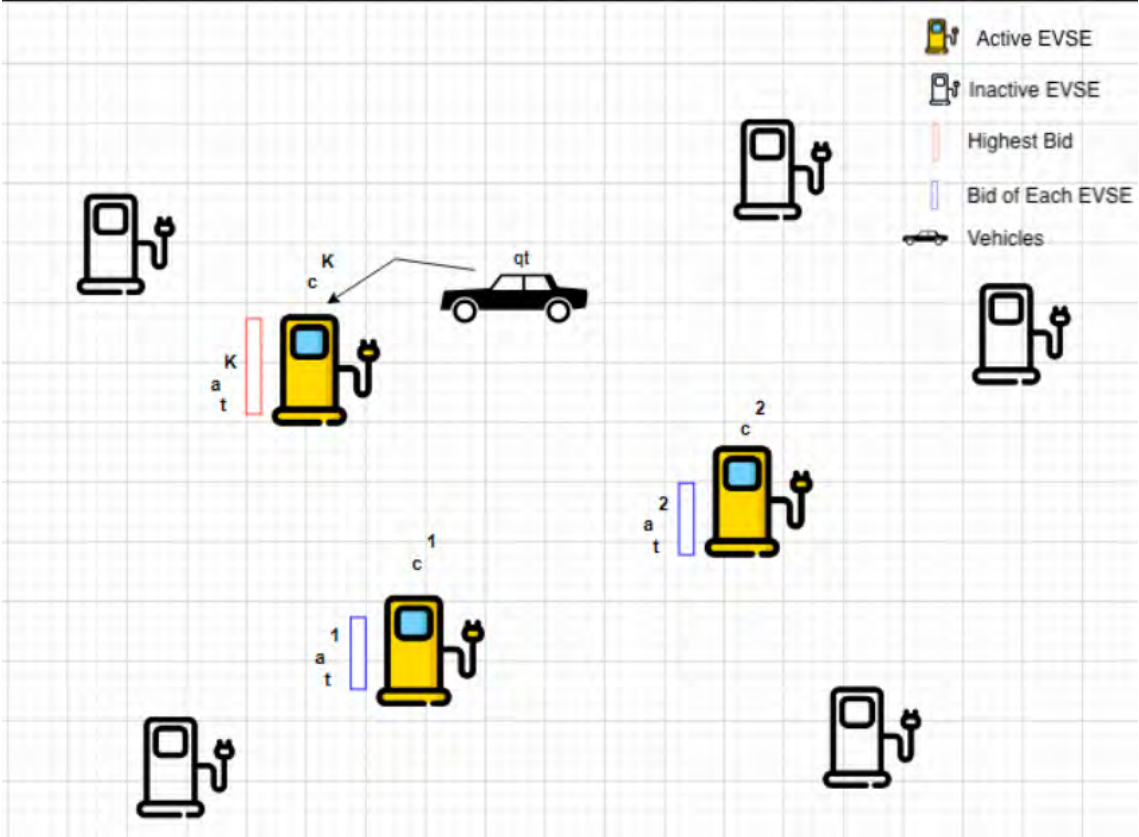


Figure 3.2: Decentralized execution of active agents

The active charging station with the best of action "bid" will receive the indication for the charging request.

Decentralized Execution

Only the learnt actor policy with its own observation is involved in the fully decentralised execution process. To be more precise, the agent $EVSE^i \in C_t^a$ responds to

a charge request q_t by taking a_t^i based on its b_θ^i :

$$a_t^i = b_\theta^i(o_t^i), \quad (11)$$

and of all the acts of Ca_t , the active agent with the biggest ai_t will be advised to use qt . Each agent is capable of light execution and is not required to be aware of the competition details for upcoming charges. Moreover, even in the event that some agents fail, the large-scale agent system is fault-tolerant.

3.0.5 Multi-Objective Enhancement

The charging indication for electric vehicles aims to minimise the average CC, the FRC, and the total WPC at the same time. These goals are joined with two additional goals provided by the reward functions outlined in Equations (1) and (2). The normalised reward distributions of r^{WPC} and r^{CC} by means of a random policy executed on a number of successfully charged requests are displayed in Figure 3.2. The allocation of various objectives might differ greatly as is shown. More importantly, the optimum solution for different objectives may differ. For example, a cheaper charging point could be invented and then become a popular trend and would require a lengthier WPC. These data imply that the policy which achieves one objective well may perform poorly at another. An indicator biased towards a few specific targets risks providing most users with an inferior experience.

A naive way to enhance multiple objectives is to maximize the total reward as a single target by averaging the rewards of the individual objectives using a set of predefined weights. Such a biased approach, however, is inadequate for dynamic adaptation to a particular target and for other learning phases. In order to ensure that the policy works well on a variety of objectives, we have created a dynamic gradient re-weighting technique that adjusts the optimization direction to various training phases. More precisely, we have extended the attentive critic which is centralized, to a number of reviewers.

Algorithm: SRL-CAVCI_model

Input: $s_{n,\nu}^t, e_{n,\nu}^{EVSE}, e_{n,\nu}^{trip}, t_{end}, p_n^t, q_{n,\nu}^t, t_{n,\nu}^{act}(t), \pi_{n,\nu}(t), R_\alpha^{CVaR}(L(x, y))$

Output: SRL-CAVCI_model

Initialization: $w_n^*, \nu^*, \gamma, V^\pi, \theta_1, \eta, \beta$

1. Randomly initialize critic networks $Q_b^{c^{wt}}, Q_b^{c^p}, Q_b$ and each actor network b^i with weights $\theta_{Q_b}^{c^{wt}}, \theta_{Q_b}^{c^p}, \theta_b^i$.
2. Initialize target networks $Q_{b'}^{c^{wt}}, Q_{b'}^{c^p}, b^i$ with weights $\theta_{Q_{b'}}^{c^{wt}} \leftarrow \theta_{Q_b}^{c^{wt}}, \theta_{Q_{b'}}^{c^p} \leftarrow \theta_{Q_b}^{c^p}, \theta_{b^i} \leftarrow \theta_{b^i}$.
3. Initialize objective-specific optimal networks $Q_{b^*}^{c^{wt}}, Q_{b^*}^{c^p}, b_{wt}^*, b_{CC}^*$ and b_{CC}^{i*} with well-trained weights $\theta_{Q_{b^*}^{c^{wt}}}, \theta_{Q_{b^*}^{c^p}}, \theta_{b_{wt}^*}, \theta_{b_{CC}^*}, \theta_{b_{CC}^{i*}}$.
4. Value of replay buffer D .
5. **For** $1t_{omax} - iterations$ **do**

6. Reset environment.

7. **For** $t = 1$ to number of requests $|\mathcal{Q}|$ **do**

(a) **For** agent $c^l \in \mathcal{C}_{qt}$ **do**

i. Take action $a_b^i = b^i(o_t^i)$ for each charging request q_{t-r} .

(b) Store the transition values $(s_t^a, u_a^a, \rho_t^a, s_{t+r}^{a'}, \rho_{t+r}^{a'}, R_{t+r}^{c^{wt}}, R_{t+r}^{c^p})$ into D .

(c) Sample a random minibatch of M transitions $(s_t^a, u_a^a, \rho_t^a, s_{t+r}^{a'}, \rho_{t+r}^{a'}, R_{t+r}^{c^{wt}}, R_{t+r}^{c^p})$ from D .

(d) Set $y_t^{c^{wt}} = R_{t+r}^{c^{wt}} + \gamma(R_{t+r}^{c^{wt}} - R_{t+r}^{c^{wt}})Q_{b'}^{c^{wt}}(s_t^a, u_t^a)|_{a_t^a=b^i(o_t^i)}$.

(e) Set $y_t^{c^p} = R_{t+r}^{c^p} + \gamma(R_{t+r}^{c^p} - R_{t+r}^{c^p})Q_{b'}^{c^p}(s_t^a, u_t^a)|_{a_t^a=b^i(o_t^i)}$.

(f) Update Critic $Q_b^{c^{wt}}$ and $Q_b^{c^p}$ by minimizing the losses:

$$L(\theta_{Q_b^{c^{wt}}}) = \frac{1}{M} \sum (Q_b^{c^{wt}}(x_t) - y_t^{c^{wt}})^2$$

$$L(\theta_{Q_b^{c^p}}) = \frac{1}{M} \sum (Q_b^{c^p}(x_t) - y_t^{c^p})^2$$

(g) Compute β_t through Eq. (13) and Eq. (14).

(h) **For** agent $c^l \in \mathcal{C}_{qt}$ **do**

i. Update actor by the sampled policy gradient:

$$\nabla_{\theta_{b^i}} J(b^i) \approx \frac{1}{M} \sum \nabla_{\theta_{b^i}} b^i(a_t|o_t^i) \nabla_{a_t^a} Q_b^{c^{wt}}(x_t)$$

$$\theta_{b^i} = (1 - \beta_t) \nabla_{\theta_{b^i}} b^i(a_t|o_t^i) \nabla_{a_t^a} Q_b^{c^p}(x_t)|_{a_t^a=b^i(o_t^i)}$$

$$\theta_{b^i} = \theta_{b^i} + \eta \nabla_{\theta_{b^i}} J(b^i)$$

ii. Update target actor networks:

$$\theta_{b^i} \leftarrow \tau \theta_{b^i} + (1 - \tau) \theta_{b^i}$$

(i) Update target critic networks:

$$\theta_{Q_{b'}^{c^{wt}}} \leftarrow \tau \theta_{Q_{b'}^{c^{wt}}} + (1 - \tau) \theta_{Q_{b'}^{c^{wt}}}$$

$$\theta_{Q_{b'}^{c^p}} \leftarrow \tau \theta_{Q_{b'}^{c^p}} + (1 - \tau) \theta_{Q_{b'}^{c^p}}$$

8. End for

End for

where each critic relates to a specific objective. We develop two centralized attentive critics in our work, referred to as Q_ϕ^e and Q_ϕ^{CC} respectively, and they correspond to the estimated returns of the reward functions r^{WPC} and r^{CC} . As the two critics share the same architecture, we only give as an illustration.

$$Q_\phi^e(x_t) = E_{s_t, u_t, p_t, r_{t:t+j}, R_{t:t+j} \sim D} \left[R_{t:t+j} + \gamma^{(T_{t+j}^c - T_t)} Q_\phi^e(x_{t+j})|_{u_{t+j}^a = b_{\theta^i}^i(o_{t+j}^i)} \right]. \quad (12)$$

where E denotes the environment, and $R_{t:t+j}^{CC}$ is the cumulative discounted reward (defined in Eq. (9)) concerning r^{CC} .

Additionally, we construct two centralized attentive critics that are linked to two objective-specific optimal policies for reward, denoted as Q_ϕ^{WPC} and Q_ϕ^{CC} , respectively. This enables us to measure the degree of convergence of different objectives. These are captured by the corresponding optimum policies $b_{\theta^{WPC}}^{i*}$ and $b_{\theta^{CC}}^{i*}$. In order to generate these objective-specific optimum policies and critiques, SRL-CAVCI can be pre-trained on a single reward. Then we compute the ratio of differences between the multi-objective policy and the optimum policy that is particular to each goal by :

$$g_t^{WPC} = \frac{Q_{\phi^{WPC*}}^{WPC}(x_{t+j})|_{u_{t+j}^a=b_{\theta^i}^i(o_{t+j}^i)} - Q_\phi^{WPC}(x_{t+j})|_{u_{t+j}^a=b_{\theta^i}^i(o_{t+j}^i)}}{Q_{\phi^{WPC*}}^{WPC}(x_{t+j})|_{u_{t+j}^a=b_{\theta^i}^i(o_{t+j}^i)}}, \quad (13)$$

The gap ratio g_t^{CC} can be derived similarly. Intuitively, a smaller gap ratio means the objective is well-optimized, which can be adjusted with a smaller step size, while a larger gap ratio means that it's poorly optimized and it needs to be bolstered by a larger update weight. Therefore, we come up with dynamic update weights, which the Boltzmann softmax function learns to adaptively modulate the step size of the two objectives.

$$\beta_t = \frac{\exp(g_t^{WPC}/\tau)}{\exp(g_t^{WPC}/\tau) + \exp(g_t^{CC}/\tau)}, \quad (14)$$

where τ is the temperature controlling the sensitivity of adjustment. Every agent of $EVSE^i \in C_t^a$ with the two critics defined above and adaptive update weights aims to learn an actor policy to maximise the following return.

$$J(b^i) = E_{s_t, u_t, p_t, r_{t:t+j}, R_{t:t+j} \sim D} \left[\beta_t Q_\phi^{WPC}(x_t) + (1 - \beta_t) Q_\phi^{CC}(x_t) |_{u_t^i=b_{\theta^i}^i(o_t^i)} \right]. \quad (15)$$

Algorithm 1 describes the full process of learning in SRL-CAVCI. Note that due to scalability reasons, we share the actor and critic network configurations across all agents.

3.1 Laxity Estimation

From our data we can get the requested energy demand, available minutes and actual energy delivered, actual charging time along with other features like user_ID, EVSE_ID etc. Based on this data we can find the requested and delivered energy rate as follows:

$$\lambda_n^{\text{req}}(t) = \frac{\sum_{v \in V} \epsilon_{n,v}^{\text{req}}}{\sum_{v \in V} \delta_{n,v}^{\text{req}}} \times 60. \quad (3.1)$$

$$\lambda_n^{\text{act}}(t) = \frac{\sum_{v \in V} \epsilon_{n,v}^{\text{act}}}{\sum_{v \in V} \delta_{n,v}^{\text{act}}} \times 60. \quad (3.2)$$

Where $\lambda_{\text{req}}(t)$ and $\lambda_{\text{act}}(t)$ represent the average energy demand rate and the energy delivery rate at EVSE $n \in \mathcal{N}$, respectively. $\epsilon_{\text{req},n,v}$ and $\epsilon_{\text{act},n,v}$ represent the energy

demand requested and actual energy delivered by EVSE $n \in \mathcal{N}$ to EV $v \in \mathcal{V}_n$. Lastly, $\delta_{\text{req},n,v}$ and $\delta_{\text{act},n,v}$ represent the session duration requested by EV $v \in \mathcal{V}_n$ at EVSE $n \in \mathcal{N}$ and the actual charging time by EVSE $n \in \mathcal{N}$ for EV $v \in \mathcal{V}_n$.

From the above equation we can find the laxity as follows:

$$L(x, y) = \min_{x \in X} E_{x \sim X} \left[\sum_{n \in \mathcal{N}} \sum_{v \in \mathcal{V}} \left| \frac{\epsilon_{n,v}^{\text{req}}}{\lambda_n^{\text{req}}(t)} - \frac{\epsilon_{n,v}^{\text{act}}}{\lambda_n^{\text{act}}(t)} \right| \right] \quad (3.3)$$

Where $\lambda_n^{\text{req}}(t)$ and $\lambda_n^{\text{act}}(t)$ are determined by the previous equations.

3.2 Conditional Value at Risk

In essence, it says that, if an investment is stable in the long run, its value at risk would be sufficient for the management of risk in a portfolio. The more un-safety of the investment, the more likely it is. And because the Value at Risk is invariant to what is outside of its own breakpoint, and it alone cannot paint a full picture of the risks. Statistical method to measure the level of financial risk in a company or investment portfolio over a given period of time is the Value at Risk (VaR) model. The deficiencies in the VaR model are meant to be corrected by Conditional Value at Risk, or CVaR. VaR is the worst case loss that is associated with a probability and time horizon; in contrast, CVaR is the expected loss in the rare case that the worst case breakpoint is ever reached. In other words, Conditional Value at Risk calculates the expected losses which occur beyond the VaR breakpoint.

We can define CVaR as follows:

$$R_\alpha^{\text{CVaR}}(L(x, y)) = -\frac{1}{\alpha(1-\omega)(\omega + \xi^2)P_\omega(\xi)\sigma_\mu} \quad (3.4)$$

We can find the $P_\omega(\xi)$ from,

$$P_\omega(\xi) = \frac{\Gamma\left(\frac{\omega+1}{2}\right)}{\Gamma\left(\frac{\omega}{2}\right)\sqrt{\pi\omega}} \left(1 + \frac{\xi^2}{\omega}\right)^{-\frac{\omega+1}{2}} \quad (3.5)$$

We consider degree of freedom by ω and μ , σ consecutively mean, and standard deviation, respectively and $L(x, y)$ where d and ξ represents a sample of laxity and a cut-off point of the laxity tail-risk respectively. Thus, the probability density function (pdf) for student t-distribution is defined as follows:

$$P(d, \omega, \mu, \sigma) = \frac{\Gamma\left(\frac{\omega+1}{2}\right)}{\Gamma\left(\frac{\omega}{2}\right)\sqrt{\pi\omega\sigma}} \left(1 + \frac{(d-\mu)^2}{\omega\sigma^2}\right)^{-\frac{\omega+1}{2}} \quad (3.6)$$

where the gamma function represents as $\Gamma(\cdot)$. We fit the t-distribution to observational laxity $d_1, d_2, \dots, d_J \in D$. This fit is computed by maximizing a log-likelihood function $l(D; \omega, \mu, \sigma)$ that is defined from

$$l(D; \omega, \mu, \sigma) = D \log \Gamma \left(\frac{\omega + 1}{2} \right) + \frac{D\omega}{2} \log(\omega) - D \Gamma \left(\frac{\omega}{2} \right) - \frac{D}{2} \log \sigma - \frac{\omega + 1}{2} \sum_{j=1}^J \log \left(\omega + \frac{(d_j - \mu)^2}{\sigma^2} \right) \quad (3.7)$$

where ω represents the degree of freedom, μ represents the mean and σ represents the standard deviation.

Chapter 4

Preliminary analysis

In the main dataset we found the requested charge and available minutes and also the actual delivered charge and charging time along with the User id which denotes each EVs and Station Id, that denotes the EVSE id. From that dataset, we used a linear model that calculated the laxity of each EVs for each EVSE. As we worked with the ACN dataset[8], we splitted time session for one hour and calculated the above function for each time slot. Finally, in order to determine the values of the DOF (Degree of Freedom), mean, and standard deviation, we computed the PDF for the student t-distributions. This parameter will help us to find the value for CVaR(Conditional Value at Risk).

4.1 Requested Charge vs Actual Delivered Charge

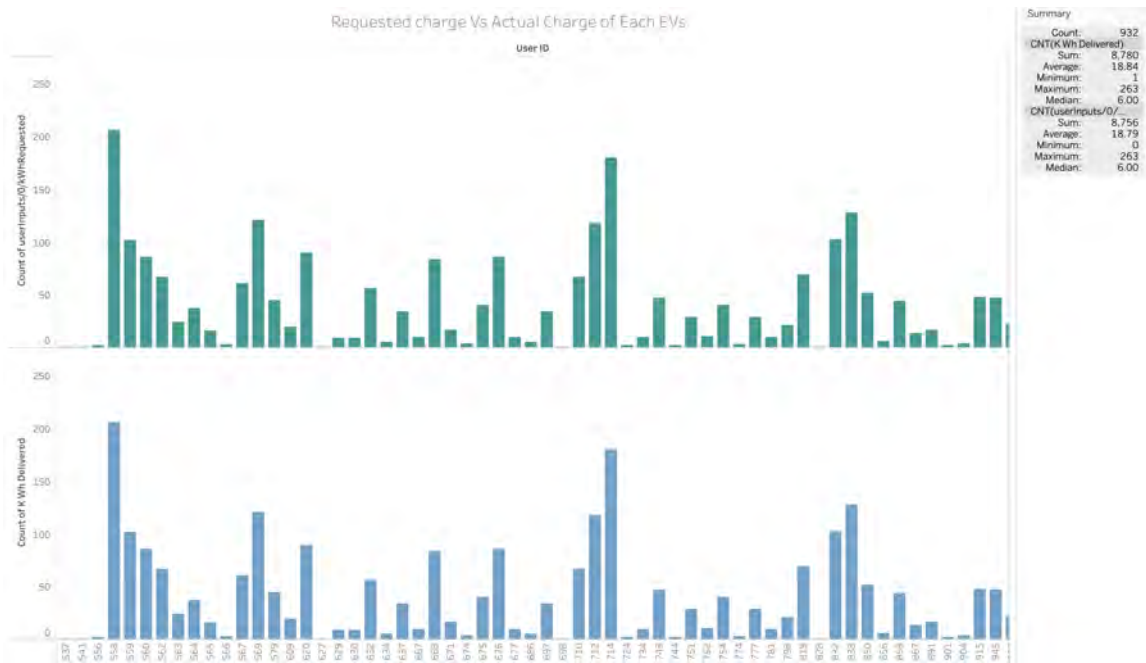


Figure 4.1: Requested Charge vs Actual Delivered Charge

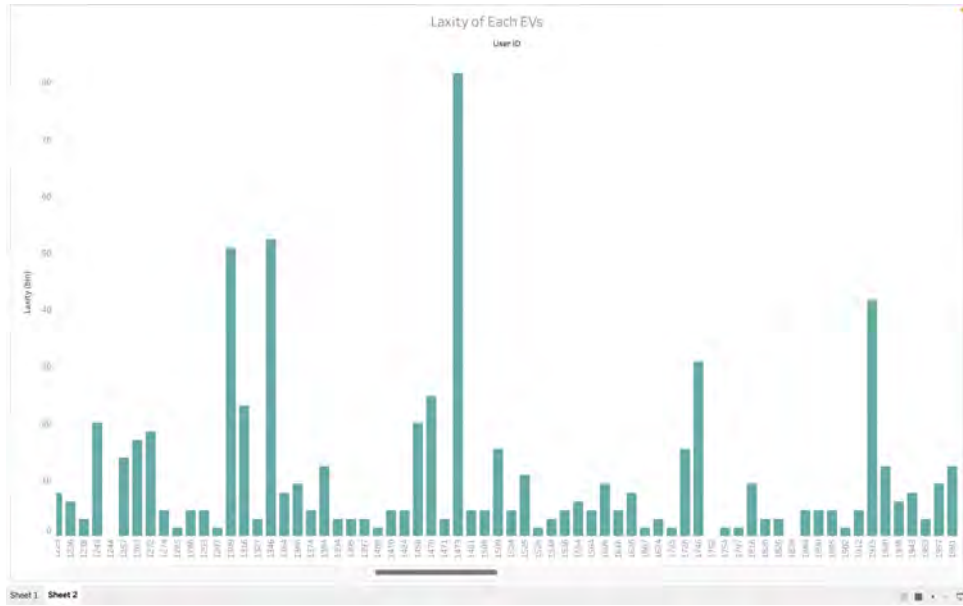


Figure 4.2: Laxity of Each EV

In these above diagrams, the irrational charging request has been shown. In the Caltech dataset, we can see that each of the electric vehicles requested for a charge, but except the autonomous vehicle, all the vehicles requested irrationally. Moreover, in some cases, the vehicles requested more than double, and that’s why the energy utilization is not being sufficient. In Figure 4.2, we can find the irrational charging request from the user.

4.2 Laxity Sum of Each EVSE

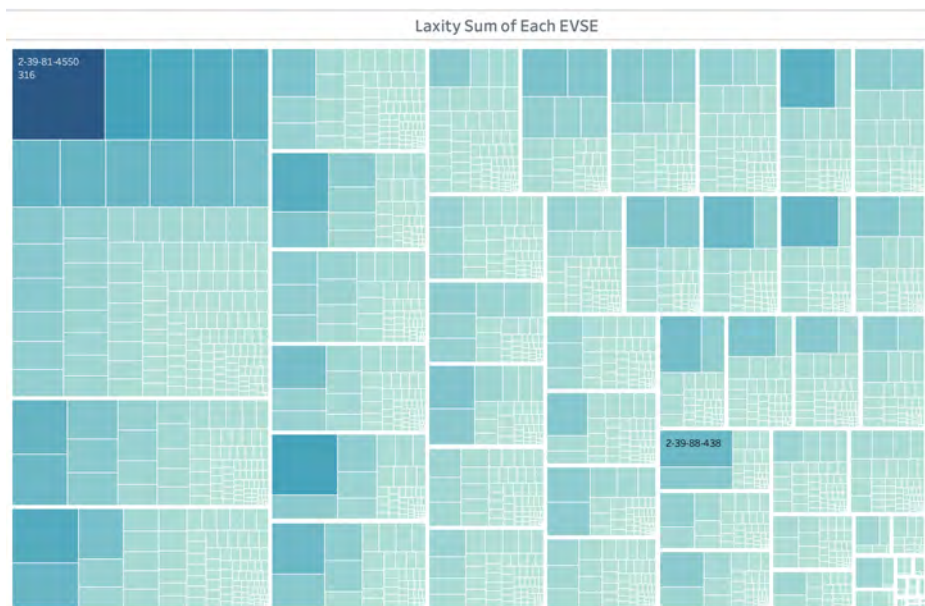


Figure 4.3: Laxity Sum of Each EVSE

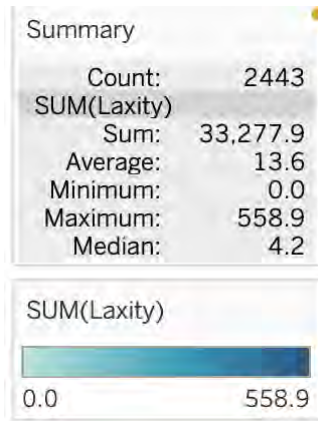


Figure 4.4: Summary of Laxity Sum of Each EVSE

In this diagram, we have represented our model outcome that calculated the laxity of each EVSE and EV. Each of the white line borders represents the EVSE, and the small blocks represent the EVs. The block size represents the laxity of each EV and EVSE, respectively.

4.3 Laxity(Sum) vs PDF(Sum) of each EVSE

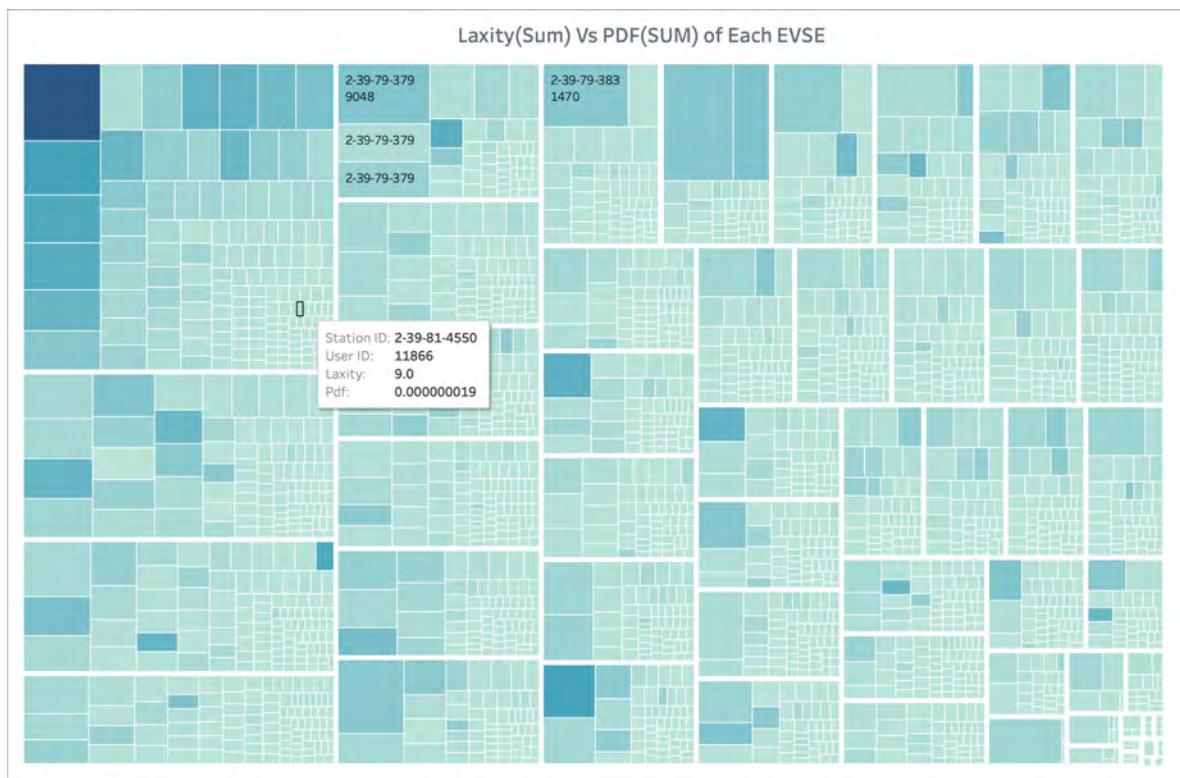


Figure 4.5: Laxity(Sum) vs PDF(Sum) of each EVSE

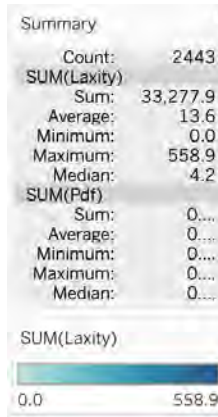


Figure 4.6: Summary of Laxity(Sum) vs PDF(Sum) of each EVSE

In the above diagram, we calculated the PDF value of the EVs and EVSEs, respectively. But here, the small block size represents the PDF value that will be required to calculate the conditional value at risk, and the white-bordered block represents the laxity of each EVSE.

Chapter 5

Performance Evaluation for SRL-CAVCI

Data description

We evaluate the SRL-CAVCI using datasets simulated for Beijing and Shanghai, which are two of the most populous cities in China. The datasets are from July 1st, 2019, and May 18, 2019. The data consists of all supply availability records, charging prices, and charging power collected from charging stations using an app available to the public that compiles real-time sensor data. Each city is further divided into grids that are 1×1 km² in size. The number of 15-minute charging requests from the grid, which consists of the station and its eight nearby grids, is summed up to calculate future demand for the related charging stations for every station. An Electronic Vehicle charging indicator simulator was used for making the real-world dataset. The training set is composed of the 28 day’s consecutive data, the validation set consists of the following three days, and the rest 14 days are used for testing.

Implementation Specifics

An 8-core M1 Mac server was used to do conduct all these experiments. We choose a discount factor $\gamma = 0.99$ to learn all the algorithms of RL, use temperature $\sigma = 0.2$ to tune the adjustment of updated weights, and set $d = 30$ minutes for modeling the charging competition. All actor and critic networks are composed of three 64-dimension linear layers and a ReLU activation for the hidden layers. The soft update of the target networks uses a $\tau = 0.001$ value, the size of the replay buffer is 1000, and the batch size is 32. In the training of our model, the learning rate is set to 5×10^{-4} , and we use the Adam optimizer for all learnable algorithms. All major hyper-parameters of each baseline are then fine-tuned through a grid search. All RL algorithms are trained for 52 iterations to recommend the top fifty nearest EVSEs; the validation set picks the best iteration to test.

Evaluation metrics

Four measurements are set up to evaluate the efficacy of our methodology and baseline indicator algorithms. We define the set of charge requests which accept our indications as Q^a . We also define the collection of charge requests which accept our

indications and which end up starting charging as $Q^e \setminus Q^a$. The cardinalities of Q^a and Q^e are denoted as $|Q^a|$ and $|Q^e|$, respectively.

We define the Mean Waiting time for Charging (MWPC) considering all the charging requests $q_t \in Q^a$ to evaluate the global waiting time for charging of our metrics:

$$\text{MWPC} = \frac{\sum_{q_t \in Q^a} \text{WPC}(q_t)}{|Q^a|} \quad (5.1)$$

where $\text{WPC}(q_t)$ is the waiting period for charging (in minutes) of charging request q_t .

We define the Mean Cost of Charging (MCC) over all charging requests $q_t \in Q^a$ to evaluate the average cost of charging, :

$$\text{MCC} = \frac{\sum_{q_t \in Q^a} \text{CC}(q_t)}{|Q^a|} \quad (5.2)$$

where $\text{CC}(q_t)$ represents the cost of charging in q_t (in CNY).

We now define the Total Saving Fee (TSF), which we use to compare our indicator method to the ground truth charging activities to compute the average daily total saving fees:

$$\text{TSF} = \frac{\sum_{q_t \in Q^a} (\text{RC}(q_t) - \text{CC}(q_t)) \times \text{CQ}(q_t)}{N_d} \quad (5.3)$$

where $\text{CQ}(q_t)$ is the electric charging quantity of q_t , N_d is the number of evaluation days, and $\text{RC}(q_t)$ is the cost of charging of the ground truth charging action. Note that the TSF, which represents the amount of fees overpaid with respect to the ground truth charging activities, can be negative.

In order to quantify the percentage of failures of the charging in our indices, we finally define the Failure Rate of Charging as follows:

$$\text{FRC} = 1 - \frac{|Q^e|}{|Q^a|} \quad (5.4)$$

Baselines

We compared our approach to the SRL-CAVCI and five baselines introduced in [13]:

- **Real:** The ground truth charging activities of the charge requests are real.
- **Random:** It randomly recommends charging stations for requests for charging.
- **Greedy-N:** recommends the closest EVSE.
- **Greedy-P:** recommends the cheapest EVSE.
- **Greedy-P-N:** recommends, for a given parameterized ratio, the top- N closest and cheapest charging stations.

Table 1: All comparative baselines on two datasets and the holistic results of our methodologies under each of our four criteria. In general, as can be shown, SRL-CAVCI is ahead of all other baselines concerning overall performance. In terms of MCC, TSF, and FRC, compared to the ground truth charging activities, SRL-CAVCI has a decrease of 16.2%, 12.5%, and 42.1%, respectively.

Algorithm	MWPC	MCC	TSF	FRC
Real	21.51	1.749	-	25.9%
Random	38.77	1.756	-447	52.9%
Greedy-N	20.27	1.791	-2527	31.3%
Greedy-P-5	23.40	1.541	9701	35.4%
Greedy-P-10	26.03	1.424	14059	39.9%
SRL-CAVCI	42.37	1.50	12304	10.1%

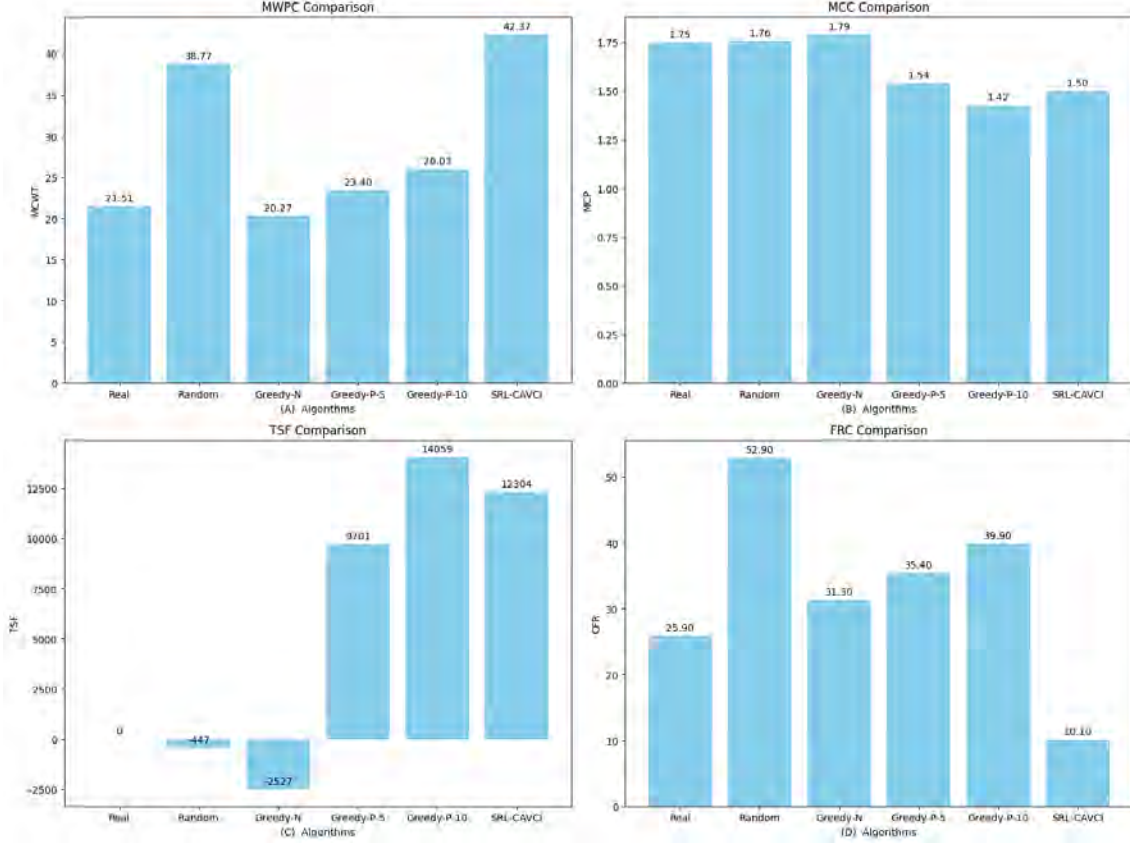


Figure 5.1: Comparison of Different Algorithms

Correlation Between Reward and Success Rate

In our SRL-CAVCI, we used the Actor-Critic Method, which is a DDPG algorithm. Although we only iterate 52 times, we observe a significant impact of reward along with the success rate.

According to the graph, we can analyze that the success rate increases along with rewards. We used negative rewards for successful charging. If EVs are successfully charged, they incur a negative reward proportional to $-WPC - CC$; thus, as WPC and CC increase, the reward becomes more negative, and if WPC and CC are minimal, the reward is minimally negative. Additionally, a penalty is imposed for charging failures.

When the iteration count was very low, the frequency of successful indicators was less, but as iterations increased, the actor-critic architecture rapidly trained using feedback from the critic network.

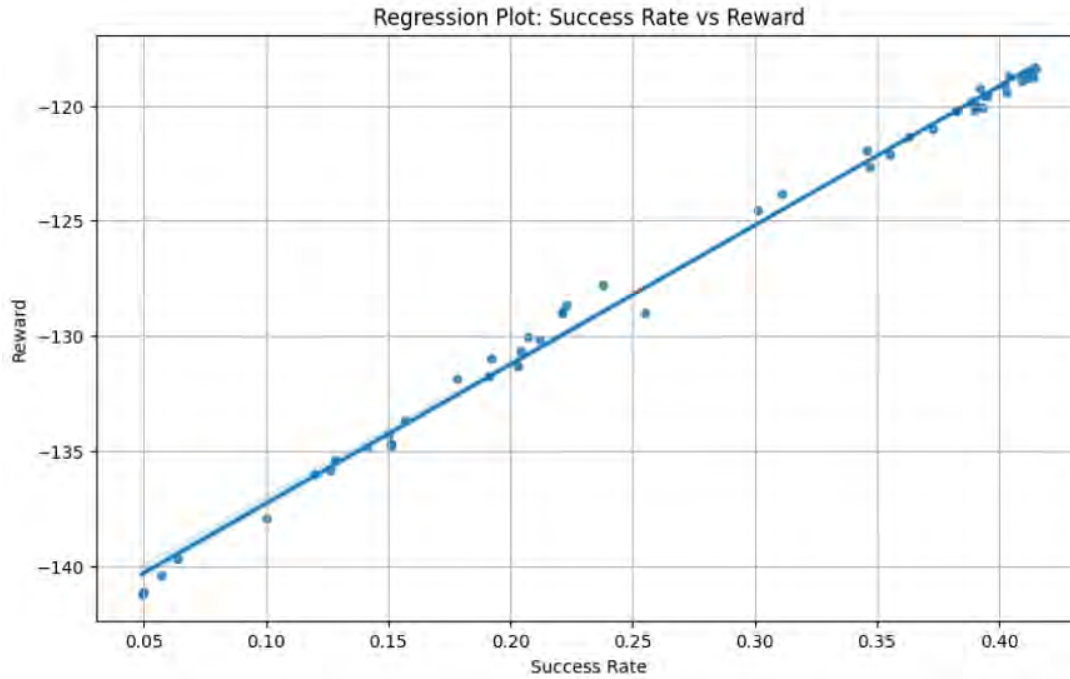


Figure 5.2: Regression Rate: Success Rate vs Reward

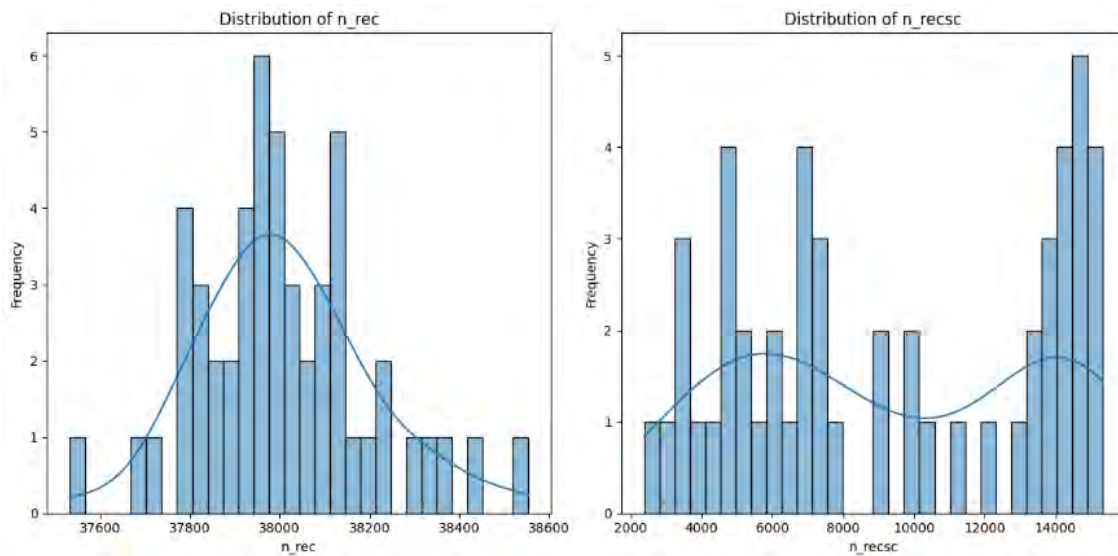


Figure 5.3: Distribution of n_rec and n_repsc

For the same number of indicators, the success rate was minimal when iterations were fewer than 10, but after 29 iterations, the success rate of the indicator increased significantly. However, with limited computational power, we only iterated 52 times, and based on this, there is no indication of a saturation stage. Moreover, the graph suggests that the success rate of the indicator will continue to increase with more iterations.

Comparison between Training and Evaluation

As DDPG works in an Actor-Critic architecture, the critic evaluates the action based on specific parameters, when the agent takes an action. After each iteration, both the critic and actor networks are updated.

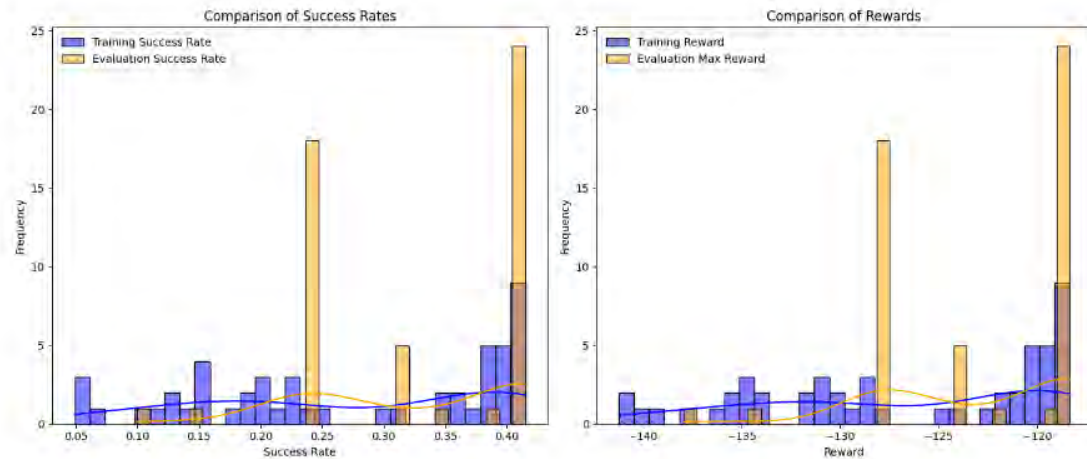


Figure 5.4: Comparison of Success Rates and Rewards

From our SRL-CAVCI model, we observe that the model performs well during execution. The frequency of reward and success rate is higher than in the training phase. Additionally, the graph shows a continuous process with no saturation stage.

Chapter 6

Conclusion

Future Work

In our research we are trying to solve the issues of charge failure and reduce the waiting period for charging and cost of charging. But compared to other models we can see the mean waiting period of charging time is significantly high but the success rate is good. Moreover, not all the vehicles were given an indication for better charging, only those who requested online got the indication. So there was much of a problem correlating the queuing success and indication success. Our aim is to bring all the vehicle charging infrastructure in one centralized system. To make this possible we need to optimize our model.

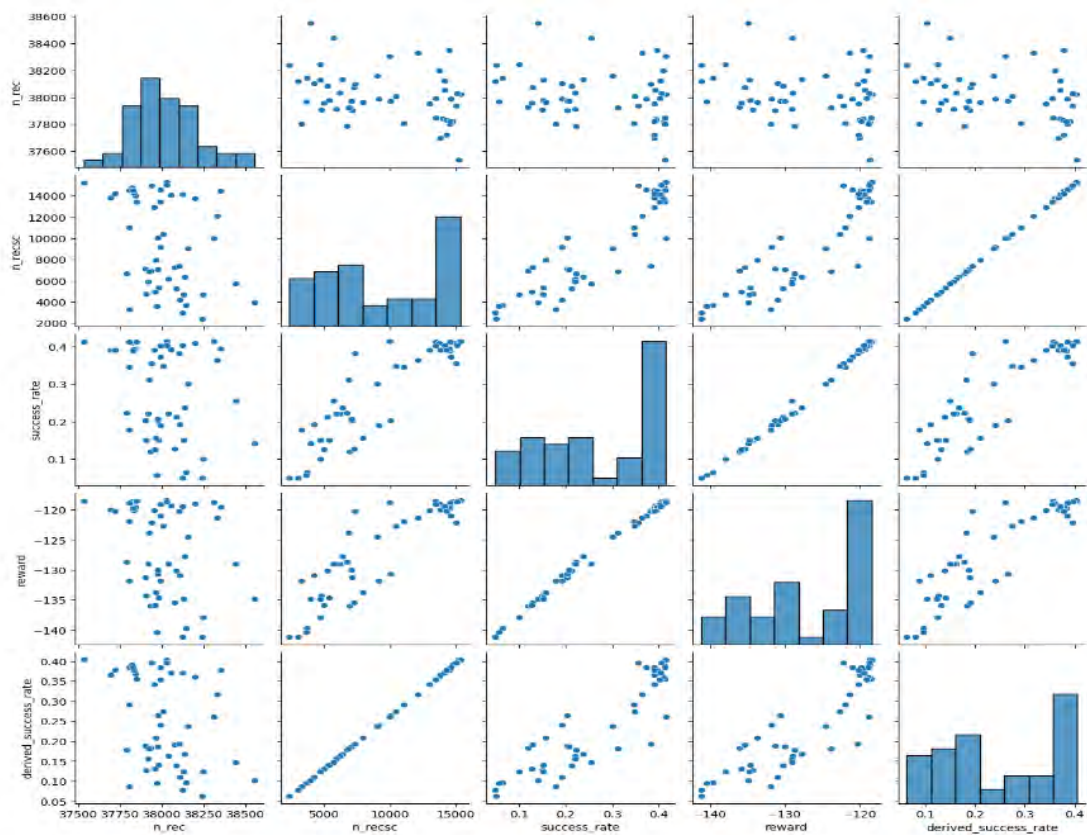


Figure 6.1: Pair plot of (Success rate, Derived success rate and Reward) Vs Indication

From the above diagram we can see that the indication of success and derived success is increasing rapidly along with rewards but rewards are still in negative. Our aim is to bring this near 0. To optimize our model, we are proposing a Safe RL and gradient descent based approach which we have already used for our reward and success rate optimization. But we didn't reduce MWPC and MCC as our initial focus was to succeed in the charging competition. To use Safe RL based optimization, we will add an extra layer in the last layer of our DDPG model which we are already working on.

Conclusion

In conclusion, this research introduces an innovative Intelligent Decision Support System (IDSS) designed to optimize the charging process for both human-driven and autonomous electric vehicles, addressing challenges related to excessive charging requests. The system encourages rational charging behaviors, excelling in identifying the nearest charging equipment and managing detailed requests. Its adaptive capability allows it to refine scheduling based on accumulated data, demonstrating superior efficiency, safety, and energy utilization in empirical tests. The study also explores a Safe Reinforcement Learning-based system to enhance charging infrastructure for connected and autonomous vehicles (CAVs). By promoting rational charging through a Rational Decision Support System (RDSS) and optimizing processes with Safe Reinforcement Learning algorithms, the research aims to overcome challenges tied to irrational energy demands. The proposed system, incorporating Risk Adversarial Agents (RAA) and local self-learning agents, contributes to intelligent decision-making, supporting the widespread acceptance of CAVs. Looking forward, upcoming research will implement the Safe RL AVCI model, expanding datasets to JPL and both Caltech sites and constructing a linear model for execution and indication function calculation. This collaborative effort signifies a significant stride towards smarter, safer, and more technologically advanced transportation networks.

Bibliography

- [1] Y. Li, J. Luo, C.-Y. Chow, K.-L. Chan, Y. Ding, and F. Zhang, “Growing the charging station network for electric vehicles with trajectory data analytics,” in *2015 IEEE 31st international conference on data engineering*, IEEE, 2015, pp. 1376–1387.
- [2] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, “Learning to communicate with deep multi-agent reinforcement learning,” vol. 29, 2016.
- [3] T. Guo, P. You, and Z. Yang, “Recommendation of geographic distributed charging stations for electric vehicles: A game theoretical approach,” in *2017 IEEE Power & Energy Society General Meeting*, IEEE, 2017, pp. 1–5.
- [4] J. K. Gupta, M. Egorov, and M. Kochenderfer, “Cooperative multi-agent control using deep reinforcement learning,” in *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Best Papers, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers 16*, Springer, 2017, pp. 66–83.
- [5] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” vol. 30, 2017.
- [6] T. Chu, J. Wang, L. Codecà, and Z. Li, “Multi-agent deep reinforcement learning for large-scale traffic signal control,” 3, vol. 21, IEEE, 2019, pp. 1086–1095.
- [7] Y. Ge, F. Zhu, X. Ling, and Q. Liu, “Safe q-learning method based on constrained markov decision processes,” *IEEE Access*, vol. 7, pp. 165 007–165 017, 2019.
- [8] Z. J. Lee, T. Li, and S. H. Low, “Acn-data: Analysis and applications of an open ev charging dataset,” in *Proceedings of the tenth ACM international conference on future energy systems*, 2019, pp. 139–149.
- [9] W. Fan, K. Liu, H. Liu, P. Wang, Y. Ge, and Y. Fu, “Autofs: Automated feature selection via diversity-aware interactive reinforcement learning,” in *2020 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2020, pp. 1008–1013.
- [10] S. Alighanbari and N. L. Azad, “Safe adaptive deep reinforcement learning for autonomous driving in urban environments. additional filter? how and where?” *IEEE Access*, vol. 9, pp. 141 347–141 359, 2021.
- [11] P. W. Shaikh and H. T. Mouftah, “Connected and autonomous electric vehicles charging reservation and trip planning system,” in *2021 International Wireless Communications and Mobile Computing (IWCMC)*, IEEE, 2021, pp. 1135–1140.

- [12] P. W. Shaikh and H. T. Mouftah, “Intelligent charging infrastructure design for connected and autonomous electric vehicles in smart cities,” in *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, IEEE, 2021, pp. 992–997.
- [13] W. Zhang, H. Liu, F. Wang, *et al.*, “Intelligent electric vehicle charging recommendation based on multi-agent reinforcement learning,” in *Proceedings of the Web Conference 2021*, 2021, pp. 1856–1867.
- [14] M. Mohammadpour, S. Kelouwani, M.-A. Gaudreau, *et al.*, “Energy-efficient local path planning of a self-guided vehicle by considering the load position,” *IEEE Access*, vol. 10, pp. 112 669–112 685, 2022.
- [15] M. S. Munir, K. T. Kim, K. Thar, D. Niyato, and C. S. Hong, “Risk adversarial learning system for connected and autonomous vehicle charging,” *IEEE Internet of Things Journal*, vol. 9, no. 16, pp. 15 184–15 203, 2022.
- [16] A. Prasetyadi, M. Y. Rezaldi, H. M. Saputra, B. Nugroho, and C. Trianggoro, “Conceptual design of charging stations for autonomous vehicle,” in *2022 5th International Conference on Networking, Information Systems and Security: Envisage Intelligent Systems in 5g//6G-based Interconnected Digital Worlds (NISS)*, IEEE, 2022, pp. 1–5.
- [17] W. Wang, X. Zhou, B. Xu, M. Lu, Y. Zhang, and Y. Gu, “A safe and self-recoverable reinforcement learning framework for autonomous robots,” in *2022 41st Chinese Control Conference (CCC)*, IEEE, 2022, pp. 3878–3883.
- [18] C. Kim, Y. Yoon, S. Kim, M. J. Yoo, and K. Yi, “Trajectory planning and control of autonomous vehicles for static vehicle avoidance in dynamic traffic environments,” *IEEE Access*, vol. 11, pp. 5772–5788, 2023.