Charging Schedule Optimization of Electric Buses Based on Reinforcement Learning

by

Wenzhuo Chen

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science
in
Energy Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

In recent years, due to the environmental concerns caused by the emissions from public transit services relying on traditional fossil fuels, the electrification of the public transit sector has attracted great attention from both automobile industry and academia. Specifically, the electric buses (EBs) potentially driven by decarbonized electricity can reduce air pollutions while achieving energy savings from regenerative braking. Yet, for a large-scale deployment of EBs in public transit services, technical challenges associated with charging schedule optimization for electricity cost and battery degradation cost reduction still need to be addressed. These challenges are further complicated by the uncertainties in EB operation related to the randomness in road and traffic conditions, passenger counts, and arrival and departure times of EBs at bus stations. In this thesis, we address these technical challenges by developing model-free reinforcement learning (RL) approaches to optimize the charging schedules of EBs. Compared with the traditional model-based approaches, the proposed RL approaches do not rely on specific models of the aforementioned uncertainties, such that they can be implemented in real-world public transit services with great flexibility. Specifically, three research topics related to EB charging schedule optimization are investigated in this thesis.

Firstly, a Markov decision process (MDP) is developed to model the operation process of EBs with in-station charging capabilities, for which the EBs are only charged at specific bus stations such as terminals and/or transit centers with pre-determined charging durations. Then, a double Q-learning algorithm is utilized to optimize the amount of power to charge each EB at each charging station. By utilizing the battery degradation cost as the reward of the RL, the optimal charging strategy for EB operation cost reduction can be obtained through an iteration process. In the case study, the performance of the proposed RL approach is evaluated based on the real-world EB operation data obtained from St. Albert Transit, AB, Canada. And the results indicate that our approach can reduce the battery degradation cost in comparison with other existing approaches.

By considering the en-route EB charging applications, for which the EBs are charged momentarily when they pick up and/or drop off passengers, an extension of the above MDP and RL approach is investigated in our second work. Specifically, a physical EB model and a battery degradation model are built to calculate the EB energy consumption and battery degradation cost, respectively. Then, a semi-Markov decision process (SMDP) is developed to characterize the operation process of EB. The main difference between the SMDP and MDP is that, for SMDP, the duration in between two adjacent charging decision-making epochs can be random, which can better characterize the real-world en-route charging operation conditions of EBs due to the randomness in road and traffic conditions, as well as the uncertainties in passenger pick-up and drop-off times. Accordingly, an average reward reinforcement learning (ARRL) approach is proposed to optimize the en-route charging strategy of EBs. The efficiency of the proposed approach is demonstrated via the real-world EB operation data provided by St. Albert Transit and the results are compared with that of the traditional charging approaches.

To further improve the efficiency of EB en-route charging, a relative value iteration reinforcement learning (RVIRL) approach is proposed in our third work. Based on the energy consumption and battery degradation models of EB operation, an extended SMDP problem is formulated to determine the charging schedule of EB on the route by considering the SoC changes, number of charging stations, maximum sojourn time at charging station, and real-time electricity pricing. Then, the RVIRL approach is utilized to obtain the optimal EB charging strategy for each en-route charging station. The convergence of the RVIRL approach is proved mathematically, which is critical to ensure the reliable operation of public transit services with EBs. The performance of the proposed approach is evaluated based on the real-world data obtained from St. Albert Transit. And the results indicate that the proposed approach can significantly reduce the electricity cost and battery lifetime degradation in comparison with other existing en-route EB charging approaches.

Preface

The material presented in this thesis is based on the original work by Wenzhuo Chen. As detailed in the following, material from some chapters of this thesis has been published or submitted for publication under the supervision of Dr. Hao Liang in concept formation and by providing comments and corrections to the article manuscript.

Chapter 2 includes the results published in the following paper:

- W. Chen, P. Zhuang, and H. Liang, "Reinforcement learning for smart charging of electric buses in smart grid," in *Proc. IEEE GLOBECOM'19*, Dec. 2019.

Chapter 3 includes the results published in the following paper:

- W. Chen and H. Liang, "Average reward reinforcement learning for optimal on-route charging of electric buses," in *Proc. IEEE VTC'20-Fall*, Nov. 2020.

Chapter 4 includes the results in following paper that has been submitted:

- W. Chen and H. Liang, "En-route smart charging for battery electric buses based on relative value iteration reinforcement learning," *IEEE Transactions on Smart Grid*, under review.

Acknowledgements

First, I would like to express my sincere gratitude to my supervisor *Prof. Hao Liang* for the continuous support of my M.Sc. study and research. All his patience, motivation, and immense knowledge have helped me so much in all the time of research and writing of this thesis.

In addition, it is an honor for me to extend my gratitude to all my M.Sc. committee members for reviewing my thesis and providing thoughtful comments to improve it. I also thank all my colleagues in my lab and friends I met at the University of Alberta during my M.Sc. program, especially Dr. Yuan Liu, who helped me so much when I started my first research work.

Finally, I would like to thank my parents for their unconditional love and support. Though we are thousands of miles apart, they are always there to encourage me to move forward.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ARRL	Average reward reinforcement learning
BSS	Battery swap station
DCM	Darken-Chang-Moody
DOD	Depth of discharge
EB	Electric buses
FQI	Fitted Q-iteration
KAIST	Korea Advanced Institute of Technology
MDP	Markov decision process
OLEV	Online electric vehicle
RIW	Random in window
RL	Reinforcement learning
RTP	Real time price
RVI	Relative value iteration
SOC	State of charge
SMDP	Semi-Markov decision process

Nomenclature

Chapter 2

ΔSOC SOC cost for capacity fade

ϵ, ζ Battery curve fitting parameter

c_{bat} Battery price

C_{DOD}^{P} DOD cost for power fade

C_{SOC}^{P} SOC cost for power fade

C_{DOD}^{Q} DOD cost for capacity fade

C_{SOC}^{Q} SOC cost for capacity fade

C_d Battery degradation cost

CF_{max} Capacity fade for 80%

$E_{T,discharge}$ Discharge Energy throughput

$E_{T,use}$ Actualm Energy throughput

E_{tp} Energy throughput

k Bus state

L Battery lifetime

L_c Lifetime degradation

L_{tot} Total battery lifetime

N Bus station index

Q Battery capacity

$SOC_{charging}^{max}$ Maximum charging SOC

SOC_{avg} Average SOC

Chapter 3

α_m Learning rate

β_e Constant parameter for EB

β_r Road fraction parameter

β'_r Road fraction parameter

β_v Constant parameter for velocity

ΔSOC_{avg} Average SOC

ΔSOC_{dev} Normalized deviation SOC

η_d Driver's efficiency

η_e Engine's efficiency

ρ Air density

Θ_2 Battery degradation paremeter

Θ_h Battery degradation paremeter for 50% SOC

A Front area

c_a Aerodynamic dragging coefficient

c_r Road fraction coefficient

$C_{battery}$ Total battery capacity

d Distance

f_θ Slope of road

F Total force

F_a Drag force

F_d Battery degradation

F_f Friction force

F_r Resultant force

F_s Gravity decomposed force

g_r Gear ratio

K_γ DOD exponent

K_σ Throughput parameter

L Battery life aging parameter

m Mass of EB

n Charging station index

N Number of throughput cycle

p_m Exploration rate

p_n Electric price

P Power consumption

r Radius of tire

S Battery stress level

SOC_n^a Arrvial SOC

SOC_n^d Depature SOC

SOC^{max} Maximum SOC

SOC_{min} Minimum SOC

t_{life} Cycle time

v Velocity of EB

v_e angular engine speed

Chapter 4

α_γ DOD exponent

α_σ Throughput constant parameter

β_e EB constant parameter

ΔE Energy consumption

ΔSOC_n^c Charging SOC

η_d Driver efficiency

η_e Engine efficiency

γ_r	Road fraction parameter
γ_r'	Road fraction parameter
ρ_a	Air density
ϖ	Battery degradation parameter
ξ_v	Velocity constant parameter
ζ_a	Aerodynamic dragging coefficient
ζ_r	Road fraction coefficient
A	Front area
$\hat{C}_b$	Total battery cost
d	Real time position
f_θ	Slope of road
F_t^a	Air drag force
F_t^f	Fraction force
F_t^r	Resultant force
F_t^s	Gravity decomposed force
F_t	Total force
g_r	Gear ratio
m_h	Mass of human
m_{bus}	Mass of bus
N	Bus station index
p_n	Real time price
$P_{n,t}$	Charging power
Q_{nom}	Battery nominal charge capacity
r_{bus}	Tire radius
S	Stress level
$\overline{SOC}^{max}$	Maximum SOC level

$\underline{SOC}_{min}$ Minimum SOC SOC level

t_c Time cycle

t_l Total expected shelf life time

T Time index

T_n^{stop} Maximum stay time

$u_{(n,n+1)}$ Passenger number

v Real time speed

1 Introduction

1.1 Background

Over the past few decades, environmental pollution and energy crisis are two of the major challenges faced by the whole world. Traditional energy sources such as fossil fuels, are major contributers to these challenges. Accordingly, renewable energy sources such as solar and wind, have been widely utilized in recent years to generate electricity and drive wheels. Aiming at reducing the metropolitan air pollution resulted from traditional fossil fuel-powered automobiles, the electrification of public transportation sector based on electric buses (EBs) has attracted great attention from both transportation industry and academia. Compared to the conventional internal combustion engine (ICE) buses powered by fossil fuels such as gasoline or diesel, the EBs can effectively alleviate the environmental concerns, reduce the speed of natural resource depletion, facilitate the generation of electricity from renewable energy resources, and offer better fuel economy and higher energy efficiency for the public transit services. For example, St. Albert was the first city in Canada to have EBs serving full-time transit routes. In 2017, St.Albert's first EB went into regular service. In the same year, a high-efficiency 301kW solar panel system was installed at the transit facility which supports one-third facility's electricity demand [1].

According to the data from CAIT Climate Data Explorer [2], the emission from the transport sectors contributes to about 14% of annual emissions as well as around 25% of the CO_2 emissions from the burning fossil fuels of the world. In terms of the transportation modes, over 72% of the global transport emissions comes from road vehicles, especially conventional fossil-fuel buses. Aiming to reduce the local air pollution resulted from gasoline or diesel-powered buses, more and more public transit services are turning to environmentally friendly alternatives such as liquefied natural gas and compressed

natural gas. In the meantime, EBs have been gradually adopted by many public transit services. The advantages of EBs can be attributed to the following points. Firstly, EBs driven by de-carbonized electricity can reduce the air pollution and noise level. Secondly, EBs can also recover electricity from regenerative braking, similar to the trolleybuses. Finally, EBs do not rely on fossil fuel resources to produce the required traction force. Multiple environmentally friendly energy sources (e.g., solar, wind, geothermal, hydropower, biomass, tidal, and wave) can be utilized to generate electricity and drive wheels.

Around the world, many countries in North America, Europe, and Asia have used EBs in their public transportation sectors for more than a decade. Specifically, the data from [3] indicates that the EB market size exceeded USD 28 billion in 2020 and is expected to grow at 11% CAGR to USD 53 billion between 2020 and 2027. And the global EB market size is projected to reach 935 thousand units by 2027 from 137 thousand units in 2019, at a compound annual growth rate (CAGR) of 27.2% [3]. The market is forecast to grow at an exponential rate due to the rapid increase in uptake of EBs as a sustainable mode of transport. The Asia-Pacific region is the largest electric bus market in the world at the moment. The growth in this region can be attributed to the dominance of the Chinese market and the presence of leading OEMs such as BYD, Yutong, Zhongtong, and Ankai in the country, resulting in the tremendous growth of the Asia-Pacific electric bus market. Middle East & Africa, which includes Egypt, South Africa, and UAE, is projected to be the fastest-growing market during the forecast period. The increasing demand for electric mass transit solutions, renowned OEMs expanding in the region, and government support are factors driving the Middle East & Africa electric bus market. UK and Norway are the domination forces of the Western Europe EB market. North America has the second-largest EB market in the world. The key companies include BYD, Volvo, Proterra, Yutong, Daimler, and Zhongtong. Other key contributors are Solaris, Ashok Leyland, Alexander Dennis, EBUSCO, and New Flyer [4].

However, compared to the conventional buses using fossil fuels, the EBs generally have a shorter driving range. Accordingly, EBs require a larger battery package and therefore there are two main challenges that need to be considered: long charging time and the high cost of large size batteries. Based on the data from BYD K9 EB [5], the battery capacity is 500 kWh, and the charging time is around 2.5 hours. The large battery packs also cost more for the initial investment. In other words, the total cost of EBs is generally more than that of the conventional buses. However, [6] indicates that the large batteries of EBs can also be utilized as generation resources for resilient emergency response, which can be considered as an extra benefit of EB applications.

Due to the challenges of the utilization of EBs, many research works are devoted to address the corresponding issues. The sizing of the energy storage system plays an important role for the EBs since the energy content and size of the energy storage system (ESS) must satisfy the requirement of the energy and power density, charging rate, safety, cycle

life, cost, etc. Based on the related research works in [7], a hybrid battery system (HBS) or hybrid ESS (HESS) can be used to replace the traditional battery of the EBs. Besides the replacement of the battery, some approaches have been proposed to optimize the battery size in order to improve the performance of the EBs. However, due to the battery degradation model and the switching nature of the HBS and HESS operation, the optimization problem of ESS sizing is nonlinear with integer or mixed-integer variables. In other words, this problem does not have a closed-form solution. In addition, power/energy management can also be used to improve the efficiency of EB motor operation and control. Generally, convex optimization can be utilized to handle the energy management issue [8]. However, due to the non-convexity caused by the binary or integer variables in some EB energy management problems, they cannot be included in the convex optimization. The limitation of the operation range of EBs can be handled by a range of remedy methods. The two common ways are battery swapping and battery charging [9], [10]. However, the cost of the battery swapping and charging stations needs to be minimized through optimal planning. Besides the installation of the battery swapping/charging stations, how to optimize the charging schedule of EBs is also a challenging issue. This technical challenge is further complicated by the uncertainties in EB operation related to the randomness in road and traffic conditions, passenger counts, arrival and departure times of EBs at bus stations. In this thesis, we address the EB charging schedule optimization by developing model-free reinforcement learning (RL) approaches. Compared with the traditional model-based approaches, the proposed RL approaches do not rely on specific models of the aforementioned uncertainties, such that they can be implemented in real-world public transit services with great flexibility.

1.2 General Terms and Definitions

In this section, the important terms used in this thesis are defined to clearly identify the scope of work done in this research.

1.2.1 Markov Decision Process

Markov decision process (MDP) is a discrete-time stochastic control process that has been used in a discrete, stochastic, and sequential environment [11]. The key point of this model is that at each state, the agent will choose appropriate action based on the current environment. After the agent selects the decision, the state will change to another state accordingly. During this process, the immediate reward generated by the agent is affected by the state of the environment as well as the probabilities of the next state transition. The goal of the MDP is to find the optimal strategy for the agent to choose the suitable action at each decision state, and accordingly, the long-term total reward during the process is maximized based on the selected actions.

1.2.2 Semi-Markov Decision Process

The semi-Markov decision process (SMDP) is an extension of the traditional MDP which can be utilized in the modeling of stochastic control problems [12], [13]. The difference between SMDP and MDP is that the decision epochs of SMDP are not restricted to discrete-time epochs like MDP. The decision epochs of SMDP are all-time epochs at which the system enters a new decision-making state. The time between two decision-makers is defined as the sojourn time. For SMDP, the sojourn time for each state is a general continuous random variable that depends on the current state and the next state. Also, the decisions are only made at specific system state change epochs based on the decision-makers. In other words, between two adjacent decision epochs, the state of the system may change several times. The continuous sojourn time of SMDP results in a distinguishing feature of the reward functions of SMDP. Besides the immediate reward generated by the agent choosing an action at an arbitrary state, the sojourn time between two adjacent decision epochs will lead to a continuous reward based on the future accrual reward rate. The total reward for SMDP is the summation of the immediate reward and continuous reward.

1.2.3 Reinforcement Learning

Reinforcement learning is used by the agents or decision-makers to obtain the optimal control policies [14]. In the RL, the rewards and punishments have been combined according to the feedback generated during the active interactions of the agent with the current environment. There are four elements in RL including the environment, agent, action, and environment feedback. During the process of RL, the agent selects an action at the current decision-making state which will lead the system to reach the next decision-making state under a unique path. Meanwhile, the agent and the system will decide on the next action. Between the transition of two adjacent states, the agent can update the corresponding information of the next state, the immediate reward, and the time spent for the state transition. Based on the information mentioned above, the agent will change the related knowledge and choose the next action. This is a whole step of the iteration process and as the process repeats, the performance of the agent improves gradually.

1.3 Literature Review

In this section, the existing research works in literature are discussed.

1.3.1 Electric Vehicle Charging Schedule Optimization

In literature, there are many research works related to the optimization of charging schedule of electric vehicles (EVs), which can shed some light on EB charging schedule optimization. In [15], a deep policy gradient-based reinforcement learning approach has been

proposed to optimize the charging strategy of EVs which can ensure the voltage security of the local grid. A multi-option charging strategy has been proposed in [16]. The corresponding charging strategy can help EVs find a suitable charging station that can reduce the charging time and charging cost. In [17], a novel random-in-window (RIW) vehicle-directed smart charging approach has been introduced. This RIW approach can reduce the peak load on the local grid and also has a benefit for reducing transformer aging. An adaptive-current charging strategy has been proposed in [18]. By using this strategy, the charging losses for the lithium-ion battery for an electric vehicle can be reduced while improving the charger efficiency. Also, in [19], an integrated algorithm has been proposed for the optimal charging schedule of the battery swap station. Based on this algorithm, the total cost can be minimized along with the potential battery damage from the utilization of high-rate chargers. In [20], an optimized charging scheme has been developed by considering the cycle battery aging effects. According to this method, the charging electricity cost and the battery aging cost can be minimized. An optimal charging policy based on fitted Q-Iteration (FQI) reinforcement learning has been introduced in [21]. In this paper, the MDP has been used to formulate the characteristics of the charging stations. Under this condition, this method is more flexible when compared with the others for cost reduction. In [22], the charging scheduling problem is investigated for EVs in a park-and-charge system in order to lessen the degradation cost during the charging process under the restriction of battery charging characteristics. An operation model of the system by considering the customers, parking garage and the battery degradation model has been proposed. Then, the algorithms based on the charging resource allocation and dynamic power adjustment are introduced to minimized the charging cost. In [23], the charging schedule of EVs has been combined with the photovoltaic system of the buildings. A smart EV charging method with PV system has been proposed in this paper by considering the charging scheduling algorithm and prototype application. In this way, the optimal schedules is designed based on the PV output and electricity consumption.

1.3.2 In-Station Charging Schedule Optimization for Electric Buses

In-station charging strategies are applicable for EBs which are charged at centralized charging stations or the depot locations after the daily service. In this way, EBs are under a centralized charging control. However, the drawbacks are the huge charging power for the charging stations, as well as a large battery pack for each EB to accommodate the daily service energy demand. Many research has been introduced in literature in this aspect.

In [24], a novel charging load prediction based battery-swap station (BSS) is proposed for EBs to extend the operating ranges. The study indicates that the stochastic characteristics of EBs due to the randomness of battery swapping and charging patterns can significantly affect the performance of the battery swap management method. To address such randomness, four main variables need to be considered, including the number of

battery swapping for an EB in one hour, starting time of the charging process, operating distance, and charging duration. In [9], an EB battery exchange system was introduced based on robots along with experimental data and advanced battery exchange technology. The research work in [10] has established a new economic model for battery replacement systems. Based on battery replacement, efficient and low-cost operation of public transportation systems can be achieved. EBs battery charging method is also studied in [25]. This article examines the impact of EB fast charging on battery lifetime and degradation, and the impact on the utility grid. As EB's fast charging will put heavy pressure on the local power distribution system, a new charging station with bidirectional external and vector control technology was developed in [26]. This article provides detailed information on charger design and its corresponding performance. In [19], an optimized control strategy has been proposed by using the retired EB batteries for the charging stations. In this paper, the number of second-use EB batteries and the charging-discharging power are optimized which can reduce the overall annual cost and the daily electricity cost for the EB charging stations. As for the fast charging of EBs, a hierarchical charging control strategy has been proposed in order to solve the charging optimization issue for the EBs. There are two layers in the method: the prediction layer and the scheduling layer. The prediction layer is used for integration and transmission of the collected data, and the scheduling layer is for the generation of the charging policy. By using this method, the voltage quality and economic cost for the system can be guaranteed. In [27], a novel multi-objective bi-level programming has been proposed to optimize the charging scheduling of electric and traditional bus fleet. This model involves two layers. The upper layer addresses vehicle selection that can reduce the operation cost and the carbon emissions by considering the operating distance and the travel time during the trips. And the lower layer is to optimize the charging strategy which can reduce the charging cost under the constraints of charging time. Also, a wireless power transfer (WPT) system is introduced in [28] for the charging of an EB. The goal is to derive a control strategy for the wireless electric vehicle charger which can be utilized to achieve fast dynamics and to control the coil current and transferred power. Accordingly, a dual-loop controller based on the generalized state space averaging approach is proposed.

1.3.3 En-Route Charging Schedule Optimization for Electric Buses

For the en-route charging strategy, the charging infrastructures are located along the operation route of the EBs. For example, the en-route charging indicates that EBs could have a fast charging during the operation process or through the wireless power transfer infrastructures which are installed under the road. EB charging schedule optimization for the en-route charging strategy is also investigated in many existing research works.

In [29], a novel EB system is introduced based on an innovative WPT technology developed by the Korea Advanced Institute of Technology (KAIST). This forms an online

electric vehicle (OLEV) based EB system with road-vehicle integration. The benefits of the OLEV are analyzed, and two methods focusing on commercial design optimization have been proposed. In particular, the first method neglects the traffic conditions except for the EBs, while the second method considers the normal traffic conditions. In [30], the optimal placement of charging stations is investigated to reduce the installation cost while satisfying the charging requirements of EBs. Two different cases, with and without considering the limitation of battery capacity, are considered and two algorithms, named electric charging station problem limited battery (ECSP_LB) and electric charging station problem (ECSP) are proposed for these two cases, respectively. Also, it is proved that the ECSP algorithm can achieve the lower bound of the performance of the ECSP_LB algorithm. In [31], a non-contact charging method for wireless power transmission systems is introduced. Some experimental results are obtained from [32]. In this paper, the structure, implementation, and applicability of the proposed online wireless charging system are presented. Besides, an analysis of wireless charging systems for online electric buses is presented in [33]. Based on the switching of the resistance circuit and the adjustment of the resonant frequency of the sensor, the phases of voltage and current are analyzed. Also, some methods are developed in literature to achieve optimal EB charging planning. In [34], the charging strategy optimization for wireless charging EB is introduced. Based on the day-ahead electricity market, the optimal reserved wholesale electricity can be determined, along with the charging strategy which can reduce the operating electricity cost of the bus system.

1.4 Thesis Motivation and Contributions

According to the discussion above, although there are many research works on EV and EB charging schedule optimization, there are still several technical challenges that need to be addressed. Firstly, the charging optimization strategy of EVs cannot be directly applied to EBs. Compared with EVs, how the key features of EBs such as fixed route and schedule, various passenger counts, and different driver habits affect the optimization model needs to be investigated. Secondly, the optimization approach which can used to determine the charging schedule with a lack of long-term historical data needs to be investigated. This is of particular importance for many public transit services with newly adopted EBs. Thirdly, how to model the operation process of EBs by considering the time inaccuracy of bus schedule due to the uncertainties in traffic conditions and passenger count needs to be investigated. Finally, for real-world applications, the convergence of the optimization approach needs to be ensured for a smooth operation of public transit services. To address these technical challenges, the main contributions of this thesis are summarized as follows.

1.4.1 Double Q-Learning for Optimal In-Station EB Charging

In this work, we investigate the proper charging demand for in-station charging of EBs, in order to minimize the operation cost of EBs during the working hours. Also, the battery lifetime can be extended based on the selection of a suitable charging SOC. A model-free double Q-learning approach is developed to achieve this objective. The main contributions of this research are as follows:

- Developing an MDP to describe the operation process of the EBs by considering the number of charging stations, the current SOC level, and the total cost;
- Developing a model-free double Q-learning approach based on the aforementioned MDP to study the iterations of EBs' charging amount at different charging stations;
- Proposing an optimized charging strategy for the EB at each charging station during its operation process while taking into account the battery degradation cost and the battery lifetime.

1.4.2 Average Reward Reinforcement Learning for Optimal En-Route EB Charging

In this work, we investigate the optimal en-route charging strategy of EBs based on SMDP and average reward reinforcement learning (ARRL). The physical model of EB is utilized to calculate the energy consumption between two adjacent charging stations based on the velocity, distance, torque, and the corresponding slope of the road. Then, based on the energy consumption of the EB, ARRL is used to generate the optimal en-route charging strategy of the EB. The main contributions of this research are as follows:

- Developing a physical battery degradation model to calculate the total cost of EB operation between two adjacent charging stations, based on which the corresponding change of SOC as well as the cost of the battery can be estimated for EB operation;
- Developing an SMDP to characterize the operation process of EB according to the physical model. The state transitions in the SMDP can be calculated based on the energy consumption estimated from the physical model;
- Developing an average-reward reinforcement learning approach to optimize the en-route charging strategy of the EB.

1.4.3 Relative Value Iteration Reinforcement Learning for Optimal En-Route EB Charging

In this work, we proposed a smart charging approach for EBs based on the RVIRL algorithm. The physical models of EB speed and torque are used to calculate the energy

consumption of a EB in between any two adjacent charging stations while the EB operating model is utilized to constrain the state of charge (SOC) level of EB on the route. After that, the battery degradation model has been implemented to calculate the battery lifetime degradation parameter which is further used to obtain the degradation cost of the EB battery. Then, the optimal en-route charging strategies of EBs can be obtained from the RVIRL algorithm by considering the total electricity cost and battery degradation cost. The performance of the proposed approach is demonstrated by utilizing the real-world data obtained from the public transit service St. Albert Transit, AB, Canada for the electricity price, related torque, and speed. The main contributions of this section are as follow:

- A physical model is developed to calculate the energy consumption of BEB between charging stations during the operation process;
- The performance evaluation based on the real-world data collect from St. Albert Transit, AB, Canada;
- An SMDP problem is formulated to modify the charging schedule of BEB on the route by considering the SOC changes, number of charging stations, maximum stop time at charging station, and electricity price;
- An RVIRL algorithm is developed to optimize the en-route charging schedule at each charging station, and the convergence of the algorithm is proved from the mathematical aspect;
- The convergence of RVIRL algorithm has been proved in this chapter.

1.5 Thesis Outline

This thesis consists of five chapters which are organized as follows:

- **Chapter 1: Introduction** - The research background and definitions are introduced in this chapter. Also, the related works for EB charging schedule optimization in recent years are reviewed. Finally, the research motivation and contributions are presented.
- **Chapter 2: Double Q-Learning for Optimal In-Station EB Charging** - This chapter presents a method to obtain the optimized charging strategy by using RL.An MDP model is developed to characterize the operation process of EB. Based on the MDP model, a double Q-learning algorithm is proposed to minimize the battery degradation cost of the EB at each charging station.
- **Chapter 3: Average Reward Reinforcement Learning for Optimal En-Route EB Charging** - This chapter presents an ARRL approach to obtain the optimal en-route charging strategy of the EB. A physical model is developed to calculate EB energy consumption between adjacent charging stations by considering the velocity, speed,

and passenger count. Then, an SMDP is developed to model the operation process of the EB, based on which an ARRL algorithm is proposed to generate the en-route charging strategies for the EB. The performance of the proposed ARRL approach is evaluated based on data from St. Alberta Transit, AB, Canada.

- **Chapter 4:Relative Value Iteration Reinforcement Learning for Optimal En-Route EB Charging** - In this chapter, a semi-Markov decision process (SMDP) based relative value iteration reinforcement learning (RVIRL) approach is proposed to optimize the en-route charging schedule of EBs. First, the energy consumption and battery degradation model are developed to characterize the electricity cost and battery degradation during the operation process of EBs, respectively. Then, the RVIRL approach is utilized to obtain the optimized charging schedule for the en-route charging stations. The performance of the proposed approach is evaluated based on the real-world data obtained from the EBs of the public transit company St. Albert Transit in Alberta, Canada.

- **Chapter 5: Conclusion and Future Works** - In this chapter, the contributions of this research and future works are summarized.

2

Double Q-Learning for Optimal In-Station EB Charging

In this chapter, we present a model-free reinforcement learning approach to optimize the charging schedules of EB when it arrives at the corresponding charging station, by considering the EB operation state estimation and battery degradation. An MDP has been developed to model the operation process of EB during the operation hours. After that, a double Q-learning approach is introduced to optimize the charging schedule of EB. The performance of the proposed approach is evaluated based on the real data collected from St. Albert Transit, AB, Canada.

2.1 System Model

A typical operating process of EB is shown in Fig. 2.1. The EB runs in a circulation trip, which will departure from the terminal station with an initial SOC in the morning. The information about the SOC of this EB and the total number of EBs at terminal station will be sent to the vehicular communication network. Then, as shown in Fig. 2.2, during the operation process, the SOC starts decreasing while this EB is running on the road. When this EB arrives at a bus station, the driver sends the current SOC value and station number to the vehicular communication network. After the vehicular communication network receiving the EB information for the corresponding bus station, it gives feedback with the charging information to the driver. The driver will follow the instruction to charge the EB with specified target SOC or not to charge the EB. During the operating process, the driver repeats the same steps at each bus station to determine whether to charge the EB or not with the instruction provided by the vehicular communication network until the EB finishes the trip and arrives at the terminal station. For EB system, there are a total of N

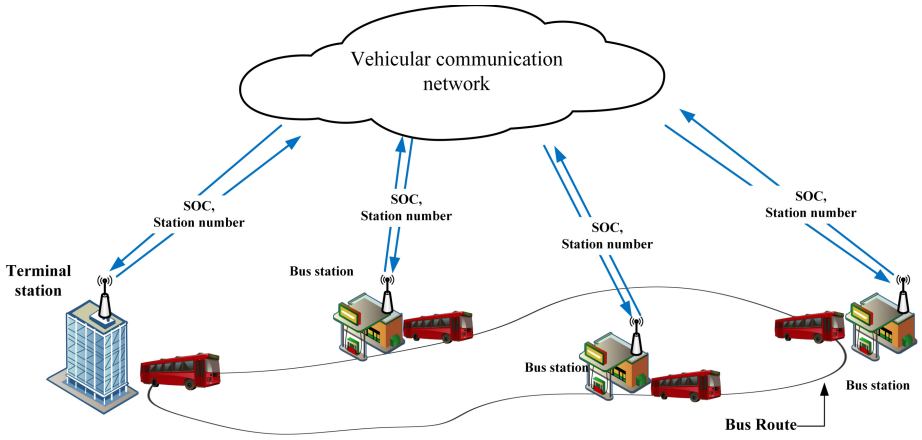

Figure 2.1: A typical operating process of EB in smart grid.

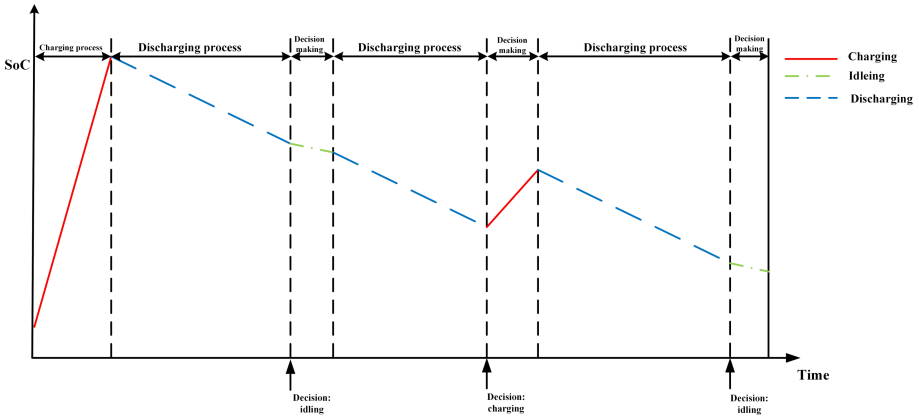

Figure 2.2: A typical SOC profile of EB in operation process.

bus stations, and each bus station can have either a charging process or idle (not charging) process. The bus stations are denoted as $\{(1,0),(2,1),\ldots,(N,k)\}$, where $k=1$ indicates that the EB is being charged at this station, while $k=0$ means that the EB is idled at this station.

Based on the discussion in [36], the battery degradation cost (C_d) can be determined as

$$C_d = c_{bat}\frac{L_c}{L_{tot}}, \tag{2.1}$$

where c_{bat} is the total cost of battery, L_c is the lifetime degradation in a unit time, which consists of the capacity fade and power fade, and L_{tot} is the total battery lifetime which can be obtained experimentally based on the repeated charging/discharging processes until the battery capacity drops below 80%. Based on [35], there are two significant factors that can affect the battery lifetime degradation, i.e., the SOC and DOD. To simplify the battery modeling process and improve the efficiency of battery degradation evaluation, it is assumed that these two factors are independent and do not affect the battery aging

process. This simplified battery model allows us to treat these two factors independently, without considering the time-variant effects for battery degradation.

Based on the assumptions made for the impacts of battery SOC and DOD on capacity fade and power fade, the battery degradation function (2.1) can be extended to as follows:

$$C_d = \max\left((C^Q_{SOC} + C^Q_{DOD}), (C^P_{SOC} + C^P_{DOD})\right), \tag{2.2}$$

where C^Q_{SOC} and C^Q_{DOD} are the costs for capacity fade and C^P_{SOC} and C^P_{DOD} are the costs for power fade. Further, by comparing with other costs, the costs caused by power fade can be neglected for simplification [37].

The function for capacity fade cost caused by SOC, C^Q_{SOC}, is formulated based on the experimental data obtained in [38]. As for the formulation of cost function of capacity fade caused by SOC, the assumption of time invariance, as discussed above, is utilized. Also, due to the assumption about the independence of effects of SOC and DOD, the battery lifetime degradation for a period is equal to the effect resulting from the battery stays at average SOC for the same period. The equation relates the capacity fade cost caused by the SOC and the average SOC of the battery is derived from the data in [38], which can be formulated as

$$C^Q_{SOC} = c_{bat} \cdot \frac{m \cdot SOC_{avg} - d}{CF_{max} \cdot \epsilon \cdot \zeta}, \tag{2.3}$$

where CF_{max} is the capacity fade when the battery capacity drops below 80%, which can be calculated as $100\% - Q_{EOL}/Q_0$ and the parameters ϵ, ζ are obtained by using the curve fitting method based on the data in [38].

The capacity fade cost caused by the DOD, C^Q_{DOD}, results from the SOC swing ΔSOC, which is the difference between the daily maximum SOC and minimum SOC. This difference is equivalent to DOD. According to [39], the relationship between ΔSOC and the battery lifetime in cycle L is given by

$$L(\Delta SOC) = \left(\frac{\Delta SOC}{\eta}\right)^{\frac{-1}{\lambda}}. \tag{2.4}$$

Here, we consider the effect on L for an arbitrary given ΔSOC is the same with average SOC that changes at ΔSOC. According to [40], a concept named energy throughput can be utilized to calculate the capacity fade cost due to a cycle with ΔSOC. From [36], the energy throughput during battery lifetime can be represented as

$$E_{tp} = L(\Delta SOC) \cdot \Delta SOC \cdot Q, \tag{2.5}$$

where Q is the battery capacity. Based on the energy throughput theory, the battery degradation cost arises from the capacity fade caused by DOD is given by

$$C^Q_{DOD} = c_{bat} \cdot \frac{(E_{T,use} - E_{T,discharge})}{E_{tp}}, \tag{2.6}$$

where $E_{T,discharge}$ is , for given $\Delta SOC_{discharge}$, defined as

$$E_{T,discharge} = L(\Delta SOC_{discharge}) \cdot \Delta SOC_{discharge} \cdot Q. \tag{2.7}$$

Also, $E_{T,charge}$ is used when the EB is being charged. Since the battery charging curve is nonlinear, the values of actual average SOC and theoretical average SOC are different. The SOC swing ΔSOC_i can change the average ΔSOC to $\Delta SOC_{avg,i}$. Then, the actual energy throughput $E_{T,use}$ of one charging cycle of battery can be written as

$$E_{T,use} = L(\Delta SOC_{avg,i}) \cdot \Delta SOC_{avg,i} \cdot Q + \Delta SOC_i \cdot Q. \tag{2.8}$$

2.2 Problem Formulation

One of the major obstacles of EBs charging scheduling optimization is the stochastic characteristics. The existence of randomness in traffic conditions, road situations, and driver's behaviors increase the complexity of this problem. In order to address the stochasticity, a model-free method is proposed based on RL. Firstly, a Markov Decision Process (MDP)-based optimization problem is formulated for the EBs charging scheduling.

The MDP problem can be defined as a tuple $(S, A, P(\cdot,\cdot), R(\cdot,\cdot), \gamma)$. More specifically, S denotes the states of the corresponding system, A is the finite set of actions, $P(\cdot,\cdot)$ is the state transition probability function, $R(\cdot,\cdot)$ is the immediate reward function, and γ is a discount factor. The details of MDP for optimal EB charging scheduling is shown below

- State: the system state at each bus station n is defined as $S_n = ((n,k), SOC_n)$, where n is the bus station in the transit system the EB stops at. Here, k indicates, at this bus station, whether the driver chooses to charge or not, and SOC_n denotes the SOC of EB at bus station n.

- Action: the action a_n represents the charging power at the current state s_n. When $a_n > 0$, this means that the EB is under charging condition and $a_n = 0$ indicates that the driver chooses not to charge the EB. The constrain of the a_n is given by

  $$0 \le a_n \le SOC^{max}_{charging}, \tag{2.9}$$

  where $SOC^{max}_{charging}$ is the maximum charging SOC of the battery of EB. The charging station of EB provides discrete charging SOC and $a_n \in \{SOC_1, SOC_2, \ldots, SOC_{max}\}$.

- State transition: the state transition probability gives the probability of transitioning from state s_n to s_{n+1} based on its corresponding action a_n, which is expressed as

  $$P(s_n, a_n, s_{n+1}) = P(s_{n+1}|s_n, a_n). \tag{2.10}$$

  Based on the EB model, the state transition can be represented as $s_{n+1} = s_n + SOC(n, n+1) + a_n$ where $SOC(n, n+1)$ is the required SOC for EB running from station n to station $n+1$.

- Reward: the reward during the operating process of EB is the summation of the cost of battery degradation between any two adjacent bus stations in one complete bus route. According to the discussion in the previous section, the cost of SOC related degradation is formulated as

$$C_n^{Q,SOC} = c_{bat} \cdot \frac{m \cdot (SOC_{avg}(n, n+1) - d}{CF_{max} \cdot \epsilon \cdot \zeta}, \tag{2.11}$$

and the degradation cost caused by DOD is

$$C_n^{Q,DOD} = c_{bat} \cdot \frac{(E_n^{T,use} - E_n^{T,discgarge})}{E_n^{tp}}. \tag{2.12}$$

For simplification of calculation, it is assumed that the SOC swing ΔSOC_i make no effort in the used energy throughput, as discussed in Section 2.1. Then (2.8) can be rewritten as

$$E_n^{T,use} = L(\Delta SOC_n) \cdot \Delta SOC_n \cdot Q + \Delta SOC_n \cdot Q, \tag{2.13}$$

where $\Delta SOC_n = SOC_n - SOC_{n+1}$. The function of energy throughput in discharging can be formulated as

$$E_n^{T,discharge} = L(SOC_{dis}(n, n+1)) \cdot SOC_{dis}(n, n+1) \cdot Q, \tag{2.14}$$

where $SOC_{dis}(n, n+1) = \min(SOC_n - SOC_{n+1}, 0)$. The energy throughput during battery lifetime can be reformulated as

$$E_n^{tp} = L(SOC_n - SOC_{n+1}) \cdot (SOC_n - SOC_{n+1}) \cdot Q. \tag{2.15}$$

Further, the reward at bus station n is defined as the battery degradation cost between bus station n and bus station $n+1$, which is expressed as

$$R_n = (C_n^{Q,SOC} + c_{bat} \cdot \frac{(E_n^{T,use} - E_n^{T,discgarge})}{E_n^{tp}}), \tag{2.16}$$

where $C_{Q,SOC}$ and $C_{Q,DOD}$ are the battery degradation costs due to capacity fade.

- Action-value function: the action-value function is defined as the expected value of the sum of future rewards based on the current state and its corresponding action for the next M bus stations. This can be formulated as:

$$Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{m=0}^{M} \gamma^m R_{n+m} | s_n = s_0, a_n = a_0\right], \tag{2.17}$$

where $Q^{\pi}(s, a)$ is the action-value function for state s and action a, while π is the policy of EB charging which maps the current state to the corresponding action based on the charging schedule.

- Discount factor: The term γ in the action-value function is the discount factor which is utilized to balance the weight between the current reward and future rewards of the remaining bus stations. For example, if $\gamma = 1$, this means that the weight of future rewards is equal to the current reward and the system is forward-looking. When $\gamma = 0$, this means that the system only takes account of the effect of current reward and the policy is myopic.

By combining the reward function and battery model of EBs, the objective function of this problem, which aims at minimizing the battery degradation cost while satisfying the EB charging requirement, can be formulated as

$$\max \sum_{n=2}^{N} (-C_n^{Q,SOC} - c_{bat} \cdot \frac{(E_n^{T,use} - E_n^{T,discgarge})}{E_n^{tp}}). \tag{2.18}$$

The constrain for this optimization problem is that the SOC at each bus station should be within the range of $20\% - 90\%$, based on the specifications of the EB manufacturer.

2.3 Double Q-Learning for EB Charging Schedule Optimization

In this section, we present an algorithm based on RL to solve this formulated problem for the optimization of daily EBs charging schedules. Based on the discussion in Section 2.2, the action-value function of the EB system is defined in this section. Generally, the optimal charging scheduling problem, formulated as an MDP, is to find an optimal policy π^* to maximize the action-value function in the form of

$$Q_\pi^*(s,a) = \max_\pi Q_\pi(s,a), \tag{2.19}$$

where $Q_\pi^*(s,a)$ is the optimal action-value function. To find the optimal action-value function, a direct method, Q-learning in RL, is introduced based on [47]. The main idea of Q-learning is that when Q_n approaches to Q^*, the policy is greedy and optimal.

For a finite MDP $\mathcal{M} = (S, A, P(s_{n+1}, s_n, a_n), R(s_n, a_n)\ , \gamma)$, Q-learning keeps an assumption of $Q_n(s,a)$ of $Q^*(s,a)$ for each state-action pair (s,a) and the action-value function is updated through the following processes:

$$\delta_{n+1}(Q) = r_{n+1} + \gamma \max_{a_{n+1} \in A} Q(s_{n+1}, a_{n+1}) - Q(s_n, a_n) \tag{2.20}$$

$$Q_{n+1}(s,a) = Q_n(s,a) + \alpha_n \delta_{n+1}(Q_n). \tag{2.21}$$

However, the performance of Q-learning in solving stochastic MDP problems may degrade for practical applications in EB charging optimization, since the action value in basic Q-learning is overestimated [41]. In other words, the maximum operator utilized in Q-learning to determine the value of the next sate could result in an overestimation for the action value. There will be a positive bias in approximating the maximum value as the

Algorithm 1 Double Q-learning for charging scheduling optimization

Input: Current state s, current action a, immediate reward R, next state Y, action-value function Q;

Output: Optimal state s^* Initialize Q^A, Q^B, s

1: Choose a, based on $Q^A(s,\cdot)$ and $Q^B(s,\cdot)$, observe r, s'
2: Randomly choose either UPDATE(A) or UPDATE(B)
3: If UPDATE(A), then
4: Define $a^* = argmax_a\, Q^A(s', a)$
5: $Q^A(s,a) \leftarrow Q^A(s,a) + \alpha(s,a)(r + \gamma Q^B(s', a^*) - Q^A(s,a))$
6: else if UPDATE(B), then
7: Define $a^* = argmax_a\, Q^B(s', a)$
8: $Q^B(s,a) \leftarrow Q^B(s,a) + \alpha(s,a)(r + \gamma Q^A(s', a^*) - Q^B(s,a))$
9: end if
10: $s^* \leftarrow s'$
11: **Return** s^*

maximum expected value which could cause a reduction in the performance of Q-learning while handling MDP problems. To address this issue, an alternative algorithm is proposed by using double estimator method to find the approximation of the maximum value for the MDP, based on the concept of Double Q-learning [41]. The details of this algorithm are illustrated in **Algorithm 1**.

In Double Q-learning, two value functions are defined and stored inside the Q-learning, which are Q^A and Q^B, respectively. These two Q functions are updated based on the value of Q function in each state for the next stage. In **Algorithm 1**, the action a^* is the optimal action with maximum value at state s' for Q function Q^A. Unlike Q learning using the optimal action to update its value function, the value of $Q^B(s', a^*)$ is used for the updating of Q^A. The similar updating process is used for Q function Q^B. It's worth noting that both Q function are learned from separate sets of experiences but for the same MDP problem. Under this condition, the excessive variance for selecting the max Q-value can be reduced, and the action of Q-value can be slightly underestimated.

For each state-action pair in Double Q-learning algorithm, the action-value functions for Q^A and Q^B can be updated, respectively, through the following equations

$$Q^A(s,a) \leftarrow Q^A(s,a) + \alpha(s,a)(r + \gamma Q^B(s', a^*) - Q^A(s,a)) \tag{2.22}$$

$$Q^B(s,a) \leftarrow Q^B(s,a) + \alpha(s,a)(r + \gamma Q^A(s', a^*) - Q^B(s,a)). \tag{2.23}$$

In these equations, $\alpha \in [0,1]$ is the learning rate which gives the threshold of convergence for the best action value function. The learning rate can be obtained by averaging the randomness of the rewards and transitions [41]. Another important factor is the discount factor $\gamma \in [0,1)$. The discount factor is the attribute indicates how critical is the immediate rewards than the future rewards in the Double Q-learning. Secondly, if the tasks are non-episodic, this parameter can guarantee each value function in Double Q-learning is finite

Figure 2.3: Tested bus route for simulation.

Table 2.1: Simulation Parameters for Double Q-learning

Parameter	Value
ϵ	15
ζ	8760
η	145.71
λ	0.6844
C_{tot}	400 \$/kWh
m	0.04717
d	7.9245
Iteration times	100
Bus station	16

and well defined. By combining **Algorithm 1** and EB model, for each iteration, the process would include the transformation from the current state to the next state, the choice of the optimal action, the calculation of the immediate rewards and future rewards, and the storage of the Q values. This process will occur between any two adjacent bus stations, and the total number of time steps is $N-1$. For each state at each bus station, the number of SOC levels and the actions of whether to charge or not need to be considered, when running the algorithm to find the optimal charging schedule, and the number of variables are n_f. Thus, for I iterations, N bus stations, and n_f variables, the computational complexity is $O(I \cdot N \cdot n_f)$.

2.4 Simulation Results

In the simulation, the data from St. Albert Transit, AB, Canada is used to build the details of MDP and Q-learning. The data used for simulation is from Sep. 1^{st}. 2018 to Sep.

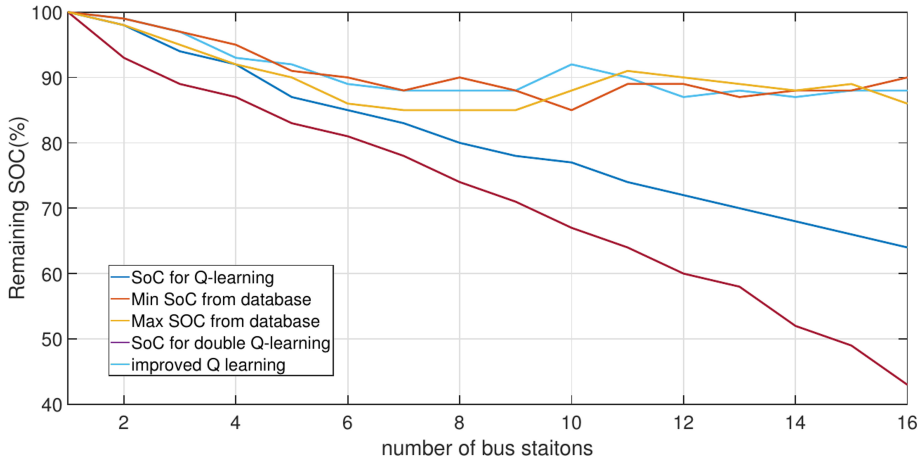

Figure 2.4: Remaining SOC for each bus station based on Q-learning.

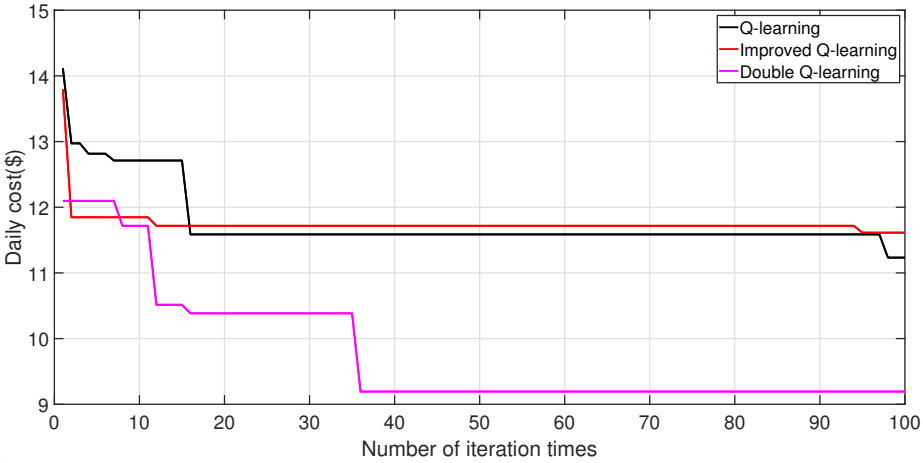

Figure 2.5: Performance among different Q-learning algorithms on daily reward.

30^{th}. 2018 for electric bus number 1402. Our case study is done in MATLAB 2021a on a desktop computer with Intel® Core™ i7-7700 CPU @3.60GHz. The average optimization time consumption for charging schedule is 762.23 seconds. The tested bus route is shown in Fig. 2.3. The red line on the map is the real bus route in the real world. There are a total of 16 bus stations on each route during the day time and the operating time of EB 1402 is from 8:00 am to 4:00 pm. The parameters used in this simulation is shown in Table I. In this simulation, the improved Q-learning method which updates the Q-values of state-action value and state-opposite action value simultaneously is used as a contrast [42]. The results of three different Q-learning algorithms, i.e., basic Q-learning, improved Q-learning, and the proposed Double Q-learning, are compared.

Fig. 2.4 shows the results of SOC under the real conditions and simulation situations. Due to the randomnesses of traffic conditions, road conditions, and driver's behavior, the remaining SOC of EB in different bus stations and days are different. Under this condition,

the SOC collected from the bus company are treated as two parts, which are the maximum SOC and the minimum SOC. According to Fig. 2.4, the blue line and red line are the maximum and minimum values of SOC, respectively. The remaining three lines are the SOC value for the three Q-learning algorithms. Since there are charging stations during the operation process, the SOC fluctuates at around 90%. The reason for not considering the on-route charging process is that StAT does not install any charging stations during the operating route.

As discussed before, the Double Q-learning is used to optimize the battery degradation cost by optimizing the charging schedule for EB. Based on the simulation result, as shown in Fig. 2.5, the battery degradation cost of different Q-learning algorithms are different, the results of Q-learning and improved Q-learning are almost the same which is close to 11.5$, while the Double Q-learning has a better performance with significantly reduce battery degradation cost. In addition, from Fig. 2.5, it can be observed that the iteration times of the Double Q-learning is much smaller than the numbers of Q-learning and improved Q-learning, which means that Double Q-learning is much more efficient. Fig. 2.6 shows that the comparison of the battery degradation cost between real situation and simulation. For each day, the battery degradation cost calculated based on the data collected from the company is used as a baseline. Then, the ratio between the simulation result and the baseline for different Q-learning algorithms are illustrated by different lines. From the Fig. 2.6, the percentages of the costs of Q-learning and improved Q-learning, in each day, are almost the same, while the percentage of the cost for Double Q-learning is relatively low. If the charging schedule is based on the Double Q-learning, according to the simulation result, the cost will decrease by about 10% for each day. Besides the lower daily battery degradation cost for Double Q-learning, another benefit for this algorithm is that the Mean Square Error (MSE) of Double Q-learning is much smaller than the Q-learning and improved Q-learning, as shown in Fig. 2.7. As a result, the Double Q-learning algorithm converges faster than the other two algorithms.

2.5 Summary

In this chapter, an optimal charging scheduling algorithm based on a model-free reinforcement learning algorithm is proposed for the minimization of EBs battery degradation cost. The battery degradation cost is analyzed by considering both the factors of battery SOC and DOD. The formulation of MDP-based EBs charging scheduling problem is presented in details. The transition probability is investigated to address the randomness of EBs charging processes. Further, the Double Q-learning which is based on reinforcement learning is introduced to optimize the charging schedule for EBs. Extensive simulation results based on real data collected from St. Albert Transit, AB, Canada are presented to evaluate the performances of our proposed Double-Q learning based optimal EBs charging scheduling algorithms. The simulation results show that our proposed algorithm can significantly

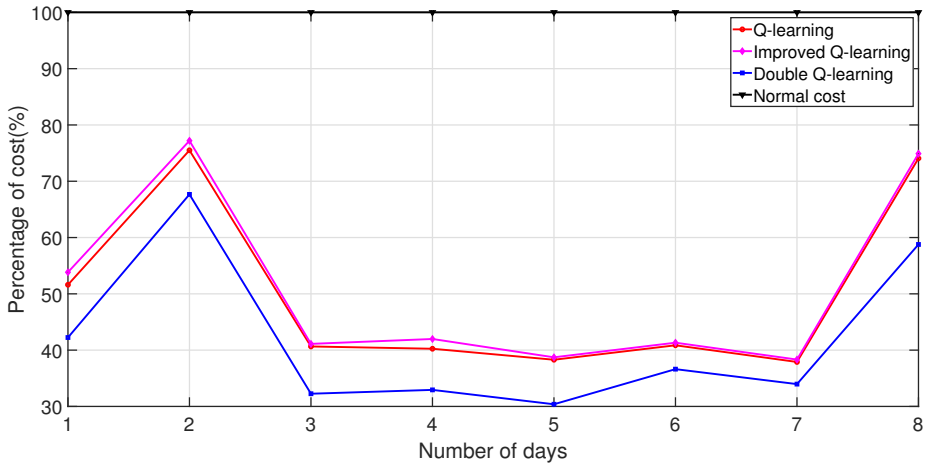

Figure 2.6: Percentage of cost for different Q-learning algorithms in 8 days.

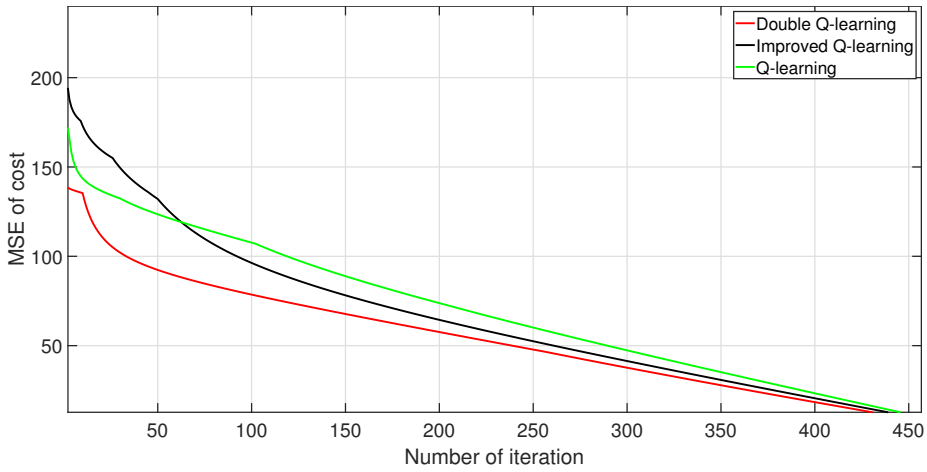

Figure 2.7: MSE for different Q-learning algorithms.

reduce the battery degradation cost during the operating time of EBs, while giving a relatively small errors when considering randomnesses of EBs charging processes. Moreover, our proposed algorithm can reduce the computational complexity significantly for practical applications.

3

Average Reward Reinforcement Learning for Optimal En-Route EB Charging

In this chapter, we develop an RL approach to optimize the en-route charging schedule for the EBs to minimize its operation cost. First, the physical model of the EBs is introduced to calculate energy consumption between adjacent charging stations on the route. Next, we develop a battery degradation model which is used to calculate the battery cost of the EBs during the charging and discharging processes. After that, an SMDP is utilized to model the operating process of the EB with the battery degradation model embedded in the value function of the ARRL algorithm to optimize the charging action for the EBs arriving at the charging stations. The performance of the proposed approach is evaluated based on real EB operation data provided by the St. Albert Transit, AB, Canada.

3.1 System Model

Fig. 3.1 shows the operating process of EBs in this chapter. According to the figure, the EB leaves the terminal bus station n_0 in the morning with an initial SOC_0 at time t_0. After the completion of the first path, EB arrives at the first en-route charging station n_1 with the current SOC level SOC_1, then EB needs to decide on charging or not. For example, if the charging decision has been made, EB will depart the station with SOC_2. By following the operational schedule, the EB leaves the charging station at t_2. And then arrives at the second charging station n_2. At this time, EB does not charge the bus, so the SOC remains the same in the charging station. During the operation process, EB will repeat several times until it runs back to the terminal station. $p_{n_1}, p_{n_2}, \ldots, p_{n_k}$ is the electric price at each charging station, n_k is the total charging station number and t_m is the total operational period of the EB.

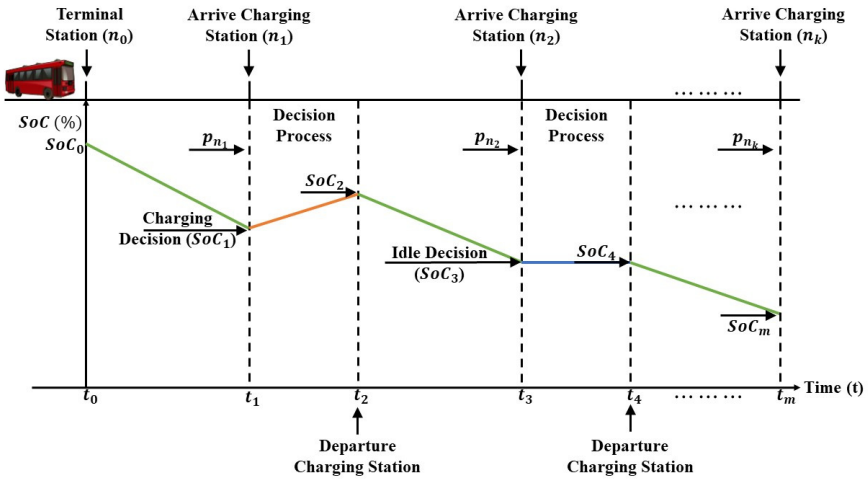

Figure 3.1: System model for electric bus

Consider the EB arrives the charging station n, the process between the adjacent charging station n and $n+1$ can be acted as a whole charging and discharging cycle for the battery. Then the battery degradation is given by:

$$F_d = F_c(SOC_n^a, SOC_n^d) + F_d(SOC_n^d, SOC_{n+1}^a). \tag{3.1}$$

where $F_c(\cdot)$ and $F_d(\cdot)$ represent the battery degradation for the charging process at charging station and discharging process between charging stations, respectively, while SOC_n^a and SOC_n^d are the arrival and departure SOC for charging station n, respectively. Also, SOC_{n+1}^a is the arrival SOC for station $n+1$.

3.1.1 Physical Model of EB Energy Consumption

The relationship between SOC and the energy consumption between the charging station can be calculated as:

$$SOC_n^d - SOC_{n+1}^a = \frac{1}{C_{battery}} \int_{T_n}^{T_{n+1}} P(t)dt. \tag{3.2}$$

where $C_{battery}$ is the total battery capacity of the EB. According to [43], power consumption can be calculated as:

$$P(t) = \frac{2\pi \cdot \tau(t) \cdot v_e(t)}{\beta_e \cdot \eta_e \cdot \eta_d}. \tag{3.3}$$

where β_e is a EB's constant parameter, η_e and η_d are the engine's efficiency and driver's efficiency respectively.

In (3.3), the angular engine speed can be calculated based on the current bus speed:

$$v_e(t) = \frac{\beta_v \cdot g_r \cdot v(t)}{2\pi r}. \tag{3.4}$$

where β_v is the constant parameter of the velocity

$$\tau(t) = \frac{r \cdot F(t)}{g_r}. \tag{3.5}$$

In (3.5), $F(t)$ is the real-time total force of the EB during the running process, r is the radius of the tire, and g_r is the gear ratio. The total force of the EB consists of four parts: friction force, aerodynamic dragging force, gravity decomposed force, and resultant force which can be indicated as:

$$F(t) = F_f(t) + F_a(t) + F_s(t) + F_r(t). \tag{3.6}$$

where $F_f(t)$ is the friction force which is calculated by the mass of the EB $m(t)$ and the velocity of the EB $v(t)$:

$$F_f(t) = c_r \cdot m(t) \cdot g \cdot (\beta_r + \beta_{r'} \cdot v(t)). \tag{3.7}$$

where c_r is the road fraction coefficient, β_r and $\beta_{r'}$ are the constant parameters corresponding to the fraction of the road.

In (4.1), $F_a(t)$ represents the drag force during the operating process caused by the air resistance for impeding the movement of the EB. In the function, ρ is the density of the air, c_a is the aerodynamic dragging coefficient, A is the front area of the bus:

$$F_a(t) = \frac{1}{2} \cdot \rho \cdot c_a \cdot A \cdot v(t). \tag{3.8}$$

where $F_s(t)$ is a part of gravity decomposed force when the EB is running up or down of the road:

$$F_s(t) = m(t) \cdot g \cdot \sin(f_\theta(d(t))). \tag{3.9}$$

where $d(t)$ is the distance according to the time and f_θ is the slope of that road corresponding to the location of the EB.

The last part for (4.1) is the resultant force of the EB $F_r(t)$:

$$F_r(t) = m(t) \cdot \frac{dv(t)}{dt}. \tag{3.10}$$

Through the previous functions, the energy consumption for EB between two adjacent charging stations can be calculated if the corresponding data is provided.

3.1.2 Battery Degradation Model

According to [48], the battery degradation between charging stations n and $n+1$ can be calculated based on the average SOC_{avg} during this process and its corresponding normalized deviation SOC_{dev} for this time interval is:

$$\Delta SOC_{dev} = 2\sqrt{3 \cdot \frac{\int_{T_n}^{T_{n+1}} (SOC(t) - \Delta SOC_{avg})^2 dt}{T_{n+1} - T_n}}. \tag{3.11}$$

where T_n and T_{n+1} are the time that the EB arrives charging station.

Then the battery degradation parameter Φ_h for the battery cycles centered on 50% SOC can be shown as:

$$\Phi_h = K_\sigma \cdot N \cdot \exp(\frac{(\Delta SOC_{dev} - 1) \cdot T_a}{K_\gamma \cdot T_b}) + 0.2 \cdot \frac{t_{cycle}}{t_{life}}. \tag{3.12}$$

where K_σ is the constant coefficient of throughput, K_γ is the exponent for depth of discharge, t_{cycle} is the time in seconds of a cycle, t_{life} is the total expected shelf life in seconds to 80% capacity at $25°C$ and 50% SOC, and N is the effective number of throughout cycles in the time interval.

Based on the Zhurkov term, the following relationship can be utilized to simplify (4.12):

$$\frac{(\Delta SOC_{dev} - 1) \cdot T_b}{K_\gamma \cdot T_a} = \frac{g \cdot S}{k \cdot T_a}. \tag{3.13}$$

where S is the stress level of the battery. In other words, it is a constant as the tensile stress of the battery.

However, in real condition, the battery charging/discharging cycle is not fully centered on 50%. In this way, the following function could extend (4.12) to a general function that matches the battery cycle for any battery level:

$$\Phi_2 = \Phi_1 \cdot \exp(\frac{K_\eta \cdot (\Delta SOC_{avg} - 0.5)}{0.25}) \cdot (1 - L). \tag{3.14}$$

where L is life aging parameter from 0 to 1.0

3.1.3 EB Operation Constraints

During the operation process of EBs, there are some constraints need to be considered. First of all, when the EB arrives an opportunity charging station n with the SOC as SOC_n^a. After the charging, the leaving SOC_n^d needs to satisfy the energy consumption for the next discharging process which can be defined as $SOC_{(n,n+1)}^d$ and it can be formulated as:

$$SOC_n^d \geq SOC_{(n,n+1)}^d. \tag{3.15}$$

Since the operating process of EB needs to follow the day-ahead schedule, the time interval that it can stay at the charging station is limited. If the EB arrives the charging station n at T_n^a, then the departure time would be limited by:

$$T_n^d - T_n^a \leq T_n^{max}. \tag{3.16}$$

where T_n^{max} is the maximum sojourn time for the EB at station n. The third is that due to the limitation of the charging time, the maximum charging SOC should satisfy the following equation:

$$SOC_n^d - SOC_n^a \leq P_{ch} \cdot T_n^{max}. \tag{3.17}$$

where P_{ch} is the charging power in the charging station. Finally, the SOC of the EB during the operation process should be in the manufacturer-specified range that can be protected

the health of the battery which is:

$$SOC_{min} \leq SOC_n \leq SOC^{max}. \tag{3.18}$$

where SOC_{min} and SOC^{max} is the minimum and maximum SOC.

3.2 Problem Formulation

In order to formulate the EBs system more preciously, the semi-Markov decision process [49] is more suitable than the traditional Markov decision process. The main difference between SMDP and MDP is the time between two adjacent decision-making epochs. As for the SMDP, this time can be random which is like the real condition of the EBs operation process. Other is that the decision epoch for SMDP is not restricted to discrete-time epoch, it is all time epochs that the state enters a new decision-making state. The decisions are only made at specific system state change that depend on the decision-maker.

When combining the SMDP with the operation process of EB, the state space in SMDP contains the number of charging station n and its corresponding battery level SOC_n which can be defined as $S = \{n, SOC_n\}$. The action space A includes the amount of charging SOC_n^{char} at each charging station. The transition probability between states would not be considered since the reinforcement learning algorithm is utilized to obtain the optimal charging schedule. The policy π in the SMDP is the optimal charging decision made by the EB in each station. When the EBs arrive the charging station, it can be acted as a decision-maker. The current state of EB is $s_n = \{SOC_n, n\}$ and the action chosen under this condition is $a_n \in A$. Based on the energy consumption model of the EB, the next state $s_{n+1} = \{SOC_{n+1}, n+1\}$ of EB can be calculated from the current state and action. And the function for calculating total expected reward of the EB is given by:

$$\begin{aligned} r(s_n, a_n) &= k(s_n, a_n) + E\left\{\int_{T_n}^{T_{n+1}} c(Y_{t,n}, s_n, a_n)dt\right\} \\ &= p_n \cdot a_n * w^{SOC} + \Phi_2(s_n, a_n, s_{n+1}) \cdot C_{battery}. \end{aligned} \tag{3.19}$$

where the first part is the immediate economic cost due to the decision of the action for charging the battery according to the current electric price p_n, while parameter w^{SOC} is used to transfer the action a_i from SOC value to power. The second part is the actual accumulative cost due to battery degradation during the charging process and the discharging process of the EB.

The sojourn time between two charging stations $T_{n,n+1}$ can be defined as:

$$\begin{aligned} \tau(s_n, a_n) = E_s^a\left\{T_{n+1} - T_n\right\} &= \int_0^\infty t \sum_{j \in S} Q(s_n, s_{n+1}, dt|a_n) \\ &= argmin(T_n^{max}, a_n \cdot w_{SOC} \cdot P^{char}) + \frac{f(i,j)}{P^{disc}}. \end{aligned} \tag{3.20}$$

where the first part for this function is the time for charging the EB and this value is constrained by the maximum stay time T_{char} for the EB can stay in the charging station and the second part is the period for the EB travel from the current bus station to the next charging

station.

Then the gain (average reward rate) of an SMDP starting at state i and continuing with policy π can be given as:

$$g^{\pi}(s_n) = \lim_{N\to\infty} \frac{\sum_0^N r^{\pi}(s_n, a_n)}{\sum_0^N \tau^{\pi}(s_n, a_n)}. \tag{3.21}$$

3.3 Average Reward Reinforcement Learning for EB Charging Schedule Optimization

Average reward reinforcement learning is model-free reinforcement learning which is a method of teaching the agent (EB) optimal policy under different states [50]. The optimal policy learned from the agent is the optimal charging schedule for the EB. The policy is aiming to find the action at each charging station that can minimize the battery cost due to the degradation and also extend the lifetime.

The iteration reward function in the algorithm is defined as:

$$\begin{aligned} R_{new}(s_n, a_n) = (1-\alpha_n)R_{old}(s_n, a_n) + \alpha_n r(s_n, s_{n+1}, a_n) \\ - \{g_n \tau(s_n, s_{n+1}, a_n) - max_b R_{old}(s_{n+1}, b)\} \end{aligned} \tag{3.22}$$

where r is the actual reward from the decision epoch n to $n+1$, and τ its corresponding time between these two epochs. g_n is the gain which is the ratio of the total reward and the total simulation time until the nth epoch and it is defined in (3.21).

The algorithm of average reward RL is shown in **Algorithm 3**. By using this algorithm, the optimal charging schedule can be obtained at each en-route charging station. In the algorithm, the energy consumption for the next decision process has been considered when the EB chooses the corresponding action at the charging station. Meanwhile, during the iteration process, the maximum charging SOC is also involved which is limited by its corresponding maximum stay time for EB at the en-route charging station.

Besides the iteration function during the operating process for the algorithm, learning rate α_m and exploration rate p_m are also two important parameters. The learning rate decides the rate of convergence in the process. The exploration rate is used to satisfy the irreducibility of the Markov chain for the system to visit any state. According to the Darken-Chang-Moody (DCM) search-then-converge procedure [51], the learning rate and exploration rate are both decreases to 0 which is shown as follows:

$$\alpha_m = \frac{\alpha_0}{1+\theta}, \theta = \frac{m^2}{\alpha_r + m}, p_m = \frac{p_0}{1+\theta}, \theta = \frac{m^2}{p_r + m} \tag{3.23}$$

where a_m and p_m are the learning rate and exploration rate for ARRL, respectively, while m is the time step for the algorithm and θ is the decreasing rate which allows the a_m and p_m to reduce from its initial value to 0. In the next section, the performance of the algorithm will be tested based on the collected historical data from St. Alberta Transit, AB, Canada. Besides, the differences of SOC at each charging station will be indicated as well as the action selected at the en-route charging station and terminal station.

Algorithm 2 Average Reward RL charging schedule optimization

Input: Current state i, current action a, immediate reward r, next state j, average reward value R;

Output: Optimal charging policy π^* Initialize action values $R_{old}(i,a) = R_{new}(i,a) = 0$, cumulative reward c_{new}, total time t_{new} and reward rate g_{new}. Also initialize the input parameters for the Darken Chang Moody scheme α_0, α_r, p_0 and p_r.

1: While $m < maxsteps$

2: Calculate p_m and a_m using DCM scheme.

3: With probability $1-p_m$, simulate an action $a \in A$, that maximizes $R_{new}(i,a)$. Otherwise choose a random (exploration) action the action space A.

4: Simulate the chosen action. Let the system at the next state is j. Also, calculate the corresponding transition time $\tau(i,j,a)$, immediate reward $r_{imm}(i,j,a)$.

5: Find $R_{new}(i,a)$ using: $R_{new}(i,a) = (1-\alpha_m)R_{old}(i,a) + \alpha_m\{r_{imm}(i,a) - g_{new}\tau(i,j,a) + max_b R_{old}(j,b)\}$

6: in case a non exploratory action was chose in step(4):

- Update total reward $c_{new} \leftarrow c_{new} + r_{imm}(i,j,a)$.
- Update total time $t_{new} \leftarrow t_{new} + \tau(i,j,a)$.
- Update average reward $g_{new} \leftarrow \frac{c_{new}}{t_{new}}$.

Else, go to next step. Set $R_{old}(i,a) \leftarrow R_{new}(i,a)$.

7: Set current state i to new state j and $m \leftarrow m+1$.

8: **Return** π^*

Table 3.1: Simulation Parameters for ARRL

Parameter	Value	Parameter	Value
β_v	50	η_e	0.8
β_e	50000	η_d	0.9
c_r	1.5	β_r	0.008
$\beta_{r'}$	0.0008	g	9.8
ρ	1.32	c_a	1.05
A	6	m	18000

3.4 Case Study

In our simulation, a specific bus route which is shown in Fig. 3.2 and schedule from St. Albert Transit, AB, Canada is used to build the bus model and optimize the charging schedule. Based on the historical data, the velocity, passenger number and road condition of the simulation route can be obtained to calculate the accurate energy consumption between two adjacent charging stations. The schedule of the EB is utilized to restrict the arrival time and departure time of the EB in the charging station. At each charging station, the maximum time that the EB can stay is 15 minutes. In one route, there is a total of 12 charging stations with 11 en-route stations and 1 terminal station. The running time interval from the schedule is 9 hours per day. The parameters used in the simulation are listed in Table 3.1. Our case study is done in MATLAB 2021a on a desktop computer with Intel®

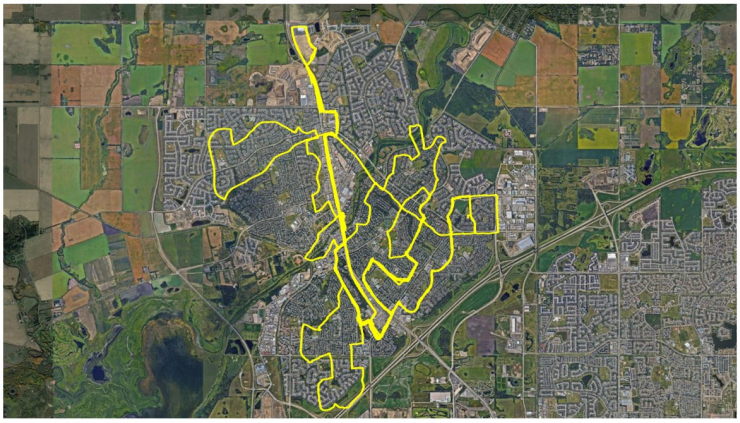

Figure 3.2: Tested bus route for simulation of ARRL.

Core^{TM} i7-7700 CPU @3.60GHz. The average optimization time consumption for charging schedule is 2437.19 seconds. In the simulation, we test the date for a week's operation process of EB. There is a total of 84 charging stations include en-route stations and terminal stations for 100000 times.

In Fig. 3.3, the SOC level at each charging station is indicated as well as the comparison of the tendency. With the terminal charging strategy, EB only charged at each terminal station with full charging. For the en-route charging policy, EB decides the charging SOC at each charging station. In this way, compared to the terminal charging condition with a large drop for SOC, en-route charging maintains the SOC in a smaller range which can decrease the cost for battery degradation. Different value of charging SOC is shown in Fig. 3.4. In this figure, the real charging strategy is that EB completes the path in the daytime and has a full charge in the terminal station. However, the drawback is that peak-load at the terminal station is higher when all EBs back to this station. Compare to the terminal charging policy, en-route charging can allocate the terminal charging process to each charging which can release the load stress in the local grid for the terminal station.

Fig. 3.5 is the R-value for the average reward RL. From the figure, it is clear that the R-value is converged to a fixed value. Since the algorithm is finding the maximum R-value during the iteration, to obtain the minimum cost of the EB, the value function should be negative. Under this condition, the R-value for each iteration result will be negative and it will accumulative to a fixed value.

3.5 Summary

In this chapter, a novel approach for optimizing the en-route charging schedule for EB is proposed to reduce the electrical cost and release the heavy-duty cased by the peak load at the terminal station when all the EBs are charging at the same time. Meanwhile, the physical model of the EB is built to calculate accurate energy consumption during the

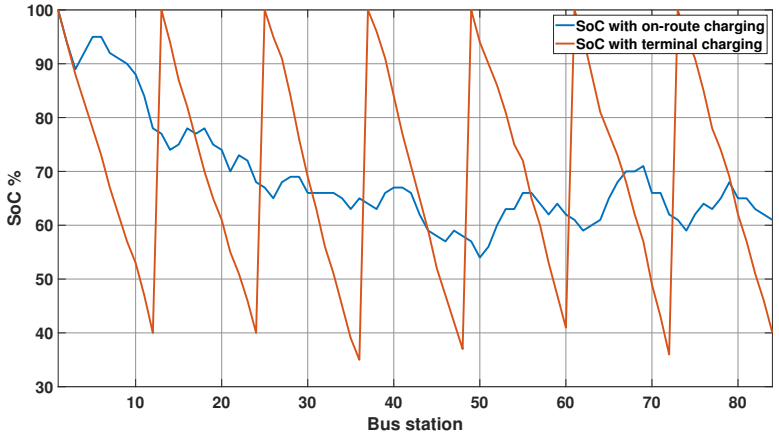

Figure 3.3: SOC level for en-route charging strategy and terminal charging strategy

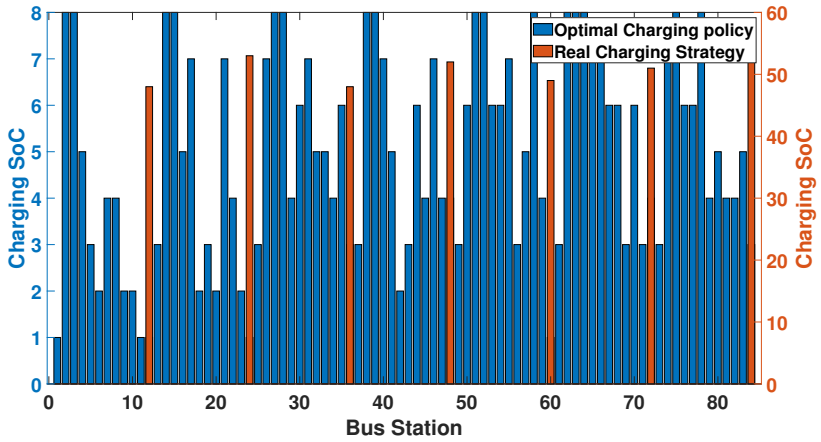

Figure 3.4: Charging SOC for en-route charging strategy and terminal charging strategy

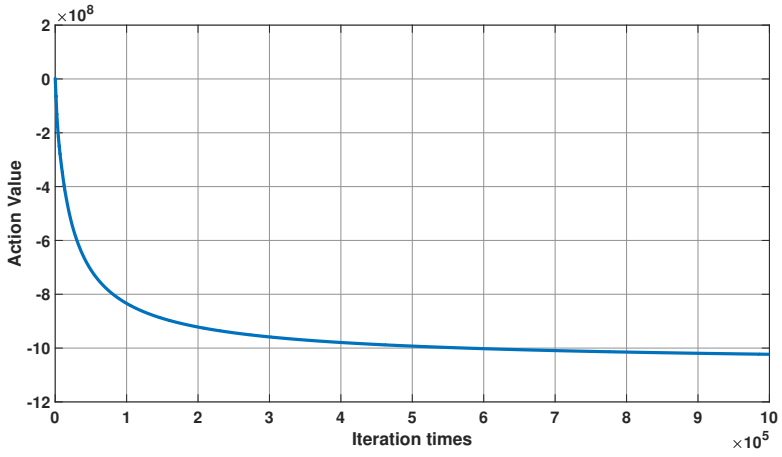

Figure 3.5: Action value for ARRL algorithm

operating process. Finally, the simulation has been done by combining the optimization approach and the physical model of the EB.

4

An RVIRL Approach En-Route EB Charging Schedule Optimization

In this chapter, we proposed a charging schedule optimization approach for EBs based on the RVIRL algorithm. The physical models of EB speed and torque are used to calculate the energy consumption of an EB in between any two adjacent charging stations while the EB operating model is utilized to constrain the SoC level of EB on the route. After that, the battery degradation model is implemented to calculate the battery lifetime degradation parameter which is further used to obtain the degradation cost of the EB battery. Then, the optimal en-route charging strategies of EBs can be obtained from the RVIRL algorithm by considering the total electricity cost and battery degradation cost, and the convergence of the algorithm is proved from a mathematical aspect. The performance of the proposed approach is demonstrated by utilizing the real-world data obtained from the public transit service St. Albert Transit, AB, Canada for the electricity price, EB torque, and EB speed.

4.1 System Model

The operation process of EB is shown in Fig. 4.1. Consider a bus trip with N bus charging stations in the set $\mathbf{N} = \{1, 2, \cdots, N\}$. Let $\mathbf{T} = \{T_1, T_2, \cdots, T_N\}$ be the set of the time moments that the EB arrives at the corresponding charging stations. At each charging station, based on the schedule of the bus, the maximum stop time for the EB is denoted as T_n^{stop}. And the real-time pricing (RTP) of electricity for EB charging can be represented by p_n [\$/kWh]. Define the passenger count from charging station n to $n+1$ as $u_{(n,n+1)}$. According to [55], the random passenger count between adjacent charging stations can be modeled based on log-normal distribution. The average mass of a EB passenger can be defined as m_h, while the mass for the empty EB is denoted by m_{bus}. Different from an

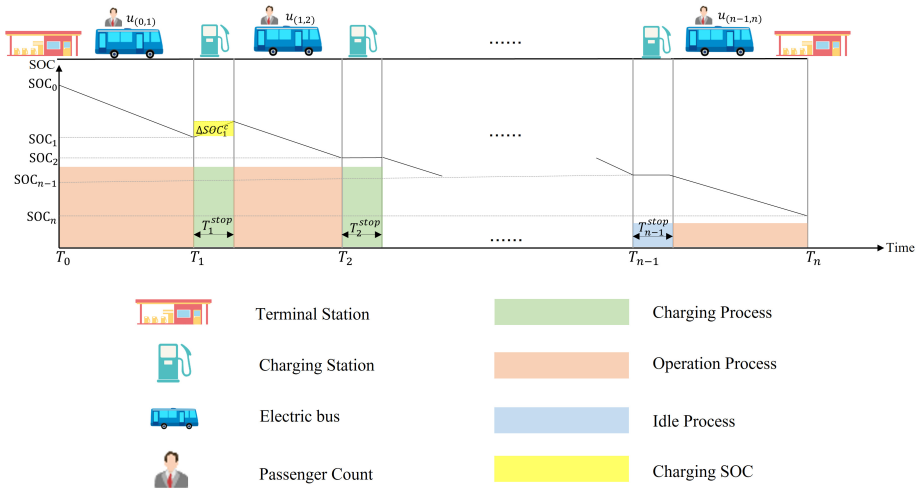

Figure 4.1: Operation process for EB.

electric vehicle, EB will follow a deterministic schedule set by the public transit service. If the EB follows the schedule closely, the time for the EB to arrive at each station is pre-determined, with very small variations due to the randomness in traffic conditions. In other words, the RTP for en-route EB charging can be determined based on the order of charging stations to be visited by each EB. During the operation, define the real-time speed and the current position of EB as v_t and d_t, respectively.

When the EB arrives at the charging station, the charging station decides whether the current EB needs to be charged or not, as well as how much energy is needed based on the available information related to RTP for EB charging, and potential energy consumption for the next trip of EB. It is worth mentioning that, in order to reduce EB charging cost and battery degradation, a charging station may not charge the EB if the current RTP is higher than that of the moment when the EB arriving at the next charging station, and the remaining SOC of the EB can satisfy the energy consumption for the next trip.

4.1.1 EB Energy Consumption Model

During the operation of a EB, its energy consumption can be calculated by using the force decomposition method. Based on the physical model of EB, the total force in EB operation process on the road can be calculated as follows:

$$F_t = F_t^f + F_t^a + F_t^s + F_t^r \tag{4.1}$$

where each force can be calculated separately as follows:

$$F_t^f = \zeta_r \cdot (m_{\text{bus}} + u_{(n,n+1)} m_h) \cdot g \cdot (\gamma_r + \gamma_{r'} \cdot v_t) \tag{4.2}$$

$$F_t^a = \frac{1}{2} \cdot \rho_a \cdot \zeta_a \cdot A \cdot v_t \tag{4.3}$$

$$F_t^s = (m_{\text{bus}} + u_{(n,n+1)} m_h) \cdot g \cdot \sin(f_\theta(d_t))) \tag{4.4}$$

$$F_t^r = (m_{\text{bus}} + u_{(n,n+1)} m_h) \cdot \frac{dv_t}{t}. \tag{4.5}$$

Specifically, the calculation process is as follows:

- F^f is the friction force with ζ_r representing the road fraction coefficient, while γ_r and $\gamma_{r'}$ are the constant parameters corresponding to the fraction of the road. Also, m_h is the average weight of each passenger;
- F^a represents the air drag force due to the air resistance of EB operation, and ρ_a is the density of the air. Also, ζ_a is the aerodynamic dragging coefficient, and A is the front area of the bus;
- F_s denotes a part of the decomposed gravity force, and f_θ is the slope of road corresponding to the location of the EB;
- F_r indicates the resultant force with v_t being the real-time speed.

Based on the real time forces on the EB, the corresponding real time torque of is given by $\tau_t = \frac{r_{\text{bus}} \cdot F_t}{g_r}$ where r_{bus} is the radius of EB tire, and g_r is the gear ratio. The angular speed of the EB can be calculated according to the EB velocity, given by

$$v_t^e = \frac{\zeta_v \cdot g_r \cdot v_t}{2\pi r} \tag{4.6}$$

where ζ_v is a constant parameter.

Then, based on the real time torque and velocity of the EB from (3.5) and (4.6), the energy consumption between charging stations can be calculated as:

$$\Delta E_{(n,n+1)}^{d,a} = \int_{T_n}^{T_{n+1}} \frac{2\pi \cdot \tau(t) \cdot v_e(t)}{\beta_e \cdot \eta_e \cdot \eta_d} dt \tag{4.7}$$

where β_e is a constant parameter, while η_e and η_d are the engine's efficiency and driver's efficiency, respectively.

4.1.2 EB Operation Model

When EB arrives at the charging station, a decision needs to be made on whether the current EB needs to be charged or not. If the charging power at the station is $P_{n,t}$, the maximum SOC increment of the EB at the station is given by

$$\Delta \text{SOC}_n^c = \frac{\int_0^{T_n^{\text{stop}}} P_{n,t} dt}{C_{\text{b}}}. \tag{4.8}$$

In order to satisfy the energy requirement of the EB for the next trip, the SOC of the EB departing from the charging station should be larger than the energy consumption for the EB to arrive at the next charging station, which can be represented as:

$$\mathrm{SOC}_n^{\mathrm{a}} + \Delta\mathrm{SOC}_n^c \geq \Delta\mathrm{SOC}_{n,n+1} \tag{4.9}$$

where SOC_n^a is the arrival SOC of the EB when it arrives the station n, $\Delta\mathrm{SOC}_n^c$ is the charging SOC of the EB in the charging station, and $\Delta\mathrm{SOC}_{n,n+1}$ is the SOC consumption of the EB traveling from station n to station $n+1$.

During the operation process of the EB, the real-time $\mathrm{SOC}(t)$ of the EB should be in the manufacturer-specified range to ensure the health of the battery:

$$\overline{\mathrm{SOC}}^{max} \geq \mathrm{SOC}(t) \geq \underline{\mathrm{SOC}}_{min}. \tag{4.10}$$

4.1.3 EB Battery Degradation Model

Between two charging stations n and $n+1$, the EB operation process can be separated into the charging process during the EB stay in station n and the discharging process from station n to $n+1$. Based on [48], the decrement of EB lifetime between two charging stations can be calculated based on the average SOC and the normalized deviation SOC for a whole charging and discharging cycle which can be defined as:

$$\begin{aligned} \mathrm{SOC}_{(n,n+1)}^v &= \frac{\Delta E_{(n,n+1)}^{d,a}}{C_b \cdot (T_{n+1} - T_n)} \\ \mathrm{SOC}_{(n,n+1)}^d &= 2\sqrt{3 \cdot \frac{\int_{T_n}^{T_{n+1}} (\mathrm{SOC}(t) - \mathrm{SOC}_{(n,n+1)}^v)^2 dt}{T_{n+1} - T_n}}. \end{aligned} \tag{4.11}$$

The battery degradation parameter $\varpi_{(n,n+1)}^h$ for the battery cycles centered on 50% SOC from charging station n to $n+1$ is given by

$$\varpi_{(n,n+1)}^h = \alpha_\sigma \cdot N \cdot \exp\left(\frac{(\mathrm{SOC}_{(n,n+1)}^d - 1) \cdot T_a}{\alpha_\gamma \cdot T_b}\right) + 0.2 \cdot \frac{t_c}{t_l} \tag{4.12}$$

where α_σ is the constant coefficient of throughput, α_γ is the exponent for depth of discharge, t_{cycle} is the time in seconds of a cycle, t_{life} is the total expected shelf life in seconds to 80% capacity at $25°C$ and 50% SOC, and N is the effective number of throughout cycles in the time interval given by $N = \int_{T_n}^{T_{n+1}} \frac{2|I(t)|}{Q_{nom}}$ where $I(t)$ is the battery current at time t and Q_{nom} is the nominal capacity of the battery.

Based on the Zhurkov term [52], the following relationship can be utilized to simplify (4.12):

$$\frac{(\mathrm{SOC}_{(n,n+1)}^d - 1) \cdot T_b}{\alpha_\gamma \cdot T_a} = \frac{g \cdot S}{k \cdot T_a} \tag{4.13}$$

where S is the stress level of the battery. In other words, it is a constant related to the tensile stress of the battery.

In order to utilize the function of battery degradation at any condition inside 50% SOC, the corresponding function can be extended to a general function as follows:

$$\varpi^{g}_{(n,n+1)} = \varpi^{h}_{(n,n+1)} \cdot \exp(\frac{K_\eta \cdot (\mathrm{SOC}^{v}_{(n,n+1)} - 0.5)}{0.25}) \cdot (1 - L) \tag{4.14}$$

where L is aging parameter from 0 to 1.0 [53].

4.2 Problem Formulation

In this paper, we utilize SMDP to formulate the charging schedule optimization of the EBs. SMDP is a generalized version of the MDP. Compared to MDP which is based on the decision making in each time slot, the decision making in SMDP can only happen at decision maker instants, and the sojourn time between decision makers can be random [54]. In this way, SMDP is more suitable for charging schedule optimization of EB in real-world applications. For EB en-route charging, it is not necessary to make charging decision at each time moment especially on the way, because EB is only charged when stopped at the charging stations. Since during the EB operating process, the longer the EB running on the road, the more energy is consumed. It is inappropriate if we only consider the total cost of the energy consumption for the decision making. Under this consideration, the average cost is more suitable for EB en-route charging schedule optimization and thus, it is considered in the following problem formulation.

As for SMDP, the five basic tuples are defined as: $\{\mathcal{S}, \mathcal{A}, \pi, \mathcal{T}(i,j,t,a), r(i,a)\}$ where $\mathcal{S}$ is the state of EB in charging stations, $\mathcal{A}$ is the charging amount of the EBs, π is the charging schedule, $\mathcal{T}(i,j,t,a)$ is the transition probability between two charging stations, and $r(i,a)$ is the total cost for the corresponding charging stations.

In detail, the state $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_n\}$ with each state $\mathcal{S}_n$ can be defined as $\mathcal{S}_n = \left\{\mathrm{SOC}_n, n, p_n, T^{\mathrm{stop}}_n, u_{(n,n+1)}\right\}$ for EB operation. The action at nth decision epoch $\mathcal{A}_{\tau_n}$ is defined as $\mathcal{A}_{\tau_n} = \{\mathcal{A}_{\tau_1}, \mathcal{A}_{\tau_2}, \cdots, \mathcal{A}_{\tau_n}\}$ where the action is $\mathcal{A}_{\tau_n} = \{\Delta\mathrm{SOC}^{c}_{n}\}$ and the decision epoch sequence $\tau = \{\tau_0, \tau_1, \cdots, \tau_n\}$ where τ_n means the time duration from charging station $n-1$ to n. π is the policy for the state-action pair at each decision epoch. Define the transition probability when the current state is i with the selected action a and the next epoch state is j within t time duration as follows:

$$\mathcal{T}(i,j,t,a) = p(X_{n+1} = j, \tau_{n+1} - \tau_n \leq t | X_n = i, \mathcal{A}_n = a)). \tag{4.15}$$

Define the probability that the next decision epoch appears within t time duration if the current state is i while action a is selected as $\mathcal{T}(i,t,a)$. Then, the probability can be calculated as:

$$\mathcal{T}(i,t,a) = \sum_{j \in \mathcal{S}} p(j,t|i,a) = \mathcal{P}(\tau_{n+1} - \tau_n \leq t | X_n = i, \mathcal{A} = a)). \tag{4.16}$$

Between two decision epochs, define $r(i,a)$ as the expected total reward which can be formulated as:

$$r(i,a) = r^{imm}(i,a) + E\left\{\int_{\tau_n}^{\tau_{n+1}} r^{acc}(Y_t,i,a)dt\right\} \tag{4.17}$$

where $r^{imm}(i,a)$ is the immediate fixed reward which can be defined as:

$$r^{imm} = p_n \cdot \Delta E_n^c. \tag{4.18}$$

Also, in (4.17), $r^{acc}(Y_t,i,a)$ is the accumulated additional reward rate which can be calculated based on the battery degradation mentioned in the aforementioned section as $r^{acc} = \hat{C}_b \varpi_{n,n+1}^g$, where $\hat{C}_b$ is the total cost of the battery of EB, while $Y_t(\tau_n \leq t \leq \tau_{n+1})$ indicates that the accumulative reward process is a natural process. For state $i \in \mathcal{S}$ and $a \in \mathcal{A}$, the mean of the stop time at state i and action a is given by

$$\mathcal{O}(i,a) = \int_0^\infty s\mathcal{T}(i,ds,a) = E\left[\tau_{n+1} - \tau_n | X_n = i, \mathcal{A}_n = a\right]. \tag{4.19}$$

Since in SMDP, besides the expected total reward between two decision epochs, the expected total time is also dependent on the policy. In this way, the average reward under policy π is defined in SMDP as follows:

$$\delta^\pi(i,a) = \frac{r(i,a)}{\mathcal{O}(i,a)}. \tag{4.20}$$

For a policy π^* and any policy $v \in \Pi$, if they satisfy $\delta^{\pi^*} > \delta^v$, we call that policy π^* is the average reward optimal policy and δ^{π^*} is the optimal average reward.

4.3 Relative Value Iteration Reinforcement Learning for Charging Schedule Optimization

In this section, an RVIRL algorithm is developed to optimize the en-route charging schedule of EB based on SMDP which does not rely on the optimal average reward in the process of learning iteration. In this way, the algorithm can converge faster since it does not need to wait the learned policy approaches to the optimal policy. The sensitivity analysis introduced in [49] is used to calculated the iteration function of RL. Based on the perturbation analysis theory, we have the following relation between the average reward and the state transition probability function under a given policy $\pi \in \Pi$:

$$\delta^\pi = \mathcal{T}^\pi(\Psi^\pi r^\pi + \Omega^\pi) \tag{4.21}$$

where Ψ is the infinitesimal generator, and Ω is the average reward. The infinitesimal generator for states i and j under action α can be defined as follows:

$$\Psi(i,j,a) = \frac{[\mathcal{T}(i,j,a) - I(i,j,a)]}{\mathcal{O}(i,a)}. \tag{4.22}$$

If we define the transition probability from i to j under the action a as $\mathcal{T}(i,j,a) = \lim_{t\to\infty} \mathcal{T}(i,j,t,a) = P\{X_{n+1} = j | X_n = i, \mathcal{A} = a\}$, then we have

$$\Omega(i,a) = \frac{\sum_{j\in\mathcal{S}} \mathcal{T}(i,j,a)\Omega(i,j,a)\mathcal{O}(i,j,a)}{\mathcal{O}(i,a)}. \tag{4.23}$$

Based on (4.21), the average reward among different policies of SMDP can be calculated as follows:

$$\delta^{\pi} - \delta^{v} = p^{\pi}\left[(\Psi^{\pi} r^{v} + \Omega^{\pi}) - (\Psi^{v} r^{v} + \Omega^{v})\right], \forall v, \pi \in \Pi. \tag{4.24}$$

The optimal condition for SMDP is that a policy π^* is optimal if and only if:

$$\Psi^{\pi^*} r^{\pi^*} + \Omega^{\pi^*} \geq \Psi^{\pi} r^{\pi^*} + \Omega^{\pi}. \tag{4.25}$$

Based on (4.25), the optimal policy π^* is given by

$$\delta^{\pi^*} e = \max_{\pi\in\Pi} \left\{ \Psi^{\pi^*} r^{\pi^*} + \Omega^{\pi^*} \right\}. \tag{4.26}$$

Accordingly, we have

$$\begin{aligned} \delta^{\pi^*} e + r^{\pi^*} &= \max\left\{ \Omega^{\pi} + \Psi^{\pi^*} r^{\pi^*} + r^{\pi^*} \right\} = \max\left\{ \Omega^{\pi} - \frac{[\mathcal{T}(i,j) - I(i,j)]}{\mathcal{O}(i)} r^{\pi^*} + g^{\pi^*} \right\} \\ &= \max\left\{ \Omega^{\pi} - (1 - \frac{I(i,j)}{\mathcal{O}(i)}) r^{\pi^*} + \frac{\mathcal{T}(i,j)}{\mathcal{O}(i)} r^{\pi^*} \right\} \\ &= \max\left\{ \Omega^{\pi} - (1 - \frac{\sum_{j\in\mathcal{S}} I(i,j)}{\mathcal{O}(i)}) r^{\pi^*} + \frac{\sum_{j\in\mathcal{S}} \mathcal{T}(i,j)}{\mathcal{O}(i)} r^{\pi^*} \right\}. \end{aligned} \tag{4.27}$$

Based on the definition of the Q factor [57], given any policy $\pi \in \Pi$ with potentials $g^{\pi}(j), j \in \mathcal{S}$, and the average reward η^{π}, for every state-action pair (i,α), the state-action value function is given by

$$Q^{\pi}(i,\alpha) = \Omega^{\pi}(i,\alpha) + \left[1 - \frac{1}{\mathcal{O}(i,\alpha)}\right] Q^{\pi}(i,\alpha) + \frac{\sum_{j\in\mathcal{S}} \mathcal{T}(i,j,\alpha) r^{\pi}(j)}{\mathcal{O}(i,\alpha)} - \delta^{\pi}, \alpha \in \mathcal{A}(i), i \in \mathcal{S}. \tag{4.28}$$

Taking the maximum on both sides over the action space $\mathcal{A}(i)$ and the optimal policy π^*, we have

$$\max_{\alpha\in\mathcal{A}_n} Q^{\pi^*}(i,\alpha) + \delta^{\pi^*} = \max_{\alpha\in\mathcal{A}_n} \{\Omega(i,\alpha) + \left[1 - \frac{1}{\mathcal{O}(i,\alpha)}\right] Q^{\pi^*}(i,\alpha) + \frac{\sum_{j\in\mathcal{S}} \mathcal{T}(i,j,\alpha) r^{\pi^*}(j)}{\mathcal{O}(i,\alpha)} \Bigg\}. \tag{4.29}$$

By comparing (4.27) and (4.29), for the optimal policy π^*, we have:

$$r^{\pi^*}(i) = \max_{\alpha\in\mathcal{A}(i)} Q^{\pi^*}(i,\alpha). \tag{4.30}$$

By combining (4.30) with (4.28), the optimality equation for Q factor under the gain-optimal policy π^* is given by

$$Q^{\pi^*}(i,\alpha) = \Omega(i,\alpha) + \left[1 - \frac{1}{\mathcal{O}(i,\alpha)}\right] Q^{\pi^*}(i,\alpha) + \frac{\sum_{j\in\mathcal{S}} \mathcal{T}(i,j,\alpha) \left[\max_{\beta\in\mathcal{A}(j)} Q^{\pi^*}(j,\beta)\right]}{\mathcal{O}(i,\alpha)} - \delta^{\pi}. \tag{4.31}$$

When the exact value of transition probability is unknown, we need to observe the state transitions on a sample path to obtain the information about the transition probabilities. Based on the aforementioned discussion, Q factors play an important role for the state-action pairs, which can be estimated by visiting all the state-action pairs for the sample path. If the action α is taken at time τ with $X_\tau = i$ and stop time $\Delta\tau$, then we have

$$\sum_{j\in\mathcal{S}} \mathcal{T}(i,j,\alpha)\left[\max_{\beta\in\mathcal{A}(j)} Q(j,\beta)\right] = E\left\{\left[\max_{\beta\in\mathcal{A}(X_{\tau+\Delta\tau})} Q(X_{\tau+\Delta\tau},\beta)\right] | X_\tau = i, A_\tau = \alpha\right\}. \quad (4.32)$$

Accordingly, the iteration functions for RVIRL can be formulated as follows:

$$\begin{aligned} Q^{k+1}(x_n,a_n) &= Q^k(x_n,a_n) + \gamma^k H^k(x_n,a_n) \\ H^k(x_n,a_n) &= \delta^k(x_n,a_n,\mathcal{O}(x_n,a_n)) + (1-\frac{1}{\mathcal{O}(x_n,a_n)}) \max_{\alpha\in\mathcal{A}_n} Q^k(x_n,\alpha) \\ &\quad + \frac{1}{\mathcal{O}(x_n,a_n)} \max_{\beta\in\mathcal{A}_{n+1}} Q^k(x_{n+1},\beta) - Q^k(x_n,a_n) - \eta \end{aligned} \quad (4.33)$$

where γ^k is the step size for RL which satisfies

$$\sum_{k=0}^{\infty}\gamma^k = \infty, \sum_{k=0}^{\infty}(\gamma^k)^2 < \infty. \quad (4.34)$$

According to [56], the optimal average reward η is not need for the EB en-route charging schedule optimization since the convergence rate might be slow or the ratio between the total reward and total time might be unexpected. Based on the definition of Q factor and function (4.31), during the process of policy iteration, the Q factor may converge to a constant which depends on the relative values of Q factor. Therefore, η can be replaced with relative Q factor, and the relative value iteration function is given by

$$\begin{aligned} H^k(x_n,a_n) &= \delta^k(x_n,a_n,\tau(x_n,a_n)) + (1-\frac{1}{\mathcal{O}(x_n,a_n)}) \max_{\alpha\in\mathcal{A}_n} Q^k(x_n,\alpha) \\ &\quad + \frac{1}{\mathcal{O}(x_n,a_n)} \max_{\beta\in\mathcal{A}_{n+1}} Q^k(x_{n+1},\beta) - Q^k(x_n,a_n) - \widetilde{Q}^k(x^n,a^n). \end{aligned} \quad (4.35)$$

However, in order to ensure the convergence of the algorithm, the relative Q factor $\widetilde{Q}^k(x^n,a^n)$ can be replaced by a reference Q factor state-action pair $Q(i^*,a^*)$, such that

$$\begin{aligned} H^k(x_n,a_n) &= \delta^k(x_n,a_n,\tau(x_n,a_n)) + (1-\frac{1}{\mathcal{O}(x_n,a_n)}) \max_{\alpha\in\mathcal{A}_n} Q^k(x_n,\alpha) \\ &\quad + \frac{1}{\mathcal{O}(x_n,a_n)} \max_{\beta\in\mathcal{A}_{n+1}} Q^k(x_{n+1},\beta) - Q^k(x_n,a_n) - \widetilde{Q}^k(i^*,a^*). \end{aligned} \quad (4.36)$$

Besides the iteration function, learning rate γ^k and exploration rate ι_k are also two important parameters. The learning rate determines the rate of convergence in the process, while the exploration rate is used to satisfy the irreducibility of the Markov chain for the

Algorithm 3 RVIRL for Charging Schedule Optimization

Input: State space $\mathcal{S}$, action space $\mathcal{A}$ and reward r
Output: Optimal charging policy π^*

1: Initialize $Q_{\text{old}}(i,a) = Q_{\text{new}}(i,a) = 0$. Also initialize the input parameters for the Darken Chang Moody scheme γ_0, γ_r, ι_0 and ι_r.
2: While $m < \text{maxsteps}$
3: Calculate ι^k and γ^k using DCM scheme.
4: With probability $1-\iota^k$, simulate an action $a \in A$, that maximizes $Q_{\text{new}}(i,a)$. Otherwise choose a random (exploration) action the action space A.
5: Simulate the chosen action. Let the system at the next state is j. Also, calculate the corresponding stop time $\tau(i,j,a)$, immediate reward $r_{imm}(i,j,a)$.
6: Find $Q_{\text{new}}(i,a)$ using:

$$Q_{\text{new}}(i,a) = (1-\gamma^k)Q_{\text{old}}(i,a) + \gamma^k((1-\frac{1}{\mathcal{O}(i,a)})\max_{\alpha\in\mathcal{A}_n} Q_{\text{old}}(i,\alpha) + \frac{r(i,a)}{\mathcal{O}(i,a)} + \frac{1}{\mathcal{O}(i,a)}\max_{\beta\in\mathcal{A}_{n+1}} Q_{\text{old}}(j,\beta) - \widetilde{Q}_{\text{old}}(i,a))$$

7: Set $Q_{\text{old}}(i,a) \leftarrow Q_{\text{new}}(i,a)$.
8: Set current state i to new state j and $m \leftarrow m+1$.
9: **Return** π^*

system to visit any state. According to the Darken-Chang-Moody (DCM) search-then-converge procedure [51], the learning rate and exploration rate are given by

$$\gamma^k = \frac{\gamma_0}{1+\theta_\gamma}, \theta_\gamma = \frac{k^2}{\gamma_r + k}, \iota^k = \frac{\iota_0}{1+\theta_\iota}, \theta_\iota = \frac{k^2}{\iota_r + k} \tag{4.37}$$

The details of the proposed RVIRL approach is shown in **Algorithm 3**. At first, initialize the state space, action space, reward as well as the parameters of the DCM scheme. During the iteration, the learning rate and exploration rate can be updated based on the iteration numbers. The selected action at each state result in the next state, the reward and stop time. Then the Q-value will be updated based on the last Q-value and the corresponding reward and stop time. When the iteration is done, the algorithm could return the optimal charging schedule for EB.

4.4 Convergence Analysis of the RVIRL Approach

In this section, of the proposed RVIRL approach is analyzed. In particular, such convergence is crucial for the reliable operation of public transit services with EBs, such that the optimal charging schedule can always be found. The basic idea for the proof of convergence is based on the stability of ordinary differential equation (ODE) [58]. Here, we first introduce three assumptions which will be used in the following convergence analysis.

Assumption 1 *Define Υ as a Lipschitz function, given by*

$$\Upsilon(Q^k) = Q^k(i_0,a_0), i_0 \in \mathcal{S}, a_0 \in \mathcal{A} \tag{4.38}$$

where i_0 and a_0 are the prescribed state and action, respectively, during the iteration process.

Assumption 2 *Define* $\|Q\|_\infty = \max_{i\in\mathcal{S},\alpha\in\mathcal{A}} |Q(i,a)|$ *and* $\|Q\|_s = \max_{i\in\mathcal{S},\alpha\in\mathcal{A}} Q(i,a) - \min_{i\in\mathcal{S},\alpha\in\mathcal{A}} Q(i,a)$ *as the max norm and span semi-norm of the corresponding Q factor.*

Assumption 3 $|\Upsilon(Q)| \leq \|Q\|_\infty$ *for all* $Q \in \mathbb{R}^d$

In order to analyze the convergence of the RVIRL algorithm for SMDP, the iteration function (4.33) can be rewritten as follows:

$$Q^{k+1}(i,a) = Q^k(i,a) + \gamma^k(\mathcal{U}(Q^k(i,a)) - \Upsilon(Q^k(i,a)) - Q^k(i,a) + \mathcal{M}^{k+1}) \tag{4.39}$$

and $\mathcal{U}$ is the function defined by:

$$\mathcal{U}(Q(i,a)) = (1-\tau(i,a)) \min_{\alpha\in\mathcal{A}} Q(i,\alpha) + \sum p(i,j,a) r(i,j,a) + \tau(i,a) \min_{\beta\in\mathcal{A}} Q(j,\beta)). \tag{4.40}$$

For any $k \geq 0$ we have

$$\mathcal{M}^{k+1}(i,a) = r(i,j,a) + (1-\tau(i,a)) \min_{\alpha\in\mathcal{A}} Q(i,\alpha) + \tau(i,a) \min_{\beta\in\mathcal{A}} Q(j,\beta) - \mathcal{U}(Q(i,a)). \tag{4.41}$$

Define $\mathcal{U}^\alpha$ and $\mathcal{U}^\beta$ as follows:

$$\begin{aligned} \mathcal{U}^\alpha(Q) &= \mathcal{U}(Q) - \kappa e \\ \mathcal{U}^\beta(Q) &= \mathcal{U}(Q) - \Upsilon(Q)e = \mathcal{U}^\alpha(Q) + (\kappa - f(Q))e \end{aligned} \tag{4.42}$$

where κ is the optimal cost which can be calculated as $\kappa = Q^k(i^*, a^*)$, with $i^* \in \mathcal{S}, a^* \in \mathcal{A}$

Since the iteration process (4.39) is in the form of a standard stochastic approximation algorithm, based on the stability of ODE, we have

$$\dot{Q}(t) = \mathcal{U}'(Q(t)) - Q(t)). \tag{4.43}$$

Then, the convergence of the iteration can be analyzed. By combining with function (4.42), the stability function can be rewritten as follows:

$$\dot{Q}(t) = \mathcal{U}^\beta(Q(t)) - Q(t). \tag{4.44}$$

Clearly, function (4.43) has a unique equilibrium point at Q^* since $\Upsilon(Q^*) = \kappa$ and $\mathcal{U}'(Q^*) = \mathcal{U}^\beta(Q^*) = Q^*$. On the other hand, if $\mathcal{U}'(Q) = Q$, we have $Q = \mathcal{U}^\beta(Q) - (\kappa - \Upsilon(Q))e$. Since the Bellman equation $Q = \mathcal{U}^\beta(Q) + ce$ has a solution if and only if $c = 0$. As a result, if $\Upsilon(Q) = \kappa$, we have $Q = Q^*$. As for function (4.44), if we define $\zeta(\cdot)$ and ϵ as the solution and an equilibrium point of the corresponding function, we can see that $\|\zeta(t) - \epsilon\|_\infty$ is non-increasing, and $\zeta(t) \to \zeta^*$ for equilibrium point ζ^* of (4.44) which depends on $\zeta(0)$.

Lemma 1 *Define two functions $\varrho(\cdot)$ and $\varsigma(\cdot)$ satisfying (4.43) and (4.44), respectively, with $\varrho(0) = \varsigma(0) = \varrho_0$. Then $\varrho(t) = \varsigma(t) + \vartheta(t)e$, where r satisfies the ODE:*

$$\dot{\vartheta}(t) = -\vartheta(t) + \kappa - \Upsilon(\varsigma(t)). \tag{4.45}$$

Proof. Based on the variation of the constant formula for $\varrho(\tau)$ and $\varsigma(\tau)$:

$$\varrho(\tau) = \varrho_0 e^{-\tau} + \int_0^\tau e(-\tau - s)\mathcal{U}^\beta(\varrho(s))ds + e\int_0^\tau e^(- \tau - s)(\kappa - \Upsilon(\varrho(s)))ds \tag{4.46}$$

$$\varsigma(\tau) = \varrho_0 e^{-\tau} + \int_0^\tau e^{(-\tau - s)}\mathcal{U}^\beta(\varsigma(s))ds. \tag{4.47}$$

Then, if we use $\mathcal{U}_n^\beta$ to represent the nth component of $\mathcal{U}^\beta$, the max and min relationship for function $\varrho(\cdot)$ and $\varsigma(\cdot)$ are:

$$\max_n(\varrho_n(\tau) - \varsigma_n(\tau)) \leq \int_0^\tau e^{(s-\tau)} \max_n(\mathcal{U}_n^\beta(\varrho(s)) - \mathcal{U}_n^\beta(\varsigma(s)))ds + e\left[\int_0^\tau e^{s-\tau}(\kappa - \Upsilon(\varrho(s)))ds\right] \tag{4.48}$$

$$\min_n(\varrho_n(\tau) - \varsigma_n(\tau)) \geq \int_0^\tau e^{(s-\tau)} \min_n(\mathcal{U}_n^\beta(\varrho(s)) - \mathcal{U}_n^\beta(\varsigma(s)))ds + e\left[\int_0^\tau e^{s-\tau}(\kappa - \Upsilon(\varrho(s)))ds\right]. \tag{4.49}$$

According to **Assumption** 2, we have:

$$\|\varrho(\tau) - \varsigma(\tau)\| \leq \int_0^\tau e^{s-\tau}\left\|\hat{\mathcal{U}}(\varrho(s)) - \hat{\mathcal{U}}(\varsigma(s))\right\| ds \leq \int_0^\tau e^{s-\tau}\|\varrho(s) - \varsigma(s)\| ds. \tag{4.50}$$

According to the Gronwall's inequality, $\|\varrho(\tau) - \varsigma(\tau)\| = 0$ for all $\tau \geq 0$. Since we have $\varrho = ce \Leftrightarrow \|\varrho\| = 0$ for some $c \in \mathbb{R}$, we can derive $\varrho(\tau) = \varsigma(\tau) + e\vartheta(\tau)$. Based on the fact that $\varrho(0) = \varsigma(0)$ and $\vartheta(0) = 0$, we have

$$\begin{aligned} \hat{\mathcal{U}}(\varrho + ce) &= \hat{\mathcal{U}}(\varrho) + ce \\ \Upsilon(\varrho + ce) &= \Upsilon(\varrho) + c. \end{aligned} \tag{4.51}$$

If $\vartheta \in \mathbb{R}$, then we have

$$\begin{aligned} \dot{\vartheta}(\tau)e &= \dot{\varrho}(\tau) - \dot{\varsigma}(\tau) = (\mathcal{U}^\beta(\varrho(\tau)) - \varrho(\tau) + \beta - \Upsilon(\varrho(\tau))e) - (\mathcal{U}^\beta(\varsigma(\tau)) - \varsigma(\tau)) \\ &= e(\epsilon - \vartheta(\tau) - \Upsilon(\varsigma(\tau))). \end{aligned} \tag{4.52}$$

So, $\dot{\vartheta}(t) = -\vartheta(t) + \kappa - \Upsilon(\varsigma(t))$ is proved. ■

Theorem 1 *Q^* is a globally asymptotically stable equilibrium point for equation (4.43).*

Proof. Based on the variation of a constant formula, function $\chi(t)$ can be defined as:

$$\chi(t) = \int_0^t e^{-(t-\tau)}(\kappa - \Upsilon(\varsigma(\tau)))d\tau. \tag{4.53}$$

If we assume that equation (4.43) has a unique equilibrium point Q^*, then $\varsigma(t) \to \varsigma^*$. Based on the above result, we have $\chi(t) \to \kappa - \Upsilon(\varsigma^*)$ so that $\varrho(t) \to \varsigma^* + (\kappa - \Upsilon(\varsigma^*))e$. In order to

prove the asymptotic stability, the Lyapunov stability condition should be satisfied by the iteration, given by

$$\begin{aligned}\|\varrho(t)-Q^*\|_\infty &\le \|\varsigma(t)-Q^*\|_\infty + |\chi(t)| \le \|\varsigma(0)-Q^*\|_\infty + \int_0^t e^{-(t-\tau)}|\kappa-\Upsilon(\varsigma(\tau))|d\tau \\ &\le \|\varrho(0)-Q^*\|_\infty + \int_0^t e^{-(t-\tau)}|\Upsilon(Q^*)-\Upsilon(\varsigma(\tau))|d\tau.\end{aligned} \tag{4.54}$$

According to **Assumption** 1, $\Upsilon(\cdot)$ is Lipschitz, such that

$$|\Upsilon(Q^*)-\Upsilon(\varsigma(\tau))| \le L\,\|\varsigma(\tau)-Q^*\| \le L\,\|\varsigma(0)-Q^*\| = L\,\|\varrho(0)-Q^*\|. \tag{4.55}$$

If we assume $L>0$, then we have

$$\|\varrho(\tau)-Q^*\|_\infty \le (1+L)\,\|\varrho(0)-Q^*\|_\infty. \tag{4.56}$$

Lemma 2 *Define $\mathcal{D}$ as an open bounded set of $\mathbb{R}^n$ and $\mathcal{C}$ be a set containing $\mathcal{D}$. We define a mapping $\Lambda_{\mathcal{D},\mathcal{C}}:\mathbb{R}^n\to\bar{\mathcal{D}}$ as follows:*

$$\Lambda_{\mathcal{D},\mathcal{C}}(x) = \lambda_{\mathcal{D},\mathcal{C}}(x)\cdot x \tag{4.57}$$

where $\lambda_{\mathcal{D},\mathcal{C}}:\mathbb{R}^k\to(0,1)$ is given by

$$\lambda_{\mathcal{D},\mathcal{C}} = \begin{cases} 0 & \text{if } x\in\mathcal{C} \\ \max\{\beta>0:\beta x\in\bar{\mathcal{B}}\} & \text{if } x\notin\mathcal{C}\end{cases} \tag{4.58}$$

where $\bar{\mathcal{B}}$ is the closure of $\mathcal{B}$. Let $\|x\|_s=\max_i x_i-\min_i x_i$ define the span semi-norm on $\mathbb{R}^n$. Then, we can denote the iteration as $x^{k+1}=G^k(x^k,\zeta^k)$, where $\{\zeta^k\}$ is a random process and $\{G^k\}$ satisfies the following condition:

$$\left\|G^k(x,\zeta)-G^k(y,\zeta)\right\|_s \le \|x-y\|_s. \tag{4.59}$$

Then the sequence $\{\widetilde{x}^k\}$ generated by the iteration $\widetilde{x}^k=G^k\left(\Lambda_{\mathcal{D},\mathcal{C}}(\widetilde{x}^k),\zeta^k\right)$ converges, based on which we can say that sequence $\{\widetilde{x}^k\}$ remains bounded.

According to [59], we consider $\mathcal{B}$ as an open neighborhood of Q^*, and $\mathcal{C}=\{x:V(x)\le c\}$ with $c>0$ is chosen to be large enough so as to ensure that $\mathcal{B}\subset interior(\mathcal{C})$. Then, we define $\mathcal{Z}:R^{d\times r}\to\mathcal{C}$ as follows:

$$\mathcal{Z}(x) = \begin{cases} x & if\, x\in\mathcal{C} \\ Q^*+\eta(x)(x-Q^*) & if\, x\notin\mathcal{C}\end{cases} \tag{4.60}$$

where $\eta(x)=\max\{a>0:Q^*+a(x-Q^*)\in\mathcal{B}\}$. Then, the scaled version of the iteration process under state i and action a can be defined as follows:

$$\bar{Q}^{k+1}(i,a)=\widehat{Q}^k(i,a)+\gamma^k(r(i,a)+(1-\tau(i,a))\min_{\alpha\in\mathcal{A}^k}\widehat{Q}^k(i,\alpha)+\tau(i,a)\min_{\beta\in\mathcal{A}^{k+1}}\widehat{Q}^k(i,\beta)-f(\widehat{Q}^k)) \tag{4.61}$$

Figure 4.2: EB data collection APP and EB routes.

where $\widehat{Q}^k = \mathcal{Z}(\bar{Q}^k)$. Based on the result of [59], iteration function (4.61) remain bounded.

Finally, based on **Lemma** 2 and the bounded result of iteration process (4.61), $\|\bar{Q}^k - Q^k\|_s \leq \mathbb{D}$ where $\mathbb{D}$ is an arbitrary constant. Then, based on the previous result, we have:

$$\sup_k \left\|Q^k\right\|_s \leq \sup_k \left\|\bar{Q}^k\right\|_s + \sup_k \left\|Q^k - \bar{Q}^k\right\|_s := \mathbb{K} < \infty. \tag{4.62}$$

Based on **Assumption** 1 and 3, the convergence of Q value, given the current state i, selected action a, and the state at next decision epoch j, can be formulated as:

$$\begin{aligned} |\min_{\varphi\in\mathcal{A}} Q^k(j,\varphi) - \Upsilon(Q^k(i,a))| =& |\Upsilon(Q^k(i,a) - e\min_{\varphi\in\mathcal{A}} Q^k(j,\varphi))| \\ \leq& \left\|Q^k(i,a) - e\min_{\varphi\in\mathcal{A}} Q^k(j,\varphi)\right\|_\infty \\ \leq& \left\|Q^k(i,a)\right\|_s \leq \mathbb{K}. \end{aligned} \tag{4.63}$$

Define $D = \max(\left\|Q^0\right\|, \max_{i,j\in\mathcal{S},a\in\mathcal{A}} |r(i,j,a)| + \mathbb{K})$. Then, the boundary of the Q factor can be derived as follows:

$$|Q^{k+1}(i,a)| \leq (1-\gamma^k)\left\|Q^k(i,a)\right\|_\infty + \gamma^k(\max_{i,j\in\mathcal{S},a\in\mathcal{A}} r(i,j,a) + \mathbb{K}) \leq (1-\gamma^k)\left\|Q(i,a)^k\right\|_\infty + \gamma^k D \tag{4.64}$$

Based on (4.64), we have $\left\|Q^k\right\|_\infty \leq D$ for all k. In other words, the value function for the proposed algorithm is always smaller than a constant as the iteration goes to infinite which means that the algorithm converges. ■

4.5 Case Study

In the case study, we use the real-world data from the public transit service St. Albert Transit from Jun. 1 to Sept. 1, with an Android APP develop by ourselves, as shown in Fig. 4.2. The parameters used for the calculation of the energy consumption are provided

Table 4.1: Simulation Parameters for RVIRL

Parameter	Value	Parameter	Value
ζ_v	50	η_e	0.8
β_e	50000	η_d	0.9
ζ_r	1.5	γ_r	0.008
$\gamma_{r'}$	0.0008	g	9.8
ρ_a	1.32	ζ_a	1.05
A	6	m_{bus}	18000
α_σ	3.66e-5	α_γ	0.717

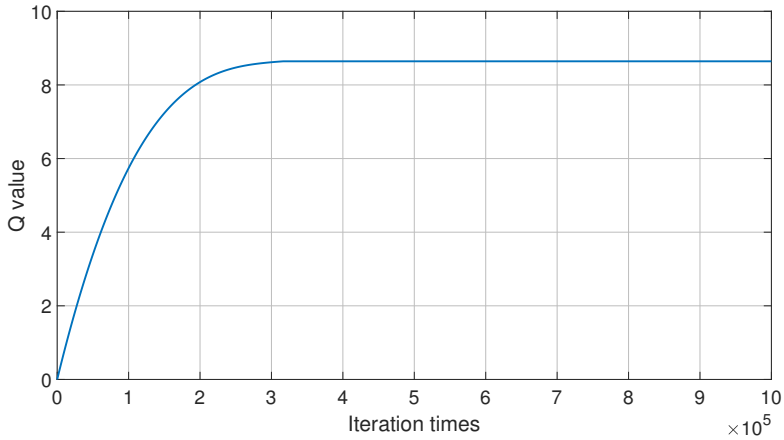

Figure 4.3: Convergence of RVIRL algorithm.

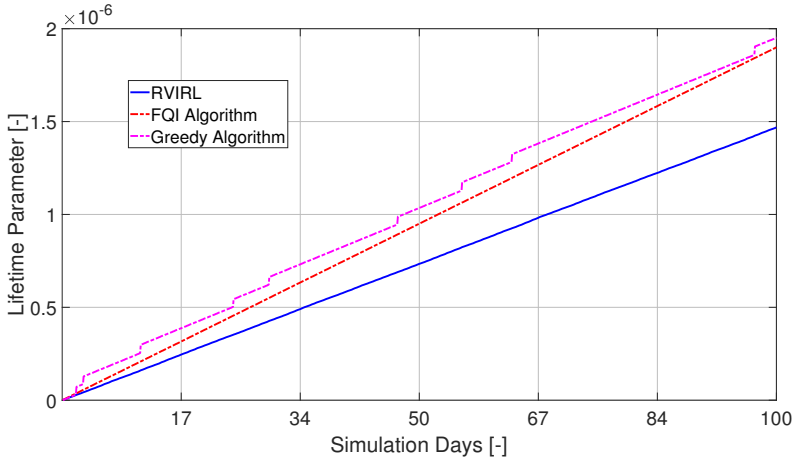

Figure 4.4: An illustration of the total battery lifetime increment.

in Table 4.1. These values are based on the EB model BYD K9 currently used by St. Albert Transit [5]. Our case study is done in MATLAB 2021a on a desktop computer with Intel® Core™ i7-7700 CPU @3.60GHz. The average optimization time consumption for charging schedule is 5869.78 seconds. According to the data, the battery capacity we used in the simulation is 300 kWh and the maximum charging power of the EB is 80 kW/h. For the EB

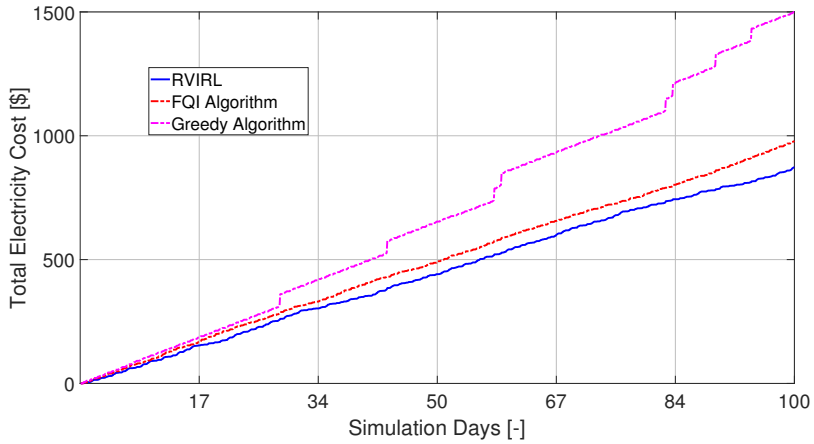

Figure 4.5: A comparison of the electricity cost among different algorithms.

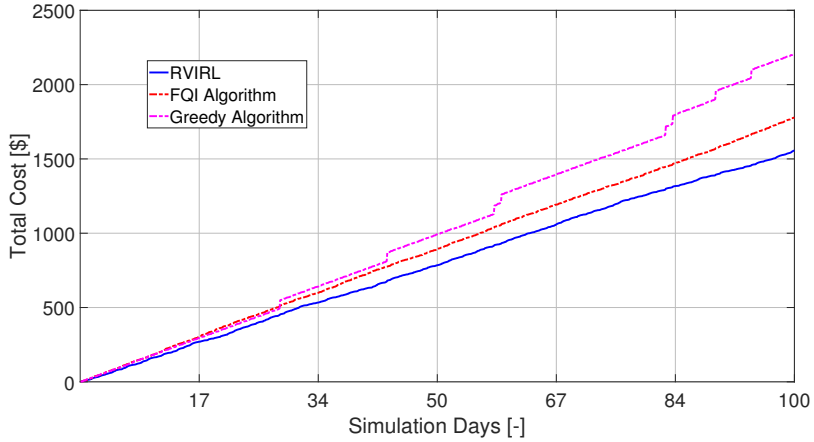

Figure 4.6: A comparison of the total cost among different algorithms.

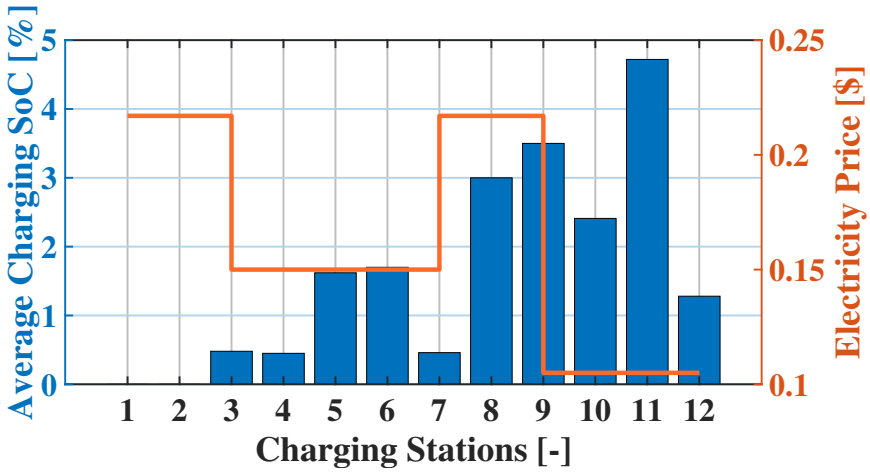

Figure 4.7: Average charging SOC and electricity price.

route schedule of the transit company, there are 12 en-route charging stations on the street. As for the algorithm in the simulation, we set the exploration rate as 0.8, while the step size is 0.2. The proposed algorithm is compared with the existing algorithms including the fitted Q iteration (FQI) RL algorithm from [21] and the greedy algorithm [27] .

Based on the aforementioned case study setup, the simulation results and the performance comparison among different algorithms are presented in the following. Fig. 4.3 indicates the convergence of the Q-value iteration for the RVIRL algorithm. Similar to our theoretical analysis, it is clear that during the iteration process, the Q-value of the algorithm can converge to a constant. The results of the total lifetime increment parameter are shown in Fig. 4.4. Since both the RVIRL algorithm and FQI algorithm utilize battery degradation parameter as part of the reward in the iteration function, the total battery lifetime loss ratio of the greedy algorithm is the largest. Compared to the FQI algorithm, the charging schedule obtained from the RVIRL algorithm has a better performance in terms of battery lifetime saving. The results for the electricity cost during the simulation period are shown in Fig. 4.5. In comparison with the charging strategies obtained from the FQI algorithm and greedy algorithm, the electricity cost based on the charging strategy from the RVIRL algorithm is the lowest. A comparison of the total cost among different algorithms is shown in Fig. 4.6. As we can see, the accumulative cost of the RVIRL for the simulation period is the lowest, and the greedy algorithm leads to the highest cost. The reason is that when the greedy algorithm selects the charging strategy, the electricity price is the only considered part. However, in RVIRL and FQI algorithms, not only the electricity price is considered, but also the degradation cost of the battery. Besides the cost comparisons among different algorithms, the relationship between the charging schedule for the RVIRL algorithm and the electricity price is shown in Fig. 4.7. According to the result from the figure, when the electricity price is high, the EB might not select a charging decision, and when the price is low, the EB might charge itself. The only difference is that when the battery SOC level is low, the EB will select a charging station in order to minimize the battery degradation cost and reduce the battery lifetime lost, even the electricity price is high. Based on the results from Fig. 4.7, the EB does not charge at the first charging stations since public transit service (St. Albert Transit) requires full charge overnight for safety concerns. However, if not full charge, our algorithm still applies. For charging station 3-7, even if the electricity price is low, EB does not charge a lot since the SOC is enough. After that, EB will select charging decision even the electricity price is high because the current SOC may not satisfy the energy consumption to the next charging stations. For lasted three charging stations, the EB charge itself since the electricity price is low.

4.6 Summary

In this chapter, an SMDP based average reward RVIRL algorithm has been proposed to optimize the en-route charging schedule of EBs. At first, the energy consumption model

has been introduced to calculate the energy consumption of EB between any two adjacent charging stations. And then, the operation model of EB on the route has been developed to characterize the changes in the SOC level and the constraints of the battery manufacturer range. After the calculation of the energy consumption of EB, the battery degradation cost of the EB can be evaluated based on the corresponding SOC level during the operation process and its related normalized deviation for a whole charging and discharging process. Based on the system model, an SMDP is used to formulated the operation process of the EB, considering the number of the charging stations, the current SOC level, the maximum stop time in the charging station, and the electricity price for the charging decision. Then, the average reward RVIRL algorithm is utilized to obtain the charging policy. Compared with the ARRL algorithm mentioned in the Chapter 3, the key feature for the RVIRL algorithm is that the reference state-action pair is used to estimate the average reward rate in ARRL algorithm. For EB en-route charging schedule optimization, due to the uncertainty of stop time (T_n^{stop}), the average reward rate could be very large if the EB needs a large charging amount within a short stop time. In this way, the rate of convergence could be slow down. Finally, a case study based on the real-world data obtained from St. Albert Transit is conducted, and the performance of the RVIRL algorithm is compared with the existing FQI algorithm and greedy algorithm. According to the results from the simulations, the total cost of RVIRL is reduced by about 17% and 33% when compared with the FQI algorithm and greedy algorithm, respectively. Besides the comparison among different algorithms, the convergence of the RVIRL algorithm is theoretically proved and verified by simulations

5 Conclusion and Future Works

In recent years, EBs have attracted much attention from transportation system agents and industry due to the economic and environment-friendly futures. However, the battery sizes of the EBs can limit their operation range, and the high charging power compare to the electric vehicles may challenge the local power system. In order to reduce the impact of EBs on the local power system and extend the operation range, RL has been utilized to optimize the charging strategies of EBs. First, a double Q-learning is used to generate the charging amount at each charging station of EB. During the learning process, the historical energy consumption and battery degradation cost are considered. Then, an ARRL approach is presented to generate the en-route charging strategies. SMDP is used to model the operation process of EBs considering the randomness of the time duration between charging stations. Finally, an RVIRL approach is developed to obtain the charging strategy of EB during its operation process. In this approach, besides the energy consumption, the real-time electricity price at the charging station is considered in the proposed method. Differences among these three works are that the first work requires the EB only charge in the terminal station which is a common charging method for the public transit services with newly adopted EBs, while the second and third works are suitable for the EB which charges during the operation process. During the operating process, if the charging stations are locating in the exchange station and the EB follows the operation schedule under less congested traffic conditions, the second work is suitable for the optimization of charging schedule. If the charging stations are combined with bus shelters with an uncertain charging time, or the EB is operating under congested traffic conditions, the charging schedule can be optimized based on the third work.

5.1 Contributions of Thesis

The main contributions of this thesis can be summarized as follows:

- An optimal charging scheduling algorithm based on a model-free reinforcement learning algorithm is proposed for the minimization of EBs battery degradation cost. The battery degradation cost is analyzed by considering the factors of battery SOC and DOD. The formulation of MDP-based EBs charging scheduling problem is presented in details. The transition probability is investigated to address the randomness of EBs charging processes. Further, the double Q-learning which is based on reinforcement learning is introduced to optimize the charging schedule for EBs. Extensive simulation results based on real data collected from St. Albert Transit, AB, Canada are presented to evaluate the performances of our proposed double-Q learning based optimal EBs charging scheduling algorithm.

- For en-route EB charging schedule optimization, specific physical model and battery degradation model are built for the calculation of the energy consumption and the cost of EB operation, respectively. The SMDP is utilized to model the operation process of the EBs, and the ARRL approach is introduced to optimize the en-route charging schedule for the EBs. The optimized charging policy and its efficiency are demonstrated based on the real EB operation data provided by the St. Albert Transit, AB, Canada.

- An extension of ARRL is further investigated for the optimization of charging strategy. Besides the physical model for the energy consumption, real-time electricity prices at charging stations as well as the corresponding charging power are also considered for the SMDP of EB. After that, an RVIRL approach with the consideration of battery cost, electricity cost, and the impact on the power system is utilized to find the optimal charging schedule. And the convergence of the RVIRL approach is proved mathematically.

5.2 Directions for Future Work

In this thesis, the charging strategy of EB can be obtained by using the RL and its extensions. However, there are still some open issues that need to have future research, such as the charging strategy of multiple EBs and the energy management for the renewable energy sources. In particular, the following topics will be investigated in our future works:

- In order to obtain a more accurate estimation of energy consumption of EBs on the road, additional stochastic parameters need to be considered such as weather conditions, traffic conditions, and passenger trip profiles. Also, the accuracy of the energy consumption can be improved by involving the influences of the region, time,

weather, and season. According to the improved EB energy consumption model, the real-time charging load on the power system can be investigated;

- Based on the previous discussion, the RL method can be utilized to obtain the optimal charging strategy for the EB. However, multiple EBs operating simultaneously might affect the charging schedules among each other. To address this issue, a multi-agent RL can be developed to jointly optimize the charging schedules of multiple EBs.

- The en-route charging station with renewable energy sources such as solar and wind as well as energy storage devices can also be investigated for the optimization of the charging strategy for each EB. Accordingly, new RL approach needs to be developed by considering the energy remaining at the station, the current charging load for the EBs as well as the electricity price for different charging stations. The MDP or SMDP based models can also be utilized to predict the amount of solar and wind energy for the charging stations.

Bibliography

[1] *Facility Solar Panels/City of St.Albert* Accessed: 2021. [Online]. Available: https://stalbert.ca/city/transit/about/solar-panels/

[2] CAIT Climate Data Explorer. 2019. Country Greenhouse Gas Emissions. Washington, DC: World Resources Institute. Available online at: http://cait.wri.org

[3] P&S Market Research, "Global electric bus market size, share, development, growth and demand forecast to 2025," Industry Insights by Technology, May, 2017. [Online]. Available: https://www.psmarketresearch.com/marketanalysis/electric-bus-market. [Accessed: 25-Apr-2019].

[4] National Academies of Sciences, Engineering, and Medicine. 2018. Battery Electric Buses—State of the Practice. Washington, DC: The National Academies Press. https://doi.org/10.17226/25061.

[5] BYD eBus Tech. Spec., [Online]. Available: http://www.byd.com/la/auto/ebus.html.

[6] H. Gao, Y. Chen, S. Mei, S. Huang, and Y. Xu, "Resilience-oriented pre-hurricane resource allocation in distribution systems considering electric buses," *Proc. IEEE*, vol. 105, no. 7, pp. 1214–1233, Jul. 2017.

[7] A. Khaligh and Z. Li, "Battery, ultracapacitor, fuel cell, and hybrid energy storage systems for electric, hybrid electric, fuel cell, and plugin hybrid electric vehicles: State of the art," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 2806–2814, Jul. 2010.

[8] P. Elbert, T. Nüesch, A. Ritter, N. Murgovski, and L. Guzzella, "Engine on/off control for the energy management of a serial hybrid electric bus via convex optimization," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 3549–3559, Oct. 2014.

[9] Q. Hui and W. Xin, "Electric bus battery-swapping system based on robots," in *Proc. IEEE CECNet'12*, April 2012.

[10] W. Choi and J. Kim, "Electrification of public transportation: Battery swappable smart electric bus with battery swapping station," in *Proc. IEEE ITEC Asia-Parcific'14*, Aug.-Sept. 2014.

[11] R. E. Bellman, "A Markovian decision process," *J. Mathematical Mechanics*, vol. 6, no. 5, pp. 679–684, 1957.

[12] J. Janssen, *Semi-Markov Models: Theory and Applications*, New York:Springer, 1999.

[13] J. Janssen and R. Manca, *Applied Semi-Markov Processes*, New York:Springer, 2005.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[15] T. Ding, Z. Zeng, J. Bai, B. Qin, Y. Yang and M. Shahidehpour, "Optimal Electric Vehicle Charging Strategy With Markov Decision Process and Reinforcement Learning Technique," *IEEE Trans. Ind Appl.*, vol. 56, no. 5, pp. 5811-5823, Sept.-Oct. 2020.

[16] Z. Moghaddam, I. Ahmad, D. Habibi and Q. V. Phung, "Smart Charging Strategy for Electric Vehicle Charging Stations," *IEEE Trans. Transport. Electrific.*, vol. 4, no. 1, pp. 76-88, March 2018.

[17] M. H. Mobarak and J. Bauman, "Vehicle-Directed Smart Charging Strategies to Mitigate the Effect of Long-Range EV Charging on Distribution Transformer Aging," *IEEE Trans. Transport. Electrific.*, vol. 5, no. 4, pp. 1097-1111, Dec. 2019.

[18] J. Ahn and B. K. Lee, "High-Efficiency Adaptive-Current Charging Strategy for Electric Vehicles Considering Variation of Internal Resistance of Lithium-Ion Battery," *IEEE Trans. Power Electron.*, vol. 34, no. 4, pp. 3041-3052, April 2019.

[19] H. Wu, G. K. H. Pang, K. L. Choy and H. Y. Lam, "An Optimization Model for Electric Vehicle Battery Charging at a Battery Swapping Station," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 881-895, Feb. 2018.

[20] A. E. Trippe, R. Arunachala, T. Massier, A. Jossen and T. Hamacher, "Charging optimization of battery electric vehicles including cycle battery aging," *Proc. IEEE ISGT'14*, Feb. 2014.

[21] N. Sadeghianpourhamami, J. Deleu and C. Develder, "Definition and Evaluation of Model-Free Coordination of Electrical Vehicle Charging With Reinforcement Learning," *IEEE Trans. Smart Grid*, vol. 11, no. 1, pp. 203-214, Jan. 2020.

[22] Z. Wei, Y. Li and L. Cai, "Electric Vehicle Charging Scheme for a Park-and-Charge System Considering Battery Degradation Costs," *IEEE Trans. Intell. Veh.*, vol. 3, no. 3, pp. 361-373, Sept. 2018.

[23] Y. Wi, J. Lee and S. Joo, "Electric vehicle charging method for smart homes/buildings with a photovoltaic system," *IEEE Trans. Consum. Electron.*, vol. 59, no. 2, pp. 323-328, May 2013.

[24] Q. Dai, T. Cai, S. Duan and F. Zhao, "Stochastic Modeling and Forecasting of Load Demand for Electric Bus Battery-Swap Station," *IEEE Trans. Power Deliv.*, vol. 29, no. 4, pp. 1909-1917, Aug. 2014.

[25] D. Lemon, P. Griffith, Z. Coffman and G. Gleason, "Electric-bus fast charging at the Santa Barbara MTD," in *Proc. IEEE Cat. No.99TH8371'99*, Jan. 1999.

[26] Xiaomin Lu, K. Lakshmi Varaha Iyer, K. Mukherjee and N. C. Kar, "Development of a bi-directional off-board level-3 quick charging station for electric bus," in *Proc. IEEE ITEC'12*, June 2012.

[27] G. Zhou, D. Xie, X. Zhao and C. Lu, "Collaborative Optimization of Vehicle and Charging Scheduling for a Bus Fleet Mixed With Electric and Traditional Buses," *IEEE Access*, vol. 8, pp. 8056-8072, 2020.

[28] R. Tavakoli, A. Jovicic, N. Chandrappa, R. Bohm and Z. Pantic, "Design of a dual-loop controller for in-motion wireless charging of an electric bus," in *Proc. ECCE'16*, 2016.

[29] Y. J. Jang, E. S. Suh, and J. W. Kim, "System architecture and mathematical models of electric transit bus system utilizing wireless power transfer technology," *IEEE Syst. J.*, vol. 10, no. 2, pp. 495-506, June 2016.

[30] X. Wang, C. Yuen, N. U. Hassan, N. An, and W. Wu, "Electric vehicle charging station placement for urban public bus systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 128-139, May 2017.

[31] T. Shijo et al., "85 kHz band 44 kW wireless power transfer system for rapid contactless charging of electric bus," in *Proc. ISAP'26*, Oct. 2016.

[32] G. Jung et al., "Wireless charging system for On-Line Electric Bus(OLEB) with series-connected road-embedded segment," in *Proc. IEEE EEEIC'13*, May 2013.

[33] Y. Kim et al., "Phase analysis of currents and voltages in the On-Line Electric Bus (OLEB) system," in *Proc. IEEE EEEIC'13*, May 2013.

[34] C. Yang,W. Lou, J. Yao, and S. Xie, "On charging scheduling optimization for a wirelessly charged electric bus system," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 6, pp. 1814-1826. June 2018.

[35] K. Smith, T. Markel, G.-H. Kim, and A. Pesaran, "Design of electric drive vehicles for long life and low cost," in *Proc. IEEE ASTR'10*, Oct. 2010.

[36] A. Hoke, A. Brissette, D. Maksimović, A. Pratt, and K. Smith, "Electric vehicle charge optimization including effects of lithium-ion battery degradation," in *Proc. IEEE VPPC'11*, Oct. 2011.

[37] J. C. Hall, A. Schoen, A. Powers, P. Liu, and K Kirby, "Resistance growth in Lithium Ion satellite cells," in *Proc. 208th ECS Meeting*, Oct. 2005.

[38] T. Markel, K. Smith, and A. Pesaran, "Improving petroleum displacement potential of PHEVS using enhanced charging scenarios," in *Proc. WEVA EVS'09*, May 2009.

[39] C. Rosenkranz, "Deep-cycle batteries for plug-in hybrid application," in *Proc. WEVA EVS'03*, Nov. 2003.

[40] V. Marano, S. Oroni, Y Guezennec, G. Rizzoni, and N. Madella, "Lithium-ion batteries life estimation for plug-in hybrid electric vehicles," in *Proc. IEEE VPPC'09*, Sept. 2009.

[41] H. van Hasselt, "Double Q-learning," in *Proc. NIPS'10*, Dec. 2010.

[42] M. Pouyan, A. Mousavi, S. Golzari and A. Hatam, "Improving the performance of Q-learning using simultanouse Q-values updating," in *Proc. IEEE ICTCK'14*, Nov. 2014.

[43] M. T. Sebastiani, R. Luders, and K. V. O. Fonseca, "Evaluating electric bus operation for a real-world BRT public transportation using simulation optimization," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2777-2786, Oct. 2016.

[44] J. Temple, "The $2.5 trillion reason we can't rely on batteries to clean up the grid," MIT Technology Review, July 2016. [Online]. Available: https://www.technologyreview.com/s/611683/the-25-trillion-reason-we-cant-rely-on-batteries-to-clean-up-the-grid/. [Accessed: 25-Apr-2019].

[45] Y. Liu and H. Liang, "An MHO approach for electric bus charging scheme optimization based on energy consumption estimation," in *Proc. IEEE PES GM'18*, Aug. 2018.

[46] X. Xu, W. Liu, X. Zhou, and T. Zhao, "Short-term load forecasting for the electric bus station based on GRA-DE-SVR," in *Proc. IEEE ISGT ASIA'14*, May 2014.

[47] C. Szepesvari, "Algorithms for reinforcement learning," in *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Morgan & Claypool Publishers, 2009.

[48] A. Millner, "Modeling Lithium Ion battery degradation in electric vehicles," in *Proc. IEEE CITRES'10*, Sept. 2010.

[49] Xi-Ren Cao, "Semi-Markov decision problems and performance sensitivity analysis," *IEEE Trans. Autom. Control*, vol. 48, no. 5, pp. 758-769, May 2003.

[50] T. K. Das, A. Gosavi, S. Mahadevan, and N. Marchalleck, "Solving Semi-Markov Decision Problems Using Average Reward Reinforcement Learning," *Manage. Sci.*, vol. 45, no. 4, pp. 560–574, 1999.

[51] C. Darken, J. Chang and J. Moody, "Learning rate schedules for faster stochastic gradient search," in *IEEE Neural Networks for Signal Processing'92*, Aug.-Sept. 1992.

[52] Zhurkov, S.N. "Kinetic concept of the strength of solids", *Int J Fract 26*, Dec.1984.

[53] G. Ning and B. N. Popov, "Cycle life modeling of lithium-ion batteries," *J. Electrochem. Soc.*, vol. 151, no. 10, pp. A1584–A1591, 2004.

[54] R. Howard, "Semi-Markovian decision processes," *Bull. Inst. Intcernat.Statist.*, vol. 40, pp. 625–652, 1963.

[55] B. Hu, S. Feng, J. Li, and H. Zhao, "Statistical analysis of passenger crowding in bus transport network of Harbin," *Physica A*, vol. 490, pp.426-438, Jun. 2018.

[56] Yanjie Li and Fang Cao, "RVI reinforcement learning for semi-Markov decision processes with average reward," in *Proc. WCICA'10*, July 2010.

[57] X. R. Cao, *Stochastic Learning and Optimization: A Sensitivity-Based Approach.* New York: Springer, 2007.

[58] V. S. Borkar and S. P. Meyn, "The ode method for convergence of stochastic approximation and reinforcement learning," *SIAM J. Control Optim.*, vol. 38, no. 2, pp. 447–469, 2000.

[59] D. P. Bertsekas, J. Abounadi, and V. Borkar, "Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms," *SIAM J. Control Optim.*, vol. 41, no. 1, pp. 1–22, 2003.