TRANSIT SIGNAL PRIORITY CONTROL USING REINFORCEMENT LEARNING BASED ON CELL TRANSMISSION MODEL

Mr. Pitipong Chanloha

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2012

การควบคุมสัญญาณที่มีลำดับความสำคัญของการเดินทางผ่านโดยใช้การเรียนรู้แบบเสริมแรงบนพื้นฐานของแบบจำลองการส่งผ่านเซลล์

นายปิติพงศ์ ชาญโลหะ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2555

Thesis Title	TRANSIT SIGNAL PRIORITY CONTROL USING REINFORCEMENT LEARNING BASED ON CELL TRANSMISSION MODEL
By	Mr. Pitipong Chanloha
Field of Study	Electrical Engineering
Thesis Advisor	Assistant Professor Chaodit Aswakul, Ph.D.
Thesis Co-advisor	Jatuporn Chinrungrueng, Ph.D.
Thesis Co-advisor	Assistant Professor Wipawee Hattagam, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

. . . . . Dean of the Faculty of Engineering
(Associate Professor Boonsom Lerdhirunwong, Dr.Ing.)

THESIS COMMITTEE

. . . . . Chairman
(Assistant Professor Tuptim Angkaew, D.Eng.)

. . . . . Thesis Advisor
(Assistant Professor Chaodit Aswakul, Ph.D.)

. . . . . Thesis Co-advisor
(Jatuporn Chinrungrueng, Ph.D.)

. . . . . Thesis Co-advisor
(Assistant Professor Wipawee Hattagam, Ph.D.)

. . . . . Examiner
(Assistant Professor Chaiyachet Saivichit, Ph.D.)

. . . . . Examiner
(Associate Professor Sorawit Narupiti, Ph.D.)

. . . . . External Examiner
(Associate Professor Poompat Saengudomlert, Ph.D.)

ปิติพงศ์ ชาญโลหะ : การควบคุมสัญญาณที่มีลำดับความสำคัญของการเดินทางผ่าน โดยใช้การเรียนรู้แบบเสริมแรงบนพื้นฐานของแบบจำลองการส่งผ่านเซลล์ (TRANSIT SIGNAL PRIORITY CONTROL USING REINFORCEMENT LEARNING BASED ON CELL TRANSMISSION MODEL)
อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร.เชาวน์ดิศ อัศวกุล
อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ดร.จตุพร ชินรุ่งเรือง
อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ผศ.ดร.วิภาวี หัตถกรรม, 97 หน้า.

วิทยานิพนธ์ฉบับนี้มีจุดมุ่งหมายเพื่อพัฒนากรอบการวิเคราะห์ใหม่ในการควบคุมสัญญาณไฟสำหรับโครงข่ายถนนที่มีรถโดยสารประจำทางด่วนพิเศษ โดยใช้การเรียนรู้อัตโนมัติแบบเป้าหมายกำกับและกระทำการตัดสินใจที่เรียกว่า การเรียนรู้แบบเสริมแรง (Reinforcement learning : RL) โดยสัญญาณไฟจราจรที่ดีที่สุดที่เป็นไปได้สามารถหาได้ขึ้นกับการเปลี่ยนแปลงของสถานะโครงข่ายที่ถูกจำลองด้วยแบบจำลองการส่งผ่านเซลล์ที่มีสัญญาณไฟ (Cell transmission model : CTM) โดยมีผลงานสามส่วนหลักที่ปรากฏอยู่ในวิทยานิพนธ์ฉบับนี้

ประการแรก การรวมเข้าของแบบจำลอง CTM เพื่อจับระบบกลศาสตร์ร่วมกับการนำ RL ไปใช้เรียกว่า การเรียนรู้แบบคิวเพื่อหาคำตอบที่ดีที่สุดที่เป็นไปได้สำหรับสี่แยกเดี่ยว แม้ว่าจะเป็นการพิจารณาเพียงแยกเดียวแต่ก็ได้มีการนำเสนอฟังก์ชันการประวิงเวลาของระบบที่เงื่อนไขขอบเขตเพื่อจับผลประสิทธิผลจากแยกข้างเคียง ด้วยการเลือกการประวิงเวลาสัญญาณไฟแดงเป็นฟังก์ชันผลรางวัล ผลการทดลองแสดงให้เห็นว่ากรอบการวิเคราะห์ที่นำเสนอ RL และ CTM ในระดับมหภาคสามารถหาคำตอบของการควบคุมได้อย่างมีประสิทธิภาพใกล้เคียงกับการเอาแต่แรงด้วยการค้นหาคำตอบแบบสัญญาณคาบที่ดีที่สุด

ประการที่สอง การเปรียบเทียบสมรรถนะของการควบคุมสัญญาณไฟจราจรที่ดีที่สุดบนพื้นฐานขอแบบจำลองอนุพันธ์เชิงทฤษฎีระบบแถวคอย M/M/1 และ D/D/1 และถูกประเมินบนพื้นฐานของการเข้าสู่ RL โดยการแบ่งที่ดีที่สุดสามารถอนุพันธ์เพื่อทำให้ค่าเฉลี่ยของการรอที่แยกเดี่ยวที่มีสองความขัดแย้งของการไหลลดลง ผลที่ได้มายืนยันมีผลใช้ได้ใน RL เพื่อคุมสัญญาณไฟ

ประการสุดท้าย การยืดออกไปเป็นโครงข่ายอันตรกิริยาแบบต่อเรียงกับระบบสัญญาณที่มีลำดับความสำคัญได้ถูกนำเสนอควบคู่ไปกับการไกลแบบทิศทางเดียวอย่างง่ายที่ไม่มีการเลี้ยวกลับ ด้วยระบบ BRT ในกรุงเทพมหานครแบบจำลองการส่งผ่านเซลล์ที่มีสัญญาณไฟแบบทั่วไปสามารถทำให้ใช้ได้กับการวางแผนพื้นที่การใช้งานล่วงหน้าของ BRT ซึ่งเหนือกว่ารถที่ไม่มีลำดับความสำคัญ ด้วยการจำลองอย่างชัดแจ้งของการมีอยู่ของ BRT ที่ถูกแยกช่องกายภาพ เช่นเดียวกับ ตำแหน่งของสถานี BRT โดยนำเสนอฟังก์ชันประวิงเวลาของทั้งผู้โดยสารที่ถูกขนส่งไปได้บน BRT และบนรถที่ไม่มีลำดับความสำคัญ และ ผู้โดยสารที่รอที่สถานี ด้วยพื้นฐานของแบบการที่จำลองสืบค้นอยู่ การวางระบบ BRT ที่ลดลง หนึ่งช่องทางด้วยตัวแบ่งช่องทาง ไม่สามารถลดการประวิงเวลาโดยสารได้เมื่อได้ลองเทียบกับถนนที่มีเงื่อนไขจราจรก่อนจะมีการวางระบบ BRT อย่างไรก็ตาม ระบบ BRT สามารถเพิ่มปริมาณงานได้ถึง 9-15% ในเงื่อนไขการติดเมื่ออย่างน้อยที่สุด 40 % ของผู้โดยสารทั้งหมด เลือกใช้ BRT ยิ่งไปกว่านั้น ระเบียบวิธีที่ที่นำเสนอให้ผลดีกว่าวิธีระเบียบวิธีการจองแบบสามัญและระเบียบวิธีควบคุมผลต่างเชิงอนุพันธ์ที่มีลำดับความสำคัญเพราะความเพิ่มขึ้นของการตระหนักของราคาการเปลี่ยนแปลงของระดับสัญญาณ จากทุกการค้นพบทั้งหมดผลงานนี้มีคุณค่าที่จะสร้างผลงานเพื่อการพัฒนาระบบขนส่งมวลชนที่ยั่งยืนต่อไป

ลายมือชื่อนิสิต .....

ภาควิชา วิศวกรรมไฟฟ้า ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

สาขาวิชา วิศวกรรมไฟฟ้า ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

ปีการศึกษา 2555 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

##5071870721: MAJOR ELECTRICAL ENGINEERING
KEY WORDS: CELL TRANSMISSION MODEL / BUS RAPID TRANSIT / Q-LEARNING
PITIPONG CHANLOHA : TRANSIT SIGNAL PRIORITY CONTROL USING REINFORCEMENT LEARNING BASED ON CELL TRANSMISSION MODEL
THESIS ADVISOR : ASST. PROF. CHAODIT ASWAKUL, Ph.D.
THESIS CO-ADVISOR : JATUPORN CHINRUNGRUENG, Ph.D.
THESIS CO-ADVISOR : ASST. PROF. WIPAWEE HATTAGAM, Ph.D., 96 pp.

This dissertation is aimed at developing a new framework to control traffic signal light for the road network with recently introduced bus rapid transit (BRT) system. By applying the automated goal-directed learning and decision making called reinforcement learning (RL), the best possible traffic signal actions can be sought upon changes of network states as modelled by the signalised cell transmission model (CTM). There are three main original contributions in this dissertation.

Firstly, the model combining CTM to capture the system dynamics together with the implementation of RL approach called Q learning has been introduced for an isolated intersection. Despite of such isolation constraint, a new external delay function has been proposed at the system boundary condition to capture the effects on the neighbourhood of that isolated intersection system. With the proper setting of red light delay as the RL reward function, reported results show that our proposed framework using RL and CTM in the macroscopic level can efficiently find the proper control solution that is close to the brute-forcely searched best periodic signal solution (BPSS).

Secondly, the performance comparison of optimal traffic signal controls based on the derivation of theoretical M/M/1 and D/D/1 models and based on the RL approach has been evaluated. In particular, based on M/M/1 and D/D/1 queuing, the optimal split has been derived to minimise the mean waiting time of an intersection with two conflicting flows. The results confirm the validity in adopting the RL approach to control the traffic signal.

Finally, an extension to a network of cascading interactions with BRT system has been proposed with simple uni-directional flows without turning movements. Motivated by the BRT system in Bangkok, the conventional signalised CTM has been generalised to cope with the preplanned space-usage priority of BRT over other non-priority vehicles by modelling explicitly the existence of BRT physical lane separator as well as the location of BRT stations. The delay function of both carried passengers on BRT and on other non-priority vehicles as well as waiting passengers at stations has been introduced. Based on the investigated scenarios, the deployment of BRT system with one lane deducted by the lane separator cannot reduce the total passenger delay in comparison with the comparable road and traffic condition before the BRT installation. However, with BRT, the passenger throughput can be greatly increased by up to 9-15% in the jamming conditions when at least 40% from the overall passengers choose the BRT for their journey. Moreover, our proposed method outperforms the conventional preemptive and differential priority control methods because of the improved awareness of signal switching cost. Based on all findings, the outstanding merit will entirely contribute towards to support the development of sustainable transportation systems.

	Student's signature .....
Department: Electrical Engineering	Advisor's signature .....
Field of Study: Electrical Engineering	Co-advisor's signature .....
Academic Year: 2012	Co-advisor's signature .....

Acknowledgements

I would like to express my sincere and grateful thanks to my advisor Asst. Prof. Dr. Chaodit Aswakul for his continued guidance, help, understanding and patience throughout the long journey of my doctoral degree course. I am also extremely indebted to his invaluable supports and outstanding comments on various aspects of my work. Furthermore, I am greatly thankful to my co-advisors, Dr. Jatuporn Chinrungrueng and Asst. Prof. Dr. Wipawee Hattagam, for their valuable suggestions, corrections and fruitful guidance along the way. Particularly, Asst. Prof. Dr. Wipawee Hattagam, I am also indebted to her for picking me up as her own Masters' student and this chance totally changed my life to be a great scholar. I would like to express my warmly thanks to Asst. Prof. Dr. Chaiyachet Saivichit for his valuable advice, constructive criticism and especially his friendship during my graduate study. It is an honor to have Asst. Prof. Dr. Tuptim Angkaew as a chair person of my thesis committee. I take this opportunity to express my sincere acknowledgement to Assoc. Prof. Dr. Sorawit Narupiti. His mentorship was substantial in providing me real experiences in the current development and implementation in the transportation nowadays. My warmest gratitude thanks to Assoc. Prof. Dr. Poompat Saengudomlert for his helpful suggestions and comments.

I am grateful to the Thailand Graduate Institute of Science and Technology (TGIST), associated with National Science and Technology Development Agency (NSTDA) for proving a scholarship, which boosted me to perform my work comfortably. I also would like to thank the 90^{th} anniversary of Chulalongkorn University fund, TRIDI, SP2 GE12 Project of the Department of Electrical Engineering, Faculty of Engineering, Chulalongkorn University for the research financial support and computing facilities.

My pronounced pleasure to give the honorary to my colleagues for sharing many interesting academic discussions, stimulations and friendship filled environment in the research laboratory. Specifically, thanks to my doctoral student colleagues, Dr.Patrachart, Dr.Kalika, Dr.My, Dr.Norrarat, Tri, Rachata and Kittisak, who always give me many insightful comments. Furthermore, my unique thanks go to the Masters' degree friends, EEPSA committee, Narisara, Adisorn, Sirawit, Malinda, Nuttida, Suwatchai, the students of the network research group, and many others that I will never forget.

I would like to pay my deepest gratitude to my parents for their priceless encouragement and precious inspiration throughout my doctoral work and in helping me to be proud of as a great scholar. This dissertation would not have been possible without the invaluable help and support from countless number of people over the past five years. These unforgettable memories have been wholeheartedly proudly shared among those people.

Contents

Chapter Page

Chapter Page

List of Figures

Figure Page

List of Tables

CHAPTER I

INTRODUCTION

Nowadays, the opportunities for further extension of physical transportation capacity within a well-established city are becoming very limited. With the continuing social, population and economic growth in metropolitan areas, many transportation facilities are being used to their full capabilities. Several strategies for traffic relief are implemented such as increasing road network capacity, demand management, and traffic control and operation. Adding new facilities are becoming difficult due to the high costs and limited spaces. Moreover, the user demand for a transportation system outgrows the system's capacity and the performance of the system degrades unavoidably. In addition, demand management is intended to balance requested traffic into the networks which have been implemented in countries. Effects of demand management using innovative local policies to match the unique demand nature of locality still remain much to be explored. To that respect, attempts to operate and control the traffic by employing the existing capacity without adding new facilities are absolutely challenging. Fortunately, the growing emphasis on information systems and communication technologies can handle the traffic problem by using advanced traffic information and control systems. The systematic application for advanced technologies to the surface transportation system has become known as Intelligent Transportation Systems (ITS).

Area traffic control (ATC) is one of the major areas in which ITS can be applied. At the local level, traffic signals are designed to manage vehicle conflicts at intersections by allocating green time among the conflicting traffic streams which must share the intersection. However, at the global level, traffic signals can be controlled from centralised servers to enhance signal control strategies to increase the throughput efficiency of road network.

Generally, traffic signal controls can be classified into three main approaches, namely, fixed-time control, actuated control and adaptive control [1]. Fixed-time traffic signals operate fixed signal timing plan regardless of the traffic demands. Actuated control employs vehicle detectors installed around an intersection to change the traffic signals of that intersection. Once vehicle detectors response for actuation, the actuated phase normally starts

with a minimal preset green time, and green time phase is automatically extended. Actuated controls can be terminated in three approaches which are gap-out, max-out and force-off [1]. Gap-out mode occurs when there is no more traffic in the corresponding approach. Max-out mode occurs when the excessive demand influences a signal phase to reach its maximum green time. Force-off mode is usually triggered when a specific approach needs to achieve an extension green band. However, the sophistication and smart idea for traffic control system is to optimise traffic networks online without being confined by a cyclic time interval or fixed time control. Adaptive signal control differs from actuated control because it incorporates decision making processes in the design of signal timing procedures. The main idea of adaptive signal control is to predict the traffic flow demands, evaluate the set of possible signal control strategies and choose the optimal feasible signal strategies with respect to current objectives.

1.1 Literature Review

Since a basic traffic signal control has been already introduced, signal control strategies with theoretical and analytical control algorithms have been proposed by researchers, e.g. Webster's model in 1958 [2], Pontryagin model in 1964 [3], discrete minimal delay model in 2000 [4]. In light of conventional theoretical models, several literatures exist and present the effective control strategies for road scenarios. For example, SCOOT (Split, Cycle and Offsets Optimization Techniques) [5] is a method that relies on cyclic timing plan and signal timing plan of SCOOT being changed periodically. There are three SCOOT parameters that can be used to influence the traffic condition, i.e., split (a green time proportion for each phase in each cycle time), cycle length (the total time for signal completion for one sequence of the signal indications) and offset (phase difference of start of green between adjacent intersections) [1].

To the aforementioned, traffic signal control becomes significantly challenging. There have been many methods reported in the literatures for controlling traffic signal at an intersection. More specifically, traffic signal control can be categorised into two main techniques being fixed-time or traffic-responsive [6]. Fixed-time control employs the historical data to estimate the signal light structures in advance. Therefore, fixed-timed control effectiveness can be worsen when the traffic is unpredictably congested. On the contrary, with the advance information systems and communication technologies, traffic-responsive control can

be used to adjust its indicated green lights according to the current observable traffic flows. The observed traffic flow can be measured directly from the sensors embedded in the road network, which results in the adjustable traffic signal control in real-time. The existing implementations for traffic signal controls are being shown as follows. Actuated control uses the demand-driven logic to control the signal timing by the installed detectors on the intersection. The cycle length and green time of actuated control may vary from cycle to cycle to response the approaching demands. All the signal phases are controlled by the detectors. Each phase can be skipped automatically if there is no present demand.

The Sydney Coordinated Adaptive Traffic System (SCATS) [7] is developed in Australia by the Road and Traffic Authority (RTA). The SCATS uses the information from vehicle detectors. The detectors are located in each lane in advance before the stop-line to adjust the signal lights. The calculation for the split is relatively proportioned to the approaching demand which is measured in terms of degree of saturation (DoS). For the signal timings, it can be automatically adjusted every cycle to aware the excessive traffic delay caused by the huge amount of demands.

In the past few decades, the management of public transportation system becomes a major concern. The application for public transportation system is generally known as Advanced Public Transportation Systems (APTS) [8]. The main objective of APTS is to improve the efficiency without a need for major infrastructure enhancements, e.g. Bus Rapid Transit (BRT). The basic definition of BRT is a flexible, high performance rapid transit mode that combines a variety of physical, operating and system elements into a permanently integrated system with a quality image and unique identity [9]. Due to the BRT flexibility, it encompasses a wide variety of applications, each tailored to a particular set of travel markets and physical environments. This flexibility is resulted from the fact that BRT vehicles, e.g. buses or specialised BRT vehicles, can travel anywhere on the pavement and the fact that a BRT basic service unit is a relatively small vehicle in comparison with rail and train based rapid transit modes. BRT applications can combine various route segments to provide a single-seat, no-transfer service that maximises customer convenience. Unlike other rapid transit modes where basic route alignment and station locations are constrained by the available right of way, BRT can be tailored to the unique origin and destination patterns of a given corridor's travel needs. As the spatial nature of transit demand changes, BRT systems can therefore adapt to these dynamic conditions. The BRT service includes several ITS

components including Automated Vehicle Location (AVL) technology, transit signal priority systems, onboard voice and digital announcements of next stop information, and real time bus arrival time information using digital countdown signs at bus stops.

AVL technology employs onboard equipment, e.g. Global Positioning System (GPS) and General Packet Radio Service (GPRS), to track a bus location. The mechanism of AVL relies on an onboard computer to calculate the location of each bus and sends data to a central control center where each bus's location is mapped on a computer screen using flag-shaped icons. Green flags represent buses ahead of schedule and red flags represent buses behind schedule. Transit management personnel monitors the bus progresses on each route and advises operators of schedule deviations. Mobile Data Terminal (MDT) facilitates voice and data communications between drivers and the transit control center and allows bus drivers to select unique messages to communicate with passengers concerning route or schedule adjustments. In addition, if a bus is behind its schedule, then the AVL system automatically links to an integrated automated signal control system and requests green signal extensions, or advanced green signals at intersections pre-authorised to provide conditional priority in predefined areas. AVL is also installed on two supervisor cars, to allow the transit control center to determine the location of supervisors relative to any buses that may need assistance [10]. This technology can reduce communication costs but requires greater intelligence [11]. However, the challenge for APTS is to reduce the congestion delays when exclusive bus lane is provided in urban areas with the limitation of spaces. Another method is to use signal control strategies and employ the current traffic signal control system to give priority for the transit vehicles such as green band extension and recall green band. This is generally known as Transit Signal Priority (TSP) [11].

Transit signal priority can be characterised into two types, namely, active and passive TSP. Passive priority is based on predefined signal timing for each intersection which is weighted and pre-optimised [11]. On the other hand, active priority employs dynamic detection and responses to transit vehicles by altering signal settings in real time to reduce the transit delays. Thus, the advantage of passive priority is a relatively low implementation cost. On the contrary, active priority needs extra hardware investment e.g. specialised detectors for transit vehicles and advanced signal controllers. Transit signal priority techniques and applications at the traffic signal in Europe have been developed and evaluated in PRISICILLA project [12].

Nowadays, transit signal priority system has been implemented in both developed and developing countries. For the transit priority system in London, priority requirements are determined at the AVL centre and transmitted to each bus through the normal polling cycle. This request is then transmitted from each bus to the downstream traffic signals via a roadside beacon, with the priority controlled by the traffic control system e.g. SCOOT [11]. SCOOT system gives priority to a bus by using predefined degree of saturation for each intersection to avoid the excessive delay to other traffics. Note that SCOOT estimates the degree of saturation from the measured ratio of average flow to maximum flows [13]. Bus priority is considered only for buses which have been delayed from bus schedule by giving corresponding extensions or recalls of green light. When a bus arrives at the end of green phase, the current green is extended with respect to the bus schedule length to allow a bus pass the stop line if the degree of saturation does not fall below the threshold limit. Likewise, when a bus arrives during a red phase, green light is recalled if the degree of saturation does fall below a specific threshold for that intersection. In addition, the highest level of communication between AVL center and ATC server is implemented in Genoa, Italy [12]. Transit signal priority system architectures according to the location of intelligence can be characterised into four types as follows [11].

- Fully centralised - traffic control and priority function operates at a central server.
- Centralised ATC and decentralised priority - the benefits from bus priority might be adversely affected by data transmission delays if the centralised priority is implemented.
- Decentralised ATC and centralised priority - wide area priority requirements take precedence over local control.
- Fully decentralised architecture - traffic control and priority operate at a local level.

In AVL system, the sustainability and improvement of public transportation are key components of area transport policy. In particular, traffic and junction delays affect the speed and reliability of bus service. AVL system lessons in Europe can be summarised as follows [11].

- Bus priority at a traffic signal is implemented with existing technologies and worthwhile benefits can be obtained.

- AVL technologies are providing new opportunities for more sophisticated form of priority, particularly differential priority [14], where bus priority levels can be awarded according to real-time need.

- GPS is becoming the preferred bus location technology; there is a diverse range of overall system architectures for delivering AVL-base bus priority.

- In-house expertise is required if these increasingly complex systems are to be managed, operated and maintained to the best effect.

The BRT system includes ITS components with AVL system together. Several researches on BRT priority (transit signal priority) and control strategies have been reported in the literatures. For example, manual strategies for traffic signal control conducted by Wilbur [15], followed by the unconditional strategy [16], conditional strategy [17] and adaptive conditional strategy [18]. Unconditional strategy grants the green extensions or recalls whenever a bus approaches at an intersection whereas the conditional strategy gives the preemptive priority whenever a bus approaches an intersection together with two constraints, i.e., the time limit for green extension and minimum elapsed time after the end of the priority period. An adaptive priority modifies signal timing plans corresponding to a performance index defined as the weighted sum of vehicle delays, bus schedule delays or delays represented by automobiles and bus passengers. In addition, the objective of transit signal priority has been changed from decreasing bus delay to enhancing the reliability of bus priority services by keeping buses on schedules [18]. Enhancement of travel time estimation becomes the objective of transit signal priority in recent research [19]. However, one problem in transit signal priority system is the signal delay variations of sequential buses. The bus signal control system is one of the sources to induce bus headway fluctuation and schedule deviation. When each bus arrives at the intersection and delay for each bus is different, the different signal status has been developed [20]. The optimal solution can be found by analyzing the effectiveness of bus average delay and deviation of the headway. For the arrival time estimation, the algorithm that uses a historical and real time vehicle local information for predicting the arrival time for the next traffic light has been proposed [21]. A few literatures have been investigated the transit signal priority various aspects e.g., arrival time estimation, green time extension [22].

Moreover, several existing literatures also have investigated for transit signal priority

at isolated intersections [23]. In [24], four architectures (preprocessor, mid-scope generator, real-time generator and final mediation) have been proposed to improve the performance of the transit priority system. However, this type of solution requires a centralised controller to calculate the optimal flows for all intersections. This raises the cost of installation and the communication mechanisms. For the centralised control, it requires an enormous amount of computation and becomes the single point of failure [25]. To alleviate the problem caused by a centralised control, an alternative method which is based on experiences gained from interacting directly with the environment has been proposed.

In the existing literatures, researchers employ the machine learning to solve for traffic signal control in a more scalable road network. For example, D. Teodorovic, V. Varadarajan, J. Popovic, M. R. Chinnaswamy and S. Ramaraj [26] use adaptive neural network (ANN) to predict the arrival of the traffic patterns. The work by Y. S. Hong, J. S. Kim, J. K. Son, and C. K. Park [27] and M. C. Choy, D. Srinivasan, and R. L. Cheu [28] has investigated the coordinated intersections for traffic signal control using neural network in a normal traffic. The work by M. Ghanim, F. Dion and G. A. Lebdeh [29] has investigated an optimisation for the transit signal priority with coordinated intersection at a microscopic level. For each arrival pattern, dynamic programming (DP) is employed to seek for the optimal splits for each direction which results in optimal performance. However, the limitation to this method is the computational burden caused by the determination of the arrival traffic patterns manually. To alleviate the problem caused by DP method, Cai [30] has proposed an adaptive dynamic programming (ADP) to avoid the curse of dimensionality (enormous amount of computation) and curse of modelling (incomplete information of the state transition). The proposed ADP overcomes both major classical problems in DP with an advanced information predicted from the future. B. G. Heydecker, C. Cai and C. K. Wong [31] has used ADP to seek for the optimal splits with complicate vehicle movement and intersection layout. However, all the reported literatures required advance traffic information patterns which may not be accurate.

A flexible approach is to learn good traffic signal control from experiences gained gradually by interacting directly with the environment. The approach, referred as reinforcement learning (RL), is a class of machine learning related to the artificial intelligence [32]. RL is a class of unsupervised learning that has potential to deal with traffic engineering problems is first proposed in [33]. Due to the characteristic of reinforcement learning, this method has

the capabilities of optimising at a local level of the network by the online learning process. Moreover, it is adaptable and does not rely on offline processing of the enormous amount of data. Reinforcement learning in traffic engineering problems has been reported in several literatures. The work by Silva in [25] has proposed reinforcement learning to control traffic lights signal at single intersection, the following work by using microscopic simulation for more realistic scenario is considered [34]. In a work by C. Jacob and B. Abdulhai [35] have addressed Q-learning which is a RL tool to deal with the highway traffic problems. For an isolated intersection control, [36], [37], [38], [39] have considered Q-learning to control with different objective functions, whereas [40] has investigated the green splits weighted by employing RL in order to minimise the number of vehicles in the system. However, the use of model proposed in [40] does not take into account a realistic situation like the change of traffic scenario and the presence of vehicles in upstream junctions. Although the existing approaches aim to find the best possible control for road traffic signal, those RL approaches have considered the individual movement of the vehicles in the microscopic level. Therefore, the computational burden becomes demanding. From the observations, an oversaturated traffic condition has not yet been considered. In the specific area e.g., Bangkok traffic becomes a major concern-particularly in terms of oversaturated conditions.

To that respect, this dissertation incorporates RL with the traffic flow behavior in the system. A basis of a well-established traffic flow model, namely, cell transmission model (CTM) is employed [41]. It should be noted that the origination of CTM has been first proposed and represented for the vehicles movements in an expressway. The enlighten work in [41] have inspired the following researchers in adopting CTM in various aspects for the uninterrupted traffic flows. A work by A. Sadek and N. Basha [42] have proposed Q-learning, which is one of the RL tools, for a traffic route guidance problem and uses a simple macroscopic model CTM to simulate the traffic flow dynamics of the system. The development of CTM to a signalised version has been first proposed by H. K. Lo, E. Chang and Y. C. Chan [43]. After the signalised version has been proposed, the inspiration leads to the other researchers in the following works. Maher and Feldman [44] have investigated the application of CTM to the optimisation of a signalised roundabout with TRANSYT technique. Work by Lin and Wang [45] have tried to optimise the traffic signal light by using CTM. The optimisation method is based on a mixed-interger linear programming for two intersections. However, the previous two works [44], [45] have considered without the transit signal pri-

ority and also neglect the enormous computation. Therefore, one of the most challenges in traffic signal operation is to find the best strategic control for road traffic networks.

To respond to that challenging issue, this dissertation is aimed at developing a new mathematical framework to control the traffic signal light for the road network with recently deployed bus rapid transit system. The road network dynamics have been captured by the state-of-the-art signalised CTM. The best possible traffic signal actions can be found by applying the automated goal-directed learning and decision making called RL. As illustrated in Figure 1.1, from the observed literatures, the traffic signal control with priority systems based on signalised cell transmission model together with an unsupervised learning have not yet been discovered. Therefore, the following chapters will provide the general knowledge, the mathematical formulations and the insightful conclusions.

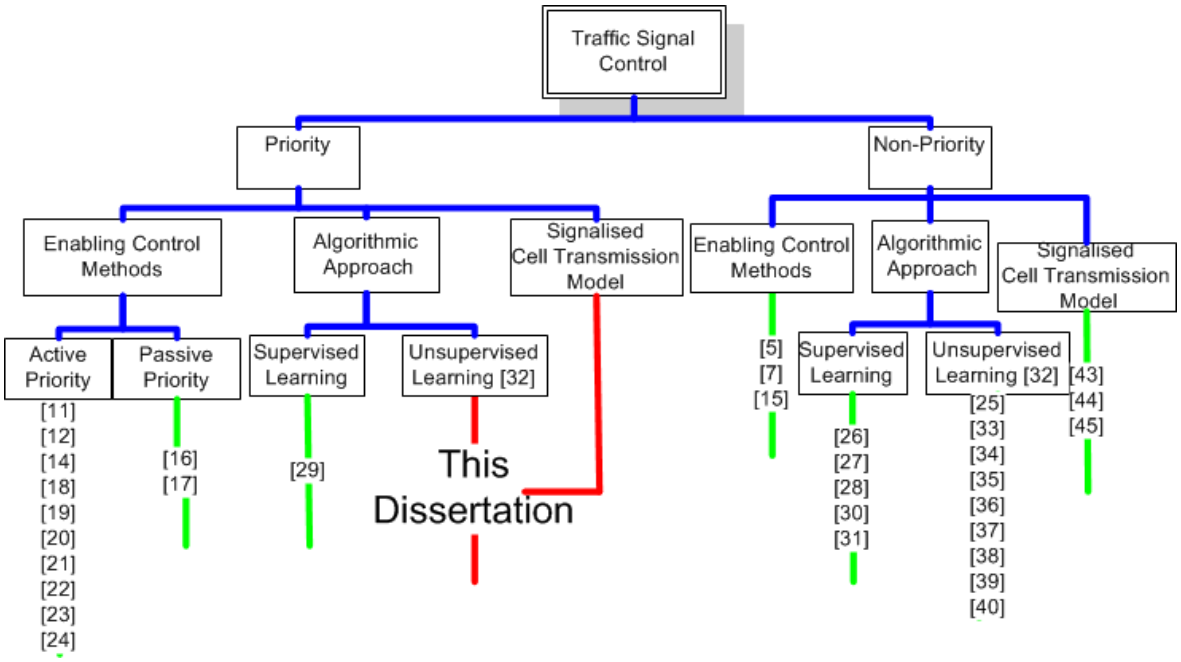

Figure 1.1 Dissertation overview

1.2 Scope of Dissertation

The scope of this work is in investigating the traffic signal control with cell transmission model for bus rapid transit network using reinforcement learning with the distinct behavior of interest.

1. Study the existing and current practice of traffic signal control with and without transit signal priority systems.

2. Develop a novel signalised CTM framework by using reinforcement learning to find the proper traffic signal for an isolated intersection without priority systems.

3. Provide a comparative control methods between the signalised CTM based reinforcement learning and the optimal control based queueing theory without priority systems.

4. Develop a novel signalised CTM framework together with the transit signal priority systems (BRT) by using reinforcement learning to find the proper traffic signal for a network scale.

5. Develop a computer program to evaluate the proposed method.

1.3 Organisation of the Dissertation

The remaining parts of this dissertation are organised as follows.

In **Chapter II**, the general background of the cell transmission model (CTM) is explained together with the signalised CTM version. The basic concept of RL has also been introduced. The current implementation of the traffic signal control method with the signal priority system will be introduced.

In **Chapter III**, the framework to control the traffic signal lights by using the Q-learning when the summation of overall traffic demand from all directions exceeds the maximum flow capacity has been investigated. In this chapter, by merely using a single intersection network scenario and with herein newly introduced boundary conditions to capture necessary vehicles backlog dynamics around the vicinity of the considered intersection, the proposed Q-learning based CTM model can help reduce computational burdens a great deal in comparison with the best periodic signal solution (BPSS). Particularly, with newly formulated RL environment using CTM parameters, by using the total network delay as a reward function, the results were not necessarily as good as initially expected. Rather, both simulation and mathematical derivation results confirm that using the newly proposed red light delay as the RL reward function gives better performance than using the total network delay as the reward function.

In **Chapter IV**, the performance comparison for optimal traffic signal controls for an isolated intersection is proposed. The optimal traffic signal controls have been analysed with two well-known M/M/1 and D/D/1 queueing models, and RL approach. The RL framework

has also integrated in conjunction with the use of the macroscopic CTM to update the vehicle state dynamics upon the change of Q-learning actions. Two approaches, the steady-state M/M/1 and D/D/1 and the Q-learning are compared in terms of the network throughput and the average vehicle delay per completed trips. The obtained control strategies are adopted and investigated in the microscopic traffic simulator AIMSUN.

In **Chapter V**, the in-depth investigations of the traffic signal control series from **Chapter III** and **Chapter IV** have been further combined to a larger network scale and by incorporating the exclusive bus rapid transit (BRT) in the systems. With the novel formulation of CTM-BRT-based RL algorithm, this chapter shows the possibilities of adopting the Q-learning to adjust the proper traffic signal light in a network scale scenario. The reported results have been divided into two main scenarios which are the comparison between non-BRT and BRT systems using Q-learning and the comparison between the distributed Q-learning and the existing distributed control methods. The reported results show the applicable range of Q-learning in controlling the traffic light on the BRT road network systems.

This dissertation concludes all the contributions in **Chapter VI**, together with the possible future research directions.

CHAPTER II

FUNDAMENTAL THEORY

This chapter provides the general knowledge being used throughout the dissertation. Theoretically, the underlying ideas are the macroscopic signalised cell transmission model (CTM), the self-automation learning system reinforcement learning (RL) and the current implementation of the transit signal priority system (BRT).

Two types of road traffic networks are considered in this dissertation: an isolated intersection in **Chapter III**, **Chapter IV** and a network with BRT systems in **Chapter V**. The general knowledge of CTM is given in Section 2.1. For the novel formulation of the CTM-BRT Q-learning, the mathematical framework has been clearly stated in each chapter. The general idea of the reinforcement learning has been introduced in Section 2.2. In Section 2.3, the existing implementation to control the road traffic network has been described. And Section 2.4 concludes this chapter and leads the idea to **Chapter III**.

2.1 Cell Transmission Model

The use of traffic simulation nowadays can be mainly classified into either microscopic or macroscopic models. Generally, the macroscopic traffic model is originally formulated as the relationship among traffic flow characteristics, e.g., flow, density, average speed of the vehicles [41]. On the contrary, the original idea for the microscopic traffic model is based on the individual tracking of the vehicle movements. Practically, the driving behaviors in real traffic situation are difficult to measure, observe and validate. By nature, the time calculation from the macroscopic model is less consumable than the microscopic model. For the traffic signal control, the decision for changing signal at intersections becomes crucial. The time calculation is considered as the first priority in choosing the traffic simulation model. Therefore, in this dissertation, the macroscopic traffic model has been chosen.

The limitations of the macroscopic CTM underlying in this dissertation are no turning movements, lane changing, ramping, lane merging and etc. The elementary of traffic parameters being considered in this dissertation are the average vehicle speed, the road length and

the calibrated wave speed coefficient. The model validation of the traffic model is not in this dissertation. Nonetheless, the real road traffic parameters can extend easily by including all traffic behavior. The employing of the CTM in this dissertation has been intended to capture the vehicle dynamics for finding the best proper traffic signal for the road traffic network with BRT. As mentioned in the beginning of this chapter, the mathematical formulation of the CTM can be found in each chapter individually.

2.2 Self-Automated Reinforcement Learning

Reinforcement learning (RL) is a class of machine learning. The classical learning methods are generally known as the supervised and the unsupervised learning. For the supervised learning, the set of all possible pairs, both the input and the expected output must be trained off-line. In an extreme circumstance, the obtainable result from the training is difficult e.g., the set of training pairs [46]. Therefore, the supervised learning has been introduced to solve this situation. The RL can learn directly from the interaction between the control agent (the decision maker) and the environment (the vicinity around the system). The core idea of the RL can be briefly explained as follows.

At the beginning, the environment has been sensed by the control agent. The agent therefore determines an action to be chosen from its own set of possible actions. The progression of the system dynamics is directly affected by the applied action from the agent. The control agent observes for the reward from the change of state. The reward value is used to tell the control agent how good or bad the previous decision is. The control agent will decide on whether to remain or change the current action depending on the immediate reward return [32]. For the traffic signal control being used in this dissertation, the agent represents the traffic light and the environment represents the vehicles in the systems. The reward corresponds to the vehicle delay. In the case of road network with BRT systems, the reward function has integrated also the passenger delay in terms of the bus priority in **Chapter V**.

2.3 Current Implementation of Transit Signal Priority Systems

The way to give the priority to the buses nowadays are manifolded. For convenience, the traffic signal can be grouped into two categories which are an isolated control and a coor-

dinated control. For an isolated control, the traffic signal has been individually controlled in each intersection. In general, the isolated control is not taken into account of the effects of the neighbourhood intersection [47]. However, in this dissertation, the newly introduced boundary condition has been introduced to capture necessary vehicle backlog dynamics around the vicinity of the considered intersection as to be shown in **Chapter IV**. The general methods to control an isolated intersection can be either fixed-time or vehicle actuated.

The fixed-time technique employs the historical data to set the pre-planned traffic control. The historical data has been used to calculate the proper signal. This system has been proven to work well in non-congested traffic scenarios [47]. For the vehicle actuated (VA) or the microprocessor optimised vehicle actuation (MOVA), this method relies on the sensors embedded in each intersection. At an intersection, the green light will be calculated based on the traffic volumes approaching an intersection. This system gives the priority by either extending or recalling the green signal for the buses.

For the coordinated control, the operation at each intersection has been sent the data through the central controller to determine the signal indication for the whole system. By using this technique, if the approaching vehicles can be served under the current road capacity, then the obtained control plan is useful. However, when the system becomes totally jammed, the proper traffic signal solution is infeasible. And moreover, the transmission of data from the local level to the central control becomes crucial. The implemented coordinated control has been widely used in various methods as mentioned in **Chapter I** e.g., SCOOT [5] and SCATS [7].

In this dissertation, the considered system has been inspired from the road network with BRT system in Bangkok, Thailand. This is an example of the road traffic network with BRT that always operates in the jammed scenarios in the rush hour period everyday. Therefore, in the following chapters, the core idea is how to control the system in the jammed scenarios.

2.4 Concluding Remarks

In this chapter, the core idea of three theories has been introduced. To the aforementioned, in Section 2.1, the CTM has been employed for updating the system dynamics. In Section 2.2, the RL has also introduced for acting as a major role in finding the best proper traffic signal control in two scenarios which are an isolated intersection and a network with

BRT systems. In Section 2.3, the current implementation of the traffic signal control method with the signal priority system will be introduced. In the next chapter, the newly formulated mathematical framework for an isolated traffic signal control with signalised CTM together with the RL technique has been introduced. The proposed reward function by using the red light delay to minimise the overall system delay will also been introduced.

CHAPTER III

ISOLATED TRAFFIC SIGNAL CONTROL WITH Q-LEARNING USING CELL TRANSMISSION MODEL

The inspiration idea behind this chapter originates from the bus rapid transit (BRT) route of road traffic networks in Bangkok, Thailand. As illustrated in Figure 3.1 [48], the U-shaped road network in Bangkok consists of two main road systems both isolated and network. The red pinned-points represents the BRT stations. Five green balloons and a blue line represents an example of the journey trip from station A to station E. Consider the green balloon "E", the intersection represents an isolated intersection. From the green balloon "A"-"B", the road link presents a network with BRT systems. From the existing literatures, the past researchers have investigated the traffic signal control with the reinforcement learning (RL) in the microscopic level only. In this dissertation, the mathematical formulation of the macroscopic signalised cell transmission model CTM together with the RL is first proposed in this chapter. For simplification, the formulation in this chapter will be firstly formulated from an isolated intersection (the green balloon "E"). To take into account of the network neighbourhood delay, this chapter employs the cascading CTM to capture the necessary vehicle backlog dynamics around the vicinity of the considered intersection.

Section 3.1 describes the problem formulation of the cascading CTM together with the Q-learning framework. In this chapter, the goal is for Q-learning to minimise the total network delay. The detailed implementation of the Q-learning algorithm can be found in Section 3.2. This chapter contributes two main ideas which are the newly proposed red light delay as the Q-learning reward function and the mathematical analysis when the summation of overall traffic demand from all directions exceeds the maximum flow capacity. Section 3.3 shows the validation of the proposed CTM-Based solution together with the applicability of Q-learning to control an isolated intersection on the microscopic AIMSUN level. Section 3.4 concludes this chapter.

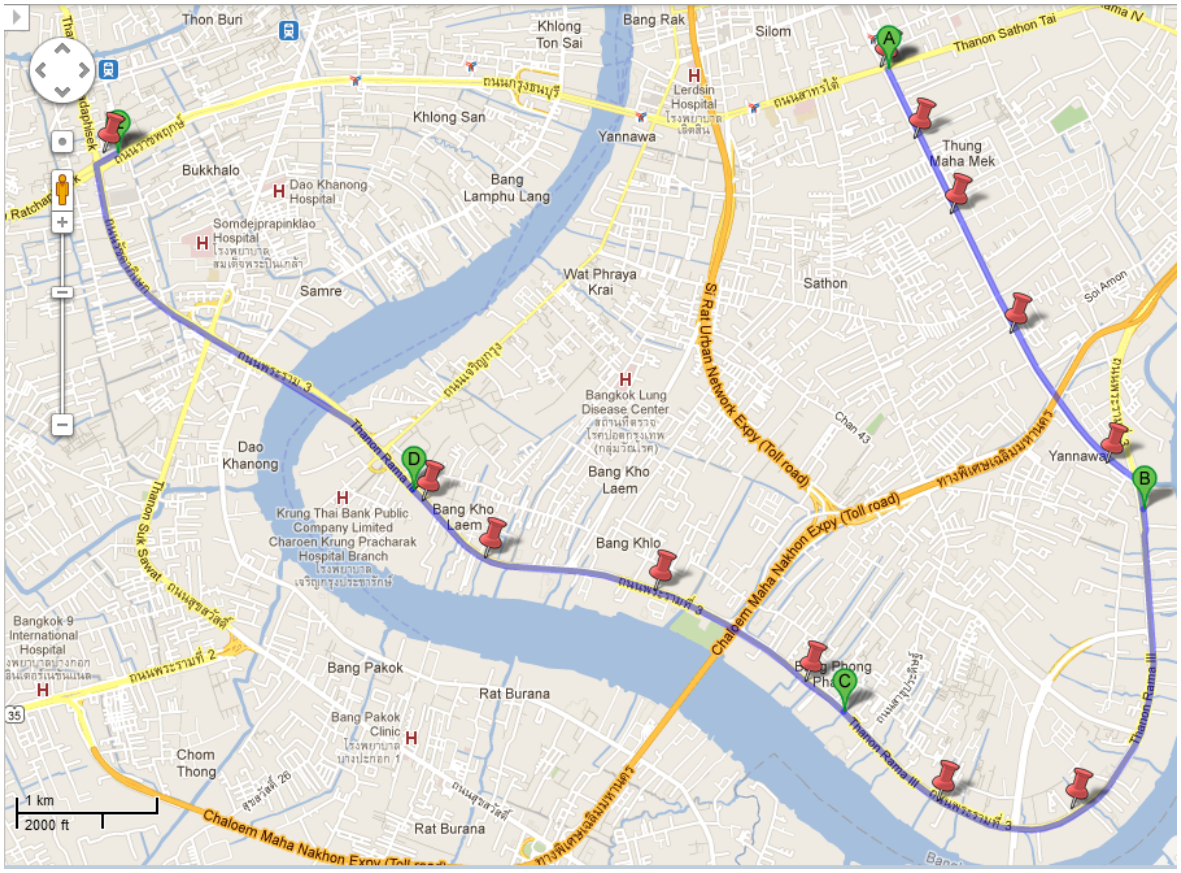

Figure 3.1 BRT route in Bangkok, Thailand

3.1 Problem Formulation

3.1.1 State Space

Suppose the vehicles in the system belong to a single class e.g. personal cars. As

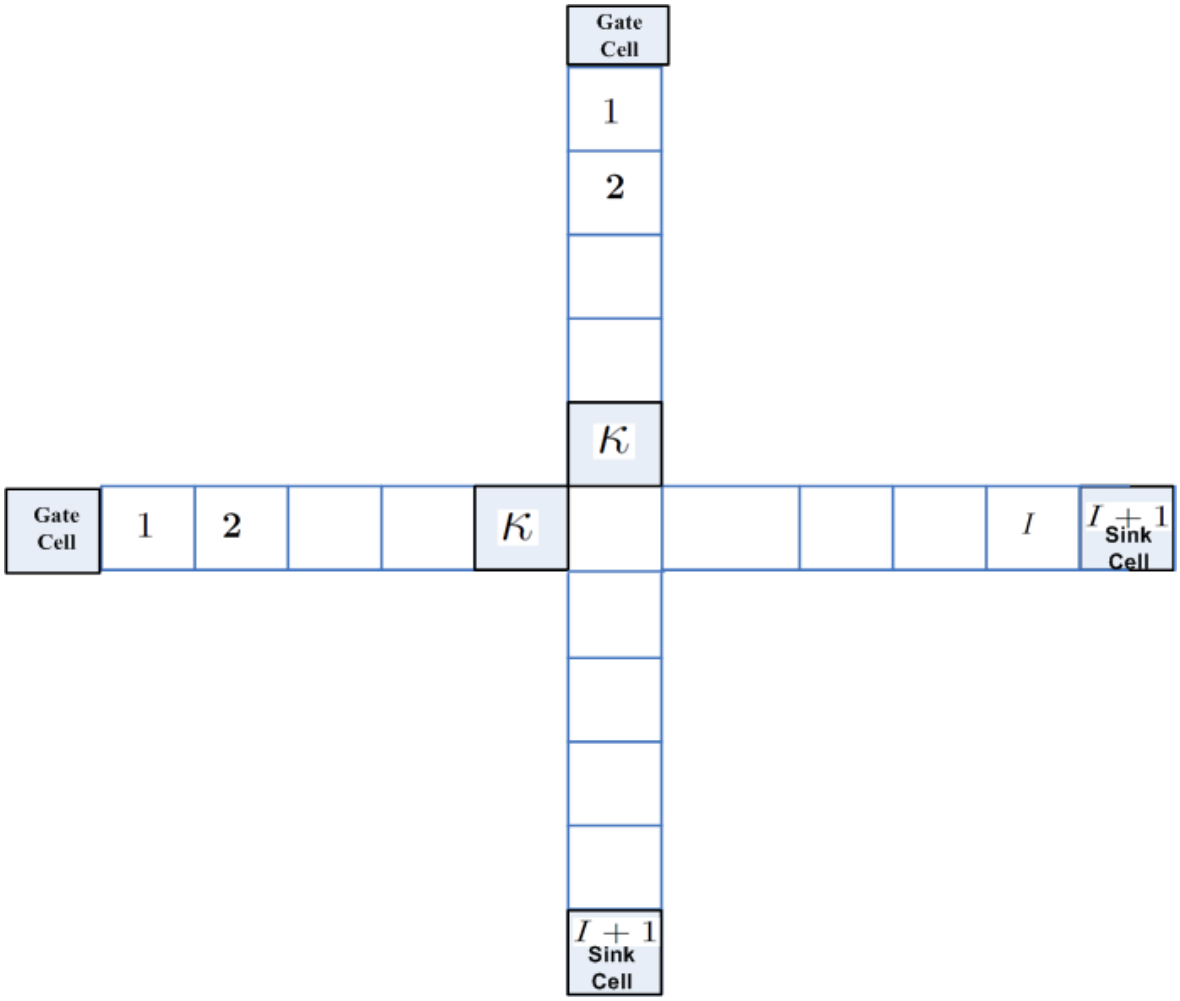

Figure 3.2 CTM boundaries and signalised cells

shown in Figure 3.2, each road is partitioned into small cells $i = 1, ..., I$. The incoming demand patterns to an intersection is classified into P directions. Let $\mathcal{S}$ be the state space of the system. For each vehicle cell i in direction p at time slot t, define $s_i^p(t)$ as the number of vehicles. Let $\mathbf{s}(t) = [s_i^p(t),\ \forall(i,p)] \in \mathcal{S}$ be the state vector which represents the total number of vehicles in the system at time slot t. Note that in a real traffic scenario, the number of vehicles can be estimated from sensors on the road. To avoid the computational burden caused by the state space explosion, the quantisation technique is employed. The

level of quantisations here can be represented by the number of deployed sensors in the road network. For simplification, let us define the quantised level of the total number of vehicles approaching the intersection from direction p at time slot t as

$$\tilde{s}^p(t) = \left\lceil \frac{\sum_{i=1}^{\kappa} s_i^p(t)}{\mathcal{C}} f \right\rceil + I(\sum_{i=1}^{\kappa} s_i^p(t) = 0), \tag{3.1}$$

where $I(.)$ is the indicator function; $\mathcal{C}$ is the maximum number of vehicles allowable in each cell $i = 1, ..., \kappa$; and f is the total number of quantisation levels. The Q-learning state can then be redefined as

$$\tilde{\mathbf{s}}(t) = [\tilde{s}^p(t),\ \forall p] \in \tilde{\mathcal{S}}. \tag{3.2}$$

3.1.2 Cell Transmission Model

To incorporate the evolution of traffic dynamics in the system, a basic macroscopic model CTM is employed. The CTM parameters can be defined as follows [41].

3.1.2.1 Sending Capability

Sending capability represents the ability to send the vehicles from cells to other cells, i.e., moving vehicles from beginning to ending cells. The sending capability can be defined as

$$\Lambda_i^p(t) = \min\left\{s_i^p(t), q_i^p(t)\right\}. \tag{3.3}$$

For cell i in direction p at time slot t, $\Lambda_i^p(t)$ is the sending capability; $s_i^p(t)$ is the number of vehicles; and $q_i^p(t)$ is the maximum number of vehicles that can flow through cell i.

3.1.2.2 Receiving Capability

Receiving capability can be calculated by considering the remaining spaces in each cell and the maximum rate of vehicles that can move through the cell. Thus, for cell i in direction p at time slot t, its receiving capability can be defined as

$$\Psi_i^p(t) = \min\{q_i^p(t), \delta_i^p\left[c_i^p(t) - s_i^p(t)\right]\}, \tag{3.4}$$

where δ_i^p is the wave speed coefficient and $c_i^p(t)$ is the maximum number of vehicles that can be present. Note that the parameter $q_i^p(t)$ is influenced by the signal phase being chosen in cell i, direction p and time slot t in the action selection.

3.1.2.3 Cell Cascading

This is the representation of the connection between two adjacent cells from the beginning cell $i-1$ and the ending cell i. The number of vehicles that flow in this cascading scenario can be calculated from the sending and receiving capability by

$$y_i^p(t) = \min\{\Lambda_{i-1}^p(t), \Psi_i^p(t)\}, \tag{3.5}$$

where $y_i^p\,(t)$ is the number of vehicles that flow into cell i in direction p at time slot t.

3.1.2.4 Flow Conservation

Flow conservation is used to update the number of vehicles for the next time slot:

$$s_i^p(t+1) = s_i^p(t) + y_i^p(t) - y_{i+1}^p(t). \tag{3.6}$$

3.1.3 Action Space

To influence the system dynamics, for each time slot, the control agent (traffic controller) must select whether it would keep the current signal indication or change it. Such decision is called action. At state vector $\tilde{\mathbf{s}}$, an action must be selected from a state dependent set $\mathcal{A}(\tilde{\mathbf{s}})$. Specifically, $\mathcal{A}(\tilde{\mathbf{s}})$ is the set of all possible actions which a traffic controller can take at state $\tilde{\mathbf{s}}$. Define action a_t as the phase of signal light to be chosen (e.g, phase 1 for the green light from West to East and phase 2 for that from North to South) at time slot t. The main goal is to optimise the traffic signal adjustment by minimising the total network delay of road network. The decision on changing actions is allowed every T time slots and is here represented by an indicator function at time slot t as follows:

$$G^p(t) \;=\; \begin{cases} 1, & \text{vehicles in direction } p \text{ get green light} \\ & \text{in the chosen action at time slot } t \\ 0, & \text{vehicles in direction } p \text{ get red light} \\ & \text{in the chosen action at time slot } t. \end{cases} \tag{3.7}$$

Note that the action space $\mathcal{A}(\tilde{\mathbf{s}})$ must be defined such that all conflicting flows are not allowed to have green light at the same time.

The system dynamics are changed according to the traffic signal lights corresponding to the action taken $a_t \in \mathcal{A}(\tilde{\mathbf{s}})$. Assume that in one time slot, vehicles can move on average to

the adjacent cells only. Let q_{max} be the maximum number of vehicles that can flow through each cell per time slot. For non-signalised cell i, the maximum number of vehicles that can flow through cell i in direction p at time slot t is given by $q_i^p(t) = q_{max}, \forall(p,t)$. For signalised cell i, the equation can be defined as follows

$$q_i^p(t) = \begin{cases} q_{max} & ; G^p(t) = 1 \quad \text{and} \quad t - \tau_i(t) > L \\ 0 & ; \text{otherwise}, \end{cases} \tag{3.8}$$

where L is the total starting/stopping loss time upon each signal change and $\tau_i(t)$ is the latest time instant where the traffic signal indication of cell i at time slot t has been changed.

3.1.4 Boundary Conditions

Figure 3.3 illustrates the boundary condition for CTM herein being used.

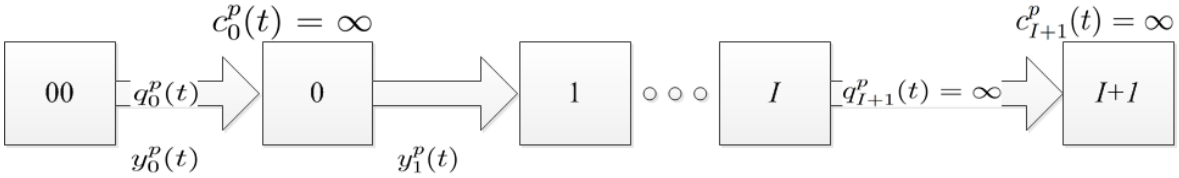

Figure 3.3 Boundary cells

3.1.4.1 Gate Cell

The boundary condition is here formulated by following [41]. At the boundary, input vehicle flows can be modelled by a cell pair ("00" and cell "0") as illustrated in Figure 3.3. A source cell "00" with an infinite number of vehicles $s_{00}^p(t) = \infty$ ready to enter an initially empty gate cell "0" of infinite size, $c_0^p(t) = \infty$. The flow capacity $q_0^p(t)$ of the gate cell "0" is set to the desired link input flow. Thus, the boundary conditions can be obtained from (3.3)–(3.6) as

$$\Lambda_0^p(t) = \min\{s_0^p(t), q_0^p(t)\} \tag{3.9}$$

$$y_0^p(t) = q_0^p(t) \tag{3.10}$$

$$y_1^p(t) = \min\{\Lambda_0^p(t), \Psi_1^p(t)\} \tag{3.11}$$

$$s_0^p(t+1) = s_0^p(t) + y_0^p(t) - y_1^p(t). \tag{3.12}$$

3.1.4.2 Sink cell

Suppose that the output cell referred as the sink cell, for all exiting traffic has infinite size $c^p_{I+1}(t) = \infty$ and $q^p_{I+1}(t) = \infty$. According to (3.4), the sink cell $I + 1$ thus has the receiving capability of

$$\Psi^p_{I+1}(t) = \infty. \tag{3.13}$$

Note that in a more realistic case, $q^p_{I+1}(t)$ can be set in accordance with the road network neighbourhood, i.e., the cell buffer of an adjacent intersection.

3.1.5 Vehicle Delay

In Q-learning, to quantify the consequence of the action taken at time slot t, an immediate reward in terms of vehicle delay is returned to the agent (traffic controller). Vehicle delay is defined as the number of vehicles that cannot move away from the present cell within each time slot. In this research, two types of vehicle delay are proposed, i.e., internal delay and external delay. At time slot t for each direction p, let $d^p_0(t)$ be the external vehicle delay and $d^p_i(t)$ be the internal vehicle delay in cell i. These delays can be expressed as

$$d^p_0(t) = s^p_0(t) - y^p_1(t), \tag{3.14}$$

$$d^p_i(t) = s^p_i(t) - y^p_{i+1}(t), \quad i = 1, 2, \ldots, I. \tag{3.15}$$

The external delay can be interpreted as the delay experienced by the vehicles that wait to enter the considered road network from its upstream neighbourhoods. The external delay value forms the boundary condition to capture necessary vehicle backlog dynamics around the vicinity of the considered intersection. The internal vehicle delay is the delay incurred within each cell along the considered road network. Combining both types of delay therefore reflects how well the action just taken by the agent (traffic controller) at state vector $\tilde{\mathbf{s}}$ is, by merely taking into account a single intersection. The next section provides the long term performance criteria in terms of these delay functions which will be optimised for the best possible traffic signal control by means of Q-learning.

3.1.6 Performance Criteria

To evaluate the optimal policy (set of actions) that minimises the total network delay, the performance criteria $\Upsilon(t)$ at time slot t is defined as

$$\Upsilon(t) = \Upsilon_{red}(t) + \Upsilon_{green}(t), \tag{3.16}$$

$$\Upsilon_{red}(t) = \sum_{p=1}^{P}\sum_{i=0}^{I}(1 - G^p(t))d_i^p(t), \tag{3.17}$$

$$\Upsilon_{green}(t) = \sum_{p=1}^{P}\sum_{i=0}^{I} G^p(t)d_i^p(t), \tag{3.18}$$

where $\Upsilon_{red}(t)$ is the "red light delay" and $\Upsilon_{green}(t)$ is the "green light delay". The red (green) light delay is the total vehicle delay from all the cells in the directions that see the red (green) light.

3.2 Signal Optimisation by Q-learning Algorithm For Isolated Intersection

Without loss of generality, let us index the signalised cells by κ as an example of CTM-based intersection model shown in Figure 3.2. Assume that no turning movement is allowed at this intersection. The signalised cells κ are used to control the traffic flows from West to East and North to South. To tackle the road traffic problem where the system always changes, a well-known method that can learn directly from experiences is employed, namely, the Q-learning method [32]. Q-learning uses the action-value function $Q(\tilde{\mathbf{s}}, a)$ to evaluate the average future reward return expressed as a function of the current state $\tilde{\mathbf{s}}$ and action a. This section explains a step-by-step implementation of Q-learning algorithm proposed in the CTM framework.

To apply Q-learning in a signalised CTM framework, a definite simulation length is used for periodically observing traffic behaviors within a study time-interval. When the current time slot of CTM reaches the simulation length, the system enters the next *episode*. In practice, episodes can represent the repeatable and non-repeatable traffic phenomena. On one hand, in a repeatable case, we can use Q-learning to tackle a recurrent congestion, e.g. during rush hours, in which traffic behaviours statistically repeat themselves from one day to another. In this case, at the beginning of each episode, our road system modelled by CTM can

be reset to the same initial-value cell density settings. On the other hand, in a non-repeatable case, Q-learning can be used to deal with a non-recurrent congestion scenario resulted from unexpected incidences like accidents or road surface maintenance. In this case, our interest is on how Q-learning would allow the signal controller to quickly learn and adapt its strategic decisions upon those unexpected changes. Consequently, the CTM state in the first time slot of next episode is defined in this case as the CTM state in the last time slot of previous episode.

Whether Q-learning is applied in the repeatable or non-repeatable cases, within each episode, the Q-learning-based traffic controller is designed to make a sequence of signal-light decisions. Let the decision epoch t_ω refer to the time instant when decision ω is made, where $\omega = 1, 2, \ldots$ and $t_\omega = t_1, t_2, \ldots$, respectively.

For each episode, the optimisation procedure of Q-learning can be summarised as follows.

1) *System Initialisation*

The number of vehicles in state vector $\mathbf{s}(0)$ can be intialised by (3.6) and (3.12) at the beginning of an episode to the latest observed state of the system in the previous episode in the non-repeatable case or to a nominal operating point of the system at the considered time period in the repeatable case. In practice, the number of vehicles $\tilde{s}^p(0)$ for all p can be measured from road traffic by counting from the sensors embedded on the road. The action value function $Q(\tilde{\mathbf{s}}, a)$ can be initialised to the latest updated value in the previous episode for both the non-repeatable case and the repeatable case. It should be noted that, different intialisations of $Q(\tilde{\mathbf{s}}, a)$ yield different results, mainly, in terms of the time convergence (the time that the algorithm needs to learn to reach the solution). Let $\omega = 1$.

2) *Action Selection*

At decision ω, with the current state observable at $\tilde{\mathbf{s}}$, the agent (traffic controller) chooses an action $a \in \mathcal{A}(\tilde{\mathbf{s}})$ to control the traffic signal by changing $G^p(t)$ in (3.7). The action can be chosen by the ϵ-greedy algorithm [32], where the greedy action is here defined as

$$a = \arg\min_{a'} Q(\tilde{\mathbf{s}}, a').$$

According to this algorithm [32], Q-learning chooses the greedy action with probability $1 - \epsilon$. And, with probability ϵ, the other actions are randomly selected according to a uniform distribution. In practice, an ϵ is a small positive value representing the explorability of learning algorithm.

3) *Update of System Dynamics*

Calculate the CTM state from time slot $t = t_\omega$ to time slot $t = t_{\omega+1} - 1$. Here, the next state vector ($\tilde{\mathbf{s}}'$) is calculated from the CTM state at time slot $t = t_{\omega+1} - 1$. In this section, three Q-functions have been compared, namely, the total network delay by considering the accumulative vehicle delays in only the directions facing red light signal, receiving the green light signal, or both. The observed reward $R(\omega)$ can then be correspondingly calculated from

$$R(\omega) = \begin{cases} \sum_{t=t_\omega}^{t_{\omega+1}-1} \Upsilon(t) & \text{in case of total network delay} \\ \sum_{t=t_\omega}^{t_{\omega+1}-1} \Upsilon_{red}(t) & \text{in case of red light delay} \\ \sum_{t=t_\omega}^{t_{\omega+1}-1} \Upsilon_{green}(t) & \text{in case of green light delay.} \end{cases} \tag{3.19}$$

4) *Update of Action Value Function*

The algorithm can learn from its past experiences accumulated in Q-function and the reward in (3.19) newly gained from the most recent action ω. By following [32], Q-function can be updated as follows

$$Q(\tilde{\mathbf{s}}, a) \leftarrow Q(\tilde{\mathbf{s}}, a) + \alpha[R(\omega) + \gamma \min_{a'} Q(\tilde{\mathbf{s}}', a') - Q(\tilde{\mathbf{s}}, a)].$$

Here, $Q(\tilde{\mathbf{s}}', a')$ represents the action value function for the next observable state vector $\tilde{\mathbf{s}}'$ and next possible action $a' \in \mathcal{A}(\tilde{\mathbf{s}}')$. Practically, $\alpha \in (0, 1]$ is the learning rate and $\gamma \in [0, 1)$ is the discount rate applied to the future expected rewards.

5) *Update of State Variable and Timing Parameter*

Update state $\tilde{\mathbf{s}} \leftarrow \tilde{\mathbf{s}}'$. And update $\omega \leftarrow \omega + 1$.

6) *Stopping Condition*

Repeat steps 2)–5) until the end of episode.

3.3 Results and Discussions

This section is aimed at reporting the findings from our series of experiments. Firstly, the convergence time and corresponding computational complexity of the proposed Q-learning algorithm has been presented. Secondly, three reward functions in (3.19) have been compared in terms of the achievable minimum total network delay values. Thirdly, with the best

choice in the reward value accounting for the vehicle delay in red-light traffic direction, Q-learning performance has been investigated in stationary/non-stationary stochastic loading scenarios. Lastly, the applicability of macroscopic CTM-based solution of the proposed Q-learning algorithm has been tested in microscopic mobility environments using AIMSUN. All the experimental results share the following common parameter settings.

1. System Parameters: As illustrated in Figure 3.2, suppose that the length of each road approaching the considered intersection is 800 metres and each road is discretised into 10 equal-length cells, i.e. $I = 10$. Each time slot has been set to 5 seconds. Each cell has the capacity $c_i^p(t)$ of 60 passenger car units (pcu) and the maximum flow rate $q_i^p(t)$ of 6.9 pcu/slot. The wave speed coefficient δ_i^p is 0.8. Note that the values of CTM parameters are based on the actual traffic data collection being calibrated for Payathai road in Bangkok, Thailand [49].

2. Control Parameters: The length of each episode is 20 minutes or 240 time slots. An action has been chosen every 3 time slots. Note that the longer the action selection is, the more outdated the decision becomes. The number of quantisation levels f has been set to 3. Practically, three levels are corresponding to the three sensors that are often deployed on the real road configuration. The first sensor at the entry of the road is used for preventing the spill-back of vehicles to upstream neighbourhoods. The second sensor is deployed in the middle of the road for the queue length estimation. The third sensor placed at the stop-line of the road is used for the wasted green prevention in an actuated signal control.

3.3.1 Q-learning Validation

This chapter proposes the newly developed version of the signalised CTM with Q-learning. The validation of the Q-learning in various traffic conditions are reported. The best periodic signal solution (BPSS) and the Q-learning solution by using the proposed framework have been compared. In each cycle time, the total network delay obtained from the BPSS can be calculated by allocating all possible splits pairs to each direction. Define λ_1 and λ_2 as the average rate of arrival traffic from West to East and North to South, respectively. Consider deterministic demand patterns with $\{\lambda_1, \lambda_2\} = \{8, 8\}, \{11, 5\}, \{13, 3\}, \{15, 1\}$

pcu/slot. Note that the other traffic conditions can be achieved by other sets of demand patterns as well, but we have analysed the example of four settings given above. From trial-and-error, the Q-learning parameters are set to $\epsilon = 0.1, \alpha = 0.01, \gamma = 0.005$ within 100 episodes. Theoretically, the learning rate (α) determines how fast the newly acquired information will override the old information. The possible value of α is in the range of $0 < \alpha \leq 1$. The discount factor (γ) determines the importance of future rewards where $0 \leq \gamma < 1$. If $\gamma = 0$, then the agent will be "opportunistic" by only considering current rewards. The parameter ϵ is a small probability, where a larger ϵ is used for a more exploration-oriented design and a smaller ϵ is used for a more exploitation-oriented design [32]. In practice, the parametric tuning for the algorithm is one of the major challenges because in different scenarios, the parameters need to be readjusted. However, the advantage of the effects of Q-learning parameters is the usable range of these parameters are wide. With the flexibility of the Q-learning parameters, the obtained solution of Q-learning can be found without readjusting as discussed in the following section of the performance in stationary/non-stationary stochastic loadings.

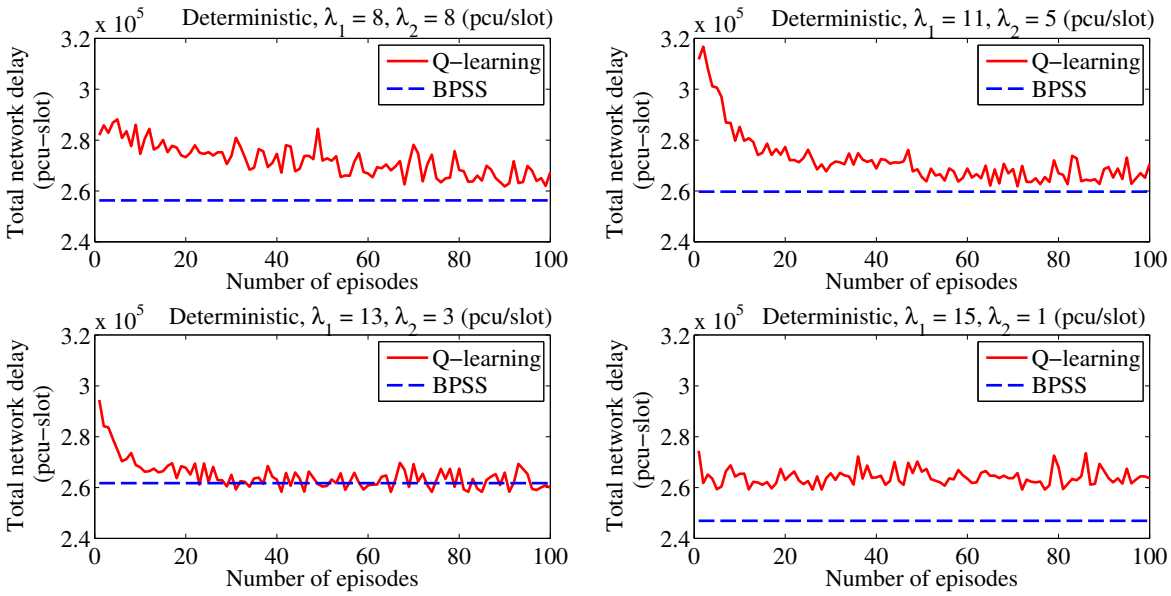

Figure 3.4 Total network delay from Q-learning vs BPSS

By using our proposed red light delay (3.17) as the reward function, Figure 3.4 and Figure 3.5 illustrate the total network delay and the allocated green time to each direction, respectively. Note that the red light delay used herein has been chosen from the following subsection focusing on the effect of reward functions. Figure 3.4 shows that the total network

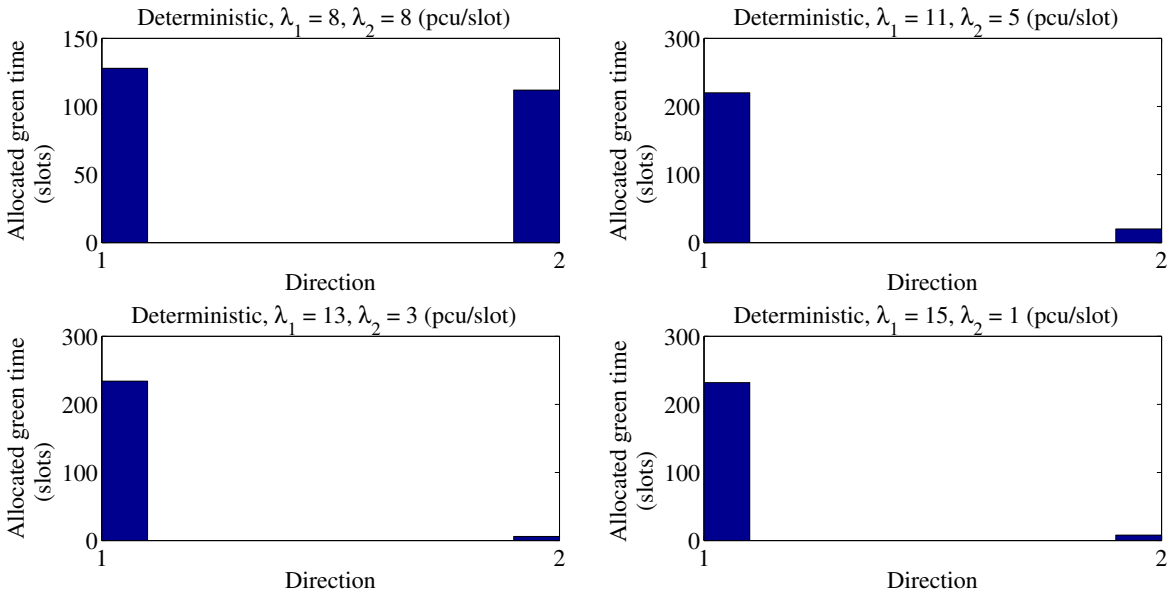

Figure 3.5 Allocated green time to each direction in last episode of Q-learning

delay from Q-learning can be found close to the solution from BPSS in most scenarios. Particularly, when $\{\lambda_1, \lambda_2\} = \{15, 1\}$ pcu/slot, Q-learning solution yields unsatisfactory result because of small traffic λ_2. Technically speaking, Q-learning requires the knowledge from its past experiences. But with a small traffic demand, the system cannot offer sufficient experiences to the Q-learning in order to achieve the solution properly. Figure 3.5 shows the number of time slots allocated to each direction. The result shows that the allocated green time in each direction is proportional to the incoming traffic demand of that direction.

To evaluate the convergence of Q-learning, an example scenario with $\{\lambda_1, \lambda_2\} = \{13, 3\}$ pcu/slot has been elaborated. Let us define the convergence criterion in terms of the average value of Q-function:

$$Q^{\text{avg}}_{\text{episode}} = \sum_{t=1}^{T} \frac{Q\left(\tilde{s}(t), a_t\right)}{T}. \tag{3.20}$$

The algorithm converges when $Q^{\text{avg}}_{\text{episode}}$ is unchanged or slightly changed with fluctuation of less than 5% in comparison with the previous episodes as illustrated in Figure 3.6.

The computational complexity has been measured in terms of the required amount of memory and the computational time to achieve the final solution. Let the number of elements in the quantised state space be denoted by $|\tilde{\mathcal{S}}|$ and that in the action space be denoted by $|\mathcal{A}|$. Note that the action space $|\mathcal{A}| = P$ where P is the total number of all road network directions. Let k be the total number of the green time pairs in the overall searching space of

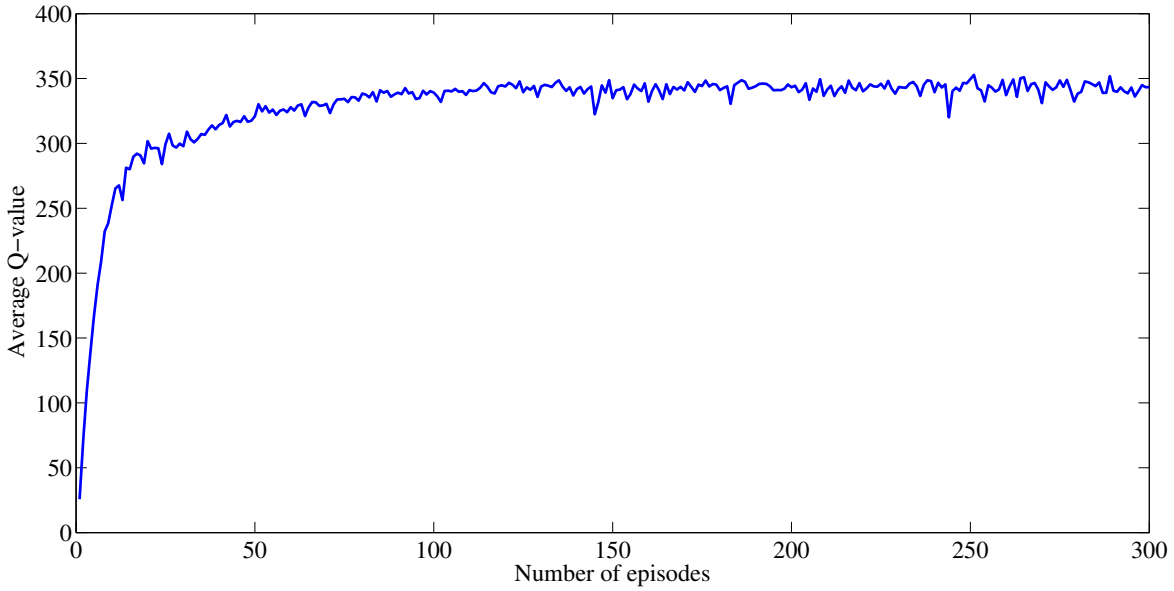

Figure 3.6 Average of Q-value for each episode

periodic signal solutions. To search for the BPSS within these k possibilities per each state, the required amount of memory is $O(a^P)$ where a is a constant. However, the amount of memory required for Q-learning is $O(|\tilde{\mathcal{S}}|P)$. The BPSS grows exponentially depending on the number of the green time pairs to be searched whereas the growth of Q-learning depends on the quantised state space and the number of actions. The memory requirement can be saved with respect to the increasing of k. From all example of our experiments, the maximum memory required for the BPSS is 748 kbytes whereas the maximum memory required for the Q-learning is 114 kbytes. Therefore, the amount of memory required in searching for the solution of Q-learning is significantly less than that for BPSS method. For the computational time, the measurement is in terms of time period from the beginning through the end of simulation in order to obtain the best corresponding signal plans. Mathematically, the computational time for the BPSS is $O(a^P)$ whereas the computational time for the Q-learning is $O(|\tilde{\mathcal{S}}|P)$. For all the experiments, the computational time has been measured by using MATLAB®with the processor Intel(R) Core(TM) i7-2630QM CPU@2.0GHz and 4GB RAM. Note that such calculation also includes the training period. The computational time of BPSS and Q-learning have been illustrated in TABLE. 3.1. The result shows that the computation of Q-learning for obtaining a control signal is significantly faster than BPSS.

Table 3.1 Computational time of Q-learning and BPSS

Pattern	Q-learning (sec)	BPSS (sec)
Loading Pattern 1	15.62	1222.63
Loading Pattern 2	15.08	1222.34
Loading Pattern 3	15.42	1223.46
Loading Pattern 4	15.21	1224.39

3.3.2 Effect of Reward Functions

The procedure to find the traffic signal solution has been illustrated in the Q-learning validation. In this subsection, three different reward functions have been investigated in both symmetric and asymmetric loading patterns. To make the experiments more realistic, the traffic demand is no longer deterministic. In this subsection, the traffic demand is a Poisson process with a constant arrival rate for each direction. For symmetric loadings, both directions have equal approaching demand from $\{1,1\},\{3,3\},...,\{15,15\}$ pcu/slot, respectively. For asymmetric loadings, λ_1 has been set to 13 pcu/slot and λ_2 is varied from $1,2,...,15$ pcu/slot. The results have been obtained with the manually fine-tuned Q-learning parameters $\epsilon = 0.1, \alpha = 0.01, \gamma = 0.005$.

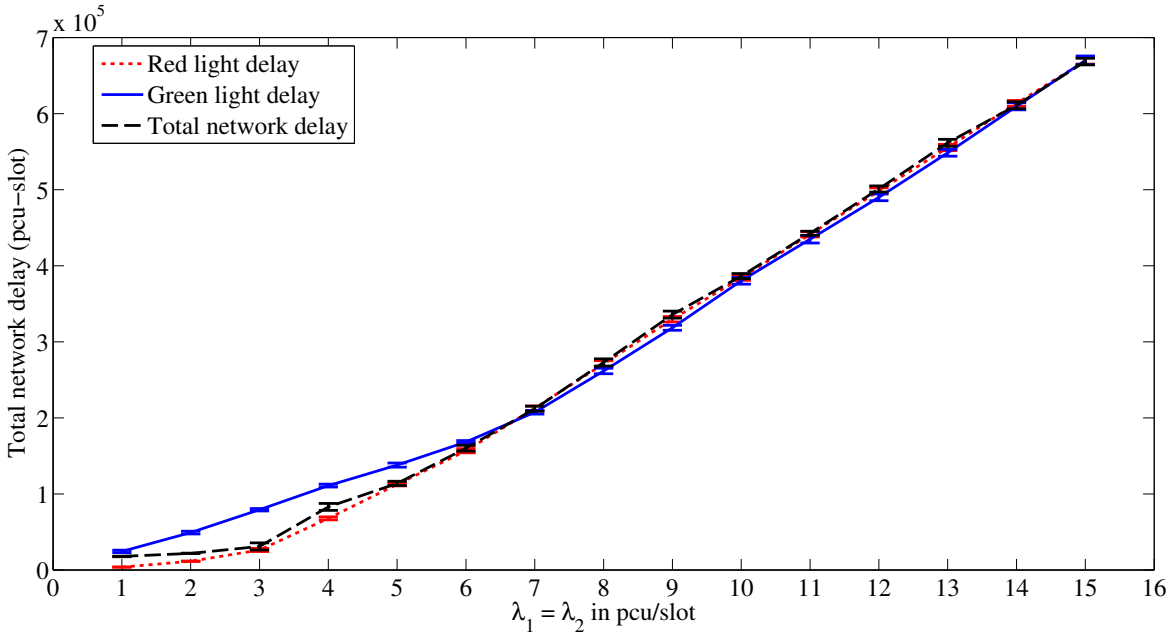

Figure 3.7 Total network delay from three reward functions on symmetric loadings

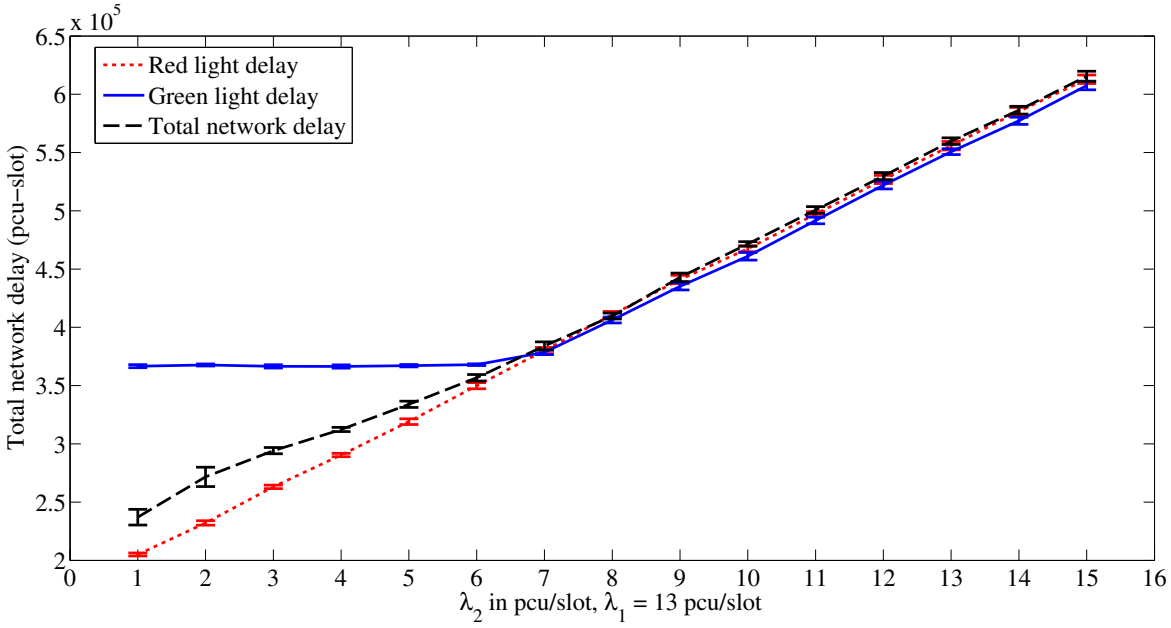

Figure 3.8 Total network delay from three reward functions on asymmetric loadings

As illustrated in Figure 3.7 and Figure 3.8, with 95% confidence interval of both symmetric and asymmetric loadings, the proposed red light delay as the reward function decreases the total network delay in comparison with the conventional case of total network delay as the reward function and greatly decreases the total network delay in comparison with the case of green light delay as the reward function. The previous statement is valid in the loading region where the summation of overall traffic demand from all directions does not exceed its maximum flow capacity ($\lambda_1 + \lambda_2 \leq 6.9$ pcu/slot). On the contrary, when the summation of overall traffic demand from all directions exceeds the maximum flow capacity ($\lambda_1 + \lambda_2 > 6.9$ pcu/slot), the case of green light delay as the reward function yields slightly low total network delay in comparison with the case using the other two reward functions. Consider the system in the case where the summation of overall traffic demand from all directions does not exceed its maximum flow capacity. In this case, any control strategy can be used because usually there is no congestion of vehicles. In such scenario, the control strategy is not complicated. However, the system in the case where the summation of overall traffic demand from all directions exceeds the maximum flow capacity, the control strategy has concerned because traffic congestion becomes a severe problem. Therefore, the following discussion will focus on the case of the summation of overall traffic demand from all directions exceeds the maximum flow capacity only.

3.3.2.1 Mathematical Analysis When Overall Traffic Exceeds The Maximum Flow Capacity

To discuss all the results under the condition when the summation of overall traffic demand from all directions exceeds the maximum flow capacity, define the major flow (minor flow) as the incoming traffic demands that exceed (does not exceed) the capacity. Two types of the road traffic phenomena have been investigated. The experiments are concerned with a major flow conflicted with a minor flow (Ma-Mi condition) and two major flows conflicted with each other (Ma-Ma condition). For the Ma-Mi condition, consider an example demand setting $\{\lambda_1, \lambda_2\} = \{13, 3\}$ pcu/slot. Our experimental results in Figure 3.9, Figure 3.10 and Figure 3.11 show the total network delay in each time slot, the delay of all cells in each direction and the action chosen in each time slot, respectively. All the results in Figure 3.9, Figure 3.10 and Figure 3.11 have been observed at the final episode at the convergence.

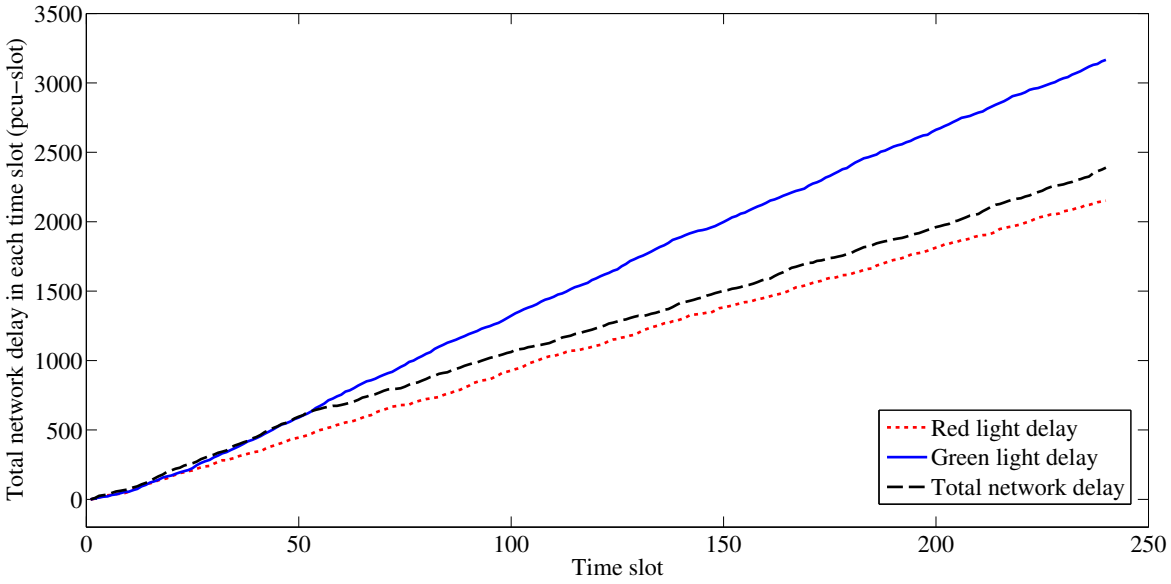

Figure 3.9 Ma-Mi: total delay in each time slot

The following discussion for the Ma-Mi conditions explains why the case of red light delay can achieve better performance than that of the other two reward functions. Recall that the Υ_{red} (Υ_{green}) denotes the red (green) light delay, which is the total vehicle delay from all the cells in the directions that see the red (green) light.

As illustrated in Figure 3.12(a) and Figure 3.12(b), the solid arrow represents the green light direction and the dash-dot arrow represents the red light direction. With a simplified

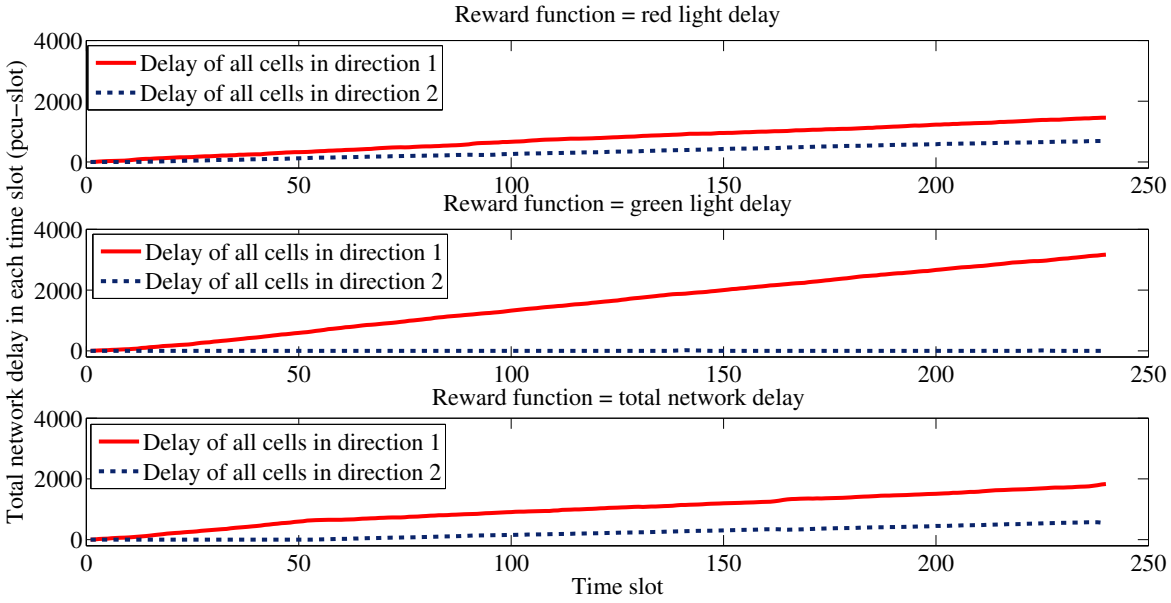

Figure 3.10 Ma-Mi: three types of reward functions and its delay in each component

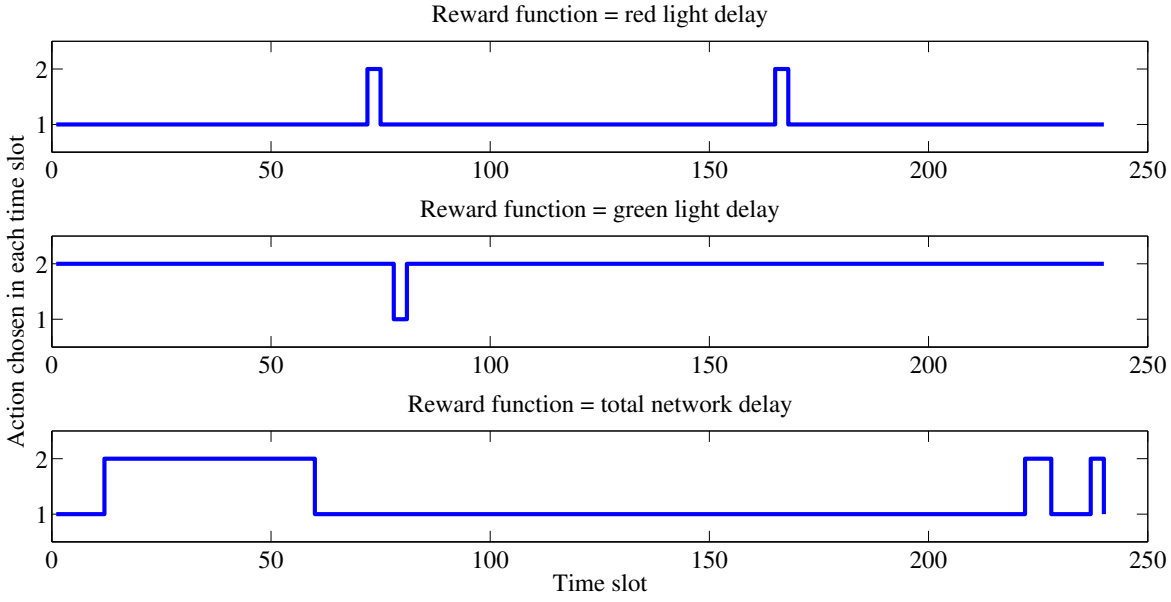

Figure 3.11 Ma-Mi: action chosen in each time slot

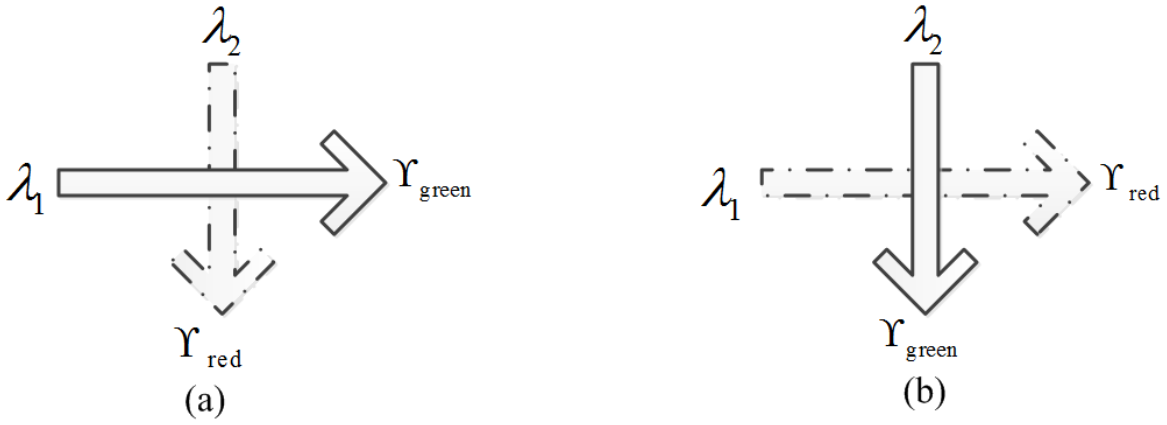

Figure 3.12: Relationship between loadings and chosen actions
(a) when the chosen action gives green light to the major flow at all time slots
(b) when the chosen action gives green light to the minor flow at all time slots

derivation, our result can be explained by using mathematical analysis as follows. Consider the derivation of accumulative delay of all cells in each direction as used in Figure. 3.9 to Figure 3.11. From (3.14) – (3.15), the accumulative delay of all cells in direction p up to time slot T can be obtained from

$$\sum_{t=0}^{T}\sum_{i=0}^{I} d_i^p(t) \;=\; \sum_{t=0}^{T}\sum_{i=0}^{I} \left(s_i^p(t) - y_{i+1}^p(t)\right). \tag{3.21}$$

At the asymptote (all the cells in overloaded direction being fully occupied), define $\bar{\Upsilon}_{red}$, ($\bar{\Upsilon}_{green}$) as the asymptotic increasing rate of expected value of the accumulative red (green) light delay. Likewise, define $\bar{\Upsilon}$ as the asymptotic increasing rate of expected value of the accumulative total network delay. The term $y_{i+1}^p(t)$ becomes zero when calculating $\Upsilon_{red}(t)$ and becomes non-zero (6.9 pcu/slot) when calculating $\Upsilon_{green}(t)$

$$\bar{\Upsilon}_{red} \;=\; \begin{cases} \lambda_2 - 0, & G^1(t) = 1 \\ \lambda_1 - 0, & G^2(t) = 1 \end{cases} \tag{3.22}$$

$$\bar{\Upsilon}_{green} \;=\; \begin{cases} \max((\lambda_1 - 6.9), 0), & G^1(t) = 1 \\ \max((\lambda_2 - 6.9), 0), & G^2(t) = 1 \end{cases} \tag{3.23}$$

$$\bar{\Upsilon} \;=\; \bar{\Upsilon}_{red} + \bar{\Upsilon}_{green},$$

$$\bar{\Upsilon} \;=\; \begin{cases} \lambda_2 + \max((\lambda_1 - 6.9), 0), & G^1(t) = 1 \\ \lambda_1 + \max((\lambda_2 - 6.9), 0), & G^2(t) = 1 \end{cases} \tag{3.24}$$

Therefore, from (3.22) – (3.24), when $\lambda_1 = 13$ and $\lambda_2 = 3$ pcu/slot,

$$\bar{\Upsilon}_{red} = \begin{cases} 3 - 0 = 3, & G^1(t) = 1 \\ 13 - 0 = 13, & G^2(t) = 1, \end{cases} \tag{3.25}$$

$$\bar{\Upsilon}_{green} = \begin{cases} 13 - 6.9 = 6.1, & G^1(t) = 1 \\ \max(3 - 6.9, 0) = 0, & G^2(t) = 1, \end{cases} \tag{3.26}$$

$$\bar{\Upsilon} = \begin{cases} 3 + 6.1 = 9.1, & G^1(t) = 1 \\ 13 + 0 = 13, & G^2(t) = 1 \end{cases} \tag{3.27}$$

From (3.25), if the reward function is $\Upsilon_{red}(t)$, then the minimum total network delay can be achieved by allocating the green light signal to the major flow (λ_1). Likewise, in (3.26), if the reward function is $\Upsilon_{green}(t)$, then the minimum total network delay can be achieved by allocating the green light signal to the minor flow (λ_2). Using $\Upsilon_{green}(t)$ as the reward function leads to the wasted green scenario (green light allocation to a particular direction without remaining vehicles) as illustrated by the term $max(3 - 6.9, 0)$. However, if $\Upsilon(t)$ is chosen as the reward function, then the minimum total network delay can be achieved by allocating the green light signal to the major flow (λ_1). The total network delay is a bit higher than the case of $\Upsilon_{red}(t)$. To explain why the total network delay from $\Upsilon(t)$ is higher than $\Upsilon_{red}(t)$. There are two concerned effects in using $\Upsilon_{red}(t)$ or $\Upsilon(t)$ as the reward function. One is the indistinguishable effect from $\Upsilon(t)$ where the agent only knows the overall network delay (3.16). Regardless of whether proper or improper action has been chosen, the value of reward in terms of total network delay is indifferent due to the summation of all vehicle delays in the system. The indistinguishable effect results in an inaccuracy (an improper action selection) and an inefficiency (an increasing of undesirable total network delay) of the action selection from Q-learning. Another is the timing effect of switched actions. In this case, the more often the action switches, the worse the total network delay is. From the discussion in the Ma-Mi condition, the recommended reward function would be the red light delay ($\Upsilon_{red}(t)$), which gives the lowest total network delay in comparison with the other two reward functions.

As an example of Ma-Ma condition, consider $\{\lambda_1, \lambda_2\} = \{13, 8\}$ pcu/slot. The values

of $\bar{\Upsilon}_{red}$, $\bar{\Upsilon}_{green}$ and $\bar{\Upsilon}$ become

$$\bar{\Upsilon}_{red} = \begin{cases} 8-0=8, & G^1(t)=1 \\ 13-0=13, & G^2(t)=1 \end{cases} \tag{3.28}$$

$$\bar{\Upsilon}_{green} = \begin{cases} 13-6.9=6.1, & G^1(t)=1 \\ 8-6.9=1.1, & G^2(t)=1 \end{cases} \tag{3.29}$$

$$\bar{\Upsilon} = \begin{cases} 8+6.1=14.1, & G^1(t)=1 \\ 13+1.1=14.1, & G^2(t)=1 \end{cases} \tag{3.30}$$

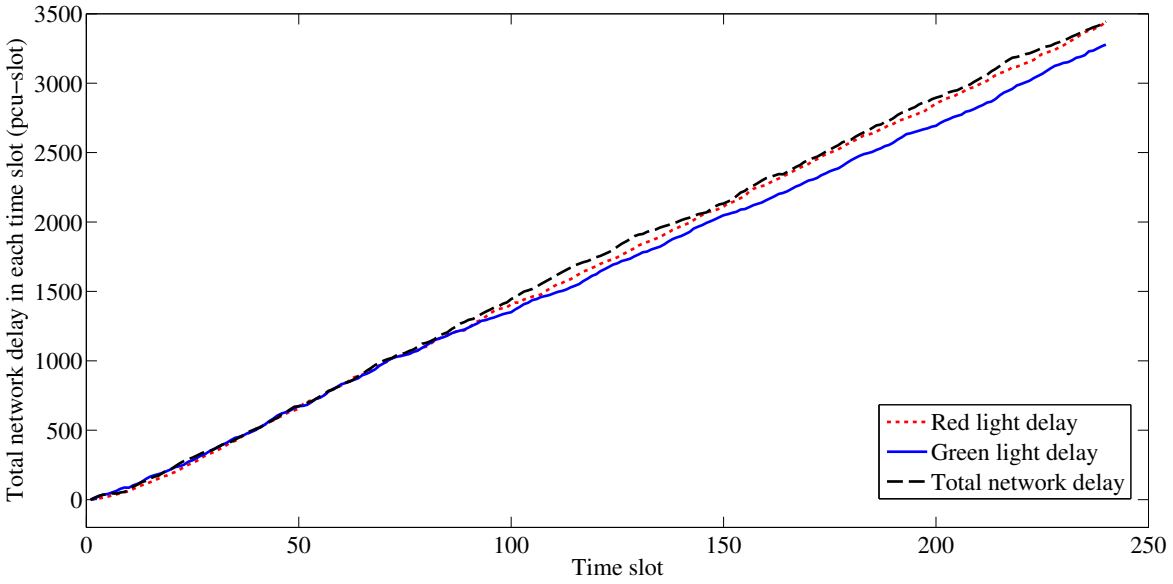

Figure 3.13 Ma-Ma: total delay in each time slot

For Ma-Ma condition, as illustrated in Figure 3.13 shows the total network delay in each time slot. Figure 3.14 and Figure 3.15 show the delay of all cells in each direction and the action chosen in each time slot, respectively. The minimum value of $\bar{\Upsilon}_{red}$ is 8 pcu-slot when the major flow receives the green light and the minimum value of $\bar{\Upsilon}_{green}$ is 1.1 pcu-slot when the minor flow receives the green light. The minimum value of $\bar{\Upsilon}$ is 14.1 pcu-slot, no matter which direction receives the green light. Consider Υ_{red} in this scenario where both directions are totally over-saturated (two major flows conflicted each other). By using $\bar{\Upsilon}_{red}$ (3.28) as the reward function, no matter what actions have been chosen, the change of total network delay becomes insignificant because the traffic is jammed. By using $\bar{\Upsilon}_{green}$ (3.29) as the reward function, the minimum total network delay can be achieved by allocating the green light to a minor flow. By using $\bar{\Upsilon}$ as the reward function, both total network delay are indifferent no matter what decisions have been taken. In Ma-Ma conditions, the proper

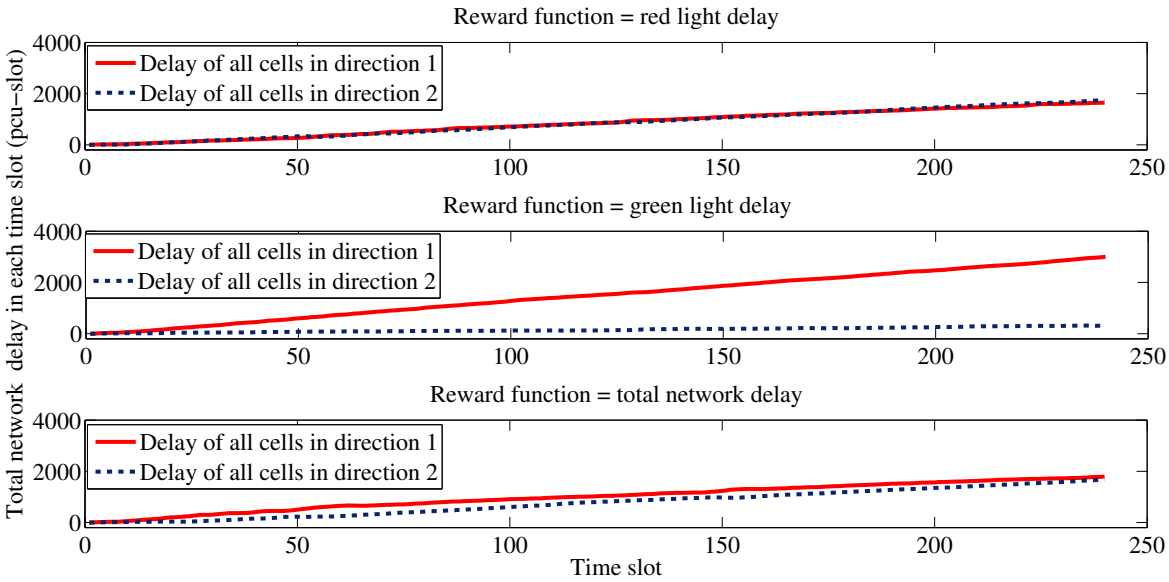

Figure 3.14 Ma-Ma: three types of reward functions and its delay in each component

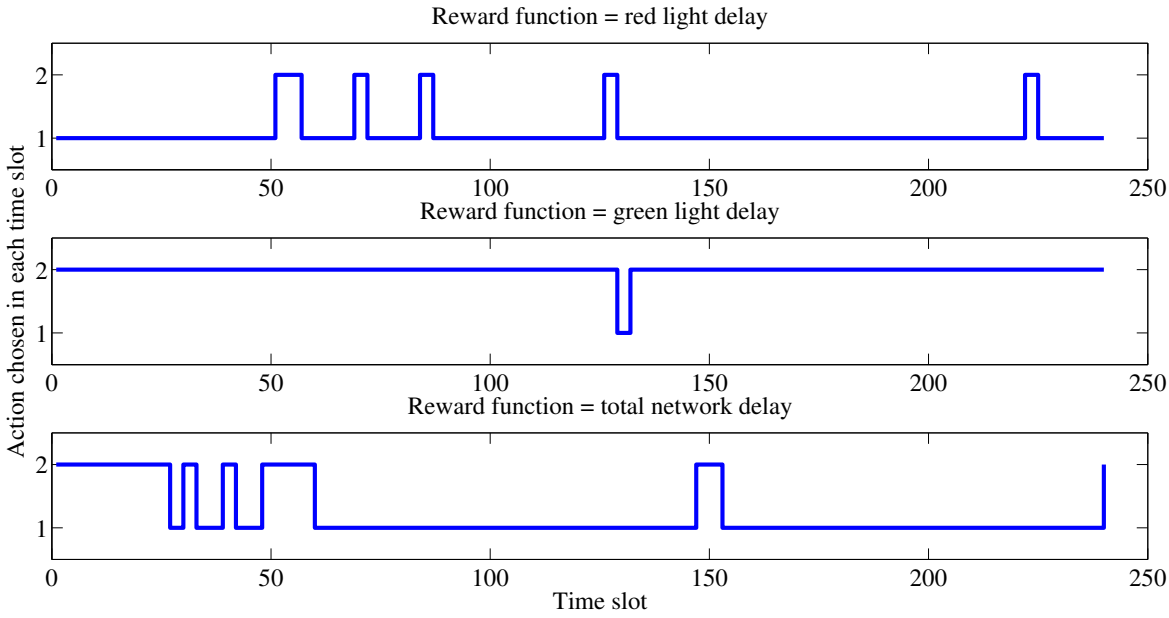

Figure 3.15 Ma-Ma: action chosen in each time slot

management of traffic signal control becomes a major concern. The next recommended traffic signal would preferably remain the same as the current traffic signal to avoid the occurrence of the system loss time. The any reward function can be used because the traffic is jammed. The reduction of total network delay becomes insignificant.

The goal is for Q-learning to minimise the total network delay. Surprisingly, by using the total network delay as a reward function, the results were not necessarily as good as initially expected. Rather, both simulation and mathematical derivation results confirm that using the newly proposed red light delay as the Q-learning reward function gives better performance than using the total network delay as the reward function. Note that a good reward function must be able to allow the algorithm to steer its instantaneous searching directions towards the final goal of minimising the total network delay. But that reward function itself needs not be the objective function i.e. the total network delay. Instead, from our numerical experiments, one should rather opt for using the red-light delay as the reward function so that the effect on future expected total network delay can be reflected within only a few time slots after an action decision has been made. On the contrary, if the total network delay is used as the reward function, then the algorithm eventually cannot find the proper solution.

3.3.3 Q-Learning Performance in Stationary/Non-Stationary Stochastic Loadings

In the Q-learning validation section, four different traffic demand patterns have been investigated. In fact, such simplification can be relaxed to more realistic case by considering on the random source probabilities. Let the traffic demand be a Poisson process with a constant arrival rate for each direction. From the previous subsection, the red light delay has been chosen as a reward function. The performance of Q-learning in adapting its solution to reach the convergence will be examined. The experiments have been set into two scenarios. Firstly, the stationary test, the change of traffic demand from a deterministic to a Poisson has been illustrated in Figure 3.16. Secondly, the non-stationary test, in reality, road network capacity changes upon time (early morning, rush hour, etc.) as illustrated in Figure 3.17. Starting from uncongested traffic condition, the 1^{st} episode until the 100^{th} episode, the traffic demand pattern is $\{\lambda_1, \lambda_2\} = \{6, 6\}$ pcu/slot. And then, the road network becomes congested (jammed) condition, the $101^{st} - 140^{th}$ episodes, traffic demand pattern is therefore changed to $\{\lambda_1, \lambda_2\} = \{13, 3\}$ pcu/slot. The congested condition returns to uncon-

gested condition, the $141^{st} - 180^{th}$ episodes, the traffic demand pattern is $\{\lambda_1, \lambda_2\} = \{6, 6\}$ pcu/slot. Finally, the congested condition happened again, the episodes 181^{st} the traffic demand pattern is $\{\lambda_1, \lambda_2\} = \{11, 5\}$ pcu/slot.

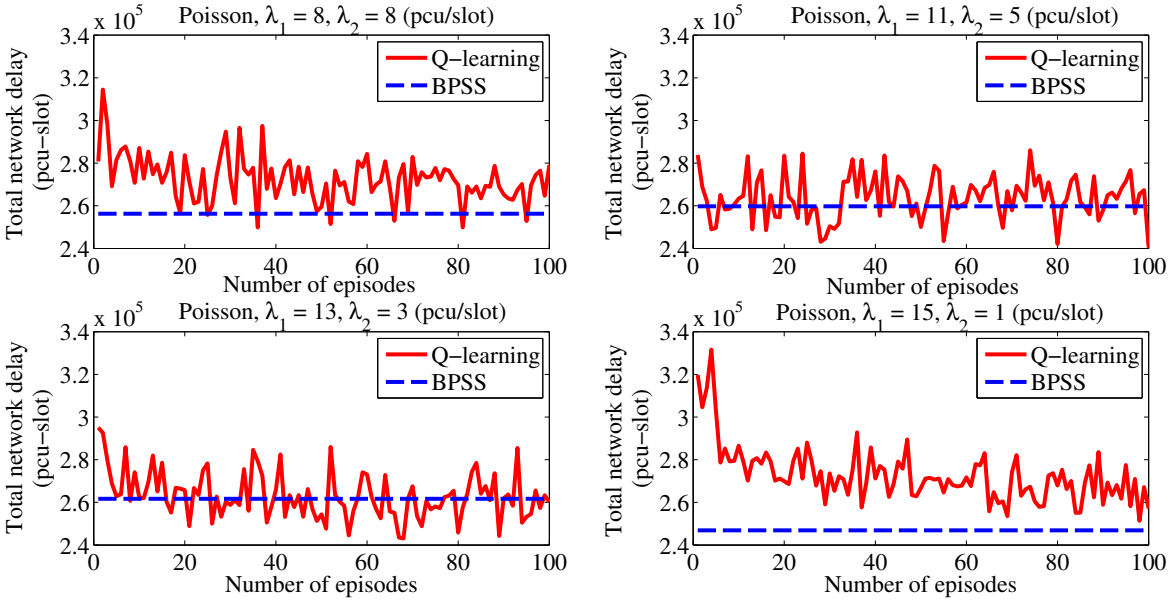

Figure 3.16 Total network delay from Q-learning with Poisson arrival

The results show the adaptability of Q-learning in reaching the solution close to the obtained solution from the BPSS method in both experiments. The abrupt change of the traffic demand patterns from uncongested to congested conditions have been imposed. However, the Q-learning still performs well in tracking closer to the BPSS solution. Therefore, with significantly less demanding computational time than BPSS, the Q-learning algorithm can be used in real-time learning-based scenarios.

3.3.4 Measure of Effectiveness Using Microscopic Traffic Simulator

To evaluate the performance between the Q-learning and the BPSS, the signal plans from both algorithm from the macroscopic level is set as the control plans in the microscopic traffic simulator AIMSUN. The traffic demand patterns have been divided into 4 loading patterns which are $\{\lambda_1, \lambda_2\} = \{8, 8\}$, $\{\lambda_1, \lambda_2\} = \{15, 1\}$, $\{\lambda_1, \lambda_2\} = \{13, 3\}$ and $\{\lambda_1, \lambda_2\} = \{11, 5\}$ pcu as patterns 1,2,3 and 4, respectively. The simulation testing in AIMSUN has been set to 2 hours. The system throughput can be calculated by

$$Throughput = \frac{V_{pass}}{V_{total}} \times 100, \tag{3.31}$$

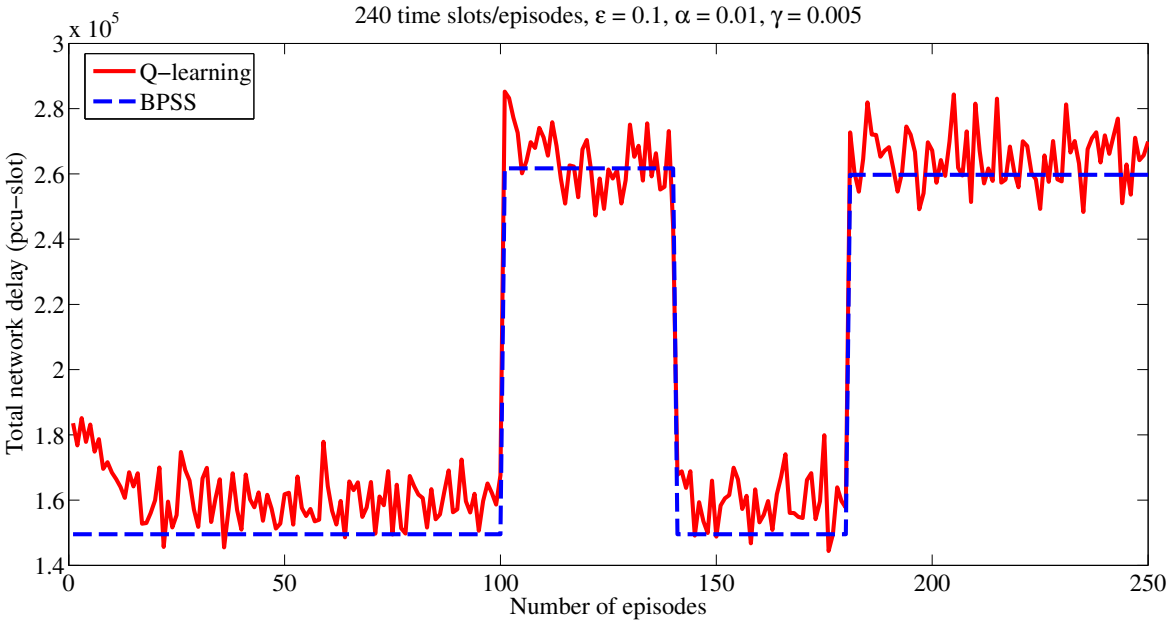

Figure 3.17 Total network delay obtained from Q-learning with varied load patterns

where V_{pass} is the number of vehicles that can pass the intersection and V_{total} is the total number of vehicles in the road system.

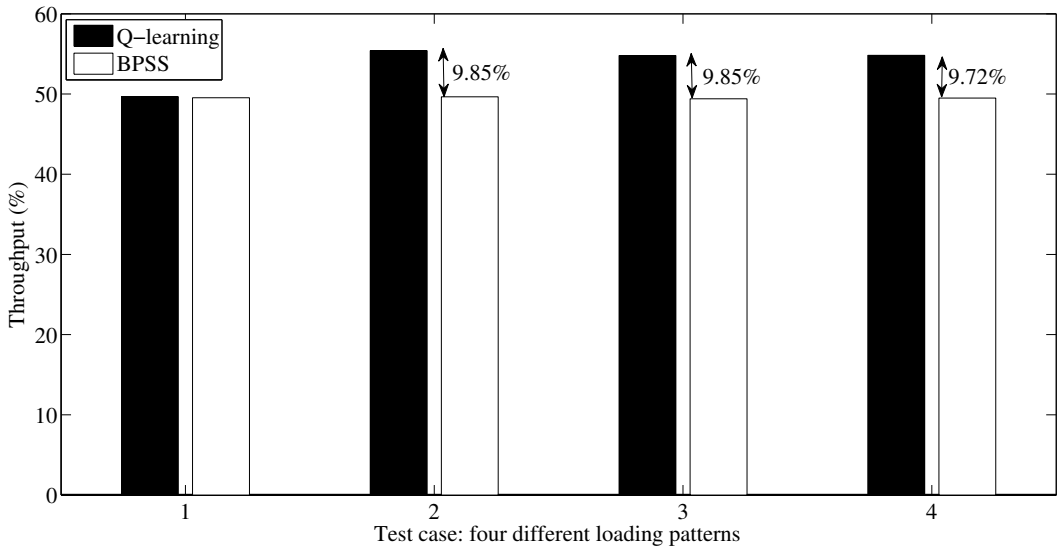

Figure 3.18 Throughput comparison between Q-learning and BPSS

As illustrated in Figure 3.18, the results obtained from the Q-learning outperform the BPSS method at a microscopic level in patterns 2, 3 and 4 (asymmetric loadings) by 9.85%, 9.85% and 9.72%, respectively. The relative improvement from BPSS by using Q-learning occurs from the change of green time allocation in each signal cycle. In particular, the Q-

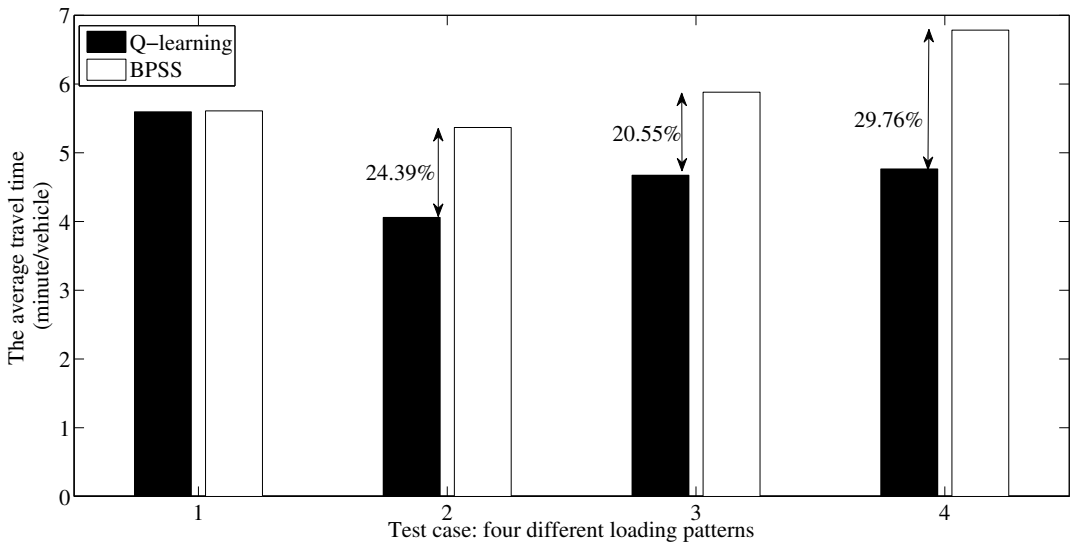

Figure 3.19 Average travel time comparison between Q-learning and BPSS

learning uses an aperiodic signal control whereas the BPSS uses a periodic signal control. The advantage of the aperiodic control to the periodic control is in terms of the adaptability that gives the solution properly.

Figure 3.19 illustrates the reduced average travel time per vehicle. The Q-learning also achieves a better performance in comparison with the BPSS in terms of reducing the average travel time per vehicle by 24.39%, 20.55% and 29.76%, respectively. With the aperiodic signal control from Q-learning, the signal can be changed upon the change of traffic conditions thereby reducing the waiting time and resulting in smoother flow of vehicles.

3.4 Summary

A new framework to control the traffic signal lights by applying one of the reinforcement learning tools, namely, the Q-learning has been proposed to seek the best possible solution to control the traffic signals where the network state has been modelled by the signalised cell transmission model. The road traffic condition is mainly focused on the situation when the summation of overall traffic demand from all directions exceeds the maximum flow capacity.

In addition, the existing works related to Q-learning have not considered the scalability issues due to the limitation in terms state space explosion. However, we attempt to alleviate the explosion by employing state space quantisation and control traffic signal in such network

scenarios.

The proposed framework is used to find the best traffic signal strategy. Surprisingly, using the newly proposed red light delay as the Q-learning reward function gives better performance than using the total network delay as the reward function. The results have been reported from the series of experiments which are the Q-learning validation, the effect of reward functions, the Q-learning performance in stationary/non-stationary stochastic loadings and the applicability of the CTM-based solution of the Q-learning algorithm in the microscopic mobility environments using AIMSUN.

The simulation results show that our proposed framework can computationally efficiently find the proper solution for road traffic systems by comparing with the best periodic signal solution (BPSS). The effect of reward functions has also been investigated and the adaptability of the Q-learning algorithm in adjusting its solution with Poisson arrival upon the change of time has also been observed. The results from the macroscopic level show that Q-learning yields the results similar to the BPSS method. However, in a microscopic level, the control strategies obtained from the CTM-based Q-learning approach outperform the BPSS in terms of the throughput and the average travel time.

With the newly proposed reward function applied to an isolated intersection, this chapter has reported the results and its applicabilities. The BPSS is no longer inapplicable due to its computational burden required. In the next chapter, our proposed CTM-based Q-learning will be compared with the classical mathematical M/M/1 and D/D/1 queuing models.

CHAPTER IV

PERFORMANCE COMPARISON OF QUEUEING THEORETICAL OPTIMALITY AND Q-LEARNING

This chapter addresses the performance comparison of optimal traffic signal controls based on two frameworks: M/M/1 and D/D/1 queueing models, and Q-learning approach. In Section 4.1, using the M/M/1 and D/D/1 models, the optimal split derivation has been obtained to minimise the mean waiting time of an intersection. In Section 4.2, the Q-learning framework has been proposed in conjunction with the use of the macroscopic cell transmission model (CTM) to update the vehicle state dynamics upon Q-learning actions. These Q-learning actions adjust the split adaptively and appropriately. In Section 4.3, the exact implementation for the Q-learning algorithm has been emphasised. The two approaches, namely the steady-state analysis of M/M/1 and D/D/1 as well as the Q-learning, have been compared in terms of the achievable network throughput and the average vehicle delay per completed trip in various loading scenarios from undersaturated towards jamming conditions. The main finding to this chapter is obtained from Section 4.4 in finding the best proper traffic signal to control road systems in different traffic patterns. Section 4.5 concludes and expresses the possibility of adapting the Q-learning in a network scale scenario with the bus rapid transit in **Chapter V**.

4.1 Queueing Traffic Model

This section introduces a simplified queueing model with two buffers and a single server, which can be mapped into two conflicting flows in an isolated intersection.

Figure 4.1 illustrates an isolated intersection which serves two flows from west to east and north to south. Figure 4.1 can be converted into a basic queueing model with two buffers and a single server as shown in Figure 4.2, where λ_p denotes the traffic arrival rate of the system for direction $p = 1, 2$. Let w_p be the ratio of green time allocated to direction p (or its split) in a signal cycle. The objective here is to find the optimal split w_p^* that minimises

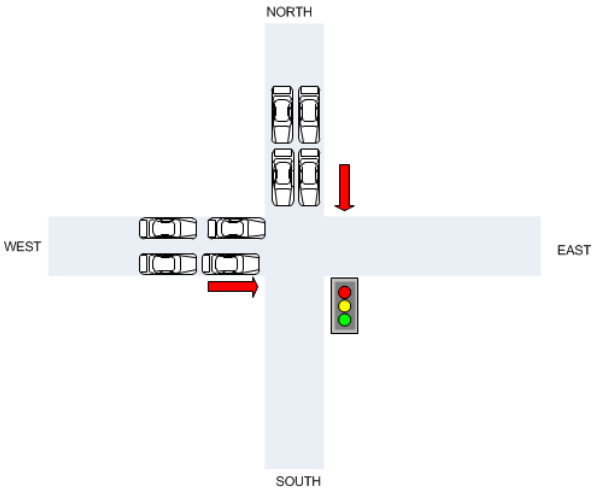

Figure 4.1 Model for two conflicting flows in isolated intersection

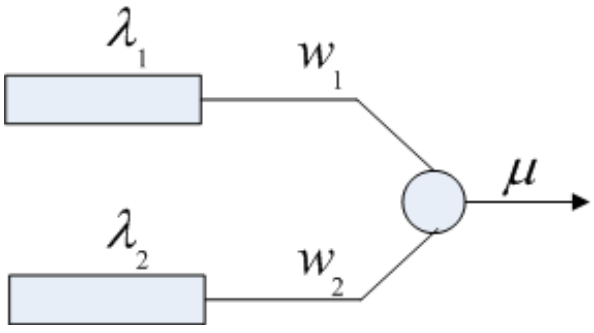

Figure 4.2 Queueing model with two incoming requests

the mean waiting time of the considered intersection system.

4.1.1 Steady State Analysis by M/M/1

The steady-state derivation is based on an M/M/1 queueing model where the vehicle arrivals in each direction are assumed to be an independent Poisson process and, during their green time period, each vehicle is assumed to spend exponentially distributed travel time through the intersection. As illustrated in Figure 4.1, an intersection has two individual conflicting flows with mean arrival rates λ_1 and λ_2, respectively. Let μ be the saturation flow rate, the flow rate at which vehicles can pass through a signalised intersection in a stable moving queue [50]. Let ρ_p be the offered load in direction p so $\rho_p = \frac{\lambda_p}{w_p \mu}$ for $p = 1, 2$. To guarantee the stability condition of the system, it is assumed that the intersection's saturation flow rate is greater than the total input flow rate from all approaching directions. Let L denote the total loss time value per signal cycle being normalised by the cycle period. Thus,

$$\sum_{\forall p} w_p + L = 1, \tag{4.1}$$

$$\sum_{\forall p} w_p < 1. \tag{4.2}$$

In the queueing steady state, the mean waiting time T_p in the system for direction p can then be obtained as follows [51]

$$\begin{aligned} T_p &= \frac{\rho_p}{1-\rho_p} \\ &= \frac{\lambda_p}{w_p \mu - \lambda_p}. \end{aligned} \tag{4.3}$$

The total network delay T is given by

$$T = \sum_{\forall p} T_p. \tag{4.4}$$

Thus,

$$w_p \mu > \lambda_p \qquad \text{for system stability} \tag{4.5}$$

$$\sum_{\forall p} w_p \mu > \sum_{\forall p} \lambda_p \tag{4.6}$$

$$\because \quad \sum_{\forall p} w_p < 1 \tag{4.7}$$

$$\therefore \quad \sum_{\forall p} \lambda_p < \mu. \tag{4.8}$$

To minimise the total network delay, differentiating T in (4.4) with respect to w_1 and equating it to zero finally gives:

$$0 = \frac{\partial}{\partial w_1}\left[\frac{\lambda_1}{w_1\mu - \lambda_1}\right] + \frac{\partial}{\partial w_2}\left[\frac{\lambda_2}{w_2\mu - \lambda_2}\right]\frac{\partial w_2}{\partial w_1}. \tag{4.9}$$

Therefore, the equation becomes

$$\frac{\lambda_1\mu}{(w_1\mu - \lambda_1)^2} = \frac{\lambda_2\mu}{(w_2\mu - \lambda_2)^2}, \tag{4.10}$$

where

$$\frac{\partial w_1}{\partial w_1} + \frac{\partial w_2}{\partial w_1} + \frac{\partial L}{\partial w_1} = \frac{\partial 1}{\partial w_1} \tag{4.11}$$

$$\frac{\partial w_2}{\partial w_1} = -1. \tag{4.12}$$

Replacing $\frac{\partial w_2}{\partial w_1} = -1$ in (4.9) gives.

$$0 = \frac{-\lambda_1\mu}{(w_1\mu - \lambda_1)^2} + \frac{-\lambda_2\mu}{(w_2\mu - \lambda_2)^2}(-1). \tag{4.13}$$

As a result, the optimal split can then be written by

$$\frac{\lambda_1\mu}{(w_1\mu - \lambda_1)^2} = \frac{\lambda_2\mu}{(w_2\mu - \lambda_2)^2}. \tag{4.14}$$

$$(w_1\mu - \lambda_1) = \varsigma(w_2\mu - \lambda_2) \tag{4.15}$$

$$w_1 - w_2\varsigma = \frac{\lambda_1 - \lambda_2\varsigma}{\mu} \tag{4.16}$$

$$w_1 + w_2 = 1 - L, \tag{4.17}$$

where $\varsigma = \sqrt{\lambda_1/\lambda_2}$. Subtracting (4.17) - (4.16) finally gives

$$w_2(1+\varsigma) = \left[1 - L - \frac{(\lambda_1 - \lambda_2\varsigma)}{\mu}\right] \tag{4.18}$$

$$w_2 = \frac{\left[1 - L - \frac{(\lambda_1 - \lambda_2\varsigma)}{\mu}\right]}{(1+\varsigma)}. \tag{4.19}$$

Replacing w_2 in (4.17) gives.

$$w_1 + \frac{\left[1 - L - \frac{(\lambda_1 - \lambda_2\varsigma)}{\mu}\right]}{(1+\varsigma)} = 1 - L \tag{4.20}$$

$$w_1(1+\varsigma) + \left[1 - L - \frac{(\lambda_1 - \lambda_2\varsigma)}{\mu}\right] = (1-L)(1+\varsigma) \tag{4.21}$$

$$w_1(1+\varsigma) - \frac{(\lambda_1 - \lambda_2\varsigma)}{\mu} = \varsigma(1-L) \tag{4.22}$$

$$w_1 = \frac{\left[\varsigma(1-L) + \left(\frac{\lambda_1 - \lambda_2\varsigma}{\mu}\right)\right]}{(1+\varsigma)} \tag{4.23}$$

Therefore, the optimal split from M/M/1 model $w^*_{1,mm1}$ and $w^*_{2,mm1}$ can be expressed finally as

$$\begin{aligned} w^*_{1,mm1} &= \frac{\left[\varsigma\,(1-L)+\left(\frac{\lambda_1-\lambda_2\varsigma}{\mu}\right)\right]}{(1+\varsigma)}, \\ w^*_{2,mm1} &= \frac{\left[(1-L)+\left(\frac{\lambda_2\varsigma-\lambda_1}{\mu}\right)\right]}{(1+\varsigma)}. \end{aligned} \tag{4.24}$$

This result from equation (4.24) represents an optimal split weighted to each individual flow.

4.1.2 Steady State Analysis by D/D/1

The arrival process and queueing service time may not be Poisson and exponential. Another model, D/D/1, has also been used, where the incoming stream of vehicles arrives at a fixed deterministic rate and their service time through the intersection is assumed constant for every vehicle. Similar to (4.1) - (4.2) of M/M/1 case, the splits $w_{p,dd1}$ of D/D/1 systems must be constrained by

$$\sum_{\forall p} w_{p,dd1} + L = 1, \tag{4.25}$$

$$\sum_{\forall p} w_{p,dd1} < 1. \tag{4.26}$$

The mean waiting time $T_{p,dd1}$ in the D/D/1 system for direction p can be obtained as follows [52]

$$T_{p,dd1} = \frac{\lambda_1}{w_{p,dd1}\mu}. \tag{4.27}$$

Recall the stability condition as given in (4.5) – (4.8). Equate $T_{p,dd1}$ in (4.27) with respect to $w_{1,dd1}$ gives:

$$0 = \frac{\partial}{\partial w_{1,dd1}}\left[\frac{\lambda_1}{w_{1,dd1}\mu}\right] + \frac{\partial}{\partial w_{2,dd1}}\left[\frac{\lambda_2}{w_{2,dd1}\mu}\right]\frac{\partial w_{2,dd1}}{\partial w_{1,dd1}} \tag{4.28}$$

$$0 = \frac{\lambda_1\mu}{(w_{1,dd1}\mu)^2} + \frac{\lambda_2\mu}{(w_{2,dd1}\mu)^2}(-1). \tag{4.29}$$

Therefore,

$$\frac{\lambda_1}{(w_{1,dd1}\mu)^2} = \frac{\lambda_2}{(w_{2,dd1}\mu)^2} \tag{4.30}$$

$$\varsigma(w_{2,dd1}) = w_{1,dd1}, \tag{4.31}$$

$$w_{1,dd1} + w_{2,dd1} = 1 - L, \tag{4.32}$$

where $\varsigma = \sqrt{\lambda_1/\lambda_2}$. Substituting (4.31) into (4.32) gives

$$\begin{aligned} \varsigma w_{2,dd1} + w_{2,dd1} &= 1 - L, & (4.33)\\ w_{2,dd1}(1+\varsigma) &= 1 - L, & (4.34) \end{aligned}$$

The optimal split can then be written as

$$w^*_{2,dd1} = \frac{1-L}{(1+\varsigma)} \quad (4.35)$$

To find $w^*_{1,dd1}$, combine (4.35) and (4.32). The optimal split $w^*_{1,dd1}$ can be obtained by

$$\begin{aligned} w_{1,dd1} + \frac{1-L}{(1+\varsigma)} &= 1 - L, & (4.36)\\ &= (1-L) - \frac{1-L}{(1+\varsigma)}, & (4.37)\\ &= \frac{(1+\varsigma)(1-L) - (1-L)}{1+\varsigma}\\ w^*_{1,dd1} &= \frac{\varsigma(1-L)}{(1+\varsigma)} \end{aligned}$$

As a summary, the optimal split of D/D/1 model becomes

$$\begin{aligned} w^*_{1,dd1} &= \frac{\varsigma(1-L)}{(1+\varsigma)}\\ w^*_{2,dd1} &= \frac{1-L}{(1+\varsigma)}. & (4.38) \end{aligned}$$

4.2 Problem Formulation for Comparing Q-learning with Queueing Models

In this chapter, the traffic signal control obtained from the CTM-based Q-learning algorithm will be compared with the optimal derivation from M/M/1 and D/D/1 queueing models. As illustrated in Figure 4.2, an intersection with two conflicting flows has been introduced. Conventionally, two traffic signal control techniques, the Q-learning and the queueing models cannot be compared directly because the difference of vehicle movements in a road system. The vehicle movements in queueing models are not taken into account. Therefore, the CTM being used throughout this chapter has been slightly modified. The vehicle movements from cell to cell have been overlooked. Unlike the previous chapter, the state space quantisation has not applied. For an intersection, a road link is not necessarily

homogeneously divided into small sub-cells but instead considered as a CTM cell at an intersection only. The following subsections show how to formulate the CTM-based Q-learning algorithm to compare with the optimal split derived from M/M/1 and D/D/1 queueing models.

4.2.1 State Space

Define $\mathcal{S}$ as the state space of the intersection system with two conflicting flows. Let $\mathbf{s} \in \mathcal{S} \subset \mathbb{Z}_+^2$ be the state vector which represents the total number of vehicles waiting for the green light at the intersection. Let $s_p(t)$ be the state variable which represents the number of vehicles in direction p at time instance t where $p = 1, 2$. Therefore, the state space $\mathcal{S}$ of all vehicle profiles in the system is given by

$$\mathcal{S} := \{\mathbf{s} = [s_1(t), s_2(t)]\}\,. \tag{4.39}$$

4.2.2 Cell Transmission Model

CTM [41] is here employed to update the Q-learning state dynamics. CTM captures the effect of control actions decided by Q-learning on the flow of vehicles in the system. The updating state depends on the green time allocated to each of approaching directions. The updating process of CTM can be summarised as follows.

4.2.2.1 Sending Capability at Intersection

Let $y_p(t)$ be the number of vehicles that can pass through the intersection in direction p at time step t:

$$y_p(t) = \min\{s_p(t), q_p(t)\}\,, \tag{4.40}$$

where $q_p(t)$ represents the maximum flow rate at which vehicles can flow from their intersection upstream to downstream road segments along each direction p at time step t.

4.2.2.2 Receiving Capability at Intersection

The receiving capability in CTM normally depends on the maximum flow rate $q_p(t)$as

$$r_p(t) = \min\{q_p(t), \varepsilon_p(t)\}\,, \tag{4.41}$$

where $\varepsilon_p(t)$ denotes the residual capacity in direction p at time step t.

4.2.2.3 Flow Conservation at Intersection

The state dynamics of CTM can then be updated in according to the chosen action in each time step as

$$s_p(t+1) = s_p(t) + x_p(t) - y_p(t), \tag{4.42}$$

where $x_p(t)$ represents the newly incoming demands in direction p at time step t.

4.2.3 Action Space

In each interval, the agent must select whether it would remain in the current signal indication or change it. The decision is referred to as an *action*. The action space, denoted by $\mathcal{A}$, is the set of all possible actions which the traffic signal controller of the considered intersection can take. Action $a \in \mathcal{A}(\mathbf{s})$ refers to the action which the agent can take at state $\mathbf{s}$.

4.2.4 Vehicle Delay

Vehicle delay is defined as the number of vehicles that cannot pass through the intersection. The vehicle delay accumulated at time step t (passenger car unit slot: pcu-slot) can be expressed as

$$d_p(t) = s_p(t) - y_p(t). \tag{4.43}$$

Note that if the allocated green time can serve all traffic in $s_p(t)$, i.e., $s_p(t) = y_p(t)$, then there is no delay happening. In each time step, dividing the total number of vehicles in (4.43), the actual delay can be found.

4.2.5 Performance Criteria

The aim of Q-learning here is to find the optimal policy that minimises the total network delay, which can be expressed in terms of the delay $d_p(t)$ at each time step t as:

$$\begin{aligned} \Upsilon(t) &= \sum_{\forall p} d_p(t) \\ &= \sum_{\forall p} \left(s_p(t) - y_p(t)\right). \end{aligned} \tag{4.44}$$

Note that $q_p(t)$ is affected by the action a, which specifies the direction that receives the green light as follows

$$q_p(t) = \begin{cases} \mu & \text{, direction } p \text{ gets green light} \\ 0 & \text{, direction } p \text{ gets red light} \end{cases}. \tag{4.45}$$

Equation (4.45) represents an action which allows the vehicles to pass through the intersection in direction p at time step t.

4.3 Signal Optimisation by Q-learning for Simplified Isolated Intersection

Table 4.1 depicts the standard Q-learning algorithm [32] which is applied to solve the problem formulated as an MDP.

Table 4.1 Psuedo-code of Q-learning algorithm

1. Initialise $Q(s, a)$ arbitrarily (here, set to zeros).
2. Repeat (for each episode):
3. Initialise s to the state of empty roads
4. Repeat (for each time step of episode):
5. Choose a from $\mathcal{A}(\mathrm{s})$ using policy derived from Q (e.g., we adopt the ϵ-greedy)
6. Take action a, observe Υ, and the next CTM state s' as the result of the taken action
7. Update the action value function: $Q(s,a) \leftarrow Q(s,a) + \alpha\,[\Upsilon + \gamma min_{a'} Q(s',a') - Q(s,a)]$
8. Update to the next CTM state: $s \leftarrow s'$;
9. until the end of simulation period.

In Table 4.1, $Q(s, a)$ represents the *action value* function representing the average future reward expected to be incurred given that the action a has been taken at the state s [32]. According to the epsilon greedy policy, the best apparent action will be selected with high probability of $1 - \epsilon$, and the other actions will be tried out randomly with a small probability of ϵ. Therefore, the best apparent action or *greedy* action is exploited most of the

time. And with probability ϵ, the concept of exploration is enabled to ensure that all of states are adequately visited. The parameter α is a small positive fraction, namely, the step-size parameter which influences the learning rate. Step-size parameter determines how much the new state action value tends towards the newly obtained reward and value of the next state-action pair. The parameter γ represents the discount rate which is used to determine the present value of future reward.

4.4 Results and Discussions

In this section, the research finding from our results will be reported. The reported results are obtained from the MATLAB® and the AIMSUN. Firstly, the optimal split obtained from the CTM-based Q-learning, the queueing model M/M/1 and the queueing model D/D/1 have been calculated from MATLAB®. Secondly, the obtained optimal split is set to the allocation of the green signal in 1 cycle time to each direction where 1 cycle time is 120 seconds. The reported results from the AIMSUN are the network throughput, the link delay, the average vehicle delay per completed trip and the mean queue length, respectively.

For the system environments, suppose the length of each road from the entry of the road to the stop line is 800 metres. The maximum flow rate has been measured from AIMSUN under the condition that the vehicles are unaffected by the red signal. From the measurement, the maximum flow rate is 2.61 pcu/s (passenger car unit per second). The results from AIMSUN have been reported from 1 hour of the simulation time. For the Q-learning environment, an action decision has been chosen every 60 seconds. Using the CTM-based Q-learning approach, the algorithm will repeat the learning process as illustrated in Table 4.1 for 50 episodes to reach the desired accuracy.

Table 4.2 illustrates the nine different sets of traffic arrival where each arrival process is Poisson. The results have been considered into two operation regions, which are the undersaturated and jamming regions, respectively. Note that the simulation settings for all nine cases are identical, except for the approaching demand to an intersection and the allocated green time. In fact, the undersaturated traffic conditions occur when the vehicle arrival rate is less than the maximum flow rate. However, if the vehicle arrival rate is greater than the maximum flow rate, then the mathematical solution cannot be solved analytically. The vehicle arrival rates have been varied to produce the offered load ratio varying from 0.2 to 1.8. Note that the optimal derivations are based on the stability condition where the all vehicles

entering the systems can be totally served. However, if all the vehicles entering the systems cannot be totally served, then the accumulative number of vehicles tends to be infinite over time. The queueing models are therefore guaranteed that there is no accumulative queue length when the stability condition is held. Sometimes, the stability is not held, the number of vehicle entering the systems will create the queue to the buffered of the systems. Moreover, the increasing of queueing length at the boundary cell is strongly not recommended. The boundary condition in this dissertation is also included the effects from neighbourhood intersections. However, this dissertation attempts to find the applicable range of Q-learning. Therefore, the jamming conditions have been investigated for the further reporting of the applicable range of Q-learning.

Table 4.2: Proportion of loading patterns corresponding to maximum service rate at considered intersection

Load type	λ_1 pcu/s	λ_2 pcu/s	Offered load ratio (μ)
1	0.435	0.087	$0.2\,\mu$
2	0.87	0.174	$0.4\,\mu$
3	1.305	0.261	$0.6\,\mu$
4	1.74	0.348	$0.8\,\mu$
5	2.175	0.435	$1.0\,\mu$
6	2.61	0.522	$1.2\,\mu$
7	3.045	0.609	$1.4\,\mu$
8	3.48	0.696	$1.6\,\mu$
9	3.915	0.783	$1.8\,\mu$

As illustrated in Figure 4.3, the results show the allocated green time to each direction for each scenario. In D/D/1 queueing model, the optimal split from (4.38) is unaffected by the service rate. Therefore, the optimal split from the D/D/1 depends on the proportion of vehicle arrival rates only.

Figure 4.4 reveals that the improvement of the network throughput in the jamming conditions can be greatly improved by up to 1.7-8.3% from the M/M/1 model and can be significantly improved up to 3.2-14.8% from the D/D/1 model. The network throughput is the ratio between completed trips and incompleted trips. The higher network throughput is, the greater performance of the system achieves. Note that in the undersaturated conditions,

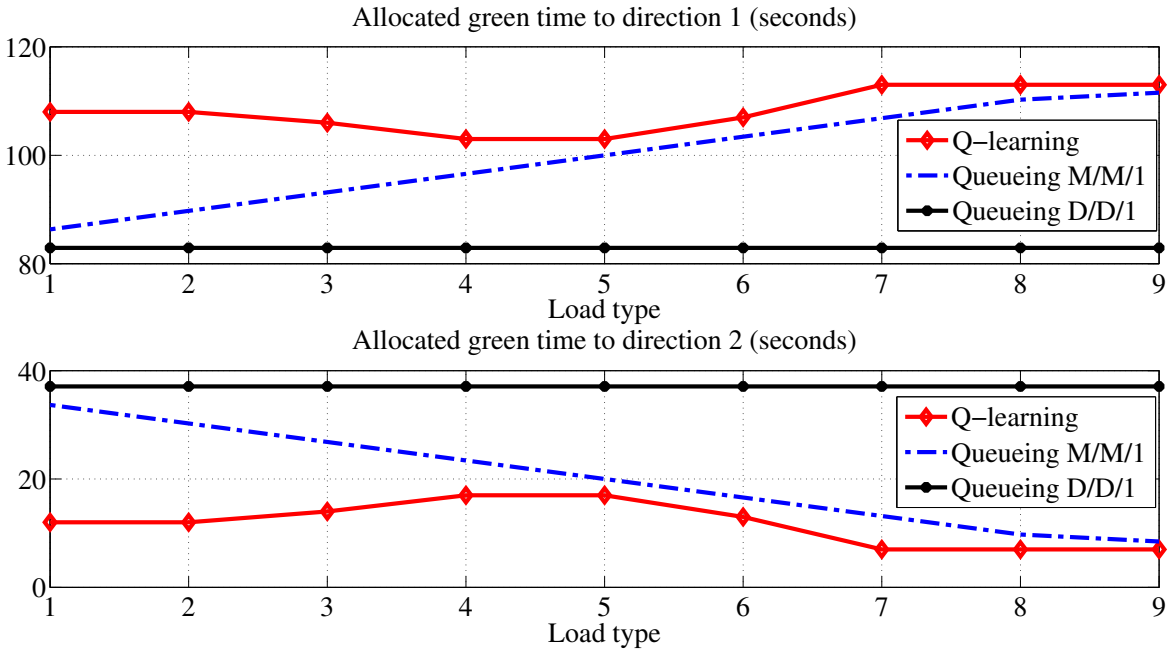

Figure 4.3 Allocated green time to each direction

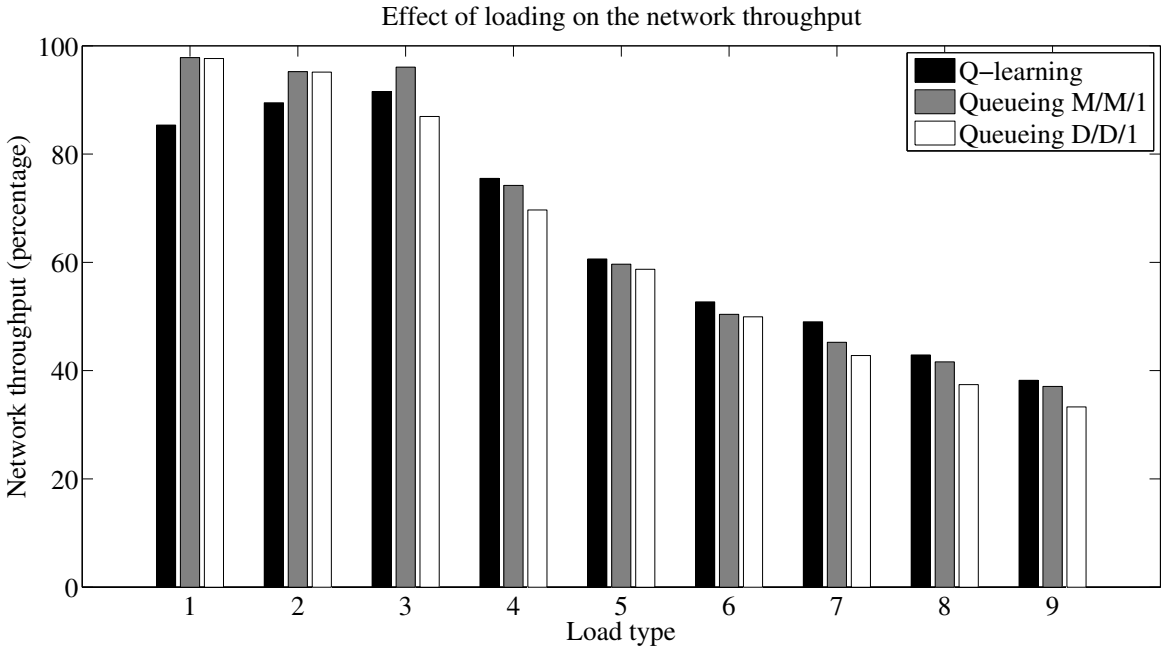

Figure 4.4: Network throughput comparison of Q-learning, Queueing M/M/1 and Queueing D/D/1 models

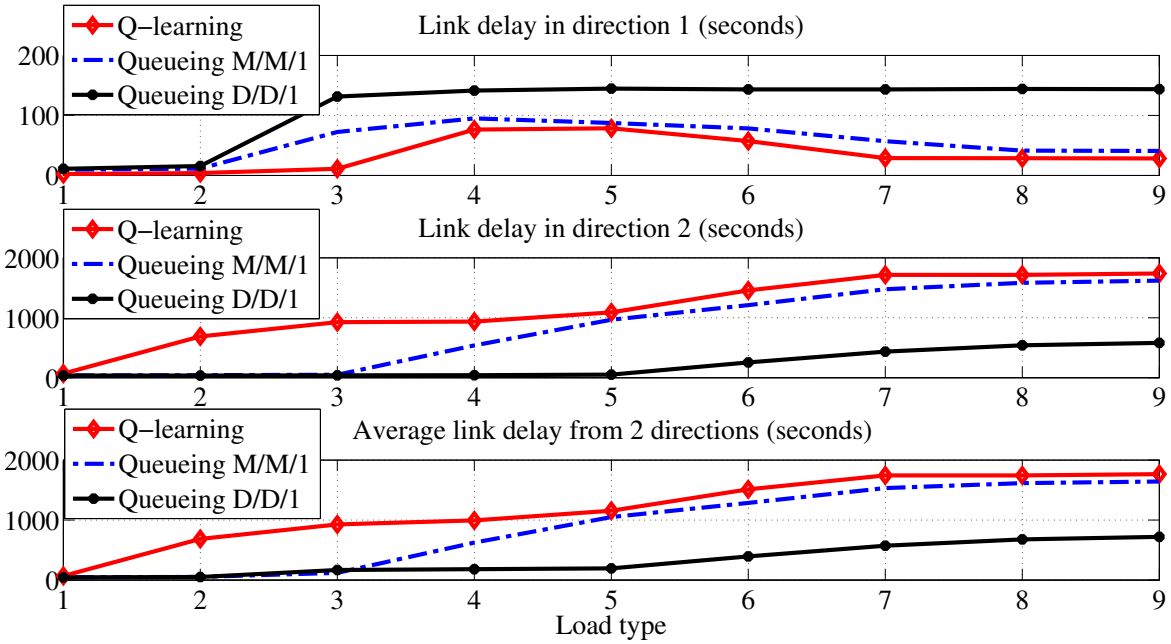

Figure 4.5 Link delay obtained from AIMSUN

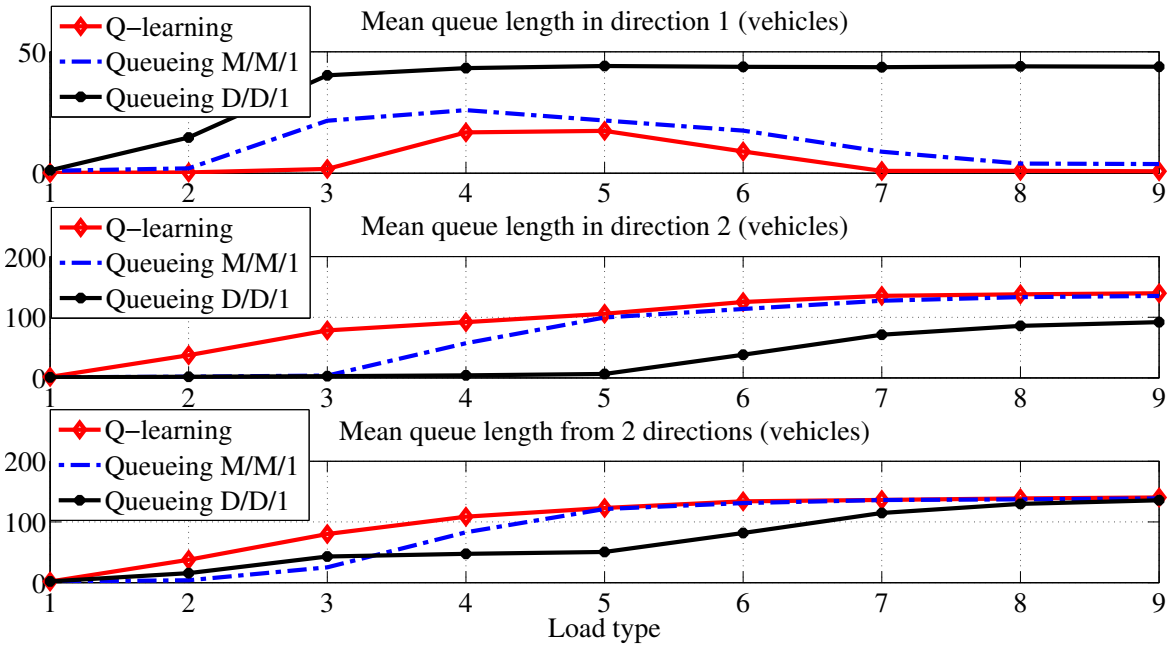

Figure 4.6 Mean queue length obtained from AIMSUN

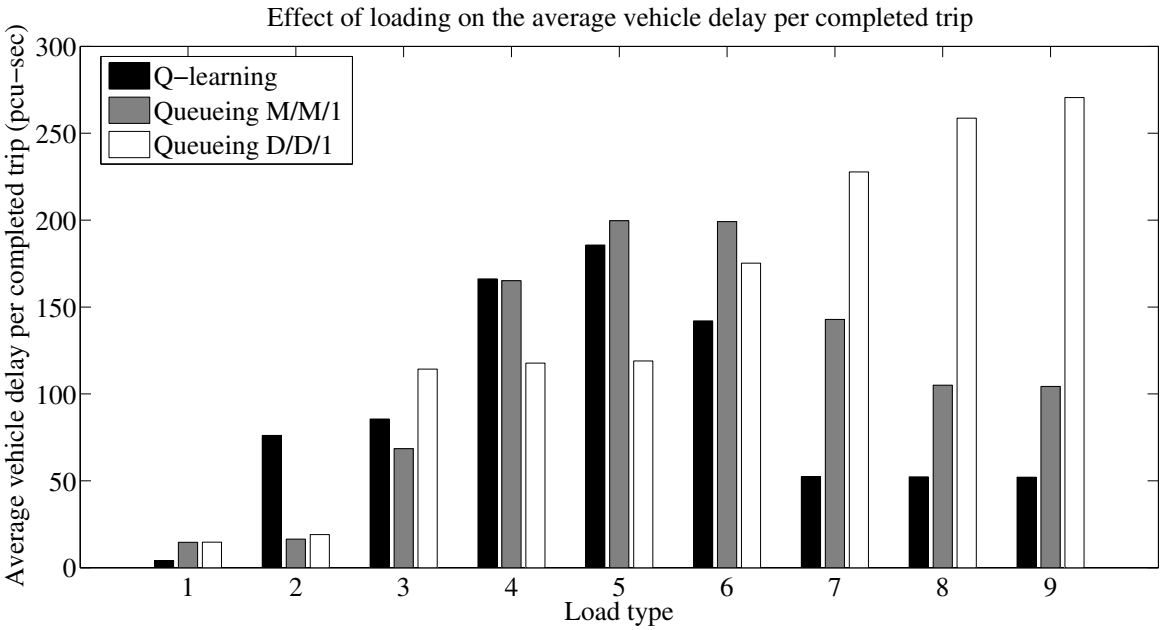

Figure 4.7: Average vehicle delay per completed trip comparison of Q-learning, Queueing M/M/1 and Queueing D/D/1 models

the network throughput by both the M/M/1 and the D/D/1 models outperform the proposed CTM-based Q-learning algorithm. For the undersaturated conditions, the exploratory from Q-learning algorithm chooses green light signals improperly, therefore, the network throughput is not good as expected.

Figure 4.5 explains why Q-learning performs well and badly in different traffic conditions. The link delay is generally known as the difference between the time spent to travel along a particular road and the free flow travel time along the road. Figure 4.5 illustrates the individual link delay for each direction and the average link delay from two directions. In each cycle time, the Q-learning algorithm has allocated the green time more often to the direction with a higher vehicle arrival rate. However, in both queueing models, the allocated green time in each direction is directly proportional to the incoming traffic demand of its direction. Therefore, by using Q-learning in the undersaturated conditions, the obtained optimal split leads the system to the wasted green scenario. However, the link delay of Q-learning performs well in the jamming conditions because Q-learning can reduce the link delay from the higher vehicle arrival rates that dominate the overall link delay of the systems. As illustrated in Figure 4.6, the results for the mean queue length can be explained with the same discussions as the link delay. However, in oversaturated traffic conditions, the network throughput is the most important than both link delay and mean queue length. For

both queueing models, optimal splits derivation applied for the undersaturated conditions perform well. However, for oversaturated conditions, Q-learning is recommended to be used because of its adaptability to the change of traffic arrival patterns.

These three approaches share the common goal of minimising the total network delay. Generally, the total network delay has been calculated from the difference of the time spent to complete a network trip and the free flow travel time along the network path. For each vehicle, the average vehicle delay per completed trip $\widetilde{AD}$ can be calculated by

$$\widetilde{AD} = \frac{\sum\limits_{\forall p}(ALD_p * CPT_p)}{\sum\limits_{p} CPT_p}, \tag{4.46}$$

where ALD_p is the average link delay in direction p and CPT_p is the number of completed trips in direction p. In Figure 4.7, for the jamming condition, the reduction of the average vehicle delay per completed trip can be greatly reduced by up to 7.0-63.4% from the M/M/1 model and can be significantly reduced up to 18.9-80.7% from the D/D/1 model. The reduction of the average vehicle delay can be explained by observing the chosen action from the Q-learning methods. Q-learning tries not to switch its action too often to avoid the system loss time. This is the main reason that is why Q-learning performs better in comparison with the other two queueing models.

4.5 Summary

This chapter evaluates an optimality analysis based on queueing models and compares with Q-learning to control the traffic signal at an isolated two-phase intersection. The Q-learning approach can improve the intersection throughput by up to 1.7-8.3% and by up to 3.2-14.8% in jamming conditions in comparison with the respective M/M/1 and D/D/1 approaches. Moreover, the average vehicle delay per completed trip can be reduced by up to 7.0-63.4% and by up to 18.9-80.7% in comparison with the respective M/M/1 and D/D/1 approaches.

In **Chapter III**, the novel mathematical framework for an isolated traffic signal control has been proposed together with the comparison of the best periodic signal solution (BPSS). This chapter, the strategic comparison between the classical two queueuing models have been proposed. The extension of the CTM-based Q-learning algorithm for a road network scale with the bus rapid transit will be reported in **Chapter V**.

CHAPTER V

TRAFFIC SIGNAL CONTROL WITH Q-LEARNING USING CELL TRANSMISSION MODEL FOR ROAD NETWORK WITH TRANSIT SIGNAL PRIORITY SYSTEM

In **Chapter III** and **Chapter IV**, the newly formulated signalised CTM-based Q-learning framework has been centred on an isolated intersection. For convenience, let us define the vehicle class of non-priority (priority) as vehicles (BRT). In this chapter, the extension to the network scale with bus rapid transit will be investigated. The detailed algorithm and calibration have been carried out in the previous two chapters. The focus of this chapter is then to shed some light on the implication and practical effectiveness of BRT road networks achievable by the Q-learning algorithm. As illustrated in Figure 3.1, three red pin-points in the middle between the pin-point "A" and the pin-point "B" will be chosen as a system example for motivating the model extension in this chapter. This BRT route has been operating in an oversaturated region during the rush hour periods every week day. The main challenge to this particular problem is then how to control the oversaturated road network with the BRT priority system.

In this chapter, Section 5.1 gives the mathematical model extension for the road network with BRT system. Section 5.2 summarises on the implementation of Q-learning in the BRT road network. Section 5.3 shows the results from the in-depth investigation of adopting Q-learning to control the BRT road network. Section 5.4 concludes this chapter.

5.1 Problem Formulation

The following CTM-BRT-based Q-learning framework is here formulated as a distributed (localised) control. For the distributed control, the individual intersection controllers are fully responsible for the change of traffic signal status at their own intersections. The

operation has been taken locally at each intersection. In fact, for the jammed traffic conditions, a centralised control system solution is infeasible due to the computational complexity, the imperfect communication infrastructures, the system overhead and the scalability problem [53]. This dissertation is therefore convinced by the concept of the distributed control to avoid such problems.

To alleviate the computation burdens caused by the curse of dimensionality, the vehicle movements of whole network systems have been considered into two major directions which are conflicting and non-conflicting with the BRT without turning movements. The curse of dimensionality refers to the arisen enormous computations when analysing in the multi-dimensional spaces. The details model of the traffics approaching at an intersection can be found in the next subsection. Moreover, the possible actions have been designed for three phases which assign the right-of-way through each intersection to flows conflicting with BRT, non-conflicting with BRT and of only BRT itself. At all intersections, the decision epochs of their control agents running the Q-learning algorithm are synchronised distributively, i.e., all the actions can be taken at the same time slots. Therefore, the following mathematical formulation can be first written for each intersection individually. However, the mathematical formulation given here in this chapter novelly includes the BRT features i.e., the BRT stations and the separated BRT lanes.

5.1.1 State Space

Suppose the vehicles in the system belong to two classes, i.e., priority (BRT) and non-priority (other vehicles). As illustrated in Figure 3.1, an example of road network with the BRT system has been chosen. The detailed intersections in the observed road segment are illustrated again in Figure 5.1 [48]. The mapping from the real road network to a simplified uni-direction CTM-BRT model has been depicted in Figure 5.2. The consideration on bi-directions has been conveniently overlooked because of no turning movement basis and no interruptions of the traffic from the opposite side of the road. By considering on uni-direction only, the size of state space has been reduced. The reduction of state space can reduce time calculation for the whole road networks systems. In Figure 5.2, the dark arrows represent the directions of vehicle movements from the upstream cells toward the downstream cells without turning movements.

Figure 5.2 illustrates the CTM-BRT model and their subnetworks. In this chapter, this

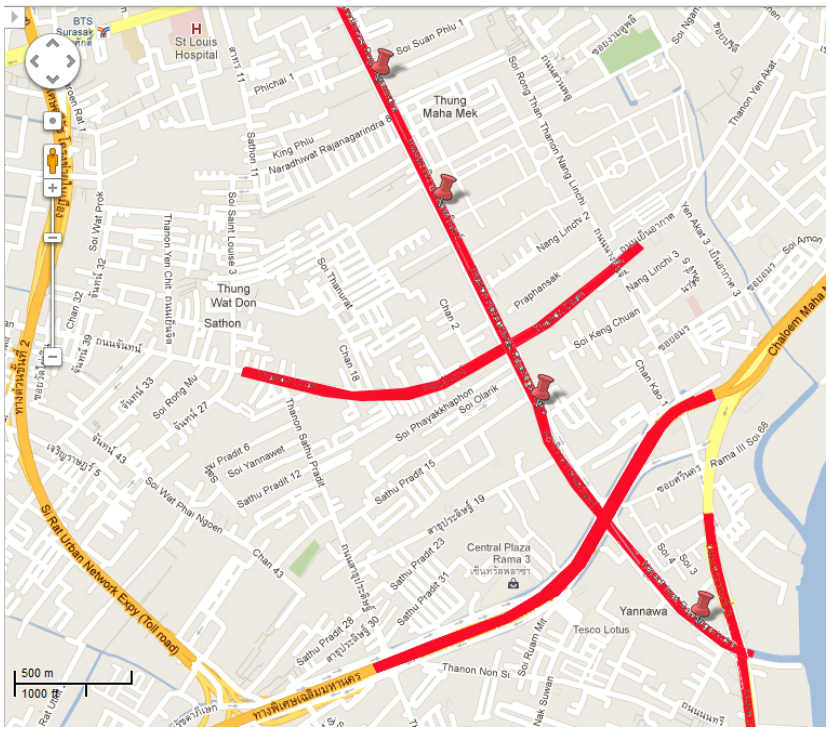

Figure 5.1 Considered BRT route segment in Bangkok

model has divided the whole BRT network into three subnetworks. Each subnetwork has its own control agent for taking the traffic signal actions locally. Figure 5.3 illustrates the control agent's viewpoint to an intersection O where lane separators for the BRT are assumed along the whole length of the road. Let us index here the non-priority cells $i = 1, \ldots, \kappa$ and priority cells $j = 1, \ldots, \kappa_b$. The incoming demand to an intersection is classified into P directions. Let $\mathcal{N}$ be the state space of the system with priority and non-priority classes. Define $\mathcal{S}$ and $\mathcal{B}$ as the state space of non-priority and priority vehicles, respectively. For each non-priority vehicle cell i and priority vehicle cell j in direction p at time slot t, define $s_i^p(t)$ as the number of non-priority vehicles and $b_j^p(t)$ as the number of priority vehicles, respectively. Let $\mathbf{s} \in \mathcal{S} \subset \mathbb{Z}_+^P$ and $\mathbf{b} \in \mathcal{B} \subset \mathbb{Z}_+^P$ be the state vectors which represent the total number of vehicles in the system. The state definition can be defined as the observable state from all the cells in the upstream road segments leading towards the considered intersection from all possible directions. Therefore, the state spaces $\mathcal{N}$, $\mathcal{S}$ and $\mathcal{B}$ of all vehicle profiles in the system are given by

$$\mathcal{N} = \mathcal{S} \times \mathcal{B}, \tag{5.1}$$

$$\mathcal{S} = \{\mathbf{s} = [s_i^p(t),\ \forall(i,p)]\}, \tag{5.2}$$

$$\mathcal{B} = \{\mathbf{b} = [b_j^p(t),\ \forall(j,p)]\}, \tag{5.3}$$

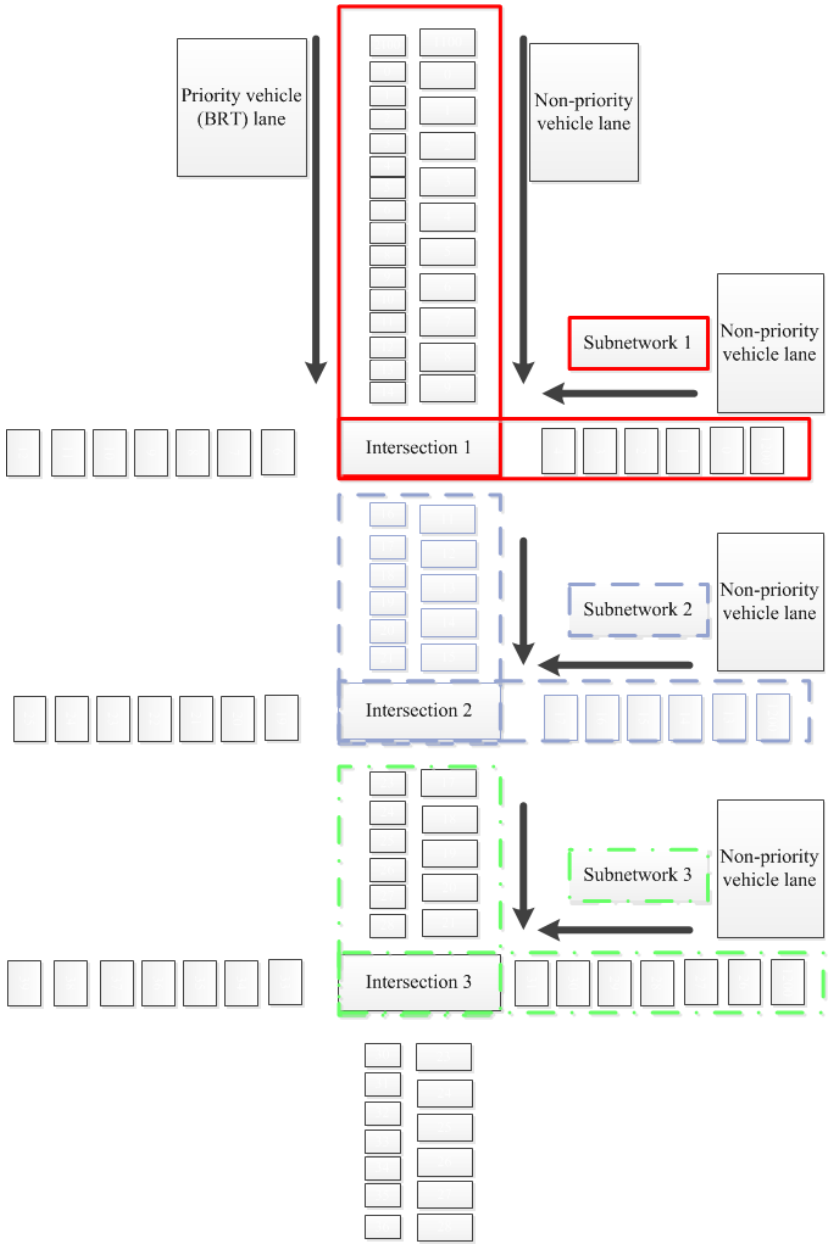

Figure 5.2 CTM-BRT model and their CTM subnetworks

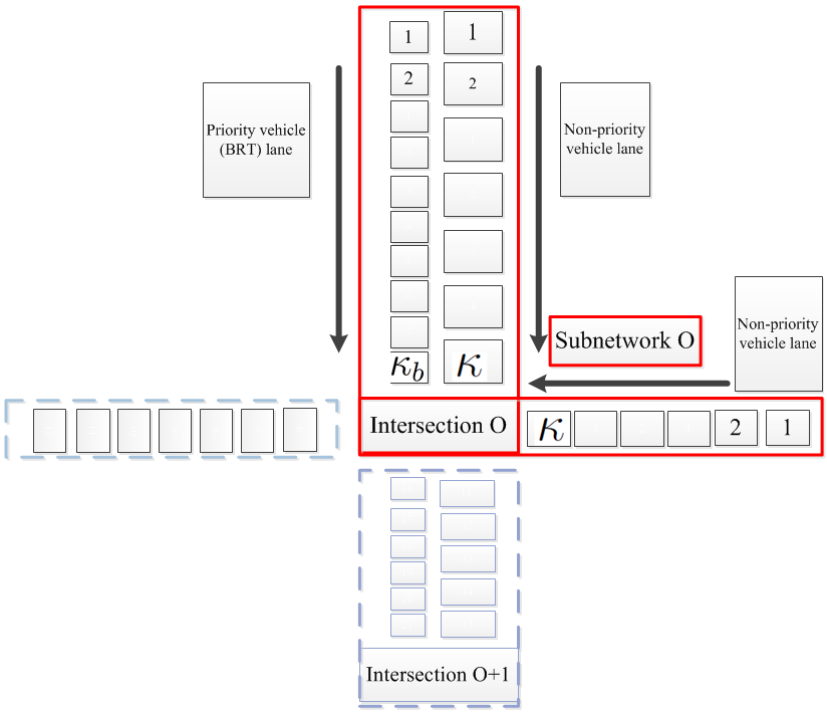

Figure 5.3 CTM cells as seen by control agent at an intersection O

where direction $p = 1, 2, ..., P$; $i = 1, 2, ..., \kappa$ and $j = 1, 2, ..., \kappa_b$. To avoid the computational burden caused by the state space explosion, in this dissertation, the quantisation technique is employed. The level of quantisations can be represented by the number of deployed sensors in the road network. Let $\tilde{s}^p(t)$ be defined as the quantised level for the total number of vehicles approaching the intersection calculated from its upstream road segment in direction p at time slot t:

$$\tilde{s}^p(t) = \left\lceil \frac{\sum_{i=1}^{\kappa} s_i^p(t)}{\mathcal{C}} f \right\rceil + I(\sum_{i=1}^{\kappa} s_i^p(t) = 0), \tag{5.4}$$

where $I(.)$ is the indicator function; $\mathcal{C}$ is the maximum number of vehicles counted from cell i to a signalised cell κ and f is the total number of quantisation levels. The RL state space can be redefined as

$$\tilde{\mathbf{s}}(t) = [\tilde{s}^p(t),\ \forall p] \in \tilde{\mathcal{S}}. \tag{5.5}$$

It should be noted that the quantisation is considered only for non-priority case. For priority vehicles, their movement can be traced individually from the equipped GPS. So, in the microscopic fashion, the quantisation is not needed for the class of priority vehicles.

5.1.2 Cell Transmission Model with Lane-Separated Transit Signal Priority Vehicle

The signalised CTM formulation with a BRT road network using Q-learning is one of the main contributions in this dissertation. The components of the model are presented as follows.

5.1.2.1 Sending Capability

Sending capability represents the ability to send the vehicles from cells to other cells, i.e., moving vehicles from the beginning to ending cells. The sending capability can be defined as

$$\Lambda_i^p(t) = \min\{s_i^p(t), q_i^p(t)\} \quad \text{for non-priority cell } i, \tag{5.6}$$

$$\Lambda_j^p(t) = \min\{b_j^p(t), q_j^p(t)\}, \quad \text{for priority cell } j. \tag{5.7}$$

Here, $q_i^p(t)$ and $q_j^p(t)$ are respectively the maximum number of non-priority and priority vehicles that can flow through their corresponding cells.

5.1.2.2 Receiving Capability

Receiving capability can be calculated by considering the remaining spaces in each cell and the maximum number of vehicles that can be present in the cell. Thus, for non-priority cell i and priority cell j in direction p at time slot t, the receiving capability respectively can be defined as

$$\Psi_i^p(t) = \min\{q_i^p(t), \delta_i^p\,[c_i^p(t) - s_i^p(t)]\}, \tag{5.8}$$

$$\Psi_j^p(t) = q_j^p(t), \tag{5.9}$$

where δ_i^p is the wave speed coefficient and $c_i^p(t)$ is the maximum number of vehicles that can be present. It should be noted that (5.9) assumes no occurrence of the temporal blockage along the cells of a priority vehicle due to the lane separators for the priority vehicle.

5.1.2.3 Cell Cascading

This is the representation of the connection between two adjacent cells, namely, the beginning cell $i-1$ and ending cell i. The number of vehicles that flow in this cascading

scenario from the sending to receiving cells can be written as

$$y_i^p(t) = \min\{\Lambda_{i-1}^p(t), \Psi_i^p(t)\}, \tag{5.10}$$

$$y_j^p(t) = \min\{\Lambda_{j-1}^p(t), \Psi_j^p(t)\}, \tag{5.11}$$

where $y_i^p(t)$ is the number of non-priority vehicles and $y_j^p(t)$ is the number of priority vehicles that flow into cell i and cell j in direction p at time slot t, respectively.

5.1.2.4 Flow Conservation

Flow conservation is used to update the number of vehicles for the next time slot:

$$s_i^p(t+1) = s_i^p(t) + y_i^p(t) - y_{i+1}^p(t), \tag{5.12}$$

$$b_j^p(t+1) = b_j^p(t) + y_j^p(t) - y_{j+1}^p(t). \tag{5.13}$$

5.1.3 Action Space

To influence the system dynamics, upon a decision epoch, the control agent at each intersection must select whether it would keep the current signal indication or change it. The decision is called action. At state vector $\tilde{\mathbf{s}}$ and $\mathbf{b}$, an action must be selected from a state-dependent set $\mathcal{A}$. Specifically, $\mathcal{A}$ is the set of all possible actions which the control agent can take. Define action as the phase of signal light to be chosen at time slot t. The control agent at each intersection must be perfectly synchronised. Otherwise, the status of the system will be totally degraded. The degradations of the system can be caused by e.g., the wasted green and temporal blockage of the conflicting vehicles. In the proposed model, the right-of-way of a priority vehicle is allowed to coexist with the non-conflicting flows of of non-priority vehicles. The main reason is to maximise the efficiency for traffic signal indication. The decision on changing action is made at every T time slots. Define indicator functions at time slot t as follows.

$$G^p(t) = \begin{cases} 1, & \text{non-priority vehicles in direction } p \text{ gets green light,} \\ & \text{in the chosen phase at time slot } t \\ 0, & \text{non-priority vehicles in direction } p \text{ gets red light,} \\ & \text{in the chosen phase at time slot } t \end{cases} \tag{5.14}$$

$$G^p_{brt}(t) \quad = \quad \begin{cases} 1, & \text{priority vehicles in direction } p \text{ gets green light,} \\ & \text{in the chosen phase at time slot } t \\ 0, & \text{priority vehicles in direction } p \text{ gets red light} \\ & \text{in the chosen phase at time slot } t. \end{cases} \tag{5.15}$$

Note that the action space $\mathcal{A}$ must be defined such that all conflicting flows are not allowed to set green light at the same time.

5.1.3.1 Signal Lights

The system dynamics can be investigated by changing the traffic signal lights according to the action which is defined in the reinforcement learning framework. In one time slot, vehicles can move to their adjacent cells only. At signalised cells of the same intersection, non-priority vehicles and priority vehicles share a common intersection capacity. For a non-signalised non-priority cell i, the maximum number of vehicles that can flow through cell i in direction p at time slot t is given by $q^p_i(t) = q_{max}, \forall_p, \forall_t$. For signalised non-priority cell i, the equation can be defined as follows

$$q^p_i(t) = \begin{cases} q_{max} & ; G^p(t) = 1 \quad \text{and} \quad t - \tau_i(t) > L, \\ 0 & ; \text{otherwise}, \end{cases} \tag{5.16}$$

where q_{max} is the maximum number of non-priority vehicles that can flow through each cell per time slot and L is the loss time of the intersection. Here, $\tau_i(t)$ denotes the latest time instant where the traffic signal indication of non-priority vehicle cell i at time slot t has been changed. The vehicle then can move to another cell according to the equation above as long as the latest signal indication is not changed. Otherwise, all the vehicles have to be stopped.

At a signalised priority cell j, the maximum number of priority vehicles that can flow through the cell is defined as

$$q^p_j(t) = \begin{cases} q_{max,brt} & ; G^p_{brt}(t) = 1 \quad \text{and} \quad t - \tau_j(t) > L_{brt}, \\ 0 & ; \text{otherwise}, \end{cases} \tag{5.17}$$

where $\tau_j(t)$ is the latest time instant where the traffic signal indication of priority vehicle cell j at time slot t has been changed and L_{brt} is the loss time of priority vehicles.

At cell m representing the location of priority-vehicle station, define parameter $q^p_j(t)$

as

$$q_j^p(t) = \begin{cases} 1 & ; \text{priority-vehicle signal is green at signalised cell } j = m \text{ in direction } p. \\ 0 & ; \text{priority-vehicle signal is red at signalised cell } j = m \text{ in direction } p \end{cases} \tag{5.18}$$

Note that for all non-signalised cell j, $q_j^p(t) = q_{max,brt}$ where $q_{max,brt}$ is the maximum number of priority vehicles that can flow through each cell per time slot. If the green light has been assigned to the priority vehicle, then the traffic light of the non-priority vehicles will be changed to green as long as the non-priority vehicles are not conflicted with the priority vehicles. In addition, if a priority vehicle reaches its station, then the priority vehicle must be held for a few time slots for picking up the passengers. This intended stop can be modelled in by setting red signal in (5.18). The value of $q_j^p(t)$ can also be changed to a value greater than 1 to model the occurrence of multiple priority vehicles in the same CTM cell at the same time slot from the bad management of traffic signal control.

5.1.4 Network Boundary Conditions

The boundary condition will be considered for the whole network, not each subnetwork, as follows.

5.1.4.1 Network Gate Cell

The boundary condition is here formulated by following [41]. At the network boundary, input vehicle flows can be modelled by a cell pair. A source cell "00" with an infinite number of vehicles $s_{00}^p(t) = \infty$ discharges into an initially empty "gate" cell "0" of infinite size, $c_0^p(t) = \infty$. The flow capacity $q_0^p(t)$ of the network gate cell is set to the desired link input flow.

$$\Lambda_0^p(t) = \min\{s_0^p(t), q_0^p(t)\}. \tag{5.19}$$

$$s_0^p(t+1) = s_0^p(t) + y_0^p(t) - y_1^p(t). \tag{5.20}$$

$$y_0^p(t) = q_0^p(t). \tag{5.21}$$

Assume the receiving capability of network gate cell is infinite. Hence, the sending capability $\Lambda_{00}^p(t)$ of source cell "00" is limited by $q_0^p(t)$ and

$$y_1^p(t) = \min\{\Lambda_0^p(t), \Psi_1^p(t)\}. \tag{5.22}$$

Let $y_1^p(t)$ be the number of non-priority vehicles that flow into cell 1 in direction p at time slot t. Thus, the desired input in each direction can be configured in parameter $q_0^p(t)$. Likewise, for the priority vehicle, the network boundary conditions can be calculated directly similar to the equations (5.19) – (5.22).

5.1.4.2 Network Sink Cell

Suppose the output cell, "sink", for all exiting traffic has infinite size $c_{I+1}^p(t) = \infty$, $c_{J+1}^p(t) = \infty$, $q_{I+1}^p(t) = \infty$ and $q_{J+1}^p(t) = \infty$. Thus, by default in this dissertation, the network sink cells $I + 1$ and $J + 1$ have the receiving capability from (5.8)

$$\Psi_{I+1}^p(t) = \infty. \tag{5.23}$$

$$\Psi_{J+1}^p(t) = \infty. \tag{5.24}$$

However, in practice, if the downstream of an intersection has been affected by network downstream back-pressure congestion or by another traffic signal light, then the network sink cell can be assigned finite value $q_{I+1}^p(t)$ and $q_{J+1}^p(t)$.

5.1.5 Passenger Delay

In Chapters III and IV, we are interested more in the vehicle delay defined as the number of vehicles that cannot move away from the present cell within each time slot. In this chapter, the vehicle delay will be redefined as the passenger delay. In fact, a single priority vehicle can carry more passengers than a single non-priority vehicle. Therefore, the comparison in terms of passenger delay would be suited for the representative of the total network delay. Two types of passenger delay are proposed, namely, the internal and the external passenger delay. At time slot t for each direction p, let $d_0^p(t)$ be the external passenger delay (if the upstream road segment of non-signalised cell $i = 1$ is outside the boundary of the considered road network) and $d_i^p(t)$ be the internal passenger delay at other non-signalised cell i. These delays can be expressed as

$$d_0^p(t) = s_0^p(t) - y_1^p(t), \tag{5.25}$$

$$d_i^p(t) = s_i^p(t) - y_{i+1}^p(t), \quad i = 1, 2, \ldots, \kappa. \tag{5.26}$$

Similar to Chapter III and Chapter IV, the external delay can be considered as the passenger delay neighbourhood outside the considered road network. The internal network delay is

considered here on roads connecting to an intersection. Combining two delay types will reflect the system behavior to be optimised for the best possible traffic signal control. For priority vehicle, let $n_j^p(t)$ be the number of carried passengers on the BRT vehicle at cell j in direction p at time slot t. If $b_j^p(t) = 0$, then $n_j^p(t) = 0$. And, let $n_0^p(t)$ be the number of passengers waiting to be carried by BRT at all the BRT stations on the considered road segment in direction p at time slot t. Assume total number of passengers $n_0^p(t)$ is assigned equally to each available BRT station m on the considered road segment. The passenger delay from the BRT can be defined as follows

$$d_{0,brt}^p(t) = \max\{0, n_0^p(t) - n_j^p(t)\}, \quad j = m, \tag{5.27}$$

$$d_j^p(t) = \left[b_j^p(t) - y_{j+1}^p(t)\right] n_j^p(t), \quad j = 1, 2, \ldots \kappa_b, \tag{5.28}$$

where $d_{0,brt}^p(t)$ is the passenger delay waiting to be carried at BRT stations.

5.1.6 Performance Criteria

To evaluate the optimal policy (set of actions) that minimises the total network delay. If the upstream road segment of non-signalised cell $i = 1$ is outside the boundary of the considered road network, then the passenger delay for non-priority vehicle $\Upsilon_{red}(t)$ at time slot t is defined as follows.

$$\Upsilon_{red}(t) = \sum_{p=1}^{P} \sum_{i=0}^{\kappa} (1 - G^p(t)) d_i^p(t) \tag{5.29}$$

Otherwise,

$$\Upsilon_{red}(t) = \sum_{p=1}^{P} \sum_{i=1}^{\kappa} (1 - G^p(t)) d_i^p(t). \tag{5.30}$$

where $\Upsilon_{red}(t)$ is the "red light delay". The red light delay is the total passenger delay from all the cells in the directions that see the red light. The validation of the red light delay can be found in **Chapter III**.

For the priority vehicle, the passenger delay at time slot t can be defined as follows.

$$\Upsilon_{brt}(t) = \sum_{p=1}^{P} \sum_{j=1}^{\kappa_b} (1 - G_{brt}^p(t)) d_j^p(t) + d_{0,brt}^p(t). \tag{5.31}$$

Therefore, the performance criteria $\Upsilon(t)$ at time slot t is defined as follows.

$$\Upsilon(t) = \Upsilon_{red}(t) + \Upsilon_{brt}(t). \tag{5.32}$$

5.2 Signal Optimisation By Q-learning Algorithm for Road Network with Transit Signal Priority

For completeness, let us elaborate the core implementation of the Q-learning algorithm. To apply Q-learning in a signalised CTM framework, a definite simulation length is used for periodically observing traffic behaviors within a study time-interval. When the current time slot of CTM reaches the simulation length, the system enters the next *episode*. The Q-learning-based traffic controller is designed to make a sequence of signal-light decisions. Let the decision epoch t_ω refer to the time instant when decision ω is made, where $\omega = 1, 2, \ldots$ and $t_\omega = t_1, t_2, \ldots$, respectively.

This section explains the brief implementation of the proposed Q-learning algorithm together with the CTM-BRT framework. For each episode, the optimisation procedure of Q-learning operated by the control agent at each intersection can be summarised as follows.

1) *System Initialisation*

The number of vehicles in state vector $\mathbf{s}(0)$ and $\mathbf{b}(0)$ can be initialised at the beginning of an episode to a nominal operating point of the system at the considered time period. The action value function $Q(\tilde{\mathbf{s}}, \mathbf{b}, a)$ can be initialised to the latest updated value in the previous episode. Let $\omega = 1$.

2) *Action Selection*

At decision ω, with the current state observable at $\tilde{\mathbf{s}}, \mathbf{b}$, the agent (traffic controller) chooses an action $a \in \mathcal{A}(\tilde{\mathbf{s}}, \mathbf{b})$ to control the traffic signal. The action can be chosen by the ϵ-greedy algorithm [32], where the greedy action is here defined as

$$a = \arg\min_{a'} Q(\tilde{\mathbf{s}}, \mathbf{b}, a').$$

According to this algorithm [32], Q-learning chooses the greedy action with probability $1 - \epsilon$. And, with probability ϵ, the other actions are randomly selected according to a uniform distribution. In practice, an ϵ is a small positive value representing the explorability of learning algorithm.

3) *Update of System Dynamics*

Calculate the CTM state from time slot $t = t_\omega$ to time slot $t = t_{\omega+1} - 1$. Here, the next state vector $(\tilde{\mathbf{s}}', \mathbf{b}')$ is calculated from the CTM state at time slot $t = t_{\omega+1} - 1$. The observed

reward $R(\omega)$ can then be correspondingly calculated from

$$R(\omega) = \sum_{t=t_\omega}^{t_{\omega+1}-1} (\Upsilon_{red}(t) + \Upsilon_{brt}(t)) \tag{5.33}$$

4) *Update of Action Value Function*

The algorithm can learn from its past experiences accumulated in Q-function and the reward in (5.33) newly gained from the most recent action ω. By following [32], Q-function can be updated as follows

$$Q(\tilde{\mathbf{s}}, \mathbf{b}, a) \leftarrow Q(\tilde{\mathbf{s}}, \mathbf{b}, a) + \alpha[R(\omega) + \gamma \min_{a'} Q(\tilde{\mathbf{s}}', \mathbf{b}', a') - Q(\tilde{\mathbf{s}}, \mathbf{b}, a)],$$

Here, $Q(\tilde{\mathbf{s}}', \mathbf{b}', a')$ represents the action value function for the next observable state vectors $\tilde{\mathbf{s}}', \mathbf{b}'$ and next possible action $a' \in \mathcal{A}(\mathbf{s}, \mathbf{b})$. Practically, $\alpha \in (0, 1]$ is the learning rate and $\gamma \in [0, 1)$ is the discount rate applied to the future expected rewards.

5) *Update of State Variable*

Update the state $\mathbf{s} \leftarrow \mathbf{s}'$, $\mathbf{b} \leftarrow \mathbf{b}'$ and $n_0^p(\omega), n_0^p(\omega) \leftarrow n_0^p(\omega+1), n_0^p(\omega+1)$. And update $\omega \leftarrow \omega + 1$.

6) *Stopping Condition*

Repeat steps 2)–5) until the end of episode.

The optimisation procedures of Q-learning above have been applied only for an intersection as presented in Chapters III. Therefore, in a road network with BRT in this chapter, this process will be updated separately and simultaneously at all intersections at the same decision epochs.

5.3 Results and Discussions

In this section, a series of experiments will be shown. For convenience, let us define the vehicle class of non-priority (priority) as vehicles (BRT). Firstly, the performance comparisons between an example of four lanes road without BRT and an example of three lanes with BRT systems will be reported. Secondly, the effects of penalty function to the number of total carried passengers will be shown. Thirdly, the total number of passengers that BRT can carry in rush hour periods will be reported. Finally, the comparison of the use of Q-learning to control the road network with BRT among two control methods: the preemptive

and the local extension recall differential priority will be shown. Finally, the formulation of the BRT station will be taken into account.

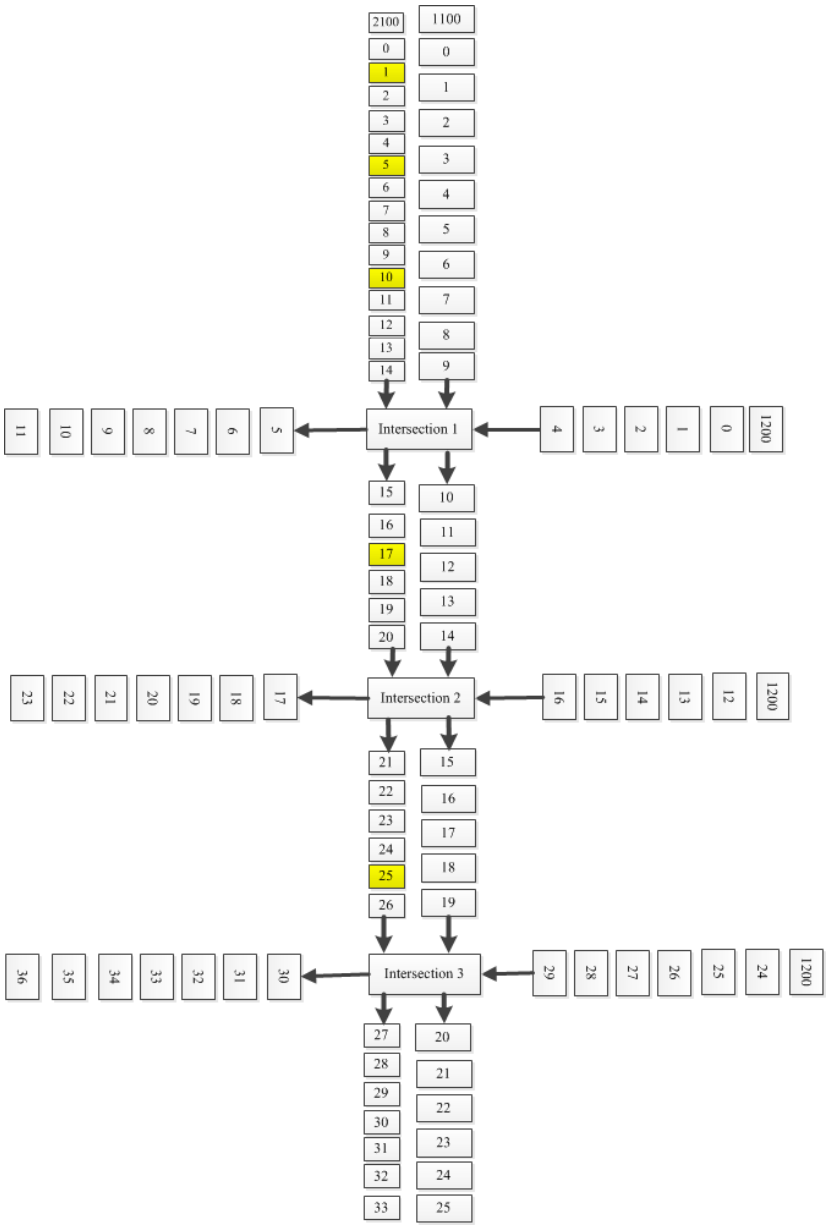

Figure 5.4 BRT model

As illustrated in Figure 5.4, suppose vehicles (BRT) can move 160 (100) metres in one-time slot on average where each time slot has been set to 10 seconds. Note that an example of the model Figure 5.4, is inspired from the U-shaped road network in Bangkok as illustrated in Figure 3.1. Each non-priority vehicle cell capacity is 120 passenger car unit (pcu). The maximum flow rate has been measured from AIMSUN under the condition that the vehicles are not affected by the red signal. The maximum flow rate $q_i^p(t)$ of 2.61 pcu/slot (passenger car unit per slot) and the maximum flow rate of BRT $q_j^p(t)$ of 1 BRT per slot. The wave speed

coefficient δ_i^p is 0.8. The wave speed coefficient has been calibrated from for Payathai road in Bangok, Thailand [49]. For the Q-learning algorithm, an action is chosen every 3 time slots. The source cells are "1100", "1200" and "2100". Each intersection has three signal phases. All the phases are phase 1 from north to south, phase 2 from west to east and BRT phase from north to south (this phase allows phase 1 to go as its in the same direction). For the shaded-cells in the middle of road segment from north to south with the cell numbered "1", "5", "10", "17" and "25", these cells are the BRT stations. The desired passengers taking BRT are assigned equally to each available BRT station m on the considered road segment. BRT waits for three time slots on average at these cells for picking up the passengers. The passenger delay being used throughout this chapter has been considered the delay at the BRT stations and the delay on the BRT.

5.3.1 Road Network with vs without Transit Signal Priority

The comparison between before and after the deployment of the BRT system is first reported here. Assume one BRT can carry 80 passengers on average and the normal service time for the BRT system is every 5 minutes. Assume the total passengers demand desired to pass the example road network has never been reduced. The simulation has been tested for two hours (720 time slots).

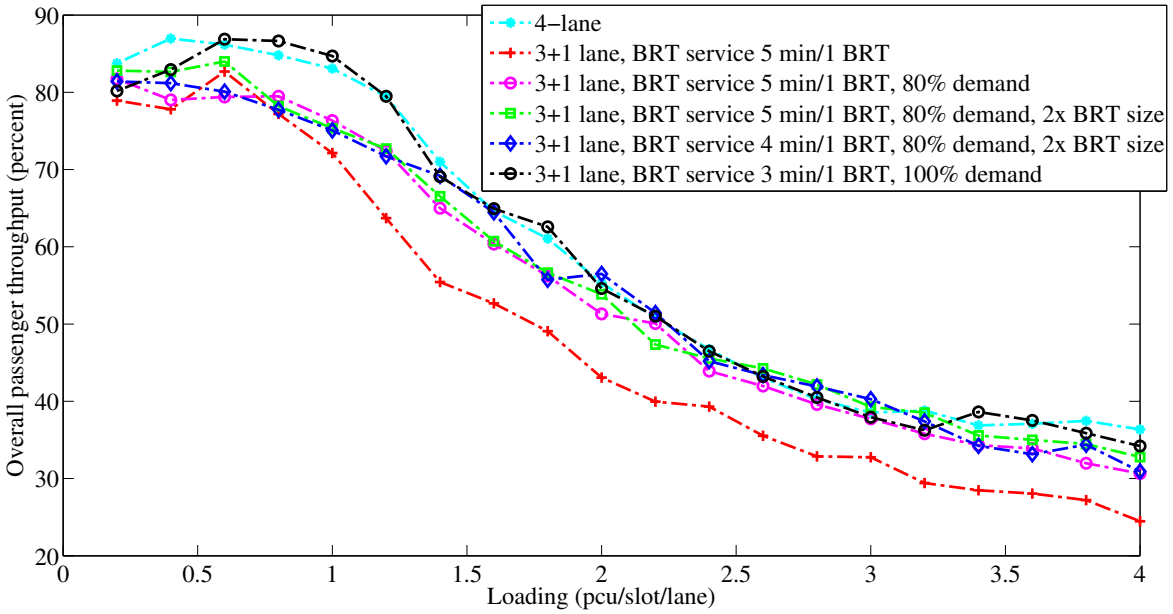

Figure 5.5 Loading vs overall passenger throughput

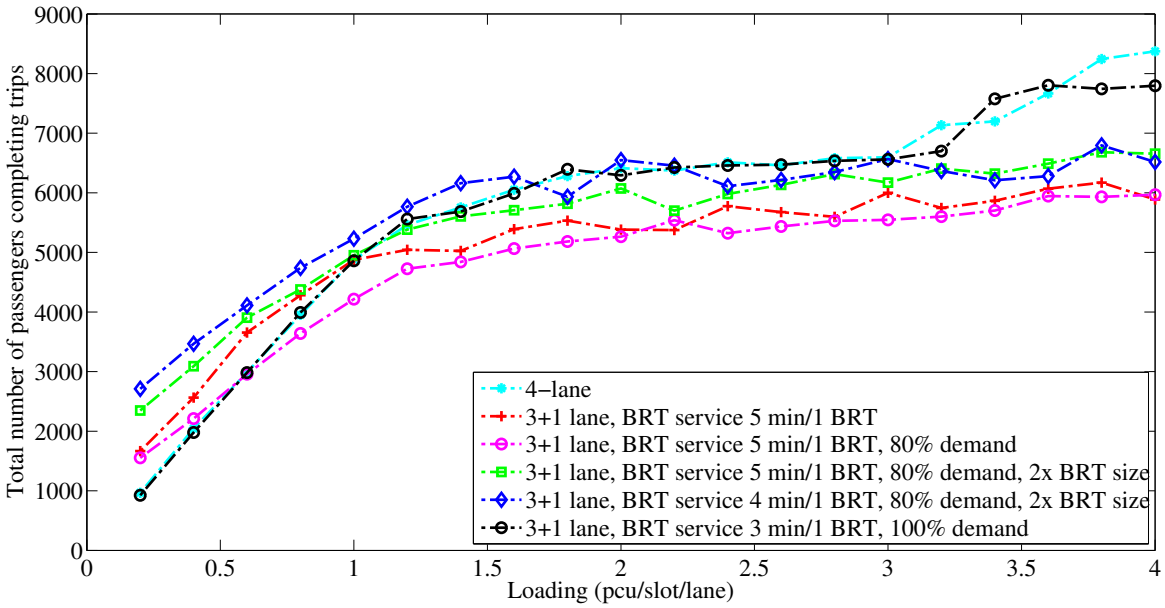

Figure 5.6 Loading vs total number of passenger completing trips

Firstly, the critical point can be found by observing when the overall system throughput (total network delay) has rapidly decreased (increased). Figure 5.5 illustrates the relationship between the change of traffic arrival rates and the overall passenger throughput. The passenger throughput is defined by the percentage between the total number of passengers who completed their journey from all directions to the total number of passengers request to pass a road network illustrated in Figure 5.4. The results have been observed in both total passenger delay and overall throughput in different scenarios which are four lanes without BRT, three lanes with BRT, reduced traffic demands for 3-lane with BRT, enlarged the BRT size and increased the BRT service frequency. The throughput has shown that between 1-1.6 pcu/slot/lane, the overall throughput has dropped to 60%. Therefore, in the next scenario, the traffic arrival rate will set to 1.6 pcu/slot/lane. Note that the maximum capacity for each intersection is 2.61 pcu/slot/lane. The desired passenger arrival rates approaching intersections from two directions are not exceeded the maximum capacity but the system performance has been operated in an oversaturated traffic conditions due to the interruptions of traffic signals, the system loss time and the wave speed coefficient. From both figures, the reported results show that the exclusive lane BRT with lane separator systems cannot reduce the overall passenger delay. However, if the BRT system must be deployed, then the advantage of implementing the BRT to the road network must be clearly shown. Theses advantages will be reported in the following subsection.

5.3.2 Performance Comparison of Transit Signal Priority and non-Transit Signal Priority Systems

From the previous results, the traffic arrival is a Poisson process with a mean arrival rate 1.61 pcu/slot/lane in each direction. As mentioned in the previous subsection, the system has been operated under oversaturated conditions. In this section, the traffic arrival will be divided into two types. First type is the class of passengers taking their own vehicles. Second type is the class of passenger taking BRT. The summation of the total traffic arrival rate must be equal to the case of overall traffic arrival rate of four lanes without BRT. For the case of three lanes with BRT, a proportion of the passengers will be weighted to use the BRT system. The observed system is operated in the rush hour periods from 06.00 am to 10.00 am. The traffic arrival rate from 06.00-07.00 am 0.8 pcu/slot/lane. From 07.00-09.00 am, the mean arrival rate is 1.6 pcu/slot/lane. From 09.00-10.00 am, the mean arrival rate is 0.6 pcu/slot/lane.

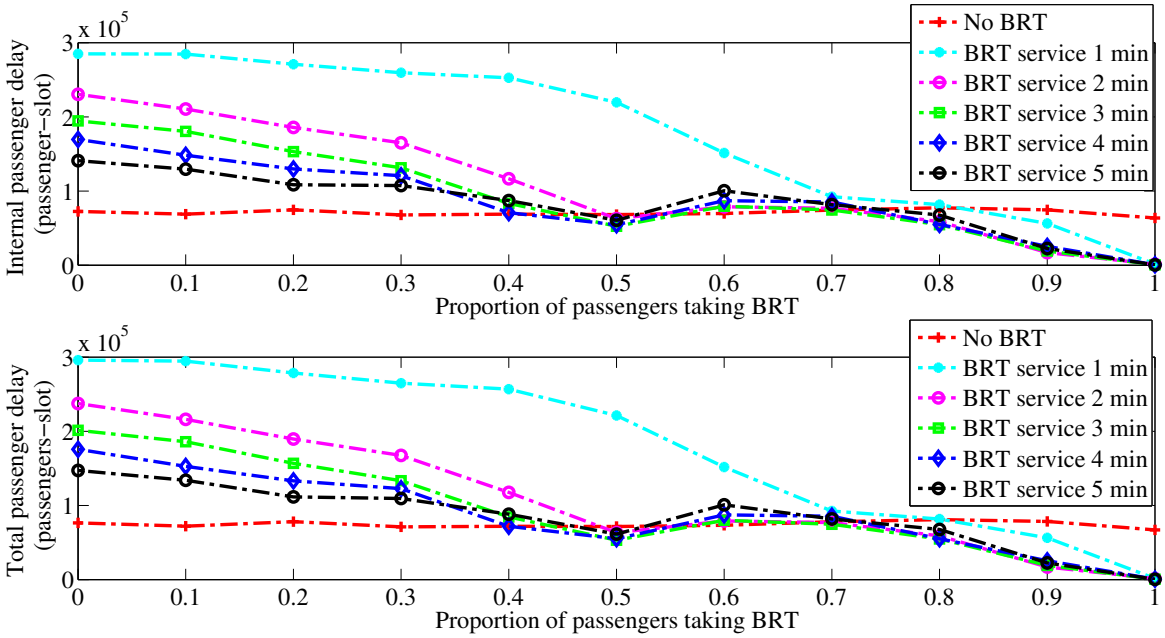

Figure 5.7 Proportion of passengers taking BRT vs internal and external passenger delay

Figure 5.7 illustrates the proportion of the passengers taking BRT versus the total internal delay and total network delay (including external delay). The result shows that increasing of the BRT service frequency causes high total passenger delay. If the BRT service frequency is too high, then the priority signal must be allocated too often e.g., the case of 1 minute of BRT service frequency. The total passenger delay is high because of the system loss time.

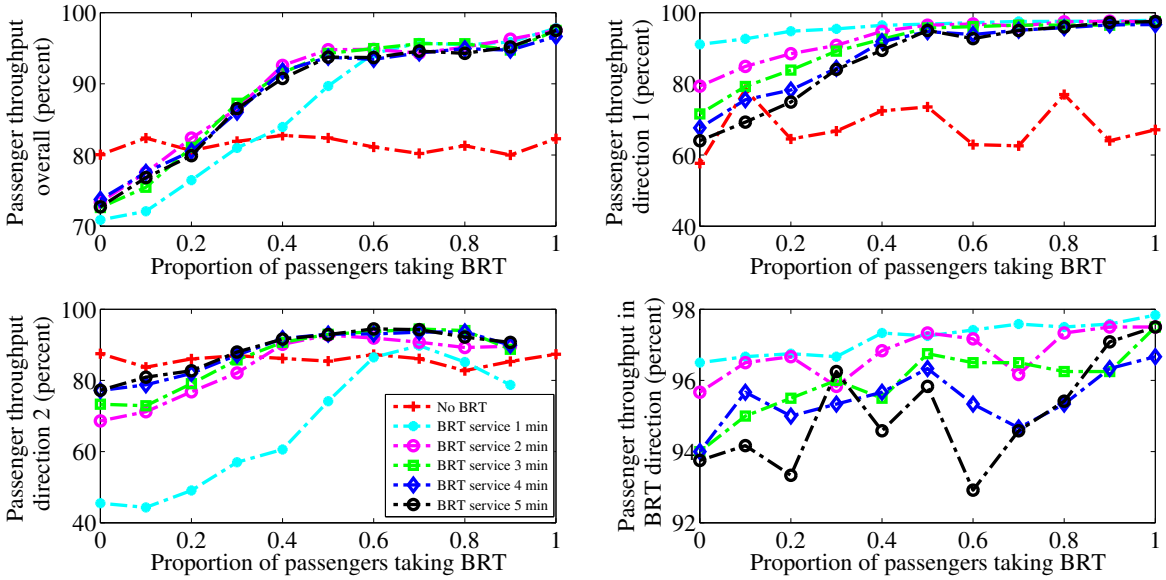

Figure 5.8 Proportion of passengers taking BRT vs passenger throughput

Figure 5.8 depicts the passenger throughput, the increasing of BRT service frequency also gives the increasing overall passenger throughput. For the other throughput subgraphs, the figure shows the throughput in each road segment. Consider the throughput approaching an intersection from west to east. If all the passengers taking BRT, the throughput cannot calculate because all the normal vehicles are waiting at BRT stations. The reported result shows that, the passenger throughput can be greatly increased by up to 9-15% in the jamming conditions when at least 40% from the overall passengers choose the BRT for their journey.

As illustrated in Figure 5.9, the result shows the percentage of passengers completing trips that can complete their journey in four hours. If the number of passengers taking the BRT increases, then the percentage of passengers completing trips decreases. The decreasing of the percentage of the passenger completing trips becomes from the fact that BRT has limited capacity.

Figure 5.10 shows the total number of passengers waiting at a BRT station. If the BRT can serve passengers more often, then the number of passengers (total passenger delay) becomes zero. The calculation of the total number of passengers at a BRT station can be found in (5.27).

Figure 5.11 illustrates the average number of passengers completing trips in one time slot. The more frequent the BRT service is, the more passengers completing trips are obtained.

From the reported results, a trade-off between the minimum total network delay and

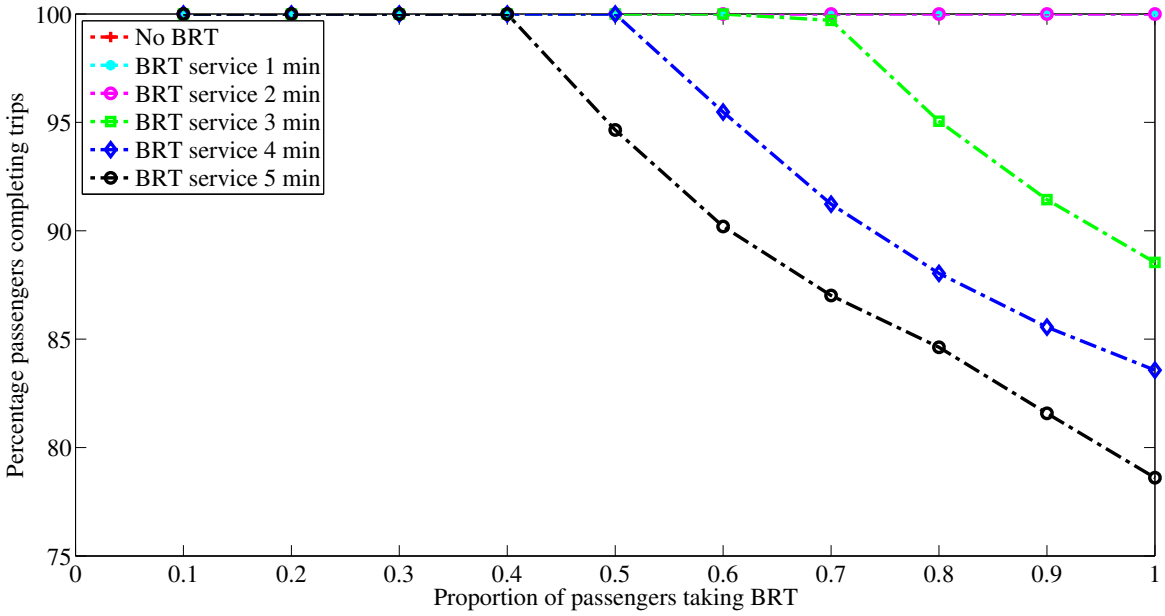

Figure 5.9 Proportion of passengers taking BRT vs percentage passengers completing trips

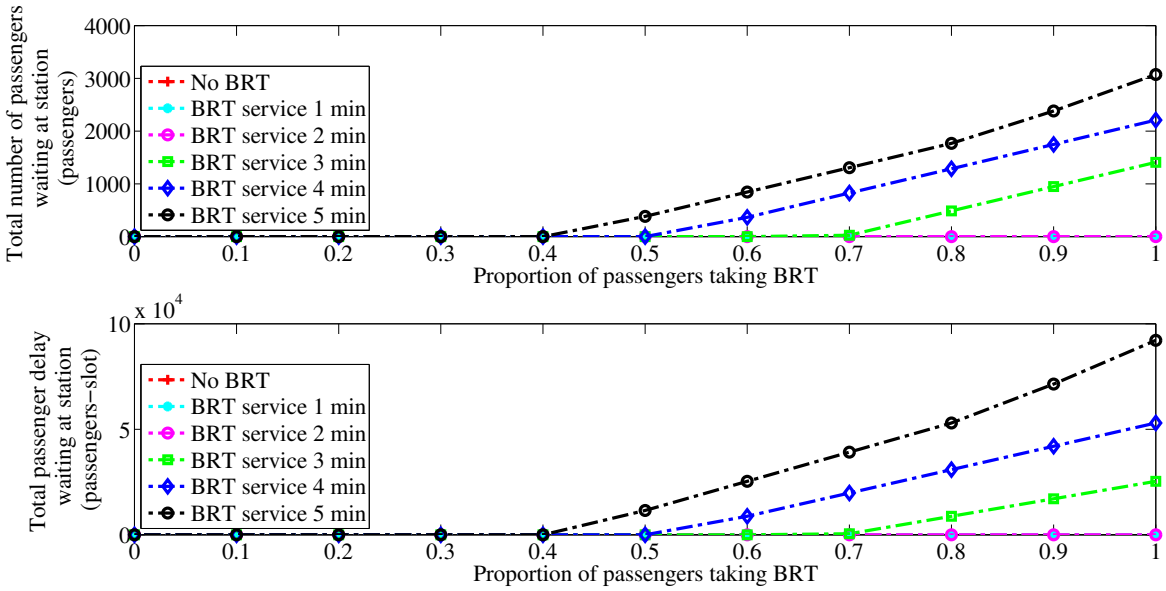

Figure 5.10: Proportion of passengers taking BRT vs total number of passengers waiting at a BRT station

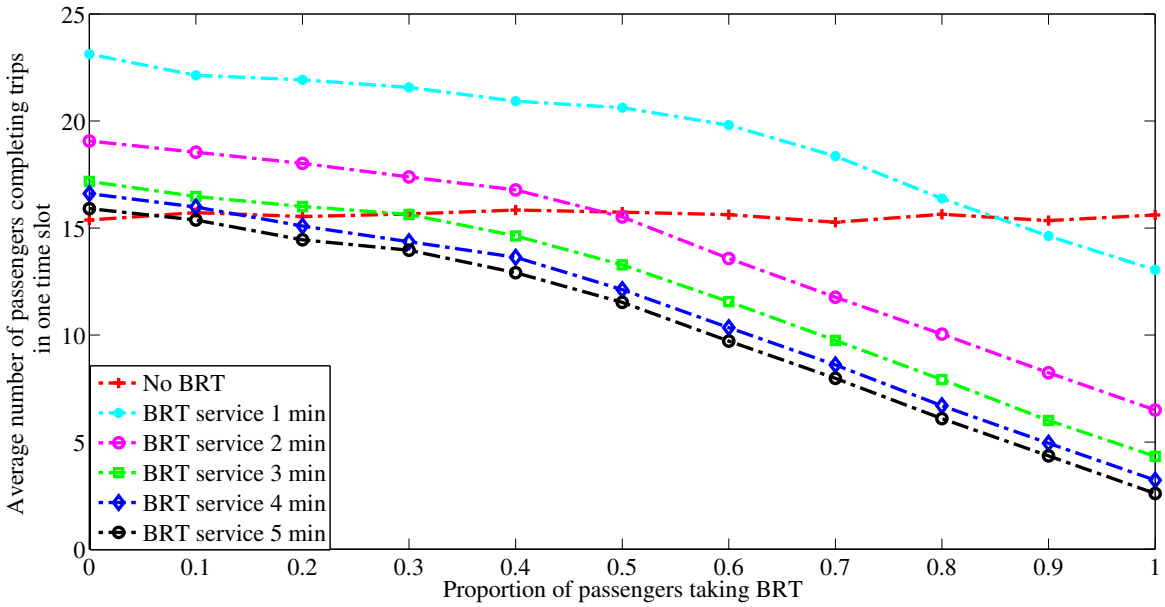

Figure 5.11: Proportion of passengers taking BRT vs average number of passengers completing trips in one time slot

the overall passenger throughput has been found. The minimum overall passenger delay may not be a good representative of the system performance when traffic is jammed. In jammed conditions, the overall passenger delay may not be easily reduced. Therefore, the overall passenger throughput is recommended for the performance metric. Practically, the possible range of the passenger taking BRT is less than 40% of the total summation of the total traffic arrival rate of four lanes without BRT.

5.3.3 Comparison of Existing Traffic Control Methods vs Q-learning

Nowadays, there are three well-known distributed traffic signal control methods deployed around the world which are fixed-time, vehicle actuated (VA) and MOVA (Microprocessor Optimised Vehicle Actuation) [47]. The fixed-time control uses the historical data to determine the green time for each approaching intersection. The VA implemented in UK has been reported that this system can give the priority to buses by either extending the current green period or shortening the other green periods. For the MOVA, this is a modernised VA version, the detected bus approaching to the intersection will be analysed individually lane-by-lane. However, the control signal is employed the bus priority concept as mentioned in the VA.

The BUS priority has been implemented in London with the BUS-SCOOT in iBUS

version [47]. In this version, the differential priority has been introduced. The differential priority gives the priority to the bus lateness only. The bus lateness refers to the outdated schedule or the minimum headway required. However, the iBUS requires centralised traffic signal control.

From the survey results of traffic signal control in London [54], the iBUS with the distributed extension and recall signal control gives the best results in terms of reduction delay. Therefore, in this subsection, the Q-learning control will be compared the results with both MOVA and iBUS distributed traffic signal control strategies. For convenience, let us rename the MOVA as the preemptive priority signal.

The comparison between the Q-learning with two existing approaches has reported. The BRT service frequency is set to every 5 minutes. The traffic arrival rate approaching from all directions is varied. The purpose of this setting is to evaluate to control methods in different loading regions. As mentioned earlier, the road network in Bangkok always operates in the jamming conditions.

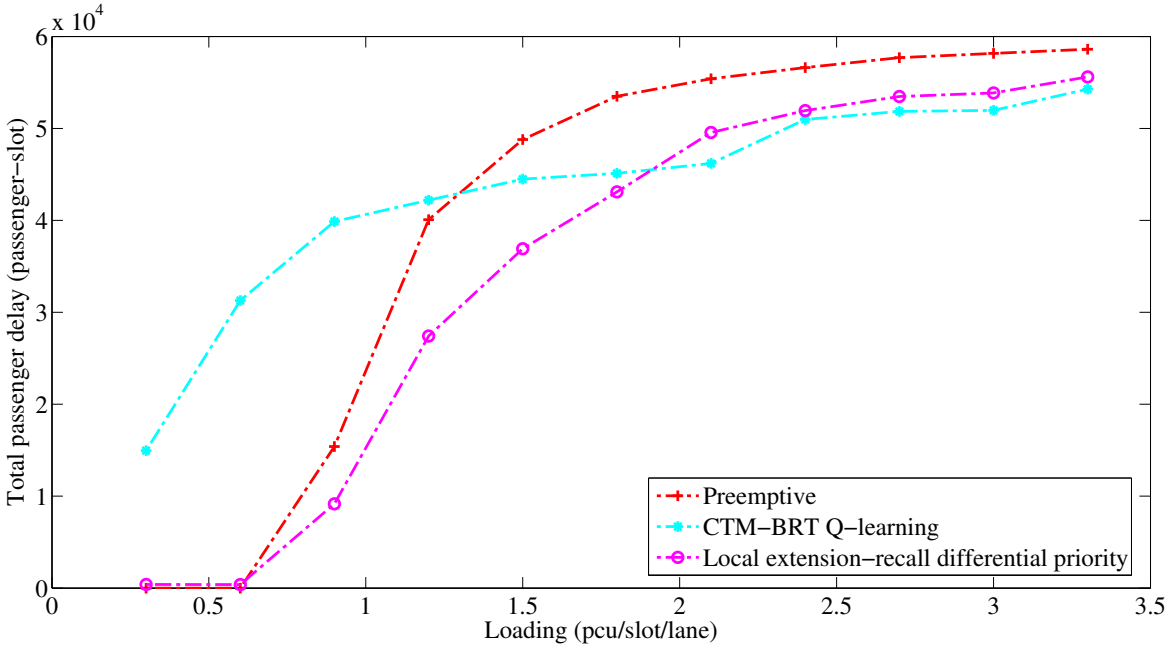

Figure 5.12 Loading vs total passenger delay

As illustrated in Figure 5.12 and Figure 5.13, the CTM-BRT based Q-learning obviously outperforms both two existing control methods in oversaturated regions (more than 50% of the maximum flow rate 2.61 pcu/slot). Note that in undersaturated regions, the Q-learning performs bad because of its exploration ability. Figure 5.14 illustrates the action

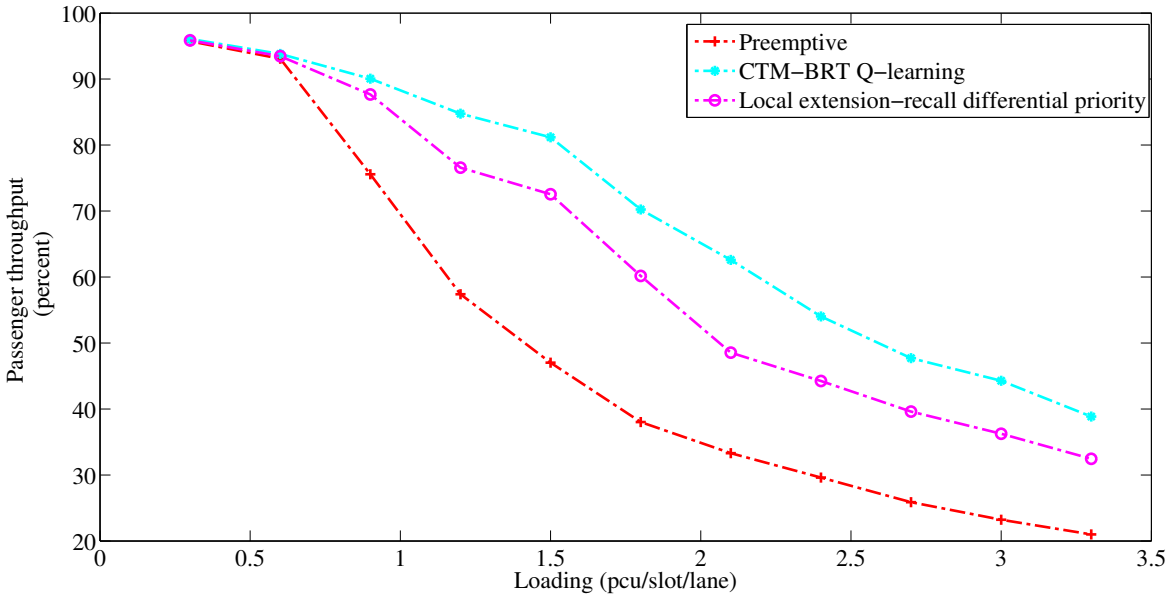

Figure 5.13 Loading vs overall passenger throughput

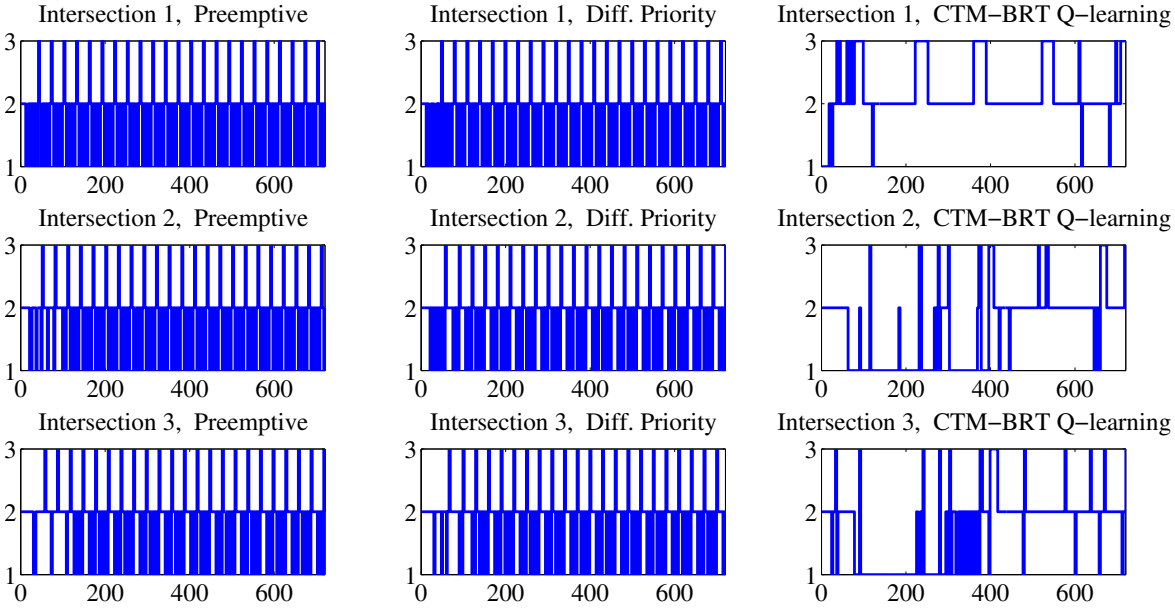

Figure 5.14 Action selection for each intersection

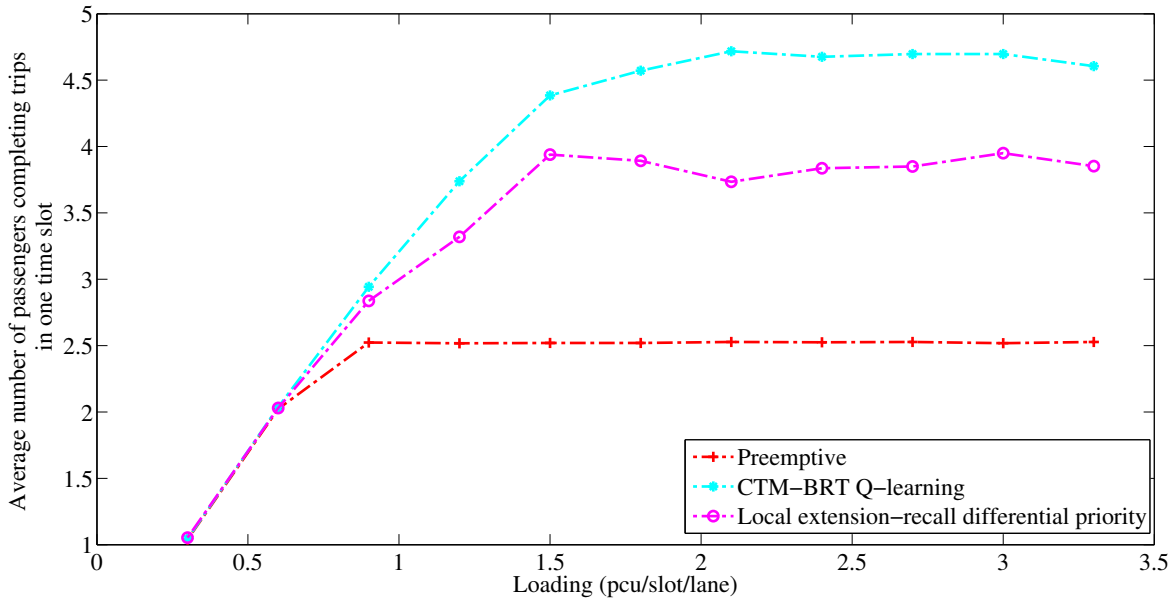

Figure 5.15 Loading vs average number of passengers by one completed trip in 1 time slot

(green light) in each intersection. The action selections in each time slot has been shown. From both existing methods, the control plans obtained from each intersection are similar to periodic signal timing control methods. However, in reality, the periodic control cannot be used due to the fluctuation of the systems. Moreover, the total passenger delay obtained from Q-learning is the lowest when the traffic arrival rate increases because Q-learning changes the action less often to reduce the loss time of the systems. The CTM-BRT based Q-learning also outperforms the two existing schemes in terms of the average number of passengers completing their journeys in one time slot as illustrated in Figure 5.15.

5.4 Summary

The original contributions in this chapter are the extension to a network of cascading interactions with transit signal priority system has been proposed with simple uni-directional flows without turning movements. Motivated by the BRT system in Bangkok, the conventional signalised CTM has been generalised to cope with the preplanned space-usage priority of BRT over other non-priority vehicles by modelling explicitly the existence of BRT physical lane separator as well as the location of BRT stations. The delay function of both carried passengers on BRT and on other non-priority vehicles as well as waiting passengers at stations has been introduced. Based on the investigated scenarios, the deployment of BRT system with one lane deducted by the lane separator cannot reduce the total passenger delay

in comparison with the comparable road and traffic condition before the BRT instalment. However, with BRT, the passenger throughput can be greatly increased by up to 9-15% in the jamming conditions when at least 40% from the overall passengers choose the BRT for their journey. Moreover, our proposed method outperforms the conventional preemptive and differential priority control methods because of the improved awareness of signal switching cost.

The in-depth investigation has been first reported from the Chapter III about the validation, discussions, reward functions and etc. The proposed red light delay as the reward function and the introduced the boundary conditions as the external delay to capture the effect from nearby neighbourhood intersections. Chapter IV investigates the performance comparison with the classical queueing model M/M/1 and D/D/1. This chapter also extends the investigation to an example of road traffic with BRT systems. Two experiments have been investigated. Firstly, the comparison between before and after the deployment of the BRT systems has been evaluated. Secondly, the comparison between the current implementations and the Q-learning has been reported.

Though an example of the comparison of the Q-learning with existing control methods has been illustrated, there are too many possibilities yet to be discovered. However, in this dissertation, the main contribution is to formulate a novel mathematical framework based on the signalised cell transmission model using Q-learning for the road network with BRT system. This dissertation has been confirmed the extension of such approach to a real implementation.

CHAPTER VI

CONCLUSION

The aim of this dissertation is to develop a new mathematical framework to control the traffic signal light for the road network traffic with the bus rapid transit system by applying the automated self learning called reinforcement learning to seek the best possible traffic signal. This dissertation has been investigated the road network in both road traffic conditions which are undersaturated and oversaturated traffic conditions. The network states have been modelled by the signalised cell transmission model (CTM). The in-depth investigation started from Chapter III to Chapter V. The summaries of the contributions of each chapter are shown in the following sections, together with suggestions for possible future works. The main emphasis of this dissertation is to show our proposed framework in finding the most proper traffic signal solution in oversaturated traffic conditions.

6.1 Contributions from Chapter III

The first model developed by using the cell transmission model (CTM) to capture the system dynamics together with the implementation of Q-learning to seek the best possible solution for an isolated intersection has been introduced. The other underlying conditions are the newly presented the external delay function in the boundary conditions to capture the effects from the road network neighbourhoods and the newly proposed red light delay as the Q-learning reward function. Both simulation and mathematical derivation results confirm that using the newly proposed red light delay as the Q-learning reward function gives better performance than using the total network delay as reward function. In addition, the existing works related to Q-learning have not considered scalability issues due to the limitation in terms state space explosion. However, we attempt to alleviate the explosion by employing state space quantisation and control traffic signal in such network scenarios.

The results have been reported from the series of experiments which are the Q-learning validation, the effect of reward functions, the Q-learning performance in stationary/non-stationary stochastic loadings and the applicability of the CTM-based Q-learning algorithm

in the microscopic mobility environments using AIMSUN. The simulation results show that our proposed framework can efficiently find the proper solution for road traffic systems by comparing with the best periodic signal solution (BPSS). The effect of reward functions has also been investigated and the adaptability of the Q-learning algorithm in adjusting its solution with Poisson arrival upon the change of time has also been observed. The results from the macroscopic level show that Q-learning can achieve the solution similar to the BPSS method. However, in a microscopic level, the control strategies obtained from the CTM-based Q-learning approach outperform the BPSS in terms of the throughput and the average travel time because the Q-learning algorithm has allocated the green time more often to the direction with a higher vehicle arrival rate.

6.2 Contributions from Chapter IV

With the newly proposed red light delay as the Q-learning reward function applied to an isolated intersection, this chapter has reported the results and their applicabilities. The BPSS is inapplicable due to its computational burden required. In this chapter, our proposed CTM-based Q-learning will be compared with the classical mathematical M/M/1 and D/D/1 queuing models.

The obtained results show that the Q-learning approach can improve the intersection throughput by up to 1.7-8.3% and by up to 3.2-14.8% in jamming conditions in comparison with the respective M/M/1 and D/D/1 approaches. Moreover, the average vehicle delay per completed trip can be reduced by up to 7.0-63.4% and by up to 18.9-80.7% in comparison with the respective M/M/1 and D/D/1 approaches. Note that the optimal derivations are based on the stability condition where the all vehicles entering the systems can be totally served. However, if all the vehicles entering the systems cannot be totally served, then the accumulative number of vehicles tends to be infinite over time. The queueing models are therefore guaranteed that there is no accumulative queue length when the stability condition is held. Sometimes, the stability is not held, the number of vehicle entering the systems will create the queue to the buffered of the systems. Moreover, the increasing of queueing length at the boundary cell is strongly not recommended.

The contribution in this chapter is to show the applicability of the Q-learning in controlling an isolated intersection by comparing with two optimal split derivation from the M/M/1 and D/D/1 models.

6.3 Contributions from Chapter V

An extension to a network of cascading interactions with transit signal priority system has been proposed with simple uni-directional flows without turning movements. Motivated by the BRT system in Bangkok, the conventional signalised CTM has been generalised to cope with the preplanned space-usage priority of BRT over other non-priority vehicles by modelling explicitly the existence of BRT physical lane separator as well as the location of BRT stations. The delay function of both carried passengers on BRT and on other non-priority vehicles as well as waiting passengers at stations has been introduced. Based on the investigated scenarios, the deployment of BRT system with one lane deducted by the lane separator cannot reduce the total passenger delay in comparison with the comparable road and traffic condition before the BRT installation. However, with BRT, the passenger throughput can be greatly increased by up to 9-15% in the jamming conditions when at least 40% from the overall passengers choose the BRT for their journey. Moreover, our proposed method outperforms the conventional preemptive and differential priority control methods because of the improved awareness of signal switching cost.

From the reported results, by operating the BRT system in our road network example, the overall passenger delay increases if none of the passengers decided to use the BRT system. The reported result is plausible because BRT requires a dedicated lane and the passenger throughput can be greatly increased by up to 9-15% in the jamming conditions when at least 40% from the overall passengers choose the BRT for their journey. One of the findings from the obtained results is that in jamming conditions, the overall passenger may not necessarily be a good performance criterion. It has also been found that the system throughput has rapidly increased because of the augmented total number of passengers completing trips. Therefore, the system throughput, especially in terms of the total number of passengers completing trips becomes a good performance criterion.

6.4 Possible Future Research on Oversaturated Traffic Conditions

From the study of application of RL with CTM to a BRT road network, there are numerous scenarios not yet covered which are worth for the possible research in the future. At the end of Chapter V, this dissertation shows the extension of the developed mathematical

framework to an example of road traffic network with BRT. However, there are many findings yet to be discovered within our proposed mathematical framework. In general, the effect of model parameters, the effect of model accuracy, the effect of real traffic data and the possibility of using our proposed framework to real road traffic situations are worthwhile studying. Moreover, if the control method from Q-learning can be applied, especially in Bangkok, then the results become insightful. The research recommendations are mainly focused on the situation when the road traffic networks have been operating in the oversaturated conditions. For the undersaturated traffic conditions, any control method can be applied.

6.4.1 Partially Observable Situation

By using our proposed CTM-BRT based Q-learning framework, the vehicles entering road segments both priority and non-priority assume to be measurable. However, for unpredictable situations, the vehicles entering road segments cannot be measured directly. The road network in Bangkok nowadays, the traffic sign boards are installed. The traffic sign boards report the relative road network density in its forward directions. Unfortunately, the road network density has been reported by colors. Based on our proposed framework, the RL state space may not know explicitly. In this situation, the partially observable reinforcement learning (PORL) [55] can be applied. Conceptually, the PORL evaluates the RL state spaces by investigating the feedback from the chosen actions and the immediate reward functions. Despite the adopting of PORL to our proposed framework, the Q-learning with state space quantisation can be directly applied to the Bangkok situation with traffic sign boards. As long as the feedback system works perfectly, the PORL may not be used.

6.4.2 Road-Space Sharing

Nowadays, the U-shaped road network with BRT systems in Bangkok consists of two road types which are road network with and without lane separators. The recommended CTM-BRT Q-learning for the case of road network without lane separators has not taken into account. However, from our proposed framework, the road network systems with lane separators can be straightforwardly applied by reducing the average speed in each individual lane e.g., the righteous lane considered as the high-occupancy vehicles (HOV). The mathematical can further extend to the multi-class cell transmission model [49]. The multi-class cell transmission model has been proven its ability in identifying two classes of vehicles e.g.,

buses and cars.

6.4.3 Signal Light

With the limitation of calculations on single machine, three signal phases have been considered. To relax our proposed assumptions, the signal phases can be further extended to a realistic scenario. Increasing of signal phases will lead to high computational burdens and state space explosions. The state space explosions have been alleviated by the quantisation techniques. However, the cardinality of the action spaces increases non-linearly depending on the total number of signal phases for all intersections. The computational burdens are inevitable.

6.4.4 Mesoscopic Traffic Model

Our proposed framework has been encompassed on the macroscopic level only. The road system parameters have been considered on the average value on each road segment. The mesoscopic traffic models consider the vehicles movements in their observed road networks. However, the mesoscopic traffic models report the output similar to the viewpoint from macroscopic models to alleviate the computational burdens. By using only the macroscopic model, the conservation of flows and equations on how traffic propagate through the systems.

6.4.5 RL Reward Functions

In this dissertation, the Q-learning reward functions have been calculated from the CTM model. Practically, the proper reward functions cannot be declared explicitly. The CTM model efficiently uses to evaluate the progressions of the vehicles' movements of the systems. Therefore, if the Q-learning has been installed at an intersection, then the Q-learning reward functions must be fine-tuned. However, with the adaptability of the Q-learning mentioned in Chapter III, the algorithm can learn and adjust to the best proper solutions as long as the system works functionally. The link throughput is not recommended because the blinding situation of the algorithm will be occurred. The blinding situation happens from the unidentified reward functions. The feedback returns to the control agents do not know explicitly how good of the previous selection is. In this situation, the feedback from the reward functions is not working due to the unidentified reward functions problems.

6.4.6 Effects of Model Parameters

The model parameters are one of the most important. For example, the wave speed coefficient, this parameter can reduce the vehicle movements. The exact value of this parameter needs to be calibrated. For the capacity, in reality, the system capacity changes upon time. Sometimes, the incidences happen on a road segment, the system capacity is therefore reduced At least, the capacity in one lane has been reduced. Therefore, the adaptability of Q-learning in solving this situations needs to be studied. The embedded sensors in the road being used in this dissertation has been chosen based on the fact that the financial supports are limited. The quantisation level has been chosen equally. If the level of quantisation is not equal, then the results would become different.

References

[1] Chen, S., Real Time Traffic Signal Control for Oversaturated Network. Ph.D. dissertation Texas Tech University, Texas, United State of America., 2007.

[2] Webster, F. V., Traffic Signal Settings. Tech. Rep. 39, Great Britain Road Research Laboratory, London, 1958.

[3] Gazis, D. C. Optimum Control of a System of Oversaturated Intersections. Operations Research 12 , 6 (1964): 815–831.

[4] Chang, T.-H. and Lin, J.-T. Optimal Signal Timing For an Oversaturated Intersection. Transportation Research, Part B 34 , 6 (2000): 471–491.

[5] Hunt, P. B., Robertson, D. I., Bretherton, R. D., and Winton, R. I., SCOOT - A Traffic Responsive Method of Coordinating Signals. Tech. Rep. 1014, TRL Laboratory Report, 1981.

[6] Chai, C., Adaptive Traffic Signal Control Using Adaptive Approximate Dynamic Programming. Ph.D. dissertation University Collage London, 2009.

[7] Sims, A. G., S.C.A.T. The Sydney Co-ordinated Adaptive Traffic System. Symposium on Computer Control of Transport 1981: Preprints of Papers, (1981): 22–26.

[8] Mimi Hwang, E. L.-L. N. N. and Okunieff, P., ADVANCED PUBLIC TRANSPORTATION SYSTEMS: THE STATE OF THE ART UPDATE 2006. Tech. Rep. FTA-NJ-26-7062-06.1, Federal Transit Administration, U.S. Department of Transportation, 2006.

[9] Levinson, H., Zimmerman, S., Clinger, J., Rutherford, S., Smith, L., Cracknell, J., and Soberman, R., Bus Rapid Transit: Case Studies in Bus Rapid Transit. TCRP Report 90, (2003): 1–62.

[10] Lambert, W., The ITS components for the Bus Rapid Transit system in the Greater Vancouver area of British Columbia, Canada costs $5.8 million (Canadian). ITE 2003 Annual Meeting.

[11] Hounsell, N. B., McLeod, F. N., and Shrestha, B. P., Bus priority at traffic signals: investigating the options. Proceedings of the 12th International Conference on Road Transport Information and Control, (2004): 287–294.

[12] Hounsell, N. B. and Wall, G. T., New Intelligent Transportation Systems Applications in.Europe to Improve Bus Services. Proceedings of 2002 Annual Meeting of Transportation Research Board, (2002): 85–91.

[13] Lee, S. S., Lee, S. H., Oh, Y. T., and Choi, K. C. Development of degree of saturation estimation models for adaptive signal systems. KSCE Journal of Civil Engineering 6 , 3 (2002): 337–345.

[14] Hounsell, N. B., Shrestha, B. P., Head, J. R., Palmer, S., and Bowen, T., The Way Ahead for London's Bus Priority at Traffic Signals. Proceedings of the 14th World Congress on Intelligent Transport Systems and Services, (2007): 193–200.

[15] Hunter, C. D., Guidelines for the Successful Implementation of Transit Signal Priority on Arterials. Ph.D. dissertation University of Washington, 2000.

[16] Lin, G., Liang, P., Schonfeld, P., and Larson, R., Adaptive Control of Transit Operations. Tech. Rep. MD-26-7002, Transportation Studies Center University of Maryland, College Park, MD, 1995.

[17] Baker, R. J., Dale, J. J., and Head, L., An Overview of Transit Signal Priority. ITS America.

[18] Vasudevan, M., Robust Optimization Model for Bus Priority under Arterial Progression. Ph.D. dissertation University of Maryland, 2005.

[19] Ma, W. and Yang, X., A Passive Signal Priority Approach for Bus Rapid Transit System. Proceedings of the 14th World Congress on Intelligent Transport Systems, (2007): 413–418.

[20] Ma, W. Development and Evaluation of a Coordinated and Conditional Bus Priority Approach. Transportation Research Record: Journal of the Transportation Research Board 3 , 6 (2010): 49–58.

[21] Tan, C., Park, S., Zhou, K., Lui, H., Lau, P., Li, M., and Zhang, W., Prediction of Vehicle Arrival Times at Signalised Intersections for Signal Priority Control. Proceedings

of the 9th IEEE Intelligent Transportation Systems Conference, (2006): 1477–1482.

[22] Ahn, K. and Rakha, H., Systems-Wide Impacts of Green Extension Transit Signal Priority. IEEE Intelligent Transport Systems Conference 2006, (2006): 91–96.

[23] Zhang, W., Lu, H., Shi, Q., and Liu, Q. Optimal signal-planning method of intersections based on bus priority. Journal of Traffic and Transportation Engineering 4 (2004): 49–53.

[24] Bao, W., Chen, Q., and Xu, X., An Adaptive Traffic Signal Timing Scheme for Bus Priority at Isolated Intersection. Proceedings of the 6th World Congress on Intelligent Control and Automation.

[25] Silva, B. C., Oliveira, D. D., Bazzan, A. L. C., and Basso, E. W., Adaptive traffic control with reinforcement learning. Proceedings of the 4th Workshop on Agents in Traffic and Transportation, (2006): 80–86.

[26] Teodorovic, D., Varadarajan, V., Popovic, J., Chinnaswamy, M., and Ramaraj, S. Dynamic programming neural network real-time traffic adaptive signal control algorithm. Annals of Operations Research 143 , 1 (2006): 123–131.

[27] Hong, Y. S., Kim, J. S., Son, J. K., and Park, C. K., Estimation of optimal green time simulation using fuzzy neural network. International Fuzzy Systems Conference Proceedings, (1999): 761–766.

[28] Choy, M. C., Srinivasan, D., and Cheu, R. L. Cooperative, hybrid agent architecture for real-time traffic signal control. Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans 33 , 5 (2003): 597–607.

[29] M. Ghanim, F. D. and Lebdeh, G. A., Integration of signal control and transit signal priority optimization in coordinated network using genetic algorithms and artificial neural networks. Transportation Research Board 88th Annual Meeting.

[30] Cai, C., An approximate dynamic programming strategy for responsive traffic signal control. Proceedings of 2007 IEEE international Symposium on Approximate Dynamic Programming and Reinforcement Learning, (2007): 303–310.

[31] Heydecker, B. G., Cai, C., and Wong, C. K., Adaptive dynamic control for road traffic signals. Proceedings of 2007 IEEE International Conference on Networking, Sensing and Control, (2007): 193–198.

[32] Sutton, R. S. and Barto, A. G. Reinforcement Learning: An Introduction. The MIT Press, Cambridge, Massachusetts, 1998.

[33] Abdulhai, B. and Kattan, L. Reinforcement learning: Introduction to theory and potential for transport applications. Canadian Journal of Civil Engineering 30 , 6 (2003): 981–991.

[34] Oliveira, D. D., Bazzan, A. L. C., Silva, B. C. D., Basso, E. W., Nunes, L., Rossetti, R., Oliveira, E. D., Silva, R. D., and Lamb, L., Reinforcement Learning-based Control of Traffic Lights in Non-stationary Environments: A Case Study in a Microscopic Simulator., 2006.

[35] Jacob, C. and Abdulhai, B., Integrated traffic corridor control using machine learning. International Conference on Systems, Man and Cybernetics, (2005): 3460–3465.

[36] Ritcher, S., Traffic light scheduling using policy-gradient reinforcement learning. The International Conference on Automated Planning and Scheduling.

[37] Wiering, M. A., Vreeken, J., Veenen, J., and Koopman, A., Simulation and optimization of traffic in a city. Intelligent Vehicles Symposium, (2004): 453–458.

[38] Li, Y., Wu, R., and Li, W., The Coordination Between Traffic Signal Control Agents Based On Q-learning. The 5th world congress on intelligent control and automation, (2004): 2690–2693.

[39] Lu, S., Liu, X., and Dai, S., Incremental multistep q-learning for adaptive traffic signal control based on delay minimization strategy. Proceeding of 7th World Congress on Intelligent Control and Automation, (2004): 2854–2858.

[40] Kaige, W., Shiru, Q., and Yumei, Z., A stochastic adaptive control model for isolated intersections. IEEE International Conference on Robotics and Biomimetics, (2008): 2256–2260.

[41] Daganzo, C. F. The cell transmission model part II: Network traffic. Transportation Research Part B: Methodological 29b , 2 (1995): 79–93.

[42] Sadek, A. and Basha, N., Self-learning intelligent agents for dynamic traffic routing on transportation networks. International Conference on Complex Systems, (2006): 503–518.

[43] Lo, H. K., Chang, E., and Chan, Y. C. Dynamic network traffic control. Transportation Research Part A: Policy and Practice 35 , 8 (2001): 721–744.

[44] Maher, M. and Feldman, O., The application of the cell transmission model to the optimisation of signals on signalised roundabouts. European Transport Conference, (2002): 1–13.

[45] Lin, W. H. and Wang, C. An enhanced 01 mixed-integer LP formulation for traffic signal control. IEEE Transactions on Intelligence Transportation Systems 5 , 4 (2004): 238–245.

[46] Xie, Y., DEVELOPMENT AND EVALUATION OF AN ARTERIAL ADAPTIVE TRAFFIC SIGNAL CONTROL SYSTEM USING REINFORCEMENT LEARNING. Ph.D. dissertation Texas A&M University, 2007.

[47] Kevin Gardner, N. H.-B. S. and Bretherton, D., AReview of Bus Priority at Traffic Signals around the World. Tech. Rep. FINAL REPORT Version 2.0, Working Program Bus Committee 2007-2009 Technical Cluster “Extra-vehicular technology”, 2009.

[48] Google Map [online]., `https://maps.google.com/maps?hl=en`.

[49] Tueprasert, K. and Aswakul, C. Multiclass Cell Transmission Model for Heterogeneous Mobility in General Topology of Road Network. Journal of Intelligent Transportation Systems 14 , 2 (2010): 68–82.

[50] Shin, C. and K.Choi Saturation flow rate estimation under rainy weather conditions for on-line traffic control purpose. ASCE Journal of Civil Engineering 2 , 3 (2008): 211–222.

[51] Kleinrock, L. Queueing Systems. Volume 1: Theory. Wiley-Interscience, 1975.

[52] Teodorovic, D. and Trani, A. A., Introduction to Transportation Engineering: Applications of Queueing Theory to Intersection Analysis Level of Service. tech. rep., Virginia Polytechnic Institute and State University, 2005.

[53] Katwijk, R., Multi-Agent Look-Ahead Traffic-Adaptive Control. Ph.D. dissertation The Netherlands TRAIL Research School, 2008.

[54] Bus Priority Survey Results [online]., http://www.scoot-utc.com/BusPriorityResults.php?menu=Results.

[55] Jaakkola, T., Singh, S. P., and Jordan, M. I., Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems. Advances in Neural Information Processing Systems 7, MIT Press, (1995): 345–352.

Appendix

Appendix

List of Publications

Pitipong Chanloha, Wipawee Usaha, Jatuporn Chinrungrueng and Chaodit Aswakul

"Traffic Signal Control with Cell Transmission Model Using Reinforcement Learning", submitted to the ASCE Journal of Transportation Engineering, under review. Content taken from Chapter III.

"Performance Comparison Between Queueing Theoretical Optimality and Q-learning Approach for Intersection Traffic Signal Control", The Forth International on Computational Intelligence, Modelling and Simulation (CIMSIM 2012), Kuantan, Malaysia. Content taken from Chapter IV

Biography

Pitipong Chanloha was born on June 16, 1983 in Bangkae District, Bangkok Province. In 2000, he began studying for his Bachelors degree at School of Telecommunication Engineering, Institute of Engineering at Suranaree University of Technology, Nakhon Ratchasima Province. After graduating, he continued to study for a Masters degree at the School of Telecommunication Engineering, Institute of Engineering, Suranaree University of Technology. He has been starting the Doctoral degree in Electrical Engineering at Chulalongkorn University, Bangkok, Thailand, since 2007. His research interests include Vehicular Networks, Intelligent Transportation Systems and Traffic Control Management.